bioRxiv preprint doi: https://doi.org/10.1101/003657; this version posted September 2, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Title: Exploring the spatially explicit predictions of the Maximum Entropy Theory of Ecology

Authors: D.J. McGlinn[1,2,3], X. Xiao[1,2], J. Kitzes[4], E.P.White[1,2]

[1]*Biology Department, Utah State University, Logan, UT 84341 USA;*

[2]*Ecology Center, Utah State University, Logan, UT 84341 USA;*

[3]*Biology Department, College of Charleston, SC 29401;*

[4]*Energy and Resources Group, University of California, Berkeley, CA 94720 USA;*

Email addresses: mcglinndj@cofc.edu; xiao@weecology.org; jkitzes@berkeley.edu; ethan@weecology.org;

Keywords: beta-diversity, community similarity, distance decay, maximum entropy, species turnover, vegetation

Short running-title: Spatially Explicit METE

Corresponding author: DJM

Number of words in abstract: 288

Number of words in main body: 5090

Number of references: 50

bioRxiv preprint doi: https://doi.org/10.1101/003657; this version posted September 2, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Abstract

Aim:

The Maximum Entropy Theory of Ecology (METE) is a unified theory of biodiversity that attempts to simultaneously predict patterns of species abundance, size, and spatial structure. The spatial predictions of this theory have repeatedly performed well at predicting diversity patterns across scales. However, the theoretical development and evaluation of METE has focused on predicting patterns that ignore inter-site spatial correlations. As a result the theory has not been evaluated using one of the core components of spatial structure. We develop and test a semi-recursive version of METE's spatially explicit predictions for the distance decay relationship of community similarity and compare METE's performance to the classic random placement model of completely random species distributions. This provides a better understanding and stronger test of METE's spatial community predictions.

Location:

New world tropical and temperate plant communities.

Methods:

We analytically derived and simulated METE's spatially explicit expectations for the Sorensen index of community similarity. We then compared the distance decay of community similarity of 16 mapped plant communities to METE and the random placement model.

Results:

The version of METE we examined was successful at capturing the general functional form of empirical distance decay relationships, a negative power function relationship between community similarity and distance. However, the semi-recursive approach consistently over-

bioRxiv preprint doi: https://doi.org/10.1101/003657; this version posted September 2, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

predicted the degree and rate of species turnover and yielded worse predictions than the random placement model.

Main conclusions:

Our results suggest that while METE's current spatial models accurately predict the spatial scaling of species occupancy, and therefore core ecological patterns like the species-area relationship, its semi-recursive form does not accurately characterize spatially-explicit patterns of correlation. More generally, this suggests that tests of spatial theories based only on the species-area relationship may appear to support the underlying theory despite significant deviations in important aspects of spatial structure.

bioRxiv preprint doi: https://doi.org/10.1101/003657; this version posted September 2, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Introduction

Community structure can be characterized using a variety of macroecological relationships such as the species-abundance, body size, and species spatial distributions. Increasingly ecologists have recognized that many of these macroecological patterns are inter-related, and progress has been made toward unifying the predictions of multiple patterns using theoretical models (Storch *et al.*, 2008; McGill, 2010). One approach to predicting suites of macrecological patterns are process-based models such as niche and neutral dispersal models, which have the potential to provide biological insight into the process structuring ecological systems (Adler *et al.*, 2007). Alternatively, a new class of constraint-based models suggest that similar patterns may be produced by different sets of processes because the form of the predicted pattern is due to the existence of statistical constraints rather than directly reflecting detailed biological processes (Frank, 2009, 2014; McGill & Nekola, 2010; Locey & White, 2013).

The Maximum Entropy Theory of Ecology (METE) is a recent attempt to explain a number of ecological patterns from the statistical constraint perspective (Harte *et al.*, 2008, 2009; Harte, 2011; Harte & Newman, 2014). METE uses the principle of entropy maximization, that the most likely distribution is the one with the least information (i.e., the one closest to the uniform distribution) subject to a set of constraints (i.e., prior information), to predict distributions of species abundance, body size, and spatial structure. A frequentist perspective on the Maximum Entropy modeling approach is that every possible configuration of a system is equally likely; therefore, the probability of a particular distribution is directly proportional to the number of configurations that distribution is compatible with (Harte, 2011; Harte & Newman, 2014). The distribution with the largest number of compatible system configurations is the predicted most likely state of the system. In contrast to detailed biological models of community

bioRxiv preprint doi: https://doi.org/10.1101/003657; this version posted September 2, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

assembly, METE has no free parameters and only requires information on total community area, total number of individuals, total number of species, and total metabolic rate of all individuals to generate its predictions.

There is strong empirical support for METE's predictions for the species abundance distribution and patterns related to the spatial distribution of individuals and species (Harte *et al.*, 2008, 2009; Harte, 2011; White *et al.*, 2012a; Xiao *et al.*, 2013; McGlinn *et al.*, 2013; Newman *et al.*, 2014). Specifically, METE has been successful at predicting spatially implicit patterns of community structure such as the species spatial abundance distribution and the species-area relationship (Harte *et al.*, 2008, 2009; McGlinn *et al.*, 2013). It has even been proposed that the METE spatial predictions yield a widely applicable universal species-area relationship (Harte *et al.*, 2009, 2013, but see Šizling *et al.*, 2011, 2013). However all of METE's spatial predictions that have been tested focus on spatially implicit patterns that ignore spatial correlations. As a result the theory has not been evaluated using one of the core components of spatial structure. This is due in part to the fact that METE's spatial correlation predictions have not been fully derived.

The most commonly studied ecological pattern that relies on these spatial correlations is the distance decay relationship (DDR) in which the similarity of species composition decreases with distance (Nekola & White, 1999). The DDR provides a spatially-explicit, community-level characterization of intra-specific aggregation patterns including correlations in space (Plotkin & Muller-Landau, 2002; Palmer, 2005; Morlon *et al.*, 2008; McGlinn & Palmer, 2011), and predicting the DDR is an important area of future development for METE because the DDR is necessary to accurately extrapolate community patterns to unsampled areas (Harte, 2011).

bioRxiv preprint doi: https://doi.org/10.1101/003657; this version posted September 2, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Here we explore METE's spatially explicit predictions for the DDR by developing analytical and simulation based solutions and comparing them to empirical data. We build on the Hypothesis of Equal Allocation Probabilities (HEAP, Harte et al. 2005, Harte 2007) using an approach that combines elements of both a non-recursive and recursive version of METE (McGlinn et al. 2013). We test those predictions using data from 16 spatially explicit plant communities and compare METE's performance to the classic Random Placement Model (RPM) in which individuals are randomly placed on the landscape (Coleman, 1981). Our approach provides a stronger evaluation of the performance of this model and whether it can explain patterns of spatial structure in the absence of detailed biological processes.

