

Maximum Entropy Theory in Ecology

April 18, 2016

Abstract

Abstract Text

Macroecology is the study of large-scale patterns of biodiversity, aggregated across many species and individuals. These phenomena encompass the distributions of organisms across space, temporal fluctuations on multiple timescales (Chisholm & O'Dwyer, 2014; O'Dwyer *et al.*, 2015), as well as multiple ways to quantify and measure biodiversity (Morlon *et al.*, 2009). Macroecological patterns take surprisingly consistent, simple forms across many different taxonomic groups and distinct habitats—for example, the increase in species richness with area is often modeled as a power law function, while the distribution of rare and abundant species can be fitted using one of a handful of common distributions [insrt classic refs]. This apparent universality, alongside the sense that it is driven by combination of high diversity and large numbers (O'Dwyer & Chisholm, 2014), has led many ecologists to draw from statistical physics to understand and predict patterns of biodiversity (Harte *et al.*, 1999; O'Dwyer & Green, 2010). Yet despite promising hints (McGill, 2010), we still currently lack the quantitative, overarching principles to explain or predict macroecological patterns.

The Maximum Entropy Theory of Ecology (Harte *et al.*, 2008, 2009; Harte, 2011; Harte & Newman, 2014), known as METE, has sought to fill this gap, providing a framework where macroecological patterns are predicted based on a handful of summary statistics for any given data set. The goal of this approach is always to identify a probability distribution, which can then be used to make predictions or compare with existing data. The principle of maximum entropy tells us that we can find a unique probability distribution that maximizes entropy, while constraining expectation values using the data we choose to feed to the algorithm. In other words, METE is a balance between maximizing entropy, and reducing entropy according to the data we provide as ecologists. The question is then simple: what data should we give to the algorithm? We *could* constrain the maximum entropy distribution with whatever data we know about a given ecological community, whether that is as specific as the number and spatial location of individuals of your favorite species, or as obscure as the skewness of the distribution of rare and abundant species. Whatever we think we know, the maximum entropy principle will then fill in the gaps adding the least possible additional information.

In contrast to this kind of generality and flexibility, METE is very specific in terms of its data requirements: the theory prescribes a set of constraints that we should measure, based on intuitive measures such as the total number of individuals in a system, the total diversity, and the total energy flux. While the specific values of these constraints will differ across ecological communities that are more or less diverse, productive, or populous, the proposal is that the appropriate constraints are identical for all systems. Predictions are then made by applying the maximum entropy principle to these constraints. This prescription has been successful, and suggests that we don't need to measure much at all to have an excellent idea of what macroecological patterns look like. But it also presents a problem. Any alternative set of choices would result in different predictions. So how and when do we know we have the 'right' set of constraints?

In this paper we introduce a principle for choosing a minimal set of constraints. Our proposed solution builds on recent work comparing the predictions of simple, neutral models of species diversity, abundance, and energy flux with maximum entropy predictions for the same distributions (Xiao *et al.*, 2015). Given a specific mechanistic model, we always need some data to fix its parameters—in the cases where the model generates distributions in an exponential family, these data are known as its sufficient statistics. Our proposal is that this minimal data is precisely what we should use to constrain the maximum entropy algorithm, thus generating a second prediction for the same data. If the model outperforms the corresponding Max Ent distribution, then our choice of mechanism as modelers was successful. If the model is outperformed by max ent, we may as well not have bothered modeling the mechanism at all—just using the sufficient statistics and maximizing entropy is a better approach. The result in effect isolates the ‘information’ coming from the choice of mechanism. We first demonstrate the approach in ecology by adapting the existing METE theory. We also show that it can be applied in other contexts.

1 Methods

1.1 Case Study: Neutral Ecology

Neutral models have provided the basic null models in fields stretching from population genetics (Kimura, 1968) and ecology (Hubbell, 2001; O’Dwyer & Chisholm, 2013) to cultural evolution and the social sciences (Shennan & Wilkinson, 2001). In common is the key assumption that selective differences are irrelevant for predicting large-scale patterns. Applying this assumption, neutral biodiversity theory has made successful predictions for species-level spatial patterns (O’Dwyer & Green, 2010), and has been tested in communities ranging from tropical forests (Volkov *et al.*, 2003), to coral reefs (Volkov *et al.*, 2007) to microbes (Woodcock *et al.*, 2007; Jeraldo *et al.*, 2012).

