Research

Ecography 35: 001–010, 2012
doi: 10.1111/j.1600-0587.2011.07103.x

Subject Editor: Miguel Araujo. Accepted 7 November 2011

How do species interactions affect species distribution models?

William Godsoe and Luke J. Harmon

W. Godsoe (godsoe@nimbios.org), Biological Sciences, Univ. of Canterbury, Private Bag 4800, Christchurch 8140, New Zealand. – L. J. Harmon and WG: Dept of Biological Sciences, Univ. of Idaho, PO Box 443051 Moscow, ID 83844-3051, USA.

One of the most promising recent advances in biogeography has been the increased interest and understanding of species distribution models – estimates of the probability that a species is present given environmental data. Unfortunately, such analyses ignore many aspects of ecology, and so are difficult to interpret. In particular, we know that species interactions have a profound influence on distributions, but it is not usually possible to incorporate this knowledge into species distribution models. What is needed is a rigorous understanding of how unmeasured biotic interactions affect the inferences generated by species distribution models. To fill this gap, we develop a general mathematical approach that uses probability theory to determine how unmeasured biotic interactions affect inferences from species distribution models. Using this approach, we reanalyze one of the most important classes of mechanistic models of competition: models of consumer resource dynamics. We determine how measurements of one aspect of the environment – a single environmental variable – can be used to estimate the probability that an environment is suitable with species distribution models. We show that species distribution models, which ignore numerous facets of consumer resource dynamics such as the presence of a competitor or the dynamics of depletable resources, can furnish useful predictions for the probability that an environment is suitable in some circumstances. These results provide a rigorous link between complex mechanistic models of species interactions and species distribution models. In so doing they demonstrate that unmeasured biotic interactions can have strong and counterintuitive consequences on species distribution models.

One of the most promising recent advances in ecological methods is the development of species distribution models (SDMs) (Peterson et al. 1999, Elith et al. 2006). These methods combine easily obtained data with sophisticated statistical methods to estimate the probability that an organism will be found in a given set of environments or locations. Despite their promise, the interpretation of SDMs is fraught with conceptual ambiguity (Soberon and Peterson 2005, Kearney 2006). A rich array of research has demonstrated that species distributions represent a complex amalgam of factors, including history, dispersal, environmental conditions and interactions with other organisms (Brown et al. 1996). However, SDMs, as currently implemented, model only a tiny subset of these factors. As a consequence, most empirical descriptions of species distributions omit important ecological details. What is needed is a rigorous conceptual understanding of how unmeasured facets of ecology affect the inferences of distribution models (Holt 2009, Soberon and Nakamura 2009, Godsoe 2010a).

Perhaps the most substantial omission from current SDMs are biotic interactions among species (Pulliam 2000, Soberon and Peterson 2005, Araujo and Guisan 2006). It has long been known that the distribution of one species can depend on interactions with other species through competition, predation, mutualism, and commensalism. But each of these processes can be difficult to measure and parameterize at the landscape scale. Given this problem, we need an intuitive understanding of how biotic interactions shape our ability to make statistical inferences with distribution data.

In response to this problem, several authors have proposed verbal models of how biotic interactions shape SDMs. Some authors have argued that most biotic interactions occur over small scales (10s to 100s of meters), while abiotic factors vary over a much broader scales (Pearson and Dawson 2003, Soberon 2007, Soberon and Nakamura 2009, Gotelli et al. 2010). As such, SDMs at large scales represent the effect of the abiotic environment on a species.

However, several lines of evidence suggest that biotic interactions might matter at larger scales. For example, ecological theory predicts that biotic interactions can produce abrupt range limits between species (Case et al. 2005) Concordantly, evolutionary biologists have documented numerous examples of sister species with parapatric distributions, where one closely related species excludes another from large areas of otherwise suitable habitat (Jordan 1905, Coyne and Orr 2004).

Alternatively, SDMs may be interpreted as little more than a model of the environments in which a species is found (the realized niche Kearney 2006, Jiménez-Valverde et al. 2008). As such, they provide little reliable information about the environments that are suitable to a species – that is, where it could be found. One can then use other information

(from experimental studies, for example) to understand the set of environments that are potentially suitable to an organism (the fundamental niche) and, from this, make conclusions about the role of biotic interactions. The hypothesis that SDMs are less reliable than inferences derived from an experimental or mechanistic understanding of ecology is, at best, incomplete. Small-scale mechanistic studies can fail to reproduce processes that are important at larger scales. For example, small-scale experimental work in the 1960s indicated that carbon was the limiting nutrient responsible for blooms of blue green algae (Lange 1967). Subsequent analyses demonstrated that, at large scales, carbon limitation did not produce blooms (Schindler 1971, Peters 1991). As a consequence, mechanistic models sometimes produce less reliable inferences than correlative models (Buckley et al. 2010).

A promising compromise position is that SDMs provide information on whether an environment is suitable to a species (Booth et al. 1988, Peterson et al. 1999, Phillips et al. 2006). Recent theoretical work formalizes this idea by showing analytically that SDMs can estimate the probability that an environment is a part of the niche (Soberon and Nakamura 2009, Godsoe 2010a). Unfortunately, this approach relies explicitly on a model of niches as sets of environments that are suitable to an organism (Hutchinson 1957). These proofs do not yet consider how the mechanisms by which organisms interact with one another shape the statistical inferences we derive from SDMs (Tilman 1977, 1982, Chase and Leibold 2003). We resolve this problem by extending the approach developed by Godsoe (2010a) to reanalyze mechanistic models of competition. To do this, we re-translate models of consumer resource dynamics (hereafter CR models) (MacArthur 1972, Chase and Leibold 2003) into the conditional probability that a species is present given complete knowledge of the environment. We then use this function to understand our ability to make predictions using limited knowledge of the environment by deriving the marginal probability that a species is present. Using this approach, we can analyze the consequences of omitting biotic interactions in SDMs. We illustrate this approach using models of competition for resources, although our results could be extended to other types of interactions, such as mutualisms or predation (Chase and Leibold 2003, Holland and DeAngelis 2010).

We focus our analyses on two distinct questions: 1) can we model the probability that an environment will be suitable for a given species using incomplete information about resources? 2) How do interactions between competitors influence the probability of presence estimated using information on the abiotic environment? Our results demonstrate that unmeasured biotic interactions have counterintuive effects, even in some of the best-known models of competition. As such, explicit mathematical analyses are needed to understand how biotic interactions shape SDMs.

The model

We focus on competition mediated by CR dynamics as such models have been analyzed for several decades and are well understood (MacArthur and Levins 1964, MacArthur 1972, Tilman 1977, 1980, 1982, Abrams 1988, Chase and Leibold 2003). CR models are mechanistic, with interactions between species being governed by the ability of each species to deplete resources. As a result of these mechanistic details, CR models are easier to interpret than more phenomenological approaches, such as Lotka–Volterra models of competition (Tilman 1980, Chase and Leibold 2003).

