Title: Maximum entropy models elucidate the contribution of metabolic traits to patterns of community assembly

Running head: Metabolic traits in maximum entropy models

Authors:

Jason Bertram[1*], Erica A. Newman[2,3], Roderick C. Dewar[4]

[1]Environmental Resilience Institute, Indiana University, Bloomington, IN 47401 USA

[2]School of Natural Resources and the Environment, University of Arizona, Tucson, AZ 85721 USA

[3]USDA Forest Service, Pacific Wildland Fire Sciences Lab, Seattle, WA 98103 USA

[4]Plant Sciences Division, Research School of Biology, The Australian National University, Canberra, ACT 2601 Australia

[*]Corresponding author email: jxb@iu.edu

Number of words in the Abstract: 268

Number of words in main text (including appendices): 4323

Number of references: 33

ACKNOWLEDGEMENTS

Funding for JB was provided by the Environmental Resilience Institute. Postdoctoral funding for EAN was provided by University of Arizona Bridging Biodiversity and Conservation Science program, and the USDA Forest Service. We thank John Harte for comments on earlier versions of this manuscript, and Dan Gruner for use of his dataset.

BIOSKETCHES

Jason Bertram is a mathematical biologist who develops ecological and evolutionary theory at Indiana University.

Erica A. Newman is postdoctoral researcher focusing on connections between disturbance ecology and biodiversity patterns, including applied topics in wildlife ecology and wildfire management.

Roderick C. Dewar is a professor at the Australian National University, researching entropy-based principles of organization in biological and physical systems.

ABSTRACT

Aim

Maximum entropy (MaxEnt) models promise a novel approach for understanding community assembly and species abundance patterns. One of these models, the "Maximum Entropy Theory of Ecology" (METE) reproduces many observed species abundance patterns, but is based on an aggregated representation of community structure that does not resolve species identity or explicitly represent species-specific functional traits. In this paper, METE is compared to "Very Entropic Growth" (VEG), a MaxEnt model with a less aggregated representation of community structure that represents species (more correctly, functional types) in terms of their per capita metabolic rates. We examine the contribution of metabolic traits to the patterns of community assembly predicted by VEG and, through aggregation, compare the results with METE predictions in order to gain insight into the biological factors underlying observed patterns of community assembly.

Innovation

We formally compare two MaxEnt-based community models, METE and VEG, that differ as to whether or not they represent species-specific functional traits. We empirically test and compare the metabolic predictions of both models, thereby elucidating the role of metabolic traits in patterns of community assembly.

Main Conclusions

Our analysis reveals that a key determinant of community metabolic patterns is the "density of species" distribution $\rho(\epsilon)$, where $\rho(\epsilon)d\epsilon$ is the intrinsic number of species with metabolic rates in the range $(\epsilon, \epsilon + d\epsilon)$ that are available to a community prior to filtering by environmental

constraints. Our analysis suggests that appropriate choice of $\rho(\epsilon)$ in VEG may lead to more realistic predictions than METE, for which $\rho(\epsilon)$ is not defined, and thus opens up new ways to understanding the link between functional traits and patterns of community assembly.

Key words: community assembly, functional traits, macroecology, metabolic requirements, resource partitioning, species-abundance distribution, statistical aggregation

1 INTRODUCTION

One of the central aims of ecology is to understand the determinants of community assembly. Many studies of community assembly involve summaries of community structure such as the species abundance distribution (SAD), species-area relationship (SAR), and analogous metabolic-rate distributions. We will refer to these summary distributions collectively as community structure distributions (CSDs). CSDs, particularly the SAD, have attracted a lot of attention because their shapes are strikingly similar across different communities, representing a rare example of "universality" in community ecology (McGill *et al.* 2007).

The existence of universal features in CSDs is intriguing because these could reflect universal aspects of the biological processes responsible for structuring communities. However, CSDs could also be universal for statistical reasons (Tokeshi 1993; Ulrich *et al*., 2010). Similar to how the normal distribution is ubiquitous because many measured quantities involve statistical averaging (the central limit theorem), CSDs could be universal simply because community-specific details disappear in aggregating patterns to the level of species counts, or other forms of averaging. This would make CSDs considerably less valuable for understanding the biological determinants of community assembly, such as how community structure depends on the

functional traits of the organisms in the community (McGill *et al.*, 2006; Díaz et al. 2013). It is therefore important to disentangle the contributions of biological versus statistical factors to CSDs. This issue is closely related to the long-running debate on the relative roles of "mechanism" and "drift" in ecology (McGill and Nekola, 2010; Vellend, 2010), and on ecosystem stability and the role of disturbance (Newman et al. 2018).

A promising recent approach for disentangling biological from statistical factors in ecological models is to use the statistical principle of maximum entropy (MaxEnt). MaxEnt models are "top-down" in that they seek to identify a minimal set of biological assumptions required to reproduce a given empirical pattern (such as a CSD). Once these assumptions have been specified, the MaxEnt principle predicts statistical patterns of community structure by effectively treating all other mechanistic details statistically as unbiased random noise. By empirically testing predictions based on different assumptions, MaxEnt provides a means to resolve the partitioning of biological versus statistical factors in driving observed ecological patterns.

MaxEnt models have had some success at predicting CSDs, but the ecological interpretation of these successes has not been straightforward. A number of MaxEnt models have appeared in the ecological literature with a variety of different assumptions and justifications (*e.g.* Shipley, Vile and Garnier, 2006; Pueyo, He and Zillio, 2007; Harte, Zillio, Conlisk and Smith, 2008; Dewar and Porté, 2008; Banavar, Maritan and Volkov, 2010; Bertram and Dewar, 2015). This had led to extensive debates about the prospects and pitfalls of this approach.

Two key issues in these debates may be identified: one conceptual, the other more technical. The conceptual issue concerns the interpretation of the MaxEnt procedure itself, including the challenge of connecting MaxEnt to familiar ecological processes such as dispersal,

105 disturbance, and interactions between organisms. This issue has been discussed at length
106 elsewhere (Dewar, 2009; McGill and Nekola, 2010; Shipley, 2010; Supp, Xiao, Ernest and
107 White, 2012; Harte and Newman, 2014; Supp and Ernest 2014; Bertram and Dewar, 2015;
108 Newman et al., 2018), and will thus not be our focus here.