The simplest flavor of neutral model is a theory of birth, death and speciation. Neutrality implies that all species have identical demographic rates, and by assumption this model is to be applied at large enough scales that immigration from outside the system is relatively rare. (At this large scale, the community is known as a metacommunity, in contrast to a ‘local’ community where immigration is a key process.) We can explore the consequences of neutrality using multiple metrics, including both the dynamics of this system (Chisholm & O’Dwyer, 2014; O’Dwyer *et al.*, 2015), and also snapshots at a single moment in time. One of the most well-explored such snapshots is the distribution of species abundances in steady state, which we think of as a probability $P(n)$ that a species picked at random from the metacommunity has abundance n . In a neutral world, this probability distribution satisfies a difference equation (Hubbell, 2001; Volkov *et al.*, 2003) and the solution takes the form of the well-known log series:

$$P_{\text{NT}}(n) = \frac{1}{n \log(1/\nu)} (1 - \nu)^n \quad (1)$$

where ν is the per capita speciation rate for the community in units of the generation time. The log series itself had been introduced much earlier by Fisher [cite Fisher] as a phenomenological candidate to fit the shape of empirical species abundance distributions, and so the appearance of the same distribution purely arising from drift was a promising result for neutral theory.

How do we determine the best choice of parameter ν for a given data set? In an ideal world, the process of speciation would be both sufficiently well-defined and easy to measure to allow an independent estimate of ν for a given system. We could then compare the form of Eq. (1) with the corresponding observed species abundance distribution. In practice, this is not possible, and ecologists testing neutral theory have interpreted ν as an effective parameter to be fitted using the species abundance data. We use the notation n_i to denote the abundance of species i , and when there are S species in a community, and a total of

$\sum_{i=1}^S n_i$ individuals, the maximum likelihood estimate for ν satisfies:

$$\frac{1 - \nu}{\nu \log(1/\nu)} = \sum_{i=1}^S n_i / S = \bar{n} \quad (2)$$

We can then use Eq. (1), with parametrization determined by Eq. (2), to compute any measure of goodness of fit, or likelihood, or comparison with alternative models, or whatever we wish—all using this point estimate of ν . Eq. (2) is informing us that to perform any of these analyses we require \bar{n} , the mean abundance per species, from the data: this is the sufficient statistic for the distribution Eq. (1).

In parallel to the simple, mechanistic, neutral model above, we can also make a prediction for the species abundance distribution using the principle of maximum entropy. In fact, we can make many such predictions, with the freedom deriving from the choice of constraints on this distribution. These constraints in turn derive from whatever quantities we have measured in a given data set. Let's suppose that we measure the mean abundance per species. Constraining the mean abundance per species \bar{n} , and maximizing the entropy of the species abundance distribution then leads to the geometric series

$$P_{ME}(n) = (e^\lambda - 1) e^{-\lambda n} \quad (3)$$

where the parameter λ is known as a lagrange multiplier and is fixed using the observed data:

$$\lambda = \log \frac{\bar{n}}{\bar{n} + 1} \quad (4)$$

Finally we note that $P_{ME}(n)$ can always be modified via a different choice of prior distribution—in Eq. (3) **we have chosen a uniform prior**. In everything that follows, it may be reasonable to choose a different prior based on additional information we may have before specifying constraints.

1.2 Base Likelihood

In summary, the case study above demonstrates two ways to make a prediction for the species abundance distribution. First, we took a simple, mechanistic model, with assumptions built in about the birth and death of individuals. Second, we made a prediction using the principle of maximum entropy. This second prediction was also neutral in a sense (in that it treats each species abundance as an independent draw from the same distribution) but makes no assumptions about species dynamics. And it was constrained (our choice) using precisely the same data needed to specify the speciation rate of the neutral model.

The neutral $P_{NT}(n)$ and max ent $P_{ME}(n)$ are (in the sense of their data overhead) on exactly the same footing. In fact, the Pitman–Koopman theorem (Pitman, 1936; Koopman, 1936; Darmois, 1945; Jeffreys, 1960) tells us that the neutral prediction is highly constrained. Any distribution with sufficient statistic \bar{n} has to belong to the same exponential family as the corresponding max ent distribution, $P_{ME}(n)$. The loophole and difference between the two distributions lies solely in the prefactor $1/n$ in the log series distribution relative to the geometric series, and we interpret this difference as the information added through specifying the dynamics of the neutral model.

This suggests a new way to test the efficacy of a mechanistic model or theory like the neutral model above:

- (i) if possible, identify the data needed to fully parametrize the distributions predicted by the model (its sufficient statistics)
- (ii) use these data to constrain a maximum entropy distribution for the same quantities
- (iii) compare the likelihood of the mechanistic model to its max ent analogue.