Here, we model the abundance of species i using the differential equation:

$$\frac{dN_i}{dt} = N_i(f_{i1}a_{i1}R_1 + f_{i2}a_{i2}R_2 - d_i) \quad (1)$$

where N_i denotes the abundance of species i. We will focus on a two species model such that $i \in 1, 2$. Each species increases in abundance as it consumes resources R_1 and R_2. A resource in such a model is a facet of the environment that increases the population growth rate of a species and that is consumed by a population of organisms (Tilman 1980). One commonly cited class of resources is nutrients, such as nitrogen, phosphorous and silicon. In aquatic systems, such resources are often measured as the amount of a resource in micro-moles per volume of water in liters μM (Tilman 1977). However, this definition of resources is sufficiently broad to apply to environmental variables as diverse as water or sunlight in other circumstances. As such, there is no common unit for all possible resources. The rate at which N_i increases depends on the per-capita feeding rate of species i on resource j, f_{ij} and the ability of species i to convert this feeding into population growth, a_{ij}. In turn, population growth is off-set by a per-capita loss (death) term, d_i.

The abundance of each resource R_j depends on additional differential equations where $j \in 1, 2$:

$$\frac{dR_j}{dt} = c(S_j - R_j) - \sum_{i=1}^{2} f_{ij}N_iR_j \quad (2)$$

Feeding by species i decreases the abundance of resource j (the term $f_{ij}N_iR_j$). However, each resource is replenished through the resource turnover term $c(S_j - R_j)$ at the rate c from a supply pool in which the abundance of the resource is S_j. As a result, in the absence of consumers, the equilibrium concentration of R_j approaches S_j. A somewhat contrived example of a system undergoing such dynamics would be a lake into which a solution with a concentration S_j of nutrients is added at rate c. The greater the difference between the concentration of resources in the lake and the concentration of resources in the solution, the faster the concentration of resources in the lake increases (hence the $S_j - R_j$ term). For convenience, we scale resource supply rates between a minimum value of 0 and a maximum of 1. Following Tilman (1980), we assume a common resource turnover rate but distinct supply pool for each resource.

Assuming that the amount of resources consumed by each species does not change substantially with a change in the availability of each resource, this model has five possible outcomes at equilibrium (Tilman 1980): 1) the supply of resources is insufficient to allow either species to persist, 2) species 1 and 2 coexist stably, 3) one species may exclude the other but the victor depends on initial conditions, 4) species 1 will exclude species 2 and exist alone regardless of initial conditions, 5) species 2 will exclude species 1, regardless of initial conditions.

In any one location, the model behavior at equilibrium depends on the supply of both resources (Fig. 1). Neither species can be present (outcome 1) if the supply of resources is insufficient. Following existing terminology, we define the Zero Net Growth Isocline for species i (hereafter $ZNGI_i$) as the line that demarcates environments with sufficient resources for species i to survive from environments in which this species is unable to survive. We will focus subsequent analyses on the species with the steepest *ZNGI*, hereafter species 1. Other model outcomes depend on two lines, hereafter L_1 and L_2 which describe how each species consumes resources. These lines pass through the point of intersection of the *ZNGI*'s of the two species ($S_{1,\text{intersection}}$, $S_{2,\text{intersection}}$) and have a slopes of $f_{i2}S_{2,\text{intersection}}/f_{i1}S_{1,\text{intersection}}$. L_1 and L_2 run parallel to the 'impact vector', which describes the impact of consumers on a given concentration of resources (I_1 and I_2 on Fig. 1). If L_1 has a shallower slope than L_2, then species 1 consumes proportionately more of R_1. When this is true, the two species coexist stably (outcome 2) so long as the supply of resources is in the region between L_1 and L_2. Conversely, when L_2 is shallower than L_1, the victor in environments between L_1 and L_2 depends on initial conditions (outcome 3). In environments above $ZNGI_1$ but below L_1 and L_2, only species 1 is present (outcome 4), and in environments above $ZNGI_2$ and above L_1 and L_2, only species 2 is present (outcome 5). Tilman (1980) provides a formal derivation of these ideas through a linear stability analysis.

If we assume that the supply for each resource is equally likely in our landscape (uniformly distributed over an interval), then the proportion of environments in a particular region of a consumer resource plots is equal to the proportion of environments in our study region that have those

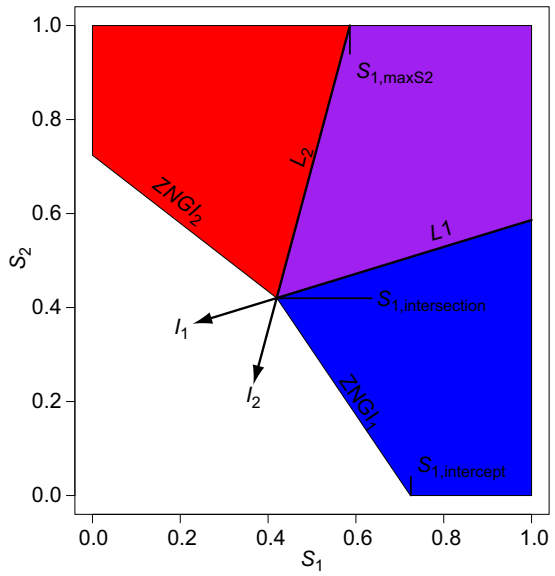

Figure 1. *ZNGI* plot including points and lines used to calculate the probability of presence with competition. In this figure, blue represents environments suitable to species 1, red represents environments suitable to species 2, purple represents environments suitable to both species and white represents environments unsuitable to both. This plot is an illustration of stable coexistence since I_1 is shallower than I_2 such that species 1 has a greater effect on resource 1.

environmental conditions. So for example, if 50% of the environments in our graph have supply points that could support coexistence, then 50% of the environments in our region can potentially support coexistence as well. As we shall see, this facilitates a simple graphical interpretation of probabilistic inferences in consumer resource models.

Next, we must model how local and regional community dynamics interact. To facilitate our graphical approach, we assume that each species is present if and only if the species would be present at equilibrium (Tilman 1982, Abrams 1988, Tilman and Pacala 1993, Chase and Leibold 2003, Holt et al. 2005 but see Abrams and Wilson 2004, Ryabov and Blasius 2011). This is akin to assuming that dynamics within a given location equilibrates much more quickly than dynamics between locations (i.e. local ecological interactions occur much more quickly than dispersal between patches). We also assume that researchers only sample presences and absences from environments to which both species can disperse (Soberon 2007). With this set of assumptions, the probability that a species is present is identical to the probability that an environment is suitable to that species. When two taxa can disperse to dramatically different environments, it is possible for models to conflate differences in suitable environments with differences in available environments (Elith and Graham 2009, Godsoe 2010b).

We analyze our ability to model the probability that an environment is suitable using an environmental variable in two scenarios: 1) what is the probability that an environment is suitable in the absence of a competitor? 2) What is the probability that an environment is suitable with a competitor?