109 Rather, our focus will be on the technical issue, which concerns the level of detail at
110 which the community is described in the model before MaxEnt is even applied (He, 2010;
111 Favretti, 2017). Changing the variables used to describe a community (*e.g.* resolving a
112 community in greater detail) can dramatically alter the predictions that MaxEnt makes about the
113 community, and yet there is no apparent *a priori* reason to prefer one choice of variables over
114 another. As a result, a variety of choices have appeared in different models, usually with little
115 justification. A comparison of these disparate approaches is required in order to better guide the
116 application of MaxEnt models in ecology.

117 Here we present a comparison of two MaxEnt-based models in ecology which have both
118 successfully reproduced observed CSDs: METE (Maximum Entropy Theory of Ecology; Harte
119 et al. 2008; Harte et al. 2009; Harte 2011; Harte and Newman, 2014) and VEG (Very Entropic
120 Growth; Dewar and Porté, 2008; Bertram and Dewar, 2013; Bertram and Dewar 2015). These
121 models are well suited for our objective of comparison because METE describes communities at
122 the same coarse-grained level of detail as the SAD, whereas VEG is more detailed in that it
123 resolves the abundance of each separate species.

124 Crucially, this difference in community description allow us to explore the biological
125 determinants of patterns of community assembly. Specifically, VEG distinguishes species by
126 their per capita metabolic traits. By contrast, METE only distinguishes separate species by their
127 abundances, and requires the total number of species present in the community as an input rather

than as a prediction. METE then predicts a distribution of metabolic rates for the individuals in a species as a function of its abundance, imparting functional traits statistically by abundance, rather than by species identity (Section 2.1). It is therefore not possible to investigate how different species-specific functional traits might modify community structure using METE. Thus, by comparing METE and VEG, we are able to investigate more transparently what sorts of functional trait assumptions are necessary for reproducing observed patterns.

2. METE AND VEG: TWO MAXENT MODELS OF COMMUNITY ASSEMBLY

2.1 METE

The central quantity predicted by METE is a joint probability distribution $R_M(n, \epsilon)$ called the "ecosystem structure function." By definition, $R_M(n, \epsilon)d\epsilon$ is the joint probability that a species selected at random from a community has abundance n, and that an individual selected at random from a species with abundance n has a metabolic requirement between ϵ and $\epsilon + d\epsilon$ (Harte et al. 2008; Harte, 2011; Appendix A). The ecosystem structure function is closely related to the SAD: if we add together all of the possible metabolic requirements ϵ we obtain the probability distribution for the abundance of a randomly selected species, $R_M(n) = \int R_M(n, \epsilon)d\epsilon$; the SAD is then simply $SR_M(n)$ where S is the total number of species present in the community. Thus, $R_M(n, \epsilon)$ is a SAD that has been extended to also incorporate information about community metabolic structure.

METE assumes that $R_M(n, \epsilon)$ satisfies two constraints

$$\sum_{n=1}^{N} \int_{\epsilon=1}^{E} n\, R_M(n, \epsilon)\, d\epsilon = N/S \qquad (1)$$

$$\sum_{n=1}^{N} \int_{\epsilon=1}^{E} n\, \epsilon\, R_M(n, \epsilon)\, d\epsilon = E/S. \qquad (2)$$

In words, these constraints say that the total number of individuals in the community $S \sum_{n=1}^{N} \int_{\epsilon=1}^{E} n\, R_M(n, \epsilon)\, d\epsilon$ is equal to N, and the total community metabolic requirement

$S\sum_{n=1}^{N}\int_{\epsilon=1}^{E} n\,\epsilon\,R_M(n,\epsilon)\,d\epsilon$ is equal to E. $R_M(n,\epsilon)$ is then obtained by maximizing the Shannon entropy $-\sum_n \int R_M \ln R_M \; d\epsilon$ subject to constraints (1) and (2), as well the constraint that $R_M(n,\epsilon)$ sums to 1 (since it is a probability distribution). This maximization procedure gives

$$R_M(n,\epsilon) \propto e^{-\lambda_1 n - \lambda_2 n\epsilon} \qquad (3)$$

where λ_1 and λ_2 are constants (Lagrange multipliers) with values chosen such that constraints (1) and (2) hold (for details, see Harte, 2011). The triplet of values *N*, *E* and *S* are the inputs to METE (note that we will not consider the area-scaling component of the full METE theory; Harte, 2011).

2.2 VEG

VEG is similar to METE in that it uses MaxEnt to infer community properties from a few constraints. Moreover, the VEG constraints are similar to METE's (see below). The major feature that differentiates VEG is that it represents community structure in more detail. In VEG, species are distinguishable, whereas METE only specifies the proportion of species with each abundance n via the ecosystem structure function (Fig. 1).

[FIGURE 1]

In contrast to METE (which uses MaxEnt to infer the ecosystem structure function directly), VEG uses MaxEnt to predict the probability $p(\boldsymbol{n})$ that, when we take a snapshot of the community, we observe the species abundances $\boldsymbol{n} = (n_1, n_2, \ldots)$ (*i.e.* the species labeled 1 has abundance n_1, and so on). In VEG, species' abundances may be zero; the number of species actually present in a snapshot is the number of nonzero elements of $\boldsymbol{n}$. VEG therefore predicts probabilities for the abundance of each species separately; consequently, VEG also predicts the expected number species that are present in the community. Species in VEG are also assigned

distinct functional traits: the individuals of species i are assumed to have a metabolic requirement of ϵ_i, where the species labels are chosen such that $\epsilon_1 \leq \epsilon_2 \leq \epsilon_3$ ..., and so on.

Similar to METE, VEG assumes total abundance and total metabolic requirement constraints

$$\sum_{\boldsymbol{n}} \sum_i n_i \, p(\boldsymbol{n}) = N \qquad (4)$$

$$\sum_{\boldsymbol{n}} \sum_i n_i \, \epsilon_i p(\boldsymbol{n}) = E \qquad (5)$$

Since $p(\boldsymbol{n})$ represents the probability of observing the "snapshots" $\boldsymbol{n}$, the probabilities can be interpreted as sample frequencies representing the proportion of time that the community spends with different abundance compositions $\boldsymbol{n}$. Consequently, constraints (4) and (5) have a clear ecological interpretation in VEG as fixing the time-averaged total abundance and total metabolic requirement of the community to have the values N and E respectively; the latter can be interpreted as an expression of the long-term steady-state ecological balance between resource use (left-hand side of Eq. (5)) and supply (E, right-hand side of Eq. (5)). In contrast, the METE constraints (Eqs. (1) and (2)), which are statements about "information", do not have a similarly straightforward ecological interpretation.