In summary: given the same data, can the mechanistic model outperform max ent? If yes, then specifying the details of the model and calculating its solution has been worthwhile. If not, whatever we have contributed to the construction of the model is only worthwhile in so much as it fixes the constraints to measure using the data—beyond that, our efforts as modelers have been futile.

Let's now consider the generalization of the neutral vs max ent case study above to an arbitrary mechanistic model prediction for a distribution $P_{\text{model}}(n)$ defined over the positive integers, with sufficient statistic $F(n)$, and S independent measurements of n , labeled n_i . This distribution must take the form:

$$P_{\text{model}}(n) = A(\alpha)h(n)e^{-\alpha F(n)} \quad (5)$$

for appropriate normalization $A(\alpha)$, and base measure $h(n)$ to be determined by solving the model. The maximum likelihood estimate of α solves:

$$\frac{\sum_{i=1}^S F(n_i)}{S} = \bar{F} = \frac{d}{d\alpha} \log A(\alpha). \quad (6)$$

The corresponding maximum entropy distribution (i.e. chosen to constrain the mean value of $F(n)$) with uniform base measure takes the form:

$$P_{\text{ME}}(n) = B(\lambda)e^{-\lambda F(n)} \quad (7)$$

for normalization $B(\lambda)$ that implicitly depends on the choice of constraint, F . Using a set of S observations labeled n_i , we can fix the parameter λ :

$$\frac{\sum_{i=1}^S F(n_i)}{S} = \bar{F} = \frac{d}{d\lambda} \log B(\lambda). \quad (8)$$

For example, in the case of Eq. (3), $B(\lambda) = e^\lambda - 1$ and $F(n) = n$, and we recapitulate Eq. (4).

Putting these two results together, we now propose that max ent in ecology should be used as a comparison for the likelihood of the model given this data. We use the point estimate given by solving Eq. (6) for parameter α , and the value of lagrange multiplier λ given by solving Eq. (8). Explicitly writing these parameters in terms of their dependence on observed \bar{F} , we propose the following expression for evaluating the quality of the base measure $h(n)$ implied by the model:

$$\begin{aligned} \log \mathcal{L}_{\text{base}}(\text{model} | \{n_i\}) &= \sum_{i=1}^S \log P_{\text{model}}(n_i | \alpha(\bar{F})) - \sum_{i=1}^S \log P_{\text{ME}}(n_i | \lambda(\bar{F})) \\ &= \sum_{i=1}^S [\log h(n_i) + F(n_i) (\lambda(\bar{F}) - \alpha(\bar{F}))] + S \left[\log \frac{A(\alpha(\bar{F}))}{B(\lambda(\bar{F}))} \right] \\ &= S \left[\overline{\log(h)} + \bar{F} (\lambda(\bar{F}) - \alpha(\bar{F})) + \log \frac{A(\alpha(\bar{F}))}{B(\lambda(\bar{F}))} \right] \end{aligned} \quad (9)$$

Is this in effect testing whether a ‘model-implied’ base measure provides a better explanation of our data than the uniform measure? Of course! If we view this as an iterative process, with increasingly refined models, we may subsequently wish to treat $h(n)$ as our base for a new maximum entropy distribution. However, for any new model one introduces, we may not have any confidence that the ‘model-implied’ $h(n)$ is correct—the model is after all a conjecture. For example, we may know nothing about the true mechanism underlying the system—only the data $\{n_i\}$. In such a case, it is reasonable to view a very basic hurdle that the model can perform better than the maximum entropy distribution with a uniform base. If it fails this test (i.e. if $\log \mathcal{L}_{\text{base}}$ defined by Eq. (9) is negative) then refining the base measure further is moot.

2 Results

2.1 Abundance and Biomass

We now demonstrate our approach using a model and probability distribution that depends on multiple independent variables. This provides a more non-trivial example of the base likelihood introduced in our methods section, but where we still have data with which to test the approach empirically. On the maximum entropy side there is no limitation—once given the constraints we can construct the corresponding max ent distribution. To provide a non-trivial mechanistic model, we turn to size-structured neutral theory (SSNT) (O'Dwyer *et al.*, 2009)[Xiao *et al* 2016]. This is an extension of the neutral ecological model introduced above, but with the addition of a new variable representing the size, mass, or energy flux of an individual. Speciation is defined in the same way as in the vanilla neutral theory, but now birth and death rates $b(m)$ and $d(m)$ can depend on the mass of an organism, m . Also, there is a new process: ontogenetic growth. Each individual grows through time with a rate $g(m)$, which may also depend explicitly on its current mass.