Methods

METE has thus far been used to derive the probability that a random cell on a landscape will be occupied by a given number of individuals (i.e., the intra-specific spatial abundance distribution). Predictions for this distribution have been based either on recursively subdividing an area in half or on predicting species abundances directly at smaller scales (Harte, 2011; McGlinn *et al.*, 2013). In addition to the spatial abundance distribution, the DDR requires a prediction for the correlations in abundance among neighboring cells, which has proven difficult to derive for METE (Harte 2011).

Developing METE's Spatially Explicit Predictions

METE's spatial predictions depend on two conditional probability distributions which are computed using independent applications of MaxEnt:

1) the species abundance distribution (SAD), $\Phi(n \mid S_0, N_0)$, the probability that a species has abundance n in a community with S_0 species and N_0 individuals, and

bioRxiv preprint doi: https://doi.org/10.1101/003657; this version posted September 2, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

2) the intra-specific spatial abundance distribution, $\Pi(n \mid A, n_0, A_0)$, the probability that n individuals of a species with n_0 total individuals are located in a random quadrat of area A drawn from a total area A_0.

The METE prediction for Φ is calculated using entropy maximization with constraints on the average number of individuals per species (N_0/S_0) and the maximum number of individuals N_0 for a given species, which yields a truncated log-series abundance distribution (Harte *et al.*, 2008; Harte, 2011). The spatially implicit Π distribution is solved for using entropy maximization with constraints on the average number of individuals per unit area (n_0/A_0) and the maximum number of individuals n_0 of a given species. Although METE requires information on total metabolic rate to derive its predictions, the exact value that this constraint takes has no influence on Φ and Π (Harte *et al.*, 2009; Harte, 2011).

Previous studies have downscaled (or upscaled) METE's predictions using recursive and non-recursive approaches. Here we develop a spatially explicit approach to downscaling METE's predictions that combines elements of both approaches and builds off an existing theoretical framework for modeling the DDR. With the recursive version of METE, Φ and Π are solved for at each successive halving or bisection of A_0 until the area of interest is reached. After each bisection, Φ and Π are calculated and used to derive predicted values of average S and N at that scale which provide updated constraints for the next bisection (Harte et al. 2009). Alternatively, a non-recursive approach can be used in which, Φ and Π at the spatial grain of interest can be solved for directly from the constraints placed at A_0 (Harte et al. 2008). A semi-recursive approach is also possible in which Π is recursively downscaled but Φ is not. The semi-recursive predictions of METE have not been previously examined but this model builds directly on the existing theoretical derivations of the DDR by Harte (2007) for the Hypothesis of Equal

bioRxiv preprint doi: https://doi.org/10.1101/003657; this version posted September 2, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Allocation Probabilities (HEAP). In Appendix A, Fig. A1 and A2 we examine how the semi-recursive formulation of METE differs from a previous examination of the METE recursive and non-recursive SARs (McGlinn et al. 2013), and in Appendix B we develop the analytical derivations of the semi-recursive formulation of the DDR.

In the semi-recursive formulation of the DDR, multi-cell correlations emerge from the spatially nested application of a recursive bisection scheme in which individuals are randomly placed in the left or right half of a cell at each bisection (Fig. 1). Biologically, this can be thought of as a sequentially dependent colonization rule in which individuals randomly choose to occupy the left or right side of an area depending on the existing number of individuals in each half (Harte et al. 2005, Harte 2007, and Conlisk et al. 2007). Our version of METE assumes that for a single bisection there is an equal likelihood for every possible spatial configuration of indistinguishable individuals (Eq. B1). Multi-cell spatial correlations emerge from this approach because the two cells that are formed from a common parent cell are adjacent to one another and are likely to be more similar in abundance than other cells on the landscape (Fig. 1). This approach has three important and inter-related limitations: 1) At each stage in the bisection algorithm, information about the cells surrounding the parent cell is ignored when determining allocations within the parent cell, 2) between-cell distance is defined in reference to an artificial bisection scheme which does not have a one-to-one correspondence with physical distance, and 3) the correlation between cells does not decrease smoothly with physical distance. Alternative approaches have been proposed for deriving the DDR for METE based on computing the single-cell Π distribution at two or more scales and then using the scaling of this marginal distribution to infer the probabilities of a given spatial configuration of abundance (Harte 2011). However, these approaches have yet to yield predictions for the DDR.

bioRxiv preprint doi: https://doi.org/10.1101/003657; this version posted September 2, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

The analytical forms of the semi-recursive formulation (Appendix B) are time-intensive to compute due to the multiple levels of recursion, ignore patterns of abundance (i.e., are formulated only in terms of presence-absence), and are not exact. An alternative approach to deriving semi-recursive METE predictions for the DDR is to use a spatially-explicit simulation.

Spatially Explicit METE Simulation

To simulate semi-recursive METE's spatial predictions, the equal probability rule (Eq. B1) that METE assumes when total area is halved is recursively applied starting at the anchor scale A_0 and progressively bisecting the area until the finest spatial grain of interest is achieved (Fig. 1). Abundance in the simulation model can be parameterized using an observed SAD or using a random realization of the METE SAD given the values of S_0 and N_0. Once the abundances of the species are assigned, each species is independently spatially distributed. Because the equal probability rule requires that there is an equal probability of 0 to n_0 individuals occurring on the left or right side of the total area A_0, the number of individuals in the left side can be set as a draw from a discrete random uniform distribution between 0 and n_0 and the remaining number of individuals are placed on the right hand side.

Datasets

We used a database of 16 spatially explicit and contiguous community datasets compiled by McGlinn et al. (2013) to evaluate the DDR predictions of recursive METE (Table 1). All of the sites were terrestrial, woody plant communities with the exception of the serpentine grassland dataset which covered a terrestrial, herbaceous plant community. In the woody plant communities, all stems were recorded that were at least 10 mm in diameter at breast height (i.e., 1.4 m from the ground) with the exception of the Oosting and Cross Timbers sites where the minimum diameter was 20 and 25 mm respectively. Recursive METE only generates predictions

bioRxiv preprint doi: https://doi.org/10.1101/003657; this version posted September 2, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

for bisections of total area; therefore, we restricted our analysis to square or rectangular areas with a length-to-width ratio of 2:1. Two of the sites had irregular plot designs: Sherman and Cocoli. At these sites we partitioned the datasets into two 2:1 rectangles and analyzed each half independently and then averaged the results (see Supplemental Information: Fig. S1 in McGlinn et al. 2013). See McGlinn et al. (2013) for additional information on site selection criteria, and in particular their Supplemental Table 1, which provides a more complete description of the datasets used in our analysis.