Again similarly to METE, $p(\boldsymbol{n})$ is obtained by maximizing the Shannon entropy $-\sum_{\boldsymbol{n}} p(\boldsymbol{n}) \ln p(\boldsymbol{n})$ subject to constraints (4), (5), and the constraint that $p(\boldsymbol{n})$ sums to 1. This maximization procedure gives

$$p(\boldsymbol{n}) \propto e^{-\sum_i (\mu_1 + \mu_2 \epsilon_i) n_i} \qquad (6)$$

where μ_1 and μ_2 are the Lagrange multipliers corresponding to constraints (4) and (5) respectively. Note that in Eq. (6), $p(\boldsymbol{n})$ depends on the spectrum of metabolic requirements present in the community $\epsilon_1 \leq \epsilon_2 \leq \cdots$. Thus the inputs of VEG are N, E and the spectrum of

values ϵ_i. In contrast to METE, the number of species (S) present in the community is an output of VEG, rather than an input.

3 COMPARING METE AND VEG

3.1 THE VEG ECOSYSTEM STRUCTURE FUNCTION

In this section we give an intuitive derivation of the ecosystem structure function implied by VEG, which will be denoted $R_V(n, \epsilon)$ (a more rigorous mathematical derivation is given in Appendix B). This will allow us to directly compare the predictions of METE and VEG.

When we sample a species at random from a community, all species present have the same probability of being selected. However, the metabolic requirement ϵ of the selected species is more likely to take some values than others due to two effects: (1) *Trait availability.* Among the species currently inhabiting the community's broader geographic region, some values of ϵ are more likely to occur than others due to intrinsic biophysical constraints on the traits determining ϵ, and the region's evolutionary history; (2) *Environmental filtering* (Shipley *et al.*, 2006). From the distribution of possible metabolic rates, some values of ϵ are more likely to be actually present in the community due to additional bias imposed by local environmental constraints (such as Eqs. (4) and (5)).

VEG represents a special case in which there is no trait variation within species: all individuals in species i have the same metabolic requirement ϵ_i (thus a VEG "species" is more appropriately interpreted as a functional type rather than a taxonomic unit; Bertram and Dewar, 2013). Thus, the first effect above (trait availability) is represented by the fact that the metabolic spectrum $\epsilon_1 \leq \epsilon_2 \leq \cdots$ may be more densely packed at some values of ϵ than at others. To represent this effect mathematically, we introduce the "density of species" distribution $\rho(\epsilon)$; $\rho(\epsilon)d\epsilon$ counts the number of metabolic requirement values ("species") contained in the interval

$(\epsilon, \epsilon + d\epsilon)$. For comparison with METE, in which ϵ is a continuous variable, we assume that the metabolic requirement spectrum is sufficiently dense that we can approximate $\rho(\epsilon)$ as a continuous function of ϵ. Intuitively, the shape of $\rho(\epsilon)$ represents the relative probabilities that a species selected at random out of all possible species that could be present in the community has a metabolic requirement within a given interval (Fig. 2).

[FIGURE 2]

Once a species has been sampled out of all possible species and its metabolic requirement has been found to be ϵ, the probability that it has abundance n, denoted $p(n|\epsilon)$, can then be straightforwardly calculated in VEG from Eq. (6) (from Appendix B, $p(n|\epsilon) \propto e^{-(\mu_1+\mu_2\epsilon)n}$). VEG also explicitly accounts for the second effect above (environmental filtering), through the Lagrange multipliers μ_1 and μ_2 that reflect the environmental constraints of Eqs. (4) and (5).

To construct $R_V(n, \epsilon)$, which only refers to species that are actually present, we restrict our attention to $n \geq 1$. Thus, the joint probability of sampling a species with abundance n from the community, and an individual from such a species with metabolic requirement ϵ, is proportional to $\rho(\epsilon)p(n|\epsilon)$, where $n \geq 1$. This gives

$$R_V(n, \epsilon) \propto \rho(\epsilon)p(n|\epsilon) \qquad (7)$$

The above argument leading to Eq. (7) for R_V (and the more rigorous argument given in Appendix B) is quite general. It can be applied to obtain the ecosystem structure function for any model in which we know the density of species $\rho(\epsilon)$ (which need not be restricted to a species-specific trait spectrum as in VEG), and which predicts abundance probabilities conditional on the trait values $p(n|\epsilon)$ (whether those probabilities are predicted using MaxEnt or by other means).

3.2 SPECIES ABUNDANCE DISTRIBUTIONS

A large number of ecological models have reproduced realistic SADs, including METE (Harte et al., 2008) and VEG (Bertram and Dewar, 2015). SAD comparisons consequently only have weak power to discriminate the predictions of different ecological theories (they are "weak tests"; McGill 2003, McGill et al. 2007). In particular, the SAD predictions of METE and VEG will not tell us much about their differences. It is interesting to demonstrate this "weak test" property of SADs explicitly in terms of the METE and VEG ecosystem structure functions.

As noted in section 2.1, the SAD is obtained by integrating the ecosystem structure function over ϵ (the SAD is proportional to $R(n) = \int R(n,\epsilon)d\epsilon$). We therefore expect that the SAD will be to some extent insensitive to the exact manner in which $R(n,\epsilon)$ depends on ϵ.

In the case of METE, the predicted SAD is almost entirely independent of the value of E/S in the metabolic constraint Eq. (2) for many of the most heavily studied SAD datasets (i.e. $R_M(n)$ is independent of λ_2; Harte et al. 2008). This behavior represents the limiting case of large *E*, corresponding to resource-rich communities. Thus, in many cases of interest, METE produces SADs that are insensitive to the value of E/S (note, however, that the existence of the metabolic constraint Eq. (2) is necessary to get a Fisher log-series form for $R_M(n) = \int R_M(n,\epsilon)d\epsilon \propto \frac{e^{-\lambda_1 n}}{n}$).