To compare our results to empirical studies of species distributions and previous representations of CR models, we then measure our ability to fit SDMs. We sample 200 environments from which we measure a single environmental variable and the presence/absence of species 1, then fit a SDM using a generalized linear model (GLM) with a binomial link function. GLM serves a familiar and convenient method for comparing our analytic results to SDMs, particularly as some of the most sophisticated SDM algorithms, such as Boosted Regression Trees, represent extensions of GLM (Friedman et al. 2000, Elith et al. 2006, 2008). We estimate the accuracy of our SDMs by calculating Area Under the receiver operating Curve (AUC). This statistic is a non-parametric estimate of a models' ability to distinguish presences from absences. It ranges from zero to one with a score of 1 representing a nearly perfect ability to distinguish between presences and absences. It should be noted that this procedure could also be applied to *ZNGI*s that are non linear, say if resources are essential (Tilman 1980).

1) The probability that an environment is suitable in the absence of a competitor

In the absence of a competitor our focal species can persist in environments above its zero net growth isocline. This mathematical observation can be used to generate the conditional probability that a species is present using only information on the resource supply points. If set of supply points are above the $ZNGI_1$, species 1 will be present, such that the conditional probability of presence is 1. Species 1 is absent otherwise, making the probability of presence equal

to zero). Mathematically, the probability that an observation X represents a presence ($X = 1$), given that the supply of resource j in an environment is s_j is the following:

$$P(X = 1 \mid S_1 = s_1, S_2 = s_2) = \begin{cases} 1 & \text{if } (S_1, S_2) \geq ZNGI_1 \\ 0 & \text{otherwise} \end{cases} \quad (3)$$

To obtain the equation for $ZNGI_1$, set $dN_1/dt = 0$ in Eq. (1), then find the solution for the non-trivial equilibrium by solving $f_{i1}a_{i1}R_1 + f_{i2}a_{i2}R_2 - d_i = 0$. Re-arranging this solution, substituting it into Eq. (3), and substituting in the supply pool of each resource gives us:

$$P(X = 1 \mid S_1 = s_1, S_2 = s_2) = \begin{cases} 1 & \text{if } S_2 \geq \dfrac{d_1}{f_{12}a_{12}} - \dfrac{f_{11}a_{11}}{f_{12}a_{12}} \\ 0 & \text{otherwise} \end{cases} \quad (4)$$

With this function, we may determine our ability to make predictions using only incomplete information, say if we could only measure S_1. Mathematically, this is a matter of using the probability of presence conditioned on measurements of S_1, S_2 to derive the marginal probability that an environment is suitable, conditioned on a measurement of S_1 (Ross 1997). This is given by:

$$P(X = 1 \mid S_1 = s_1) = \frac{\int_0^1 P(X = 1 \mid S_1 = s_1, S_2 = s_2) f_{S_1,S_2}(s_1, s_2) dS_2}{f_{s1}(s_1)}$$

where $f_{S_1}(s_1)$ is the probability for observing a particular value of S_1, and $f_{S_1,S_2}(s_1, s_2)$ is the joint probability of observing a given set of values for S_1 and S_2. In the absence of any other information, we assume that S_1 and S_2 are independent and uniformly distributed over (0,1), $f_{S_1}(s_1) = f_{S_1,S_2}(s_1, s_2) = 1$. This premise is equivalent to assuming we are equally likely to observe any combination of resource supply points. It also simplifies the mathematical problem of calculating the probability of presence to one of integrating out the effect of S_2. Appendix 1 provides the formal derivation as for this probability as:

$$P(X = 1 \mid S_1 = s_1) = \begin{cases} 1 & \text{if } 1 < h(S_1) \\ h(S_1) & \text{if } 0 < h(S_1) < 1 \\ 0 & \text{if } h(S_1) < 0 \end{cases} \quad (5)$$

where

$$h(S_1) = 1 + \frac{f_{11}a_{11}}{f_{12}a_{12}} S_1 - \frac{d_1}{f_{12}a_{12}}$$

After noting that $0 \leq P(X = 1 \mid S_1 = s_1) \leq 1$. The mechanics of computing the probability of presence for other distributions of the supply points of resources – say normal or exponential – are well developed (Ross 1997) and similar to those we have presented here. So for example, if S_1 and S_2 are non-independent, or S_2 has some distribution other than uniform over (0,1), the joint density function $f_{S_1,S_2}(s_1, s_2)$ changes. Similarly, if S_1 is non-uniformly distributed over the interval (0,1), both $f_{S_1,S_2}(s_1, s_2)$ and $f_{S_1}(s_1)$ will change. In Supplementary material Appendix 3, we simulate SDMs for the competition model that we present below using variables that are correlated, exponentially distributed, or uniformly distributed over an interval other than (0,1).

Using independent, uniformly distributed supply points facilitates a simple graphical interpretation of our model. Figure 2 presents a consumer resource diagram. To derive the probability of presence given measurements of S_1, start by considering a given supply point of S_1, say environments where $S_1 = 0$. The probability an environment will be suitable given $S_1 = 0$ is the proportion of environments where $S_1 = 0$ that are above $ZNGI_1$. Graphically, this is a matter of drawing a vertical line at $S_1 = 0$ and measuring the portion of this line where S_2 is sufficiently high to allow our species to persist. To develop the marginal probability, we must extend this calculation through all possible supply points of S_1. Figure 2B illustrates the marginal probability of presence (black). In this figure, environments with high S_1 values invariably support presences. Environments with lower resource supply points sometimes support species 1.

A correlative SDM using only information on S_1 fit to the simulated data set in Fig. 2 closely matches the marginal probability of presence and produces strong predictions, with an AUC score of 0.93 (blue dotted line Fig. 2B).

2) The probability that an environment is suitable given the presence of a competitor

When a competitor (species 2) is present, it will be able to exclude species 1 from some environments. To calculate the probability that an environment is suitable when coexistence is stable, we start with the probability of presence given the supply of both resources:

$$P(X = 1 \mid S_1 = s_1, S_2 = s_2) = \begin{cases} 1 & \text{if } (S_1, S_2) \geq ZNGI_1 \text{ and} \\ & \text{Species}_2 \text{ does not exclude Species}_1 \\ 0 & \text{otherwise} \end{cases}$$

Graphically, we can obtain the marginal probability of presence by considering the proportion of the plot in Fig. 3A where species 1 is present at each possible value of S_1. Appendix 2 provides a formal derivation of the marginal probability of presence in this model as:

$$P(X = 1 \mid S_1 = s_1) = \begin{cases} 1 & \text{if } S_{1,\text{intercept}} < S_1 \\ h(S_1) & \text{if } S_{1,\text{max}S_2} < S_1 < S_{1,\text{intercept}} \\ g(S_1) - h(S_1) & \text{if } S_{1,\text{intersection}} < S_1 < S_{1,\text{max}S_2} \\ 0 & \text{if } S_1 < S_{1,\text{intersection}} \end{cases} \quad (6)$$

where:

$$g(S_1) = \frac{f_{22}(f_{21}a_{21}d_1 - f_{11}a_{11}d_2)}{f_{21}(f_{12}a_{12}d_2 - f_{22}a_{22}d_1)} S_1 + \frac{(f_{21} - f_{22})(a_{11}d_2f_{11} - a_{21}d_1f_{21})}{f_{21}(a_{11}a_{22}f_{11}f_{22} - a_{12}a_{21}f_{12}f_{21})}$$

This function describes the probability of presence based on a few points on the S_1 axis (Fig. 1). $S_{1,\text{intersection}}$ represents the intersection of the $ZNGI$s for species 1 and species 2.