Data Analysis

We compared the fit of METE with and without the observed SAD and the random placement model (RPM) to the empirical DDRs. The METE predictions represented averages of the abundance-based Sørensen index across 200 simulated communities. The abundance-based RPM predictions were generated by distributing the observed number of individuals of each species randomly in space and then computing the average abundance-based Sørensen index across 500 permutations (Morlon *et al.*, 2008).

The DDR is sensitive to the choice of the spatial grain of comparison (Nekola & White, 1999); so, we examined the DDR at several spatial grains for each dataset. We examined spatial grains resulting from 3-13 bisections of A_0. To ensure that the samples at a given grain were square we only considered odd numbers of bisections when A_0 was rectangular and even numbers of bisections when A_0 was square. To ensure the best possible comparison between the observed data and METE and to avoid detecting unusual spatial artefacts in the METE predicted patterns we employed the "user rules" of Ostling et al. (2004) such that samples at a specific grain (i.e., level of bisection) were only compared if they were separated by a specific line of bisection (i.e., a given separation order, Fig. 1 and Appendix A, Fig. A3). This approach was

bioRxiv preprint doi: https://doi.org/10.1101/003657; this version posted September 2, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

taken rather than the standard method of constructing the DDR from all possible pairwise sample comparisons without reference to an imposed bisection scheme. We computed geographic distance by averaging the distance between all the compared samples compared at a given separation order. For the Crosstimbers study site we were not able to examine the DDR based on the METE SAD because of difficulty in generating random realizations of the METE SAD needed for the community simulator when S_0 is less than approximately 10. Typically averages of community similarity are used to examine the geometry of the DDR; however, in some cases the distribution of the similarity metric may be strongly skewed and therefore we computed both averages and medians of community similarity at each separation order.

We used weighted least squares (WLS) regression to account for differences in the number of pairwise comparisons at different spatial lags (there are many more comparisons at short lags) when fitting the power and exponential models of the DDR (Venables & Ripley, 2002). We examined the power model and exponential models because they are the simplest statistical models of the DDR, and it was recently suggested that at fine spatial scales the DDR should be best approximated by a power model (Nekola & White, 1999; Nekola & McGill, 2014).

We checked that our results were consistent with the results provided in previous studies (Harte, 2007, Fig. 6.7 and 6.8, 2011, Fig. 4.1), and that the DDR generated by the community simulator closely agreed with the analytical solution Eq. B5 (Appendix B, Fig. B1). The code to recreate the analysis is provided as Appendix D and at the following publicly available repository: http://dx.doi.org/10.6084/m9.figshare.978918.

bioRxiv preprint doi: https://doi.org/10.1101/003657; this version posted September 2, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Results

In general, the semi-recursive METE distance decay relationship (DDR) provided a poor fit to the empirical DDR (Figs. 2 and 3). The average and median community similarity results were highly correlated (r = 0.98) and generated qualitatively similar results (Appendix A Figs. A5 and A8); therefore, we focus on the results based on averaging similarity. While the METE DDRs exhibited the general functional form of the empirical DDRs, an approximately power-law decrease in similarity with distance, they typically had lower intercepts and steeper slopes than the empirical DDRs (Fig. 2, Appendix A, Fig. A4 and A6). Both the empirical and METE predicted DDR were better approximated by power rather than exponential models (Appendix A, Fig. A6). METE converged towards reasonable predictions at fine spatial grains; however, this is to be expected because at these scales similarity in both the observed and predicted patterns must converge to zero due to low individual density (grey points in Fig. 3A,B). This is because when individual density is low the probability of samples sharing species decreases rapidly simply due to chance. The RPM is known to be a poor model for distance decay because it does not exhibit a decrease in similarity with distance. However, it fit the empirical DDR slightly better than METE (Figs. 2 and 3).

The METE DDR was not strongly influenced by the choice of using the observed or the METE SAD (Figs. 2 and 3A,B). The METE SAD typically yielded a DDR with a slightly lower intercept with the exception of the four tropical sites where it produced DDRs with slightly higher intercepts. In general, we did not observe strong consistent differences between the habitat types (Fig. 2, Appendix A, Fig. A7).

Our formulation of a semi-recursive METE produced SARs that generally agreed (i.e., within the 95% CI) with the recursive and non-recursive formulations of METE (Harte et al.

bioRxiv preprint doi: https://doi.org/10.1101/003657; this version posted September 2, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

2009); however, it did appear that the semi-recursive approach systematically deviated towards lower richness at fine spatial scales which is consistent with predicting stronger patterns of spatial aggregation compared to the other formulations of METE (Appendix A, Fig. A1 and A2).

Discussion

The semi-recursive METE distance decay relationship (DDR) was well approximated by a decreasing power function, and thus consistent with the general form of empirical DDRs, but it provided a poor fit to empirical data. Specifically, the slope and the intercept of this power function deviate substantially from empirical data resulting in a poor fit. These deviations contrast with a number of studies showing that the theory successfully predicts both the Π distribution and the SAR (Harte *et al.*, 2008, 2009; Harte, 2011; McGlinn *et al.*, 2013; but see Šizling *et al.*, 2011). Both Π and the SAR are influenced by the spatially explicit pattern of intraspecific aggregation but neither pattern reflects inter-quadrat correlations and therefore they represent coarse metrics of spatial structure. The combination of a well fit SAR and a poorly fit DDR suggests that the current version of METE accurately characterizes average occupancy, but fails to characterize the spatial relationships among cells (McGeoch & Gaston, 2002; Storch *et al.*, 2003; McGlinn & Hurlbert, 2012; Nekola & McGill, 2014).