VEG allows us investigate the "weak test" property in greater depth because we can independently change the form of $\rho(\epsilon)$ and check if this appreciably changes the VEG SAD. Suppose for illustrative purposes that the metabolic requirement spectrum has a power law form (Dewar and Porté, 2008)

$$\rho(\epsilon) \propto \epsilon^{\alpha} \qquad (8)$$

where α is a free parameter. Our motivation for Eq. (8) is to have a simple one-parameter function in which we can control the relative density of species at low versus high ϵ ($\alpha = 0$ corresponds to a uniformly spaced spectrum; Fig. 2). Using Eq. (8), it can be shown that

$$R_V(n) = \int R_V(n,\epsilon)d\epsilon \propto \frac{e^{-\mu_1 n}}{n^{\alpha+1}} \quad (9)$$

for all but the lowest abundance species (see Appendix C). Thus, although the exact quantitative shape of $R_V(n)$ does depend on α (both explicitly in Eq. (9) and implicitly via the fact that $\mu_1$and μ_2 depend on α), $R_V(n)$ will qualitatively have the familiar "hollow curve" shape (McGill et al. 2007) regardless of the particular choice of α. In particular, $\alpha = 0$ gives the Fisher log-series (similar to METE). Thus, since $\rho(\epsilon)$ represents the spectrum of functional traits, we can conclude that the shape of the VEG SAD is only marginally sensitive to the metabolic trait values of the species present.

However, recall that VEG predicts the total number of species/functional types S in the community (Sec. 2.2). This predicted S is more sensitive to the assumed metabolic trait values than the SAD shape, and could differ from the observed value of S for given observed values of N and E. By contrast, METE uses the empirically observed value of S to construct the METE SAD.

3.3 METABOLIC-RANK DISTRIBUTIONS

In this section we compare the metabolic dependence of the two structure functions $R_V(n,\epsilon)$ and $R_M(n,\epsilon)$. We do this in two ways: via the marginal distribution for individual metabolic rates $R(\epsilon) = \sum_n R(n,\epsilon)$, and via the individual-level energy distribution (IED) defined by $\Psi(\epsilon) = \frac{S}{N}\sum_n nR(n,\epsilon)$ (Harte 2011; Newman *et al.*, 2014). $R(\epsilon)$ is the probability that a species sampled at random from the community has metabolic rate ϵ, while $\Psi(\epsilon)$ is the probability that an individual sampled at random from the community has metabolic requirement

ϵ. In contrast to the SAD (Section 3.1), $R(\epsilon)$ and $\Psi(\epsilon)$ are both sensitive to the shape of $\rho(\epsilon)$ in VEG. We can thus ask, what shape does $\rho(\epsilon)$ need to be to match metabolic data, and how does this $\rho(\epsilon)$ compare to the predictions of METE?

Following Harte et al. (2017), we calculate and plot $\Psi(\epsilon)$ cumulatively such that log metabolic rate appears on the vertical axis, and the horizontal axis is the proportion of the population with metabolic rate greater than or equal to a given ϵ (*i.e.* the rank of the corresponding individual). We assumed a power law spectral density as in Eq. (8), taking α as a free parameter to be fitted, and then minimized the least-squares difference between measured log metabolic rates and the predictions of VEG. We repeated this procedure for the three datasets considered in Harte et al. (2017): Barro Colorado Island trees (Hubbell et al. 2005), Hawaiian island arthropods (Gruner, 2007), and Rocky Mountain subalpine meadow plants (Newman et al. 2014).

In all three datasets we found values of α that give superior $\Psi(\epsilon)$ fits to METE (bottom three panels of Fig. 3; note the logarithmic horizontal axis). This is no great victory given that we have introduced a free parameter α that is not available to METE, but it confirms that the power law form for $\rho(\epsilon)$ gives plausible metabolic predictions. METE and VEG both track the middle and higher ranks closely, but at lower ranks the VEG metabolic rates are too low whereas the METE predictions are too high. The corresponding marginal metabolic distributions $R_M(\epsilon)$ and $R_V(\epsilon)$ (upper panels in Fig. 3) confirm that METE assigns higher probabilities to the highest values of ϵ ($R_M(\epsilon)$ has a longer tail).

[FIGURE 3]

4 DISCUSSION

A key insight of the above analysis is that the ecosystem structure $R(n,\epsilon)$ is sensitive to the shape of the density of species distribution represented mathematically by $\rho(\epsilon)$. In the case of VEG, Eq. (7) implies $R_V(\epsilon) \propto \rho(\epsilon)\sum_{n\geq 1} p(n|\epsilon) = \rho(\epsilon)[1-p(0|\epsilon)]$. This expression clearly shows the two effects introduced at the start of Sec. 3.1: $R(\epsilon)$ is the density of species $\rho(\epsilon)$ multiplied by the probability $1-p(0|\epsilon)$ that a species with metabolic rate ϵ is actually present in the community.

Whereas VEG requires us to specify the form of $\rho(\epsilon)$, METE infers $R_M(n,\epsilon)$ using only MaxEnt and the constraint equations (1) and (2). In this sense METE is a null model for the contribution of functional traits to community patterns, treating functional traits as "random noise" within the community constraints imposed by S, N, and E. However, METE only infers the trait distribution as would be observed in already-assembled communities. METE refers only to species that are already present in the community, and does not give an expression for $p(0|\epsilon)$; it is therefore not possible to compute the density of species $\rho(\epsilon)$ implicitly inferred by METE. Nonetheless, observed ecological communities generally have a large proportion of individuals with low metabolic requirement. This implies $p(0|\epsilon) \approx 0$ and thus $R(\epsilon) \approx \rho(\epsilon)$ for low ϵ (see the convergence of $R(\epsilon)$ and $\rho(\epsilon)$ in VEG in the upper panels of Fig. 3), giving us a glimpse of the $\rho(\epsilon)$ predictions of METE.