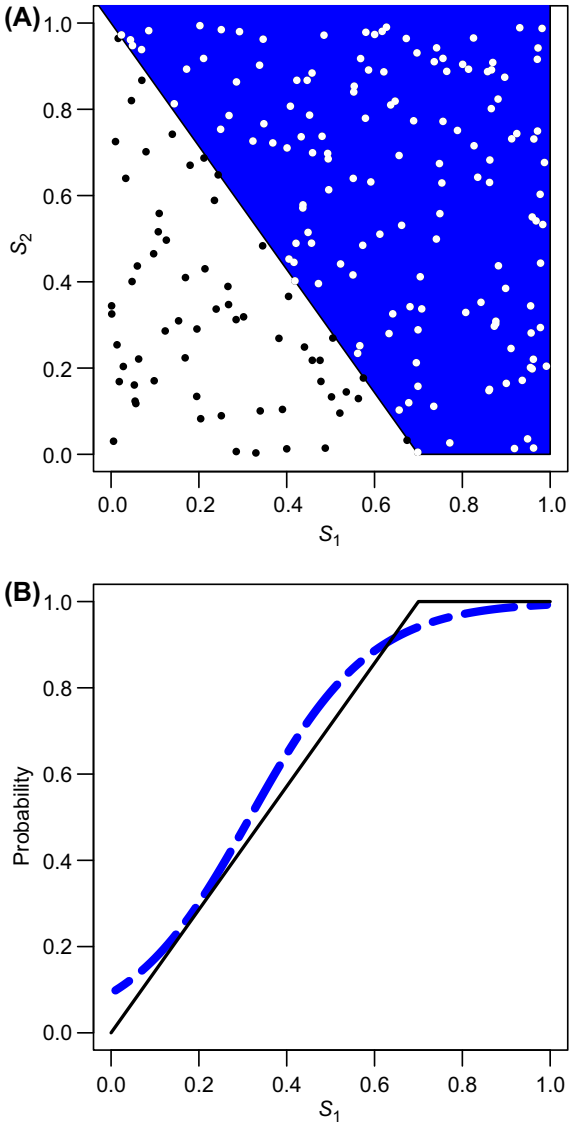

Figure 2. A graphical interpretation of the probability that a species will be present in an environment given the supply point of resource 1. Panel (A) highlights areas above the $ZNGI_1$ (blue). Unsuitable environments (below the *ZNGI*) are white. Dots represent a sample of two hundred environments including absences (black) and presences (white). Panel (B) indicates the marginal probability of presence conditioned on supply points for species two (black line). The blue dotted line represents an estimate of the probability of presences from a GLM using the observed presence/absence data displayed in panel (A). This plot uses the following parameter values: $a_{11} = 0.1$, $a_{12} = 0.021$, $f_{11} = 0.0047$, $f_{12} = 0.016329$, $d_1 = 0.00034$. AUC score 0.900.

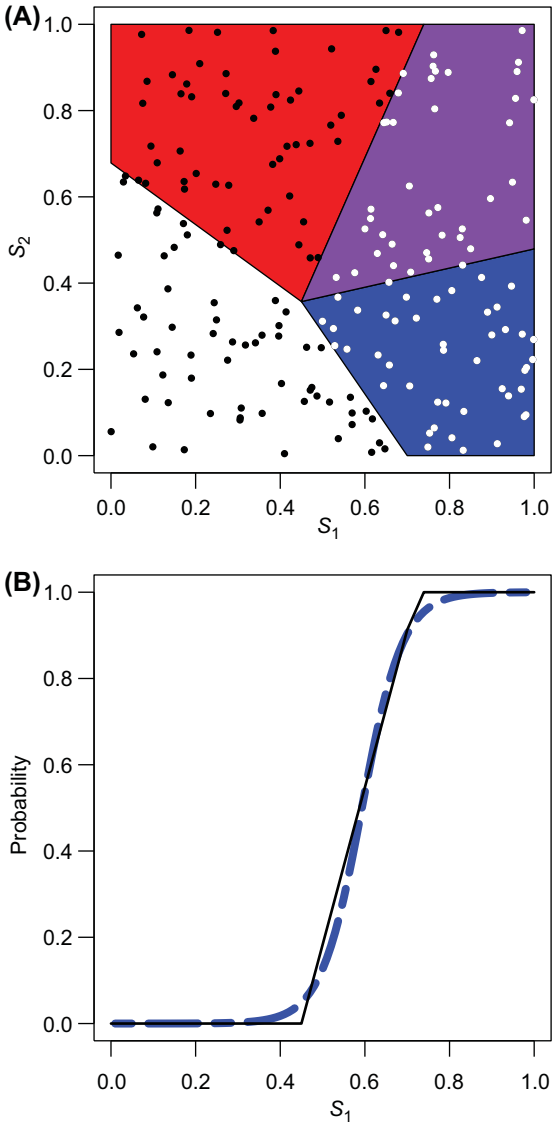

Figure 3. A graphical interpretation for the probability a species will be present given the supply point for resource 1 and assuming that our focal species interacts with a competitor. In panel (A), species 1 will be present in environments where it coexists with its competitor (purple), or environments where it is competitively dominant (blue). It is absent from environments where it loses to its competitor (red), or environments with insufficient resources for either species (white). Dots represent a sample of two hundred environments including absences (black) and presences (white). Panel (B) shows the marginal probability of presence conditioned on S_1 (black line), along with the probability of presence estimated from a GLM fit using the presence/absence observations portrayed in panel (A). This plot uses the following parameter values: $a_{11} = 0.2$, $a_{12} = 0.5$, $a_{21} = 0.4$, $a_{22} = 0.2$, $f_{11} = 0.25$, $f_{12} = 0.07$, $f_{21} = 0.025$, $f_{22} = 0.07$, $d_1 = 0.035$, $d_2 = 0.0095$.

$S_{1,\max S_2}$ represents the point where a line collinear with the depletion vector for species 1 reaches the maximum possible supply for S_2. $S_{1,\text{intercept}}$ is the point where the *ZNGI* for species 1 intersects the S_1 axis. $h(S_1)$ is the proportion of environments above the *ZNGI* for species 1 at each value of S_1, and $g(S_1)$ is the proportion of environments below a line along the depletion vector for species 1. For simplicity, we present results assuming that $S_{1,\max S_2} < S_{1,\text{intersection}}$. We present a slightly modified version of this function in Appendix 2 assuming $S_{1,\text{intersection}} < S_{1,\max S_2}$. A GLM predicting the probability of presence given s_1 values from the simulation results presented Fig. 3 produces exceptionally strong predictions, with an AUC score of 0.99.