These results only apply directly to the particular HEAP-based semi-recursive version of the spatial METE theory, which represents a middle ground in terms of approach between Harte et al. (2008) and Harte et al. (2009). Other approaches to deriving the METE DDR may perform better than the semi-recursive approach if they can be developed. It has been suggested that there is no *a priori* reason to prefer one version of the theory and that the best way to choose among the different versions is empirically (Haegeman & Etienne, 2010; Harte, 2011). However, the traditionally defined recursive and non-recursive versions of METE have shortcomings with

bioRxiv preprint doi: https://doi.org/10.1101/003657; this version posted September 2, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

respect to how their assumptions and predictions are scaled, and the semi-recursive approach we defined is limited by its dependence on an artificial bisection scheme. Specifically the recursive approach predicts that the SAD has the same functional form, a truncated log-series, at all scales. This is problematic because SADs are typically not scale-invariant if, as METE predicts, species display intraspecific spatial aggregation (Green & Plotkin, 2007; Šizling *et al.*, 2009). The non-recursive approach does not suffer from this problem because the SAD is only solved for at the anchor scale; however, Haegemann and Etienne (2010) found that the non-recursive predictions for a multi-cell generalization of the Π distribution were scale-inconsistent. The semi-recursive approach does not suffer from this shortcoming because its multi-cell form (see Eq. 2.2 in Conlisk *et al.*, 2007) is only defined over the set of bisections that are consistent with a landscape in which n_0 individuals are distributed (see Appendix C for proof). However, the set of bisections is artificial and multi-cell correlations only emerge from this approach in reference to bisection distance rather than directly to physical distance between cells such that cells have equal magnitude of correlation regardless of their physical distance if they have equivalent separation orders (see Conlisk *et al.*, 2007 for a critique of distances defined by separation indices). An important future direction for METE is to attempt to develop spatial multi-cell predictions using approaches that avoid these shortcomings and the two approaches suggested by Harte (2011) for deriving the METE DDR may provide a useful starting point for future development.

Our results suggest that semi-recursive METE differs from spatial patterns observed in nature. This deviation could indicate that the emergent statistical approach to modeling spatial structure is incorrect, with specific biological processes such as dispersal limitation or environmental filtering directly controlling spatial correlation (Condit *et al.*, 2002; Gilbert & Lechowicz, 2004; Karst *et al.*, 2005; Seidler & Plotkin, 2006; Chase, 2007; McGlinn & Palmer,

bioRxiv preprint doi: https://doi.org/10.1101/003657; this version posted September 2, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

2011). Alternatively it could mean that while the general idea underlying the theory is valid, the specific formulation is wrong. For example it could be that the approaches outlined by Harte (2011) that are more sophisticated in how they handle spatial correlations will be more appropriate or that a generalized version of this kind of recursive approach like that developed by Conlisk et al. (2007) in which the degree of aggregation is a tunable parameter will capture the reality of biological systems more precisely. However, process-, and constraint-based models should not necessarily be treated as mutually exclusive. For example other process-based theories make power-law like predictions for the form of the DDR. In fact, it has recently been suggested that at fine spatial scales most theories will make predictions that are approximately power-law in nature (Nekola & McGill, 2014). This means that simply noting power-law like DDR relationships does not provide a strong method for differentiating among theories. In fact, had we simply looked for power-law like behavior we would have concluded that the semi-recursive METE was consistent with empirical data. However, one of the properties that makes METE such a strong theory is that it makes specific predictions for precise parameters as well as general forms of empirical relationships. This allows it to be more rigorously compared to data and to other theories that predict different parameters values for a similar general form of the DDR (e.g., neutral theory).

Our results mirror those of Xiao et al. (2013) and Newman et al. (2014) evaluating the non-spatial aspects of METE. All three studies show that when evaluating the theory using multiple patterns simultaneously some of the predictions perform well and some perform poorly. It is inherently difficult for theories to predict large numbers of patterns simultaneously, which is why evaluating theory in this way provides stronger tests than evaluating single patterns (McGill, 2003; McGill *et al.*, 2006). General theories like METE that make multiple predictions are

bioRxiv preprint doi: https://doi.org/10.1101/003657; this version posted September 2, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

therefore both easier to evaluate and also more broadly useful since they allow a large number of patterns to be predicted from a relatively small amount of information. Because there are many patterns to evaluate it is also more likely that deviations from theory will be identified (White *et al.*, 2012). In some cases these deviations may indicate that the theory is fundamentally unsound, but in others it may suggest modifications to the theory to address the observed deviations (White *et al.*, 2012). Whether METE can be modified to address the observed deviations from empirical data remains to be seen. In the case of the DDR, despite its generality, there are a limited number of models that attempt to predict the DDR from first principles (Chave & Leigh, 2002; Condit *et al.*, 2002; Zillio *et al.*, 2005; Harte, 2007, 2011; Nekola & McGill, 2014), which means that it may be worth pursuing the METE approach further.

METE is one of several general theories in ecology that make many predictions for many aspects of ecological community structure based on only a small amount of information. Our analysis of the semi-recursive formulation of METE's spatially explicit prediction for the DDR suggests that this form of the theory over-predicts the strength of spatial correlation. These results coupled with studies of the species-area relationship suggest that semi-recursive METE accurately predicts the scaling of species occupancy but not spatial correlation. More generally, our results demonstrate that tests of spatial theories that focus solely on the species-area relationship and related patterns are only evaluating part of the spatial pattern, the distribution of occupancy among cells. Evaluating these theories using the DDR in addition to the SAR will help identify cases where the theories are correctly identifying some aspects of spatial structure, but not others, and thus yield stronger tests of the underlying theory. In some cases this will require extending the theory to make additional predictions, but this effort will provide both more testable and more usable theories.

bioRxiv preprint doi: https://doi.org/10.1101/003657; this version posted September 2, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Acknowledgements

We thank John Harte for extensive conversations on the use of maximum entropy in ecology. Robert K. Peet provided data for the oak-hickory and old field pine forests. Jessica Green provided data for the serpentine grassland data. José Arévalo provided data for the oak woodland data. The BCI forest dynamics research project was made possible by National Science Foundation grants to Stephen P. Hubbell: DEB-0640386, DEB-0425651,DEB-0346488, DEB-0129874, DEB-00753102, DEB-9909347, DEB-9615226, DEB-9615226, DEB-9405933, DEB-9221033, DEB-9100058, DEB-8906869, DEB-8605042,DEB-8206992, DEB-7922197, support from the Center for Tropical Forest Science, the Smithsonian Tropical Research Institute, the John D. and Catherine T. MacArthur Foundation, the Mellon Foundation, the Small World Institute Fund, and numerous private individuals, and through the hard work of over 100 people from 10 countries over the past two decades. The plot project is part the Center for Tropical Forest Science, a global network of large-scale demographic tree plots. The Luquillo Experimental Forest Long-Term Ecological Research Program, supported by the U.S. National Science Foundation, the University of Puerto Rico, and the International Institute of Tropical Forestry. This research was supported by a CAREER grant from the U.S. National Science Foundation to EPW (DEB-0953694).

References

Adler, P.B., HilleRisLambers, J. & Levine, J.M. (2007) A niche for neutrality. *Ecology Letters*, **10**, 95–104.