There is now also tremendous freedom in defining this model, because the functions $b(m)$, $d(m)$ and $g(m)$ are arbitrary. For this analysis, we choose the simplest case, where all three are independent of mass, m . Even in this case, the combination of birth, death and growth still introduces variation in individual masses, as well as variation in the average size and total biomass across different species. There are multiple distributions that can be computed in this size-structured theory, and we now explore two. First, we consider the joint distribution that a species chosen at random will have abundance n and total biomass (summed across all n individuals) M . Under the rules of SSNT, this distribution is (see Supplementary Information):

$$P_{\text{SSNT}}(n, M) = \frac{1}{m_0 \log(1/\nu)} \frac{(1 - \nu)^n}{n!} \left(\frac{M}{m_0} \right)^{n-1} e^{-\frac{M}{m_0}}. \quad (10)$$

where n takes values in the positive integers and M is a continuous variable ≥ 0 . (The latter definition is straightforward to generalize to account for a finite initial mass of new individuals). ν is the speciation rate in units of the generation time, while m_0 is a mass scale and is equal to the ratio of rates g/d . Finally, we note that marginalizing over total biomass M returns us to the simpler result for the log series species abundance distribution given in Eq. (1)—if you choose not to measure species biomass, you are back to the regular neutral theory.

The two sufficient statistics of the joint distribution $P_{\text{SSNT}}(n, M)$ given by Eq. (10) are mean biomass per species $\sum_i M_i/S = \bar{M}$ and mean abundance per species, $\sum_i n_i/S = \bar{n}$. More explicitly, the maximum likelihood estimates of parameters ν and m_0 are given by:

$$\begin{aligned} \frac{1 - \nu}{\nu \log(1/\nu)} &= \bar{n} \\ m_0 &= \frac{\bar{M}}{\bar{n}} \end{aligned} \quad (11)$$

We now carry out our strategy of constructing a max ent distribution with uniform base measure to provide a baseline for the performance of $P_{\text{SSNT}}(n, M)$. Constraining \bar{M} and \bar{n} , we arrive at the following max ent distribution for M and n in a size-structured community:

$$P_{\text{SSME}}(n, M) = (e^{\lambda_1} - 1) \lambda_2 e^{-\lambda_1 n} e^{-\lambda_2 M} \quad (12)$$

where the Lagrange multipliers impose the constraints on \bar{M} and \bar{n} and take the values:

$$\begin{aligned} \lambda_1 &= \log \frac{\bar{n}}{\bar{n} + 1} \\ \lambda_2 &= 1/\bar{M} \end{aligned} \quad (13)$$

We now have an explicit, multivariate example of our proposed entropic correction. In Figure 1 we evaluate the performance of the size-structured neutral theory by computing its log likelihood (with parameters set by Eq. (11)), with an entropic correction given by subtracting $\log P_{\text{SSME}}(n, M)$ with Lagrange multipliers fixed using the data. Across XYZ forest plots we see that SSNT almost always outperforms [???] its max ent baseline. In other words, if all you know about a forest plot is its mean abundance per species \bar{n} and mean biomass per species \bar{M} , you can still do better assuming neutral dynamics than you would by choosing a maximum entropy distribution with a uniform base and the same data.

[FIGURE HERE. Test against same data Xiao used?]

2.2 Abundance and Biomass: Take Two

We next consider a fine-grained way to test the size-structured neutral theory. In addition to the measuring each species' abundance and its total biomass, we also measure the mass of each of its individuals. Replacing the joint distribution above for n and M , we can make a neutral prediction for the precise distribution of masses within a species:

$$P_{\text{SSNTI}}(n, m_1, \dots, m_n) = \frac{1}{\log(1/\nu)} \frac{(1-\nu)^n}{n} \frac{1}{m_0^n} \prod_{j=1}^n e^{-m_j/m_0}. \quad (14)$$

We have labeled this distribution 'SSNTI', so that the I stands for individual-level. The sufficient statistics for the parameters ν and m_0 are again given by mean abundance per species and mean total biomass per species:

$$\begin{aligned} \frac{1-\nu}{\nu \log(1/\nu)} &= \bar{n} \\ m_0 &= \frac{\sum_{j=1}^n m_j}{\bar{n}}. \end{aligned} \quad (15)$$