Figure 4 illustrates the consequences unstable coexistence. Panels A, B and C simulate the dynamics of 100 locations for 100 000 yr assuming identical feeding rates, conversion efficiency, death rates, resource parameters and the initial

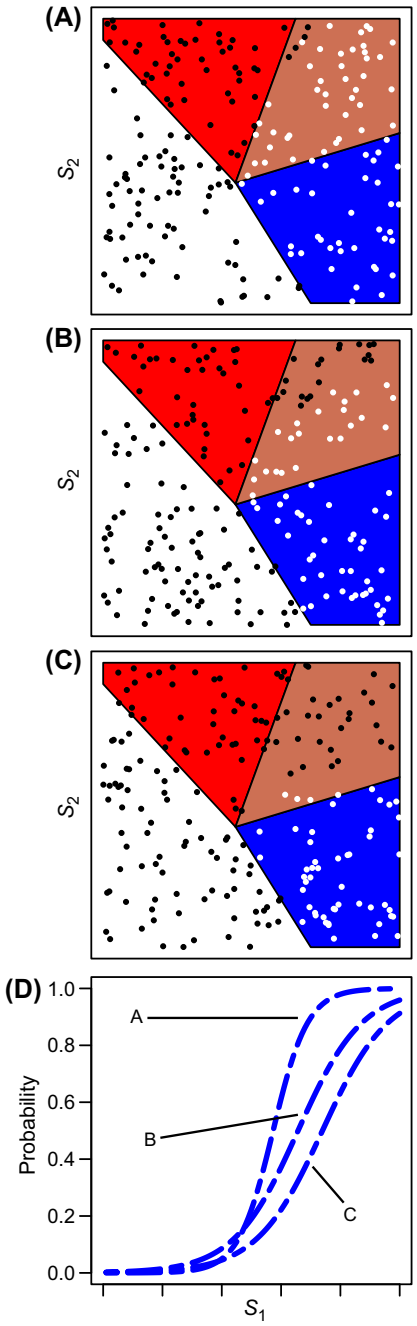

Figure 4. An example of how unstable coexistence can alter the outcome of SDMs. Panels (A–C) use identical parameter values with one exception, the initial abundance of species 1 which in (A) is equal to that of species 2, in (B) is one half the abundance of species 2 and in (C) is one tenth the abundance of species 2. In large sections of the plot the outcome of the model is similar to that in Fig. 3, however in the region between L_1 and L_2 (brown) the outcome of competition changes with initial conditions. As the initial abundance of species 1 decreases it becomes less able to outcompete species 2 in environments with high values of S_2. As a result species 1 is nearly always present in (A) and nearly always absent in (C). Decreasing the initial abundance of species 1 thus weakens our ability to use S_1 to predict the presence of species 1. (D) Plots SDMs fit to each panel which produce slightly different predictions and have different AUC scores A = 0.985, B = 0.924, C = 0.914. This plot uses the following parameter values: $a_{11} = 0.1$, $a_{12} = 0.02$, $a_{21} = 0.15$, $a_{22} = 0.4$, $f_{11} = 0.1$, $f_{12} = 0.3$, $f_{21} = 0.3$, $f_{22} = 0.1$, $d_1 = 0.007$, $d_2 = 0.037$, initial conditions: $N_2 = 0.01$, $R_1 = S_1$, $R_2 = S_2$.

abundance of species 2. The only difference between each panel is the initial abundance of species 1. In panel A the initial abundance of species 1 is 0.01, equal to the abundance of species 2. In this scenario, species 1 excludes species 2 from most of the environments between L_1 and L_2. In panel B, species 1 is initially half as abundant as species 2. As a result, species 2 excludes species 1 for some environments between the two lines. In panel C, species 1 is initially one tenth as abundant as species 2. As a result, species 2 frequently excludes species 1. In Fig. 4D, we illustrate GLM models fit to each of these three scenarios. The AUC scores associated with these models varied from 0.914 to 0.985. Figure 4A produced the highest score.

Discussion

Our research goal is to understand how unmeasured biotic interactions affect SDMs. We have both a general response to this problem and specific observations from models of CR dynamics. SDMs use information on the environment to predict where a species is present. For many SDM algorithms, this process can be formalized in the following way – the methods estimate the probability of presence conditioned measurements of the environment. Unmeasured biotic interactions affect SDMs by altering this conditional probability. We can calculate the effect of unmeasured interactions using tools from probability theory. When we apply this general approach to CR models, we obtain counter-intuitive results: unmeasured biotic interactions can improve our ability to make predictions using SDMs. The ecology of biotic interactions is a vast topic, and no single set of models can reflect this topic in its entirety. Our point is that we need explicit theory to study specific questions about the role of biotic interactions, and that when we apply this theory to existing models, it is easy to find cases where biotic interactions have substantial and counter-intuitive effects.

We return to the questions that guided our research. First, we asked: can we model the probability that an environment will be suitable for a given species using incomplete information about resource dynamics? Our mathematical results, including analytic formulae, show that we can, in fact, do this. Even using relatively complex consumer-resource models, the presence of species 1 in certain locations and measurements of only one resource can be used to produce useful predictions.

We also asked how complex interactions between competitors shape species distributions. It is easy to show that, contrary to some suggestions in the literature, species ranges can be shaped by biotic interactions in CR models. As with other consumer resource models, competition can easily exclude an organism from some local environments. At a regional scale, some environments have a sufficiently large supply of resources to allow the persistence of species 1 on its own, but are unsuitable to species 1 in the face of a competitor. As a result, competition changes the relationship between resource measurements and the probability that an environment is suitable.

Given this information, we may develop a graphical understanding of one of the most vexing problems for interpreting SDMs – whether correlative models provide

information on the fundamental or realized niche. The example developed in Eq. 5 and Fig. 2 can be thought of as an estimate of the probability that an environment is a part of the fundamental niche of species 1, given data on the environment. Authors have argued that this is what we must learn to garner information on species distributions (Kearney 2006, Jiménez-Valverde et al. 2008). Others argue that correlative SDMs provide information on fundamental niche (Soberon and Peterson 2005, Soberon 2007). However, the marginal probability that an environment is a part of the fundamental niche is distinct from the probability that an environment is potentially suitable given unmeasured competition, as in Eq. 6 and Fig. 3. As a result, SDMs using data on the abiotic environment cannot typically be interpreted as estimates of the fundamental niche. Indeed if competitive exclusion is common, then we should not expect models of the probability that an environment is a part of the fundamental niche to offer the best predictions of the environments that are suitable.

Counter-intuitively, in Fig. 3 the presence of competition actually strengthens our ability to use measurements of the abiotic environment to predict the environments that are suitable. To see this, note that in our examples a distribution model fit to the probability of presence given competition has a higher AUC score than a model for the probability of presence in the absence of competition. This implies that, under some circumstances, biotic interactions make it easier to use measurements of the abiotic environment to model distributions.