VEG explicitly separates the trait values that are possible from the trait values that are actually observed post-assembly. Since $\rho(\epsilon)$ is an input, VEG represents an explicit model for the contribution of functional traits to CSDs. This begs the question of what then determines $\rho(\epsilon)$ as the appropriate choice in VEG. There are at least two answers:

(i) On short timescales, $\rho(\epsilon)$ may simply express the mix of potential species that are available to the community at any given time, as in biodiversity manipulation experiments where a given restricted set of species is thrown together and left to self-organize. This short-term $\rho(\epsilon)$ could be highly contingent on the community's recent history, and could have a strong effect on metabolic patterns following disturbance.

(ii) On longer timescales, $\rho(\epsilon)$ may express the totality of conceivable species that might be available to the community. In this case $\rho(\epsilon)$ would depend on how we define species in the first place. With reference to Eq. (7), the choice $\alpha = 0$ corresponds to *defining* "species" by their metabolic requirement, *i.e.* discretize ϵ–space into equal intervals of width $\Delta\epsilon$ and define species i to be the set of individuals whose metabolic requirement e lies between $(i-1)\Delta e$ and $i\Delta e$. Alternatively, "species" could be defined via biomass (in which case the value of α in Eq. (8) may reflect metabolic scaling as in Dewar & Porté 2008); or via other individual traits (t) on which metabolic requirement depends, $\epsilon(t)$.

In either case, VEG opens up ways to understanding the link between functional traits and CSDs that are simply not available to METE.

One of METE's great strengths is that it only requires three parameters S, N and E for all of its predictions. How might the above insights be used to improve the predictions of METE without damaging this exceptional parsimony? The answer may lie in the inclusion of a prior distribution for ϵ representing a contribution from the density of species $\rho(\epsilon)$, which plays a role in ecology analogous to the "density of states" in physics describing the distribution in energy-space of available quantum-mechanical particle states. The upper panels of Fig. 3 suggest that

352 this trait distribution should give less weight to the higher values of ϵ, and also to the lower

353 values of ϵ in the Barro Colorado and Hawaiian communities.

REFERENCES

Banavar, J. R., Maritan, A., & Volkov, I. (2010). Applications of the principle of maximum entropy: from physics to ecology. Journal of Physics: Condensed Matter, 22(6), 063101.

Bertram, J, and Dewar, R. C. (2015). Combining mechanism and drift in community ecology: a novel statistical mechanics approach, *Theoretical ecology*, 8: 419-35.

Bertram, J, and Dewar, R. C. (2013). Statistical patterns in tropical tree cover explained by the different water demand of individual trees and grasses, *Ecology*, 94: 2138-44.

Díaz, S., Purvis, A., Cornelissen, J. H., Mace, G. M., Donoghue, M. J., Ewers, R. M., Jordano, P. & Pearse, W. D. (2013). Functional traits, the phylogeny of function, and ecosystem service vulnerability. *Ecology and evolution*, 3(9), 2958-2975.

Dewar, R. C., and Porté A. (2008). Statistical mechanics unifies different ecological patterns, *Journal of theoretical biology*, 251: 389-403.

Dewar, R. C. (2009). Maximum entropy production as an inference algorithm that translates physical assumptions into macroscopic predictions: Don't shoot the messenger. Entropy, 11(4), 931-944.

Favretti, M. (2017). Remarks on the maximum entropy principle with application to the maximum entropy theory of ecology. Entropy, 20(1), 11.

Gruner, D.S. (2007) Geological age, ecosystem development, and local resource constraints on arthropod community structure in the Hawaiian Islands. *Biological Journal of the Linnean Society*, **90**, 551-570.

Harte, J., (2011) Maximum entropy and ecology: a theory of abundance, distribution, and energetics. Oxford University Press

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

Harte, J, T Zillio, E Conlisk, and AB Smith. (2008). Maximum entropy and the state-variable approach to macroecology, *Ecology*, 89: 2700-11.

Harte, J., Smith, A.B. and Storch, D., 2009. Biodiversity scales from plots to biomes with a universal species–area curve. Ecology letters, 12(8), pp.789-797.

Harte, J., Newman, E. A., & Rominger, A. J. (2017). Metabolic partitioning across individuals in ecological communities. Global Ecology and Biogeography, 26(9), 993-997.

He, F. (2010). Maximum entropy, logistic regression, and species abundance. Oikos, 119(4), 578-582.

Hubbell, S., Condit, R., & Foster, R. (2005). Forest census plot on Barro Colorado Island. Retrieved from http://ctfs.arnarb.Harvard.edu/webatlas/datasets/bci

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

McGill, B. J., Enquist, B. J., Weiher, E., & Westoby, M. (2006). Rebuilding community ecology from functional traits. Trends in ecology & evolution, 21(4), 178-185.

McGill, Brian J, Rampal S Etienne, John S Gray, David Alonso, Marti J Anderson, Habtamu Kassa Benecha, Maria Dornelas, Brian J Enquist, Jessica L Green, and Fangliang He. (2007). Species abundance distributions: moving beyond single prediction theories to integration within an ecological framework, *Ecology letters*, 10: 995-1015.

McGill, B. J., & Nekola, J. C. (2010). Mechanisms in macroecology: AWOL or purloined letter? Towards a pragmatic view of mechanism. Oikos, 119(4), 591-603