Arévalo, J.R. (2013) Spatial analysis and structure of a cross-timber stand in the TallGrass Prairie Preserve (Pawhuska, Oklahoma). *Journal of Forestry Research*, **24**, 47–52.

bioRxiv preprint doi: https://doi.org/10.1101/003657; this version posted September 2, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Chase, J.M. (2007) Drought mediates the importance of stochastic community assembly. *Proceedings of the National Academy of Sciences of the United States of America*, **104**, 17430–17434.

Chave, J. & Leigh, E.G. (2002) A spatially explicit neutral model of beta-diversity in tropical forests. *Theoretical Population Biology*, **62**, 153–168.

Coleman, B.D. (1981) On random placement and species-area relations. *Mathematical Biosciences*, **54**, 191–215.

Condit, R. (1998) *Tropical Forest Census Plots*, Springer-Verlag and R. G. Landes Company, Berlin, Germany and Georgetown, Texas.

Condit, R., Pitman, N., Leigh, E.G., Chave, J., Terborgh, J., Foster, R.B., Nunez, P., Aguilar, S., Valencia, R., Villa, G., Muller-Landau, H.C., Losos, E. & Hubbell, S.P. (2002) Beta-diversity in tropical forest trees. *Science*, **295**, 666–669.

Condit, R., Salomon Aguilar, Hernandez, A., Perez, R., Lao, S., Angehr, G., Hubbell, S.P. & Foster, R.B. (2004) Tropical Forest Dynamics across a Rainfall Gradient and the Impact of an El Niño Dry Season. *Journal of Tropical Ecology*, **20**, 51–72.

Conlisk, E., Bloxham, M., Conlisk, J., Enquist, B. & Harte, J. (2007) A new class of models of spatial distribution. *Ecological Monographs*, **77**, 269–284.

Frank, S.A. (2014) Generative models versus underlying symmetries to explain biological pattern. *Journal of Evolutionary Biology*, **27**, 1172–1178.

Frank, S.A. (2009) The common patterns of nature. *Journal of Evolutionary Biology*, **22**, 1563–1585.

bioRxiv preprint doi: https://doi.org/10.1101/003657; this version posted September 2, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Gilbert, B. & Lechowicz, M.J. (2004) Neutrality, niches, and dispersal in a temperate forest understory. *Proceedings of the National Academy of Sciences of the United States of America*, **101**, 7651–7656.

Gilbert, G.S., Howard, E., Ayala-Orozco, B., Bonilla-Moheno, M., Cummings, J., Langridge, S., Parker, I.M., Pasari, J., Schweizer, D. & Swope, S. (2010) Beyond the tropics: forest structure in a temperate forest mapped plot. *Journal of Vegetation Science*, **21**, 388–405.

Green, J.L., Harte, J. & Ostling, A. (2003) Species richness, endemism and abundance patterns: tests of two fractal models in a serpentine grassland. *Ecology Letters*, **6**, 919–928.

Green, J.L. & Plotkin, J.B. (2007) A statistical theory for sampling species abundances. *Ecology Letters*, **10**, 1037–1045.

Haegeman, B. & Etienne, R.S. (2010) Entropy maximization and the spatial distribution of species. *The American Naturalist*, **175**, E74–E90.

Harte, J. (2011) *Maximum Entropy and Ecology: A Theory of Abundance, Distribution, and Energetics*, Oxford University Press Inc., New York.

Harte, J. (2007) *Toward a mechanistic basis for a unified theory of spatial structure in ecological communities at multiple spatial scales*. *Scaling Biodiversity* Ecological Reviews. (ed. by D. Storch, P.A. Marquet, and J.H. Brown), pp. 101–126. Cambridge University Press, Cambridge.

Harte, J., Kitzes, J., Newman & Rominger, A.J. (2013) Taxon Categories and the Universal Species-Area Relationship: A Comment on Šizling et al., "Between Geometry and Biology: The Problem of Universality of the Species-Area Relationship." *The American Naturalist*, **181**, 282–287.

bioRxiv preprint doi: https://doi.org/10.1101/003657; this version posted September 2, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Harte, J. & Newman, E.A. (2014) Maximum information entropy: a foundation for ecological theory. *Trends in Ecology & Evolution*, **29**, 384–389.

Harte, J., Smith, A.B. & Storch, D. (2009) Biodiversity scales from plots to biomes with a universal species-area curve. *Ecology Letters*, **12**, 789–797.

Harte, J., Zillio, T., Conlisk, E. & Smith, A.B. (2008) Maximum entropy and the state-variable approach to macroecology. *Ecology*, **89**, 2700–2711.

Hubbell, S.P., Condit, R. & Foster, R.B. (2005) Barro Colorado Forest Census Plot Data. URL http://ctfs.arnarb.harvard.edu/webatlas/datasets/bci.

Hubbell, S.P., Foster, R.B., O'Brien, S.T., Harms, K.E., Condit, R., Wechsler, B., Wright, S.J. & de Lao, S.L. (1999) Light-gap disturbances, recruitment limitation, and tree diversity in a neotropical forest. *Science*, **283**, 554–557.

Karst, J., Gilbert, B. & Lechowicz, M.J. (2005) Fern community assembly: The roles of chance and the environment at local and intermediate scales. *Ecology*, **86**, 2473–2486.

Locey, K.J. & White, E.P. (2013) How species richness and total abundance constrain the distribution of abundance. *Ecology Letters*, **16**, 1177–1185.

McDonald, R.I., Peet, R.K. & Urban, D.L. (2002) Environmental correlates of oak decline and red maple increase in the North Carolina piedmont. *Castanea*, **67**, 84–95.

McGeoch, M.A. & Gaston, K.J. (2002) Occupancy frequency distributions: patterns, artefacts and mechanisms. *Biological Reviews*, **77**, 311–331.

McGill, B. (2003) Strong and weak tests of macroecological theory. *Oikos*, **102**, 679–685.

McGill, B.J. (2010) Towards a unification of unified theories of biodiversity. *Ecology Letters*, **13**, 627–642.

bioRxiv preprint doi: https://doi.org/10.1101/003657; this version posted September 2, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

McGill, B.J., Maurer, B.A. & Weiser, M.D. (2006) Empirical evaluation of neutral theory. *Ecology*, **87**, 1411–1423.

McGill, B.J. & Nekola, J.C. (2010) Mechanisms in macroecology: AWOL or purloined letter? Towards a pragmatic view of mechanism. *Oikos*, **119**, 591–603.

McGlinn, D.J. & Hurlbert, A.H. (2012) Scale dependence in species turnover reflects variance in species occupancy. *Ecology*, **93**, 294–302.

McGlinn, D.J. & Palmer, M.W. (2011) Quantifying the influence of environmental texture on the rate of species turnover: evidence from two habitats. *Plant Ecology*, **212**, 495–506.