Using these as constraints, we can in parallel construct the corresponding individual-level maximum entropy distribution to use as a baseline for the performance of $P_{\text{SSNTI}}(n, m_1, \dots, m_n)$:

$$P_{\text{SSMEI}}(n, m_1, \dots, m_n) = (e^{\lambda_1} - 1) \lambda_2^n e^{-\lambda_1 n} \prod_{j=1}^n e^{-\lambda_2 m_j} \quad (16)$$

where the Lagrange multipliers impose the constraints on $\sum_{j=1}^n m_j$ and \bar{n} and take the values:

$$\begin{aligned} \lambda_1 &= \log \frac{\bar{n}}{\bar{n} + 1} \\ \lambda_2 &= \frac{\bar{n}}{\sum_{j=1}^n m_j} \end{aligned} \quad (17)$$

In Figure 2 we evaluate the performance of the individual-based size-structured neutral theory by computing its log likelihood (with parameters set by Eq. (11)), with an entropic correction given by subtracting $P_{\text{SSMEI}}(n, m_1, \dots, m_n)$ with Lagrange multipliers fixed using the data. Across XYZ forest plots we see that ? SSNTI is still a good explanation of the data relative to the corresponding maximum entropy distribution, but the difference is not as striking as in Figure 1. What changed? Well, the individual-based SSNTI neutral model has a larger number of independent variables than its aggregated counterpart SSNT, but when conditioned on a fixed total biomass, $\sum_{j=1}^n m_j = M$, $P_{\text{SSNTI}}(n, m_1, \dots, m_n)$ becomes equal to $P_{\text{SSNT}}(n, M)$: if you blur your eyes and only pick up on total biomass, the two neutral predictions are identical, as they should be. The same is not true of the two max ent distributions, labeled SSME and SSMEI. What changed

is that we implicitly told P_{SSMEI} that total biomass M is comprised of a set of indistinguishable individuals of masses $\{m_j\}$ —in effect, there are $n - 1$ degrees of freedom. The entropic correction therefore provides a stricter baseline the more we ask of the neutral model.

[FIGURE TWO HERE. Test against same data as Figure One]

Where does this leave the successful Maximum Entropy Theory of Ecology (Harte, 2011), which prescribes a particular set of constraints, and makes predictions of the same types of distributions as the above? Our approach here has been to let a mechanistic model choose the max ent constraints for us. In fact, we don't know of any flavor of neutral theory whose independent variables and sufficient statistics match the METE degrees of freedom and constraints. Does this mean METE is somehow wrong? Not at all. But it does mean that we have to justify the constraints chosen in METE post hoc, rather than motivating them by a choice of mechanistic model, as we have done here. In due course, we may identify another model of abundance and biomass that matches these constraints, and which we can then compare directly with METE.

3 Discussion

- Recapitulation
- I don't see any case where I can apply exactly John's Max ent. Perhaps that is exactly the point to make.
- What does being better than the appropriate max ent tell you? Maybe not much. Many models will be in the same equivalence class in that they produce e.g. a Poisson rather than geometric distribution, or a log series rather than geometric. Do these equivalence classes become smaller the higher dimensional the distribution, e.g. $P(n, M)$ rather than $P(n)$ alone?
- So I am not sure that being better than max ent is meaningful necessarily. It certainly doesn't tell you about goodness of fit. But on the other hand, being *worse* than max ent does seem important. If your model can't outperform max ent, that's a problem to be explained!
- Bias-variance trade-off in modeling. In some sense we are evaluating the performance of the parameter-free part of a model's predicted distribution. For this to perform well across multiple examples of similar data, it would suggest there is something truly universal in that data. On the other hand, a given $h(n)$ can't perform equally well against all possible distributions. So the statement that a model is a 'good one' has to be with reference to a specific kind of data.
- No penalization here for model complexity (unless it is harder to better than ME the more complex the model is, and I don't know this). Related to the point about multiple independent variables above.
- Models predict dynamics typically. In neutral case, if we knew abundances for all species *of the same age*, this would actually be geometric. It is the aggregation across ages that produces the log series. We don't typically look at this as species age is not well-defined or easy to measure if it were. But it is interesting—**looking more deeply at the dynamics may produce a result which is the same as the max ent, while aggregating over the data produces a result different from max ent (and better). Same with aggregating over individual body-sizes in the abundance/mass models. So is there a key lesson I am missing here? Every kind of aggregation (temporal or individual identity, or...) that the neutral model undergoes, means that the corresponding max ent distribution is different (and worse) Important?**