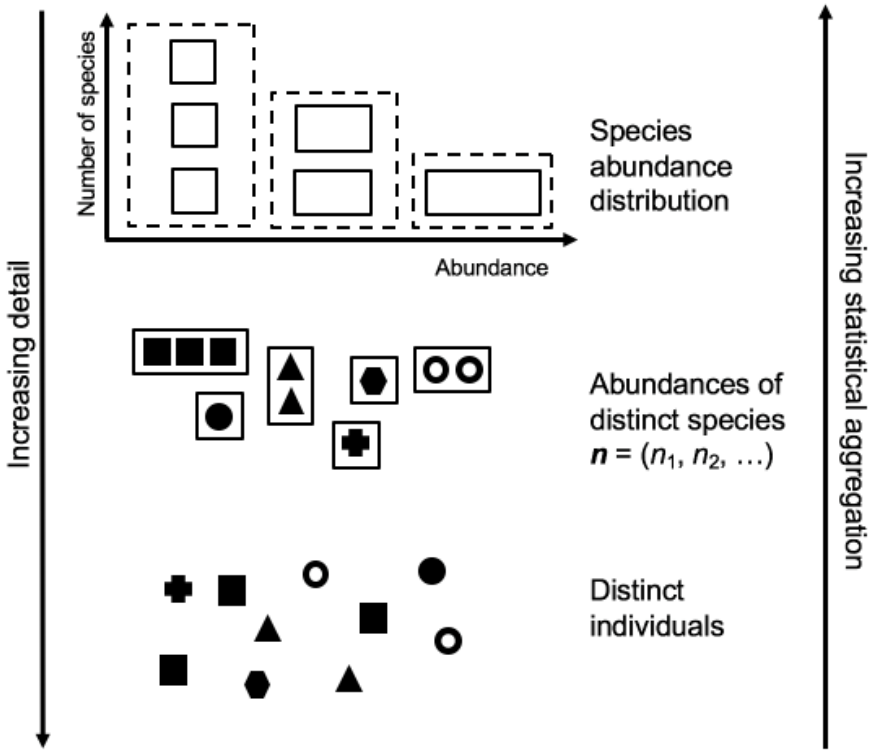
Muller-Landau, H.C., Condit, R.S., Harms, K.E., Marks, C.O., Thomas, S.C., Bunyavejchewin, S., Chuyong, G., Co, L., Davies, S., Foster, R. & Gunatilleke, S. (2006). Comparing

399 tropical forest tree size distributions with the predictions of metabolic ecology and
400 equilibrium models. *Ecology Letters*, **9**, 589-602.
401 Newman, E.A., Harte, M.E., Lowell, N., Wilber, M. & Harte, J. (2014) Empirical tests of within-
402 and across-species energetics in a diverse plant community. *Ecology*, **95**, 2815-2825.
403 Newman, E.A., Wilber, M.Q., Kopper, K.E., Moritz, M.A., Falk, D.A., McKenzie, D. and Harte,
404 J., 2018. Disturbance macroecology: integrating disturbance ecology and macroecology
405 in different-age post-fire stands of a closed-cone pine forest as a case study. bioRxiv,
406 p.309419.
407 O'Dwyer, J.P., Rominger, A. & Xiao, X. (2017). Who Constrains the Constraints? Reinterpreting
408 maximum entropy in ecology: a null hypothesis constrained by ecological
409 mechanism. *Ecology letters*, **20**, 832-841.
410 Pueyo, S., He, F., & Zillio, T. (2007). The maximum entropy formalism and the idiosyncratic
411 theory of biodiversity. *Ecology Letters*, 10(11), 1017-1028.
412 Rominger, A.J. and Merow, C. (2016) meteR: an R package for testing the maximum entropy
413 theory of ecology. *Methods in Ecology and Evolution*, **8**, 241-247.
414 Shipley, B., Vile, D., & Garnier, É. (2006). From plant traits to plant communities: a statistical
415 mechanistic approach to biodiversity. *Science*, 314(5800), 812-814.
416 Shipley, B. (2010). Community assembly, natural selection and maximum entropy models.
417 Oikos, 119(4), 604-609.
418 Supp, S.R., Xiao, X., Ernest, S.K.M. and White, E.P., 2012. An experimental test of the response
419 of macroecological patterns to altered species interactions. Ecology, 93(12), pp.2505-
420 2511.

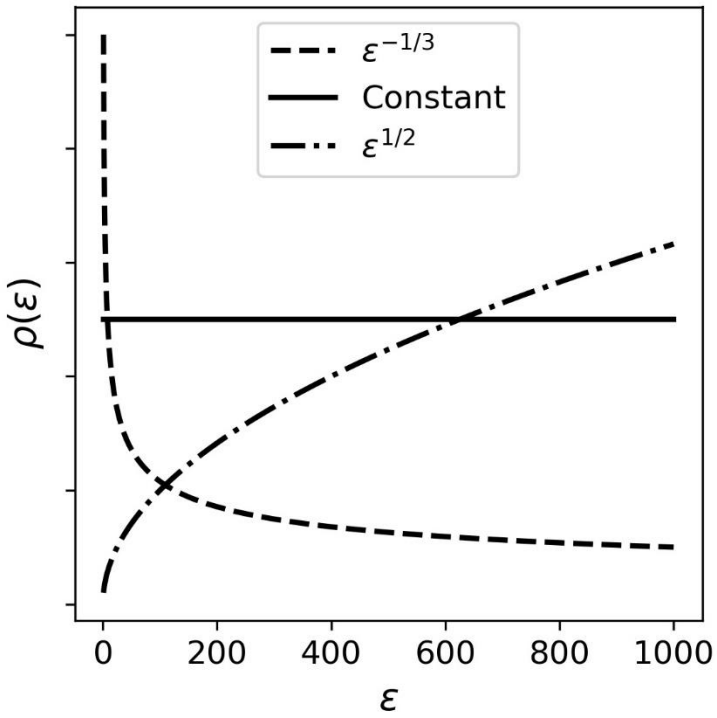
421 Supp, S.R. and Ernest, S.M., 2014. Species-level and community-level responses to disturbance:
422 a cross-community analysis. Ecology, 95(7), pp.1717-1723.
423 Tokeshi, M. (1993). Species abundance patterns and community structure. *Advances in*
424 *ecological research* **24**, 111-186.
425 Ulrich, W., Ollik, M., & Ugland, K. I. (2010). A meta-analysis of species–abundance
426 distributions. Oikos, 119(7), 1149-1155.
427 Vellend, M. (2010). Conceptual synthesis in community ecology. *The Quarterly review of*
428 *biology*, 85(2), 183-206
429 Xiao, X., McGlinn, D. & White, E. (2015) A strong test of the Maximum Entropy Theory of
430 Ecology. *American Naturalist*, **185**, E70-E80.
431
432

FIGURES

Figure 1. Three levels of detail commonly used for describing ecological communities. At the greatest level of detail (bottom), the distinct identities of individuals and their spatial locations are known. At the intermediate levels of detail found in many well-mixed models such as Lotka-Volterra models (middle), the abundance of each distinct species is known. At the lowest level of detail and highest level of statistical aggregation (top), species identities are lost, and the SAD provides the only description of species diversity.