McGlinn, D.J., Xiao, X. & White, E.P. (2013) An empirical evaluation of four variants of a universal species–area relationship. *PeerJ*, **1**, e212.

Morlon, H., Chuyong, G., Condit, R., Hubbell, S., Kenfack, D., Thomas, D., Valencia, R. & Green, J.L. (2008) A general framework for the distance-decay of similarity in ecological communities. *Ecology Letters*, **11**, 904–917.

Nekola, J.C. & McGill, B.J. (2014) Scale dependency in the functional form of the distance decay relationship. *Ecography*.

Nekola, J.C. & White, P.S. (1999) The distance decay of similarity in biogeography and ecology. *Journal of Biogeography*, **26**, 867–878.

Newman, E.A., Harte, M.E., Lowell, N., Wilber, M. & Harte, J. (2014) Empirical tests of within- and across-species energetics in a diverse plant community. *Ecology*.

Ostling, A., Harte, J., Green, J.L. & Kinzig, A.P. (2004) Self-similarity, the power law form of the species-area relationship, and a probability rule: Reply. *American Naturalist*, **163**, 627–633.

bioRxiv preprint doi: https://doi.org/10.1101/003657; this version posted September 2, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Palmer, M.W. (2005) Distance decay in an old-growth neotropical forest. *Journal of Vegetation Science*, **16**, 161–166.

Palmer, M.W., Peet, R.K., Reed, R.A., Xi, W. & White, P.S. (2007) A multiscale study of vascular plants in a North Carolina piedmont forest. *Ecology*, **88**, 2674.

Peet, R.K. & Christensen, N.L. (1987) Competition and Tree Death. *BioScience*, **37**, 586–595.

Plotkin, J.B. & Muller-Landau, H.C. (2002) Sampling the species composition of a landscape. *Ecology*, **83**, 3344–3356.

Seidler, T.G. & Plotkin, J.B. (2006) Seed dispersal and spatial pattern in tropical trees. *PLoS Biology*, **4**, 2132–2137.

Šizling, A.L., Kunin & Storch, D. (2013) Taxon-and-Area Invariances, Maximum Entropy, and the Species-Area Relationship: A Reply to Harte et al. *The American Naturalist*, **181**, 288–290.

Šizling, A.L., Kunin, W.E., Šizlingová, E., Reif, J. & Storch, D. (2011) Between Geometry and Biology: The Problem of Universality of the Species-Area Relationship. *The American Naturalist*, **178**, 602–611.

Šizling, A.L., Storch, D., Reif, J. & Gaston, K.J. (2009) Invariance in species-abundance distributions. *Theoretical Ecology*, **2**, 89–103.

Storch, D., Šizling, A.L. & Gaston, K.J. (2003) Geometry of the species-area relationship in central European birds: testing the mechanism. *Journal of Animal Ecology*, **72**, 509–519.

Storch, D., Šizling, A.L., Reif, J., Polechová, J., Šizlingová, E. & Gaston, K.J. (2008) The quest for a null model for macroecological patterns: geometry of species distributions at multiple spatial scales. *Ecology Letters*, **11**, 771–784.

Venables, W.N. & Ripley, B.D. (2002) *Modern Applied Statistics with S*, 4th edn. Springer.

bioRxiv preprint doi: https://doi.org/10.1101/003657; this version posted September 2, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

White, E.P., Katherine M. Thibault & Xiao, X. (2012a) Characterizing species abundance distributions across taxa and ecosystems using a simple maximum entropy model. *Ecology*, **93**, 1772–1778.

White, E.P., Xiao, X., Isaac, N.J.B. & Sibly, R.M. (2012b) *Methodological tools*. *Metabolic Ecology*, Wiley-Blackwell.

Xi, W., Peet, R.K., Decoster, J.K. & Urban, D.L. (2008) Tree damage risk factors associated with large, infrequent wind disturbances of Carolina forests. *Forestry*, **81**, 317–334.

Xiao, X., McGlinn, D.J. & White, E.P. (2013) A strong test of the Maximum Entropy Theory of Ecology. *arXiv*.

Zillio, T., Volkov, I., Banavar, J.R., Hubbell, S.P. & Maritan, A. (2005) Spatial scaling in model plant communities. *Physical Review Letters*, **95**.

Zimmerman, J.K., Everham, III, E.M., Waide, R.B., Lodge, D.J., Taylor, C.M. & Brokaw, N.V.L. (1994) Responses of tree species to hurricane winds in subtropical wet forest in Puerto Rico: implications for tropical tree life histories. *Journal of Ecology*, **82**, 911–922.

Biosketch

Weecology is founded on the belief that better communication and collaboration between empirical and quantitative scientists is necessary for tackling many of the big scientific challenges in ecology. The purpose of Weecology is to facilitate collaborative research through a variety of mechanisms including shared resources, expertise, and web-based collaborative tools. Our hope is help scientists (and ourselves) collaborate and communicate across disciplinary divides and generate higher-quality novel research as a result.

bioRxiv preprint doi: https://doi.org/10.1101/003657; this version posted September 2, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Tables:

Table 1. Summary of the habitat type and state variables of the vegetation datasets. The state variables are total area (A_0), total abundance (N_0) and total number of species (S_0). A_{min} and A_{max} are the finest and coarsest areas (m^2) examined. Data were collected on woody forest plants with the exception of the serpentine site which contained herbaceous grassland plants.

Site name	**Habitat type**	**Ref**	A_{min}	A_{max}	A_0	N_0	S_0
BCI	tropical	[1-3]	61.0	62500	500000	205096	301
Sherman	tropical	[4]	2.4	625	20000	7623	175
Cocoli	tropical	[4]	2.4	625	20000	4326	139
Luquillo	tropical	[5]	15.3	15625	125000	32320	124
Bryan	oak-hickory	[6-8]	2.1	535	17113	3394	48
Big Oak	oak-hickory	[6-8]	2.4	625	20000	5469	40
Oosting	oak-hickory	[9]	16	4096	65536	8892	39
Rocky	oak-hickory	[6-8]	3.5	900	14400	3383	37
Bormann	oak-hickory	[6-8]	4.8	1225	19600	3879	30
Wood Bridge	oak-hickory	[6-8]	1.2	315	5041	758	19
Bald Mtn.	oak-hickory	[6-8]	2.4	156	5000	669	17
Landsend	old field, pine	[6-8]	1.0	264	8450	2139	41
Graveyard	old field, pine	[6-8]	2.4	625	10000	2584	36
UCSC	mixed-evergreen	[10]	5.4	1406	45000	5885	31
Serpentine	serpentine	[11]	0.3	4	64	37182	24
Cross Timbers	oak woodland	[12]	9.8	2500	40000	7625	7
Ranges			0.3-61.0	4-62500	64-500000	669-205096	7-301