An important consideration is how robust are these conclusions to parameter values? Models of CR dynamics have been influential particularly as conceptual descriptions of biotic interactions. However, they are difficult to parameterize directly in empirical systems. See, for example, Miller et al. (2005) who review 1333 citations for Tilman (1980, 1982), but finds only 26 well-designed tests of some facet of CR theory. Several studies do measure components of CR models. Tilman and Wedin (1991) for example makes empirical observations of species *ZNGI*s. This work focuses on competition among plants for a single nutrient, nitrogen. In their analysis, they use the concentration of nitrogen after three years of growth by grass species as a surrogate for the equilibrium concentration of nitrogen in the presence one consumer species – a point on the *ZNGI*. The impact of organisms on a resource (f_{ij}'s) can be measured as the difference between the amount of the resource supplied and the amount of resource present after consumption by the organisms. For example, Goldberg and Miller (1990) measure percentage of incident light that makes its way through a canopy of plants – a surrogate for the feeding rate of plants on sunlight. Experimental work frequently manipulates the inflow rate and supply point of resources through nutrient additions (Tilman 1977, Goldberg and Miller 1990, Tilman and Wedin 1991). It is however difficult develop parameter values for multiple resources and multiple species at the landscape scale.

In spite of these limitations we can show that biotic interactions would change the results of SDMs over a variety of parameter values. In Supplementary material Appendix 3, we simulate the ecological dynamics of our model with the deSolve package in R (Soetaert et al. 2009, R Development Core Team 2009). In many of these simulations, SDMs fit from data generated in the presence of biotic interactions are significantly different from SDMs fit from data generated in the absence of biotic interactions. Moreover, biotic interactions have the potential to both increase and decrease the AUC score derived from an SDM. The simplest way to encapsulate these results is to consider the conditions under which information on S_1 is most informative for predicting the probability of presence for species 1. When the probability of presence changes gradually with changes in S_1, then AUC scores are relatively low as is the case in Fig. 2. Under these parameter values, the presence of a competitor excludes species 1 from environments with low values for S_1 Fig. 3. As a result, the AUC score is higher for models fit to data in the presence of a competitor. Conversely, in the largest parameter value tested for a_{22} in Supplementary material Appendix 3 species 2, excludes species 1 from environments with all but the lowest concentration of S_2. The probability of presence changes little with values of S_1 in this scenario and as a result the presence of a competitor lowers AUC scores.

These conclusions must be tempered with caution because we have used a relatively simple framework to model the interactions between local environments and the landscape as a whole. Notably, we have assumed that ecological dynamics within patches equilibrate rapidly relative to dispersal between patches. It is, of course, possible that more complex dispersal mechanisms may degrade our ability to make predictions. A full treatment of this problem is beyond the scope of this paper. However, previous analyses of the interaction between dispersal and CR dynamics provide an outline of what we may expect to find. Abrams and Wilson (2004) analyze a two patch model with two competing species that migrate from one patch to another. As the level of dispersal approaches zero in Abrams and Wilson (2004), only the species with the lower resource requirement persists in either patch. In other words, in the limit of low dispersal, a species can only be present in an environments that are suitable and in which the species posses a competitive advantage. This result does not hold across all values of migration. Specifically, a species with higher resource requirements but a low rate of dispersal can outcompete a species with lower resource requirements and a high rate of dispersal. If we were to fit an SDM naively under such a scenario, we would erroneously infer that environments are most suitable to one species, when in fact that species only persists because it loses fewer individuals to migration into unsuitable patches.

Similarly, it is possible that one or both species are unable to disperse to some patches. If our focal species cannot disperse, this will lead to erroneously exaggerating the number of unsuitable environments. Conversely, if the other species cannot disperse to all patches, it is possible that our focal species will be present in environments that would be unsuitable in the face of competition.

We have assumed that the only thing to change from one location to another is the supply of resources. This omits other sources of uncertainty, say if the the *ZNGI*s changed from location to location due to other unmeasured facets of the environment. In addition, theory on multiple species that consume three or more essential resources indicates that complex dynamics are possible in larger ecological

communities, including oscillations and chaos (Huisman and Weissing 1999, 2001a, b). In principle, we can compute the marginal probability of presence given resource measurements in such models. However, the analytical and computational challenges are much greater. Transient dynamic cannot be ignored in such models, so analyzing equilibria is insufficient. For some parameters, it is difficult to predict the outcome of competition using measurements of the initial abundance of each species (Huisman and Weissing 2001b). Though multispecies CR models offer serious analytical challenges, these problems are hardly unique to the role of biotic interactions in species distributions. Indeed, one of the major conclusions of studies of multispecies CR models is the need to better develop probabilistic inferences: 'Despite knowledge of all species traits and species interactions, it is impossible to predict in advance which species will become dominant. Only predictions in terms of probabilities make sense' (Huisman and Weissing 2001b). As such, we believe that there are substantial opportunities to extend calculations of the probability of presence in two species CR models we have developed here to study multi-species interactions.

Finally, we asked how SDMs that incorporate only measurements of the abiotic environment can be reliable in the face of complex biotic interactions. We have a two-part answer to this question. First, though SDMs typically ignore the mechanisms that underlie species distributions, they accurately estimate the probability that an environment is suitable conditioned on the environmental variables we do measure. To see this, note that we started from a mechanistic model with 11 parameters in addition to four variables. Using only information on presences and measurements of one parameter – the supply point for a single resource, we generated SDMs that furnish strong predictions. Second, the marginal probability of presence provides valuable information of the set of environments that are suitable to an organism.

Our findings also have implications for the variables we must use to model species distributions. Current literature emphasizes modeling distributions with variables that capture the mechanisms that govern an organisms ability to thrive in an environment. This is not the conclusion of our results. Instead, models that include only a single variable provide valuable clues about which environments will be suitable to an organism. Indeed, in some circumstances, the relationship between an environmental variable and the suitability of an environment is strengthened by unmeasured biotic interactions between species. This counterintuitive result arises because of the conflicting goals of SDMs. Though we may hope to garner a complete mechanistic understanding of how nature works, we may still use incomplete information to generate predictions. The usefulness of a particular variable depends only indirectly on its mechanistic role, and we must use probability theory to realistically assess its ability to generate predictions. Though we have articulated this conclusion using assumptions frequently evoked in the community ecology literature such as linear functional responses and substitutable resources, the probabalistic methods we have used could potentially be applied more realistic models.

This final insight underscores a significant but underexploited role for mathematical theory in the field of SDMs. Our understanding of ecology is complex and parameter rich. Though our ability to model species distributions is constantly improving, it seems unlikely that any existing approach will be able to fully encapsulate the interactions between organisms and the environment. If this is true, we need an understanding of the best way to use our limited knowledge of the natural world to make predictions about distributions. Our results demonstrate the uncertainty generated by biotic interactions can be readily modeled. As such, mathematical theory can provide a natural link between our mechanistic understanding of nature and our incomplete observations of species distributions. However, this will require ecologists and biogeographers to use the tools of mathematics rather than the free tools available to run SDMs.