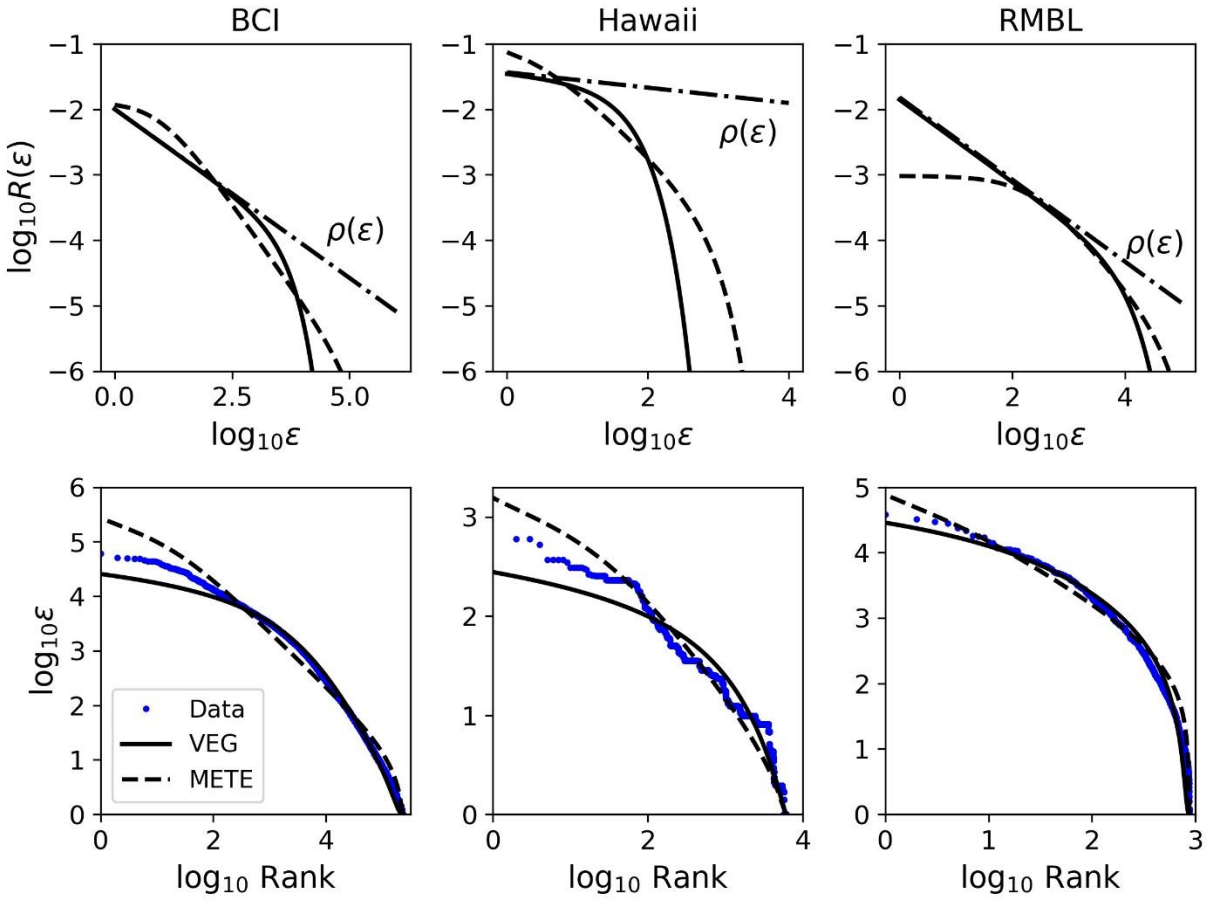


445 **Figure 2.** The function $\rho(\epsilon)$ counts the local density of metabolic rates in the assumed spectrum
446 of possible rates $\epsilon_1 \leq \epsilon_2 \leq \cdots$ in VEG. It represents the relatively probability that a randomly
447 selected VEG "species" has metabolic rate ϵ when sampled from all possible "species".
448



449
450

Figure 3. Comparison of METE and VEG rank-metabolism relationships in multiple communities: Barro Colorado Island trees (BCI; $\alpha = -0.52, \rho(1) = 10^4$), Hawaiian island arthropods (Hawaii; $\alpha = -0.12, \rho(1) = 10^2$), and Rocky Mountain subalpine meadow plants (RMBL; $\alpha = -0.63, \rho(1) = 10^2$) communities. Each upper/lower panel pair shows the $R(\epsilon)$ and rank-ϵ curves for the same METE and VEG ecosystem structure functions.



APPENDIX A: NOTE ON THE DEFINITION OF THE ECOSYSTEM STRUCTURE FUNCTION

Our definition of the ecosystem structure function differs slightly from that given in Harte et al. (2008), which reads "[The ecosystem structure function] is the probability that if a species is picked at random [...], then it has abundance *n* and if an individual is picked at random *from that species*, then its metabolic requirement is in the range $\epsilon, \epsilon + d\epsilon$" (our italics). The Harte et al. (2008) definition suggests that METE keeps track of species identity. In fact, $R_M(n, \epsilon)$ depends only on n and ϵ, and not species identity. Thus, $R_M(\epsilon|n) = R_M(n, \epsilon)/\sum_n R_M(n, \epsilon)$ is the probability of picking an individual with metabolic requirement ϵ conditional on it coming from a species with abundance n. There is no way within METE to distinguish between different species with the same abundance n, and therefore there is no reason to specify which species the individual is selected from in the definition of the ecosystem structure function (Favretti, 2017).

APPENDIX B: DERIVING THE ECOSYSTEM STRUCTURE FUNCTION FROM DISTINGUISHABLE SPECIES

In section 2.1, the METE structure function R_M was inferred directly from community-level constraints. Here we derive an analogous VEG structure function $R_V(n, \epsilon)$.

We start by defining the probability distribution $P(n, \epsilon, i, \boldsymbol{n})$ as follows: $P(n, \epsilon, i, \boldsymbol{n})d\epsilon$ is the joint probability that the community has species abundances $\boldsymbol{n}$, that a species picked at random from the community has species label i, that this chosen species has abundance $n_i = n$, and that an individual from this chosen species has metabolic requirement in the interval $(\epsilon, \epsilon + d\epsilon)$. $R_V(n, \epsilon)$ is then obtained by marginalizing with respect to i and $\boldsymbol{n}$, i.e. $R_V(n, \epsilon) = \sum_i \sum_{\boldsymbol{n}} P(n, \epsilon, i, \boldsymbol{n})$ where $n \geq 1$.