[1.] Condit (1998), [2.] Hubbell et al. (1999), [3.] Hubbell et al. (2005), [4.] Condit et al. (2004), [5.] Zimmerman et al. (1994), [6.] Peet and Christensen (1987), [7.] McDonald et al. (2002), [8.] Xi et al. (2008), [9.] Palmer et al. (2007), [10.] Gilbert et al. (2010), [11.] Green et al. (2003), [12.] Arévalo (2013)

bioRxiv preprint doi: https://doi.org/10.1101/003657; this version posted September 2, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Figures

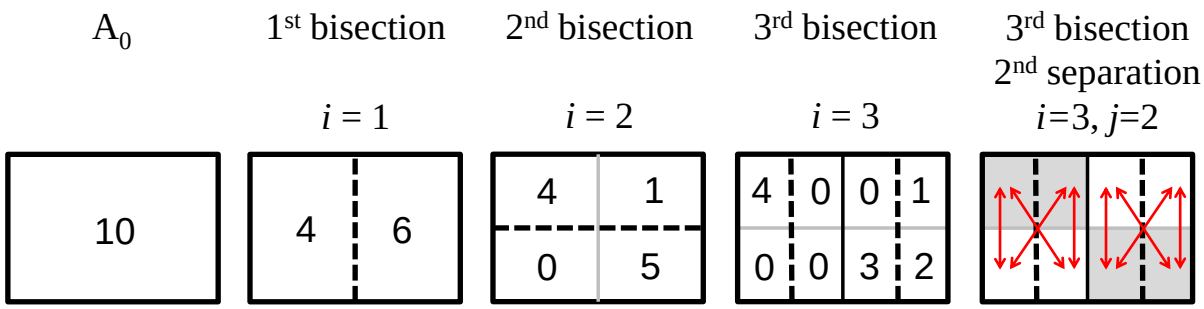
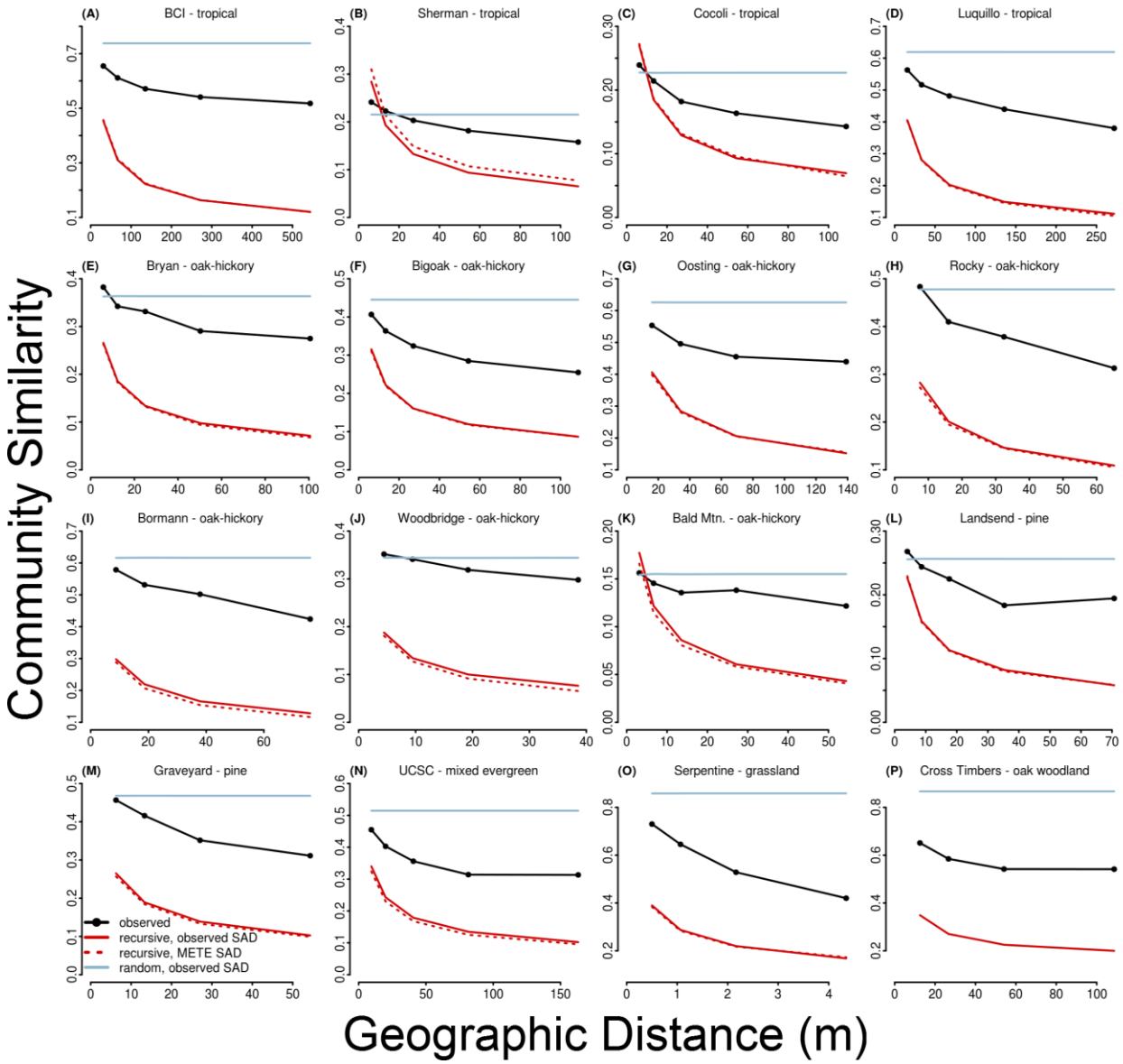


Fig 1. This diagram illustrates the "user rules" of how a landscape is bisected and how samples are compared for a given separation order. In this specific example, three bisections are used to generate a spatially explicit distribution of 10 individuals. In the last panel, the eight pairwise comparisons (arrows) at separation order of 2 for a scale of $A_0/2^3$ (i.e., $A_{i=3}$, $D_{j=2}$) are illustrated. When simulating random bisections the number of individuals distributed to the left or right of the bisection line is a random draw from a discrete uniform distribution.

bioRxiv preprint doi: https://doi.org/10.1101/003657; this version posted September 2, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.