References

- Albert, R. & Barabási, A.-L. (2002). Statistical mechanics of complex networks. *Reviews of modern physics*, 74, 47.
- Barabási, A.-L. & Albert, R. (1999). Emergence of scaling in random networks. *science*, 286, 509–512.
- Chisholm, R. & O'Dwyer, J. (2014). Species ages in neutral biodiversity models. *Theoretical Population Biology*, 93, 85–94.
- Darmois, G. (1945). Sur les limites de la dispersion de certaines estimations. *Revue de l'Institut International de Statistique*, 9–15.
- Harte, J. (2011). *Maximum entropy and ecology: a theory of abundance, distribution, and energetics*. Oxford University Press.
- Harte, J., Kinzig, A. & Green, J. L. (1999). Self-similarity in the distribution and abundance of species. *Science*, 284, 334–336.
- 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. (2008). Maximum entropy and the state-variable approach to macroecology. *Ecology*, 89, 2700–2711.
- Hubbell, S. P. (2001). *The Unified Neutral Theory of Biodiversity and Biogeography*. Princeton Univ. Press, Princeton.
- Jeffreys, H. (1960). An extension of the pitman–koopman theorem. In: *Mathematical Proceedings of the Cambridge Philosophical Society*, vol. 56. Cambridge Univ Press.
- Jeraldo, P., Siposa, M., Chia, N., Brul, J., Dhillon, A., Konkel, M., Larson, C., Nelson, K., Qua, A., Schook, L., Yang, F., White, B. & Goldenfeld, N. (2012). Quantification of the relative roles of niche and neutral processes in structuring gastrointestinal microbiomes. *Proc Natl Acad Sci*, 109, 692–698.
- Kimura, M. (1968). Evolutionary rate at the molecular level. *Nature*, 217, 624–626.
- Koopman, B. O. (1936). On distributions admitting a sufficient statistic. *Transactions of the American Mathematical Society*, 39, 399–409.
- McGill, B. J. (2010). Towards a unification of unified theories of biodiversity. *Ecology Letters*, 13, 627–642.
- Morlon, H., White, E. P., Etienne, R. S., Green, J. L., Ostling, A., Alonso, D., Enquist, B. J., He, F., Hurlbert, A., Magurran, A. E. *et al.* (2009). Taking species abundance distributions beyond individuals. *Ecology Letters*, 12, 488–501.
- O'Dwyer, J. & Chisholm, R. (2013). Neutral Theory and Beyond. In: *Encyclopedia of Biodiversity*. Elsevier.
- O'Dwyer, J. & Chisholm, R. (2014). A mean field model for competition: From neutral ecology to the red queen. *Ecology Letters*, 17, 961–969.
- O'Dwyer, J. & Green, J. (2010). Field theory for biogeography: a spatially-explicit model for predicting patterns of biodiversity. *Ecology Letters*, 13, 87–95.
- O'Dwyer, J., Lake, J., Ostling, A., Savage, V. & Green, J. (2009). An integrative framework for stochastic, size-structured community assembly. *Proc Natl Acad Sci*, 106, 6170–6175.

- O'Dwyer, J., Sharpton, T. & Kembel, S. (2015). Backbones of Evolutionary History Test Biodiversity Theory in Microbial Communities. *Proc Natl Acad Sci*, 112, 8356–8361.
- Pitman, E. J. G. (1936). Sufficient statistics and intrinsic accuracy. In: *Mathematical Proceedings of the Cambridge Philosophical Society*, vol. 32. Cambridge Univ Press.
- Shennan, S. & Wilkinson, J. (2001). Ceramic Style Change and Neutral Evolution : A Case Study from Neolithic Europe. *American Antiquity*, 66, 577–593.
- Volkov, I., Banavar, J. R., Hubbell, S. P. & Maritan, A. (2003). Neutral theory and relative species abundance in ecology. *Nature*, 424, 1035–1037.
- Volkov, I., Banavar, J. R., Hubbell, S. P. & Maritan, A. (2007). Patterns of relative species abundance in rainforests and coral reefs. *Nature*, 450, 45–49.
- Woodcock, S., Van Der Gast, C. J., Bell, T., Lunn, M., Curtis, T. P., Head, I. M. & Sloan, W. T. (2007). Neutral assembly of bacterial communities. *FEMS Microbiology Ecology*, 62, 171–180.
- Xiao, X., O'Dwyer, J. P. & White, E. P. (2015). Comparing process-based and constraint-based approaches for modeling macroecological patterns. *Ecology*.