To marginalize P, we first write it as a product of conditional distributions

$$P(n, \epsilon, i, \boldsymbol{n}) = P(\epsilon|n, i, \boldsymbol{n})P(n|i, \boldsymbol{n})P(i|\boldsymbol{n})P(\boldsymbol{n}). \quad \text{(B1)}$$

Here $P(\boldsymbol{n}) = p(\boldsymbol{n})$ is the probability that the community has abundance vector $\boldsymbol{n}$, $P(i|\boldsymbol{n}) = \left(1 - \delta^0_{n_i}\right)/S(\boldsymbol{n})$ is the probability that a species picked from a community with abundances $\boldsymbol{n}$ has species label i (i.e. 0 if species i is absent, $1/S(\boldsymbol{n})$ if present), $P(n|i, \boldsymbol{n}) = \delta^n_{n_i}$ is the probabilty that species i has abundance n given the species abundances are $\boldsymbol{n}$, and $P(\epsilon|n, i, \boldsymbol{n})d\epsilon$ is the probability that an individual picked from species i has metabolic requirement in the interval $(\epsilon, \epsilon + d\epsilon)$ given species i has abundance n and the community abundances are $\boldsymbol{n}$. We thus obtain

$$R_V(n, \epsilon) = \sum_i \sum_{\boldsymbol{n}} P(\epsilon|n, i, \boldsymbol{n})\delta^n_{n_i}\left(1 - \delta^0_{n_i}\right)p(\boldsymbol{n})/S(\boldsymbol{n})$$

$$= \sum_i \sum_{\boldsymbol{n}} P(\epsilon|n, i, \boldsymbol{n})\delta^n_{n_i}p(\boldsymbol{n})/S(\boldsymbol{n}) \quad \text{(B2)}$$

where we have used the fact that $\delta^n_{n_i}\left(1 - \delta^0_{n_i}\right) = \delta^n_{n_i}$ for $n \geq 1$.

494 In the case of VEG, all individuals in species i have the same metabolic requirement ϵ_i,
495 and so $P(\epsilon|n,i,\boldsymbol{n}) = \delta(\epsilon - \epsilon_i)$ where δ is the Dirac delta function (i.e. the probability that a
496 randomly selected individual from species i has metabolic requirement ϵ is 1 in the immediate
497 vicinity of ϵ_i, and is 0 otherwise). Thus, from (B2) we have

503 $$R_V(n,\epsilon) = \sum_i \delta(\epsilon - \epsilon_i) \sum_{\boldsymbol{n}} \frac{\delta_{n_i}^n p(\boldsymbol{n})}{S(\boldsymbol{n})} \qquad \text{(B3)}$$

498 An ecologically important special case of (B3) occurs when the variation in the number of
499 species present from one snapshot to the next is small relative to the expected number of species,
500 such that $S(\boldsymbol{n})$ is approximately constant with value given by $S = \sum_{\boldsymbol{n}} S(\boldsymbol{n})\, p(\boldsymbol{n})$. This occurs,
501 for example, if most of the species present have large expected abundances. Eq. (B3) then
502 simplifies to

504 $$R_V(n,\epsilon) = \frac{1}{S} \sum_{\{i|n_i \geq 0\}} \delta(\epsilon - \epsilon_i)\, P(n_i = n) \qquad \text{(B4)}$$

505 where $P(n_i = n) = \sum_{\boldsymbol{n}} \delta_{n_i}^n p(\boldsymbol{n})$ is the probability that species i has abundance n.

506 In VEG, we have from Eq. (6) (Bertram and Dewar, 2015)

507 $$P(n_i = n) = (1 - e^{-(\mu_1 + \mu_2 \epsilon_i)}) e^{-(\mu_1 + \mu_2 \epsilon_i) n}.$$

508 To make it explicit that this probability depends on the metabolic requirement of species i, we
509 use the notation $p(n|\epsilon_i) \equiv P(n_i = n)$ (that is, $P(n_i = n)$ in VEG is the probability that a species
510 has abundance n given that its metabolic requirement is ϵ_i).

511 Since the ecosystem structure function is a probability density in the continuous variable
512 ϵ, we can introduce a spectral density $\rho(\epsilon)d\epsilon$ that counts the number of metabolic requirement
513 levels ϵ_i in each interval $(\epsilon, \epsilon + d\epsilon)$. From Eq. (B4), $R_V(n,\epsilon)$ can then be written in the form

514 $$R_V(n,\epsilon) = \frac{\rho(\epsilon) p(n|\epsilon)}{S}$$

APPENDIX C: THE VEG SPECIES ABUNDANCE DISTRIBUTION

Assuming a power–law density of states $\rho(\epsilon) \propto \epsilon^{\alpha}$, we have from Eq. (7)

$$R(n) = \int_0^{\infty} R(n,\varepsilon)d\varepsilon \propto \int_0^{\infty} \varepsilon^{\alpha}(1 - e^{-(\mu_1+\mu_2\epsilon)})e^{-(\mu_1+\mu_2\epsilon)n}d\varepsilon$$

By making the substitution $x = \mu_2\varepsilon n$, the integral is found to be:

$$\int_0^{\infty} \varepsilon^{\alpha}(1 - e^{-(\mu_1+\mu_2\epsilon)})e^{-(\mu_1+\mu_2\epsilon)n}d\varepsilon = \Gamma(\alpha+1)\frac{e^{-\mu_1 n}}{(\mu_2 n)^{\alpha+1}}\left[1 - e^{-\mu_1}\left(\frac{1}{1+1/n}\right)^{\alpha+1}\right]$$

where

$$\Gamma(\alpha+1) = \int_0^{\infty} x^{\alpha}e^{-x}dx$$

is the gamma function.

For large n we have $1/(1+1/n) \approx 1$ so that:

$$R(n) \propto \frac{e^{-\mu_1 n}}{(\mu_2 n)^{\alpha+1}}$$