515

516

bioRxiv preprint doi: https://doi.org/10.1101/003657; this version posted September 2, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

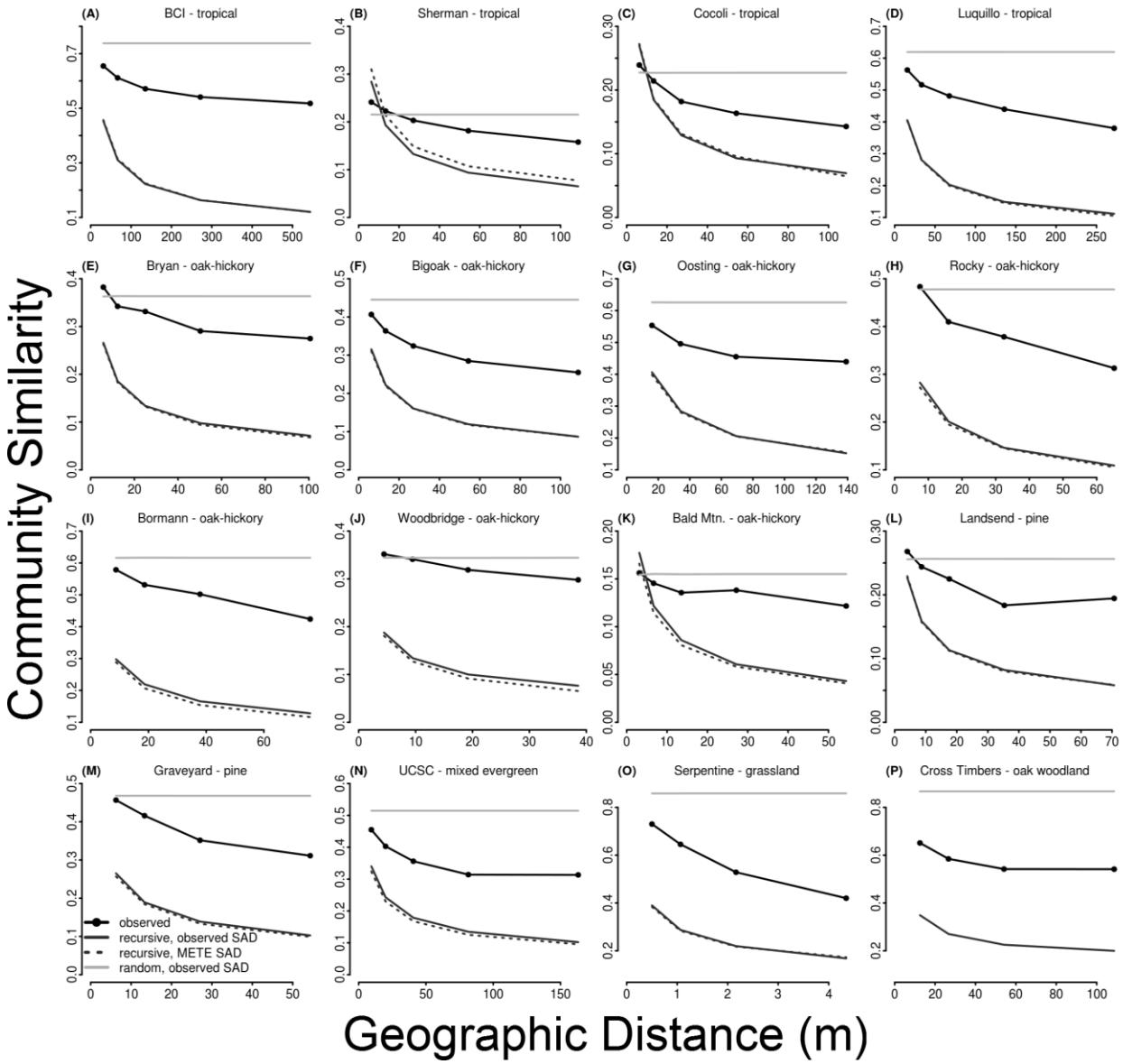


Fig 2. The observed (black line with dots) and predicted distance decay relationships (METE: dark grey lines, solid for the observed SAD, dashed for the METE SAD; random placement: light grey line) for each site at a single spatial grain. Community similarity represents the average of the abundance-based Sørensen index for each spatial lag. The spatial grain displayed was taken at either 8 or 9 bisections of the total area depending on whether the total extent was a square or a rectangle respectively. Geographic distance was calculated as the average physical

bioRxiv preprint doi: https://doi.org/10.1101/003657; this version posted September 2, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

distance between the samples compared at given separation order (see Methods and Fig.1 for additional information).

bioRxiv preprint doi: https://doi.org/10.1101/003657; this version posted September 2, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

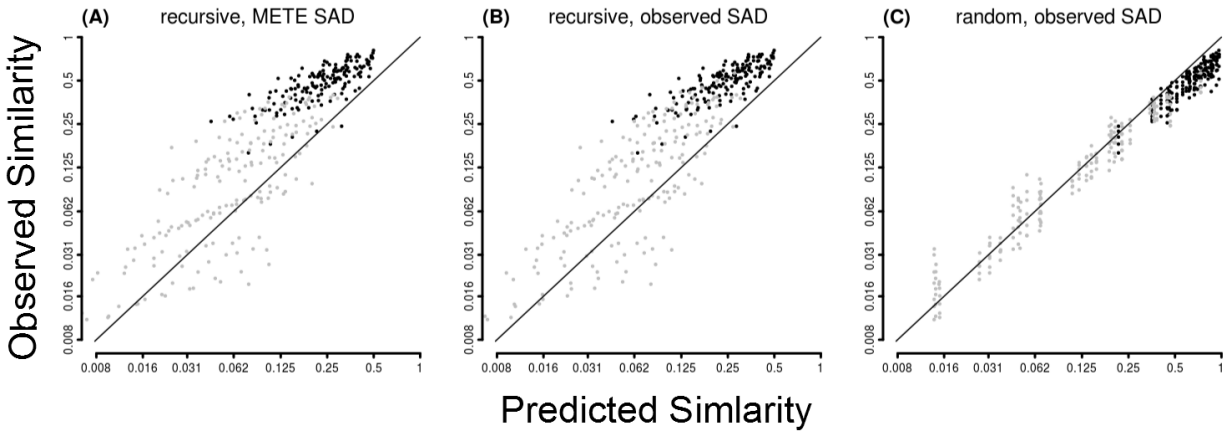


Fig 3. The log-log transformed one-to-one plots of the predicted and observed abundance-based Sørensen similarity values for the three models across all distances and spatial grains. The solid line is the one-to-one line. The grey points represent values from spatial grains in which the average individual density was low (i.e., less than 10 individuals) and thus both the observed and predicted similarities must be close to zero simply because of a sampling effect.