A Journal of Macroecology

Global Ecology and Biogeography

MACROECOLOGICAL METHODS

Delineating probabilistic species pools in ecology and biogeography

Dirk Nikolaus Karger[1,2*], Anna F. Cord[3], Michael Kessler[2], Holger Kreft[4], Ingolf Kühn[5,6,7], Sven Pompe[5,8], Brody Sandel[9