Acknowledgements – Helpful comments were provided by N. Sanders, L. Gross, C. Crawley, S. Bewick, J. Hughes, M. Leibold and D. Simberloff. WG was supported by a post-doctoral fellowship at the National Inst. for Mathematical and Biological Synthesis, an institute sponsored by the National Science Foundation, the U.S. Dept of homeland security, the U.S. Dept of Agriculture through the National Science Foundation Award EF-0832858, with additional support from the Univ. of Tennessee, Knoxville. LJH was funded by NSF grant DEB 0919499.

References

Abrams, P. 1988. Resource productivity-consumer species diversity: simple models of competition in spatially heterogeneous environments. – Ecology 69: 1418–1433.

Abrams, P. A. and Wilson, W. G. 2004. Coexistence of competitors in metacommunities due to spatial variation in resource growth rates; does R* predict the outcome of competition. – Ecol. Lett. 7: 929–940.

Araujo, M. B. and Guisan, A. 2006. Five (or so) challenges for species distribution modeling. – J. Biogeogr. 33: 1677–1688.

Booth, T. H. et al. 1988. Niche analysis and tree species introduction. – For. Ecol. Manage. 23: 47–59.

Brown, J. H. et al. 1996. The geographic range: size, shape, boundaries, and internal structure. – Annu. Rev. Ecol. Syst. 27: 597–623.

Buckley, L. B. et al. 2010. Can mechanism inform species' distribution models. – Ecol. Lett. 13: 1041–1054.

Case, T. J. et al. 2005. The community context of species' borders: ecological and evolutionary perspectives. – Oikos 108: 28–46.

Chase, J. M. and Leibold, M. A. 2003. Ecological niches – linking classical and contemporary approaches. – Univ. of Chicago Press.

Coyne, J. A. and Orr, H. A. 2004. Speciation. – Sinauer.

Elith, J. and Graham, C. H. 2009. Do they? How do they? WHY do they differ? On finding reasons for differing performances of species distribution models. – Ecography 32: 66–77.

Elith, J. et al. 2006. Novel methods improve prediction of species' distributions from occurrence data. – Ecography 29: 129–151.

Elith, J. et al. 2008. A working guide to boosted regression trees. – J. Anim. Ecol. 77: 802–813.

Friedman, J. et al. 2000. Additive logistic regression: a statistical view of boosting. – Ann. Stat. 28: 337–407.

Godsoe, W. 2010a. I can't define the niche but I know it when I see it: a formal link between statistical theory and the ecological niche. – Oikos 119: 53–60.

Godsoe, W. 2010b. Regional variation exaggerates ecological divergence in niche models. – Syst. Biol. 59: 298–306.

Goldberg, D. and Miller, T. E. 1990. Effects of different resource additions on species diversity in an annual plant community. – Ecology 71: 213–225.
Gotelli, N. J. et al. 2010. Macroecological signals of species interactions in the danish avifauna. – Proc. Natl Acad. Sci. USA 107: 5030–5035.
Holland, J. N. and DeAngelis, D. L. 2010. A consumer-resource approach to the density-dependent population dynamics of mutualism. – Ecology 91: 1286–1295.
Holt, R. D. 2009. Bringing the hutchinsonian niche into the 21st century: ecological and evolutionary perspectives. – Proc. Natl Acad. Sci. USA 106: 19659–19665.
Holt, R. D. et al. 2005. Theoretical models of species' borders: single species approaches. – Oikos 108: 18–27.
Huisman, J. and Weissing, F. J. 1999. Biodiversity of plankton by species oscillations and chaos. – Nature 402: 407–4010.
Huisman, J. and Weissing, F. J. 2001a. Biological conditions for oscillations and chaos generated by multispecies competition. – Ecology 82: 2682–2695.
Huisman, J. and Weissing, F. J. 2001b. Fundamental unpredictability in multispecies competition. – Am. Nat. 157: 488.
Hutchinson, G. E. 1957. Concluding remarks. – Cold Spring Harbor Symp. Quant. Biol. 22: 415–427.
Jiménez-Valverde, A. et al. 2008. Not as good as they seem: the importance of concepts in species distribution modelling. – Divers. Distrib. 14: 1472–4642.
Jordan, D. S. 1905. The origin of species through isolation. – Science 22: 545–562.
Kearney, M. 2006. Habitat, environment and niche: what are we modeling. – Oikos 115: 186–191.
Lange, W. 1967. Effect of carbohydrates on the symbiotic growth of planktonic blue-green algae with bacteria. – Nature 215: 1277–1278.
MacArthur, R. H. 1972. Geographical ecology: patterns in the distribution of species. – Harper and Row.
MacArthur, R. H. and Levins, R. 1964. Competition, habitat selection, and character displacement in a patchy environment. – Proc. Natl Acad. Sci. USA 51: 1207–1210.
Miller, T. E. et al. 2005. A critical review of twenty years' use of the resource-ratio theory. – Am. Nat. 165: 439–448.
Pearson, R. G. and Dawson, T. P. 2003. Predicting the impacts of climate change on the distribution of species: are bioclimate envelope models useful? – Global Ecol. Biogeogr. 12: 361–371.
Peters, R. H. 1991. A critique for ecology. – Cambridge Univ. Press.
Peterson, A. T. et al. 1999. Conservatism of ecological niches in evolutionary time. – Science 285: 1265–1267.
Phillips, S. J. et al. 2006. Maximum entropy modeling of species geographic distributions. – Ecol. Model. 190: 231–259.
Pulliam, R. 2000. On the relationship between niche and distribution. – Ecol. Lett. 3: 349–361.
Ross, S. 1997. A first course in probability. – Prentice Hall.
Ryabov, A. B. and Blasius, B. 2011. A graphical theory of competition on spatial resource gradients. – Ecol. Lett. 14: 220–228.
Schindler, D. W. 1971. Carbon, nitrogen and phosphorous and the eutrophication of freshwater lakes. – J. Phycol. 7: 321–329.
Soberon, J. 2007. Grinnellian and eltonian niches and geographic distributions of species. – Ecol. Lett. 10: 1115–1123.
Soberon, J. and Peterson, A. T. 2005. Interpretation of models of fundamental ecological niches and species' distributional areas. – Biodivers. Inf. 2: 1–10.
Soberon, J. and Nakamura, M. 2009. Niches and distributional areas: concepts, methods, and assumptions. – Proc. Natl Acad. Sci. USA 106: 19644–19650.
Soetaert, K. et al. 2009. deSolve: general solvers for initial value problems of ordinary differential equations (ODE), partial differential equations (PDE) and differential algebraic equations (DAE). – R package ver. 1.5.
Tilman, D. 1977. Resource competition between plankton algae: an experimental and theoretical approach. – Ecology 58: 338–348.
Tilman, D. 1980. A graphical-mechanistic approach to competition and predation. – Am. Nat. 116: 362–393.
Tilman, D. 1982. Resource competition and community structure. – Princeton Univ. Press.
Tilman, D. and Wedin, D. 1991. Dynamics of nitrogen competition between successional grasses. – Ecology 72: 1038–1049.
Tilman, D. and Pacala, S. 1993. The maintenance of species richness in plant communities. – In: Ricklefs, R. E. and Schluter, D. (eds), Species diversity in ecological communities: historical and geographical perspectives. Univ. of Chicago Press, pp. 13–25.

Supplementary material (Appendix E7103 at <www.oikosoffice.lu.se/appendix>). Appendix 3.

Appendix 1: the probability that an environment is suitable without competition

Start with the function for the probability that a species will be present in Eq. 4. To obtain the probability of presence conditioned only on S_1 we must integrate out S_2 over the possible values of this variable; 0 to 1. This is equivalent to calculating the portion of the graph above $ZNGI_1$ for each value of S_1:

$$\begin{aligned} P(X=1 \mid S_1=s_1) &= \int_0^1 P(X=1 \mid S_1=s_1,\, S_2=s_2)ds_2 \\ &= 1-\int_0^1\left(\frac{d_1}{f_{12}a_{12}} - \frac{f_{11}a_{11}}{f_{12}a_{12}}S_1\right)dS_2 \\ &= 1+\frac{f_{11}a_{11}}{f_{12}a_{12}}S_1 - \frac{d_1}{f_{12}a_{12}} \end{aligned} \quad (7)$$

Note that the probability of presence ranges from zero to one giving us Eq. 5. Note that this calculation and the calculation in the following section assume that resources are uniformly distributed over the interval (0,1) as discussed in the main text.

Appendix 2: the probability that an environment is suitable in the face of competition

To compute the probability of presence, we must calculate several values along the S_1 axis. First, we compute the intersection of the two *ZNGI*s to find the minimum value of S_1 at which species 1 can live. We must then determine the point where L_2 reaches the maximum possible value for S_2, and the point at which $ZNGI_1$ crosses the S_1 intercept Fig. 1.

The equations describing the two *ZNGI*s are:

$$ZNGI_1\text{: } R_2 = \frac{d_1 - f_{11}a_{11}R_1}{f_{12}a_{12}}$$

$$ZNGI_2\text{: } R_2 = \frac{d_2 - f_{21}a_{21}R_1}{f_{22}a_{22}}$$

By finding the solution for this system of two equations we obtain the point of intersection of the two *ZNGI*s:

$$S_{1,\text{intersection}} = \frac{f_{12}\,a_{12}d_2 - f_{22}a_{22}d_1}{f_{12}a_{12}f_{21}a_{21} - f_{11}a_{11}f_{22}a_{22}}$$

$$S_{2,\text{intersection}} = \frac{f_{21}a_{21}d_1 - f_{11}a_{11}d_2}{f_{12}a_{12}f_{21}a_{21} - f_{11}a_{11}f_{22}a_{22}}$$

since the slope is:

$$\frac{f_{22}R_2}{f_{21}R_1}$$

L_2 a line collinear with impact vector for species 2, and passing through the intersection of the the two *ZNGI*s can be described with the equation:

$$S_2 = \frac{f_{22}S_{2,\text{intersection}}}{f_{21}S_{1,\text{intersection}}}S_1 + b$$

we may obtain the value of b by substituting in the intersection point for the two *ZNGI*s ($S_1 = S_{1,\text{intersection}}$,$S_2 = S_{2,\text{intersection}}$) giving us:

$$b = \frac{(f_{21} - f_{22})(a_{11}d_2f_{11} - a_{21}d_1f_{21})}{f_{21}(a_{11}a_{22}f_{11}f_{22} - a_{12}a_{21}f_{12}f_{21})}$$

This produces a final line of:

$$S_2 = \frac{f_{22}(f_{21}a_{21}d_1 - f_{11}a_{11}d_2)}{f_{21}(f_{12}a_{12}d_2 - f_{22}a_{22}d_1)}S_1 + \frac{(f_{21} - f_{22})(a_{11}d_2f_{11} - a_{21}d_1f_{21})}{f_{21}(a_{11}a_{22}f_{11}f_{22} - a_{12}a_{21}f_{12}f_{21})} \quad (8)$$

We must find S_{1,maxS_2}. To do this we must find the S_1 value when $S_2 = 1$, given by:

$$S_{1,\text{maxS}_2} = (1 - b)\left(\frac{f_{21}(f_{12}a_{12}d_2 - f_{22}a_{22}d_1)}{f_{22}(f_{21}a_{21}d_1 - f_{11}a_{11}d_2)}\right)$$

We can solve for the S_1 intercept for the $ZNGI_1$:

$$S_{1,\text{intercept}} = \frac{d_1}{f_{11}a_{11}}$$

Using Eq. 8 we calculate the proportion of environments below the line describing the impact vector as:

$$\int_0^1 \left(\frac{f_{22}(f_{21}a_{21}d_1 - f_{11}a_{11}d_2)}{f_{21}(f_{12}a_{12}d_2 - f_{22}a_{22}d_1)}S_1 + \frac{(f_{21} - f_{22})(a_{11}d_2f_{11} - a_{21}d_1f_{21})}{f_{21}(a_{11}a_{22}f_{11}f_{22} - a_{12}a_{21}f_{12}f_{21})}\right)dS_2$$

where:

$$g(S_1) = \frac{f_{22}(f_{21}a_{21}d_1 - f_{11}a_{11}d_2)}{f_{21}(f_{12}a_{12}d_2 - f_{22}a_{22}d_1)}S_1 + \frac{(f_{21} - f_{22})(a_{11}d_2f_{11} - a_{21}d_1f_{21})}{f_{21}(a_{11}a_{22}f_{11}f_{22} - a_{12}a_{21}f_{12}f_{21})}$$

Assuming $S_{1,\text{maxS}_2} < S_{1,\text{intercept}}$, we can now derive Eq. 6 in the main text. When $S_1 < S_{1,\text{intersection}}$, species 1 is invariably absent and so the probability of presence is zero. When S_1 is between $S_{1,\text{intersection}}$ and S_{1,maxS_2}, the only suitable environments are in between L_2 and $ZNGI_1$. The proportion of environments lying between these points is given by $g(S_1) - h(S_1)$. In the subsequent line segment between S_{1,maxS_2} and $S_{1,\text{intercept}}$, the proportion of suitable environments is given by $h(S_1)$. When $S_{1,\text{intercept}} < S_1$ the probability of presence is 1. Combining the results from each of these line segments gives us Eq. 6 in the main text.

Alternatively, it is possible that $S_{1,\text{intercept}} < S_{1,\text{maxS}_2}$, in which case we reverse the order of these two terms recognizing that in the segment between $S_{1,\text{intercept}}$ and S_{1,maxS_2}, the probability of presence is the proportion of environments below L_2 i.e. $g(S_1)$. This produces the conditional probability:

$$P(X = 1 \mid S_1 = s_1, S_2 = s_2) = \begin{cases} 1 & \text{if } S_{1,\text{maxS}_2} < S_1 \\ g(S_1) & \text{if } S_{1,\text{intercept}} < S_1 < S_{1,\text{maxS}_2} \\ g(S_1) - h(S_1) & \text{if } S_{1,\text{intersection}} < S_1 < S_{1,\text{intercept}} \\ 0 & \text{if } S_1 < S_{1,\text{intersection}} \end{cases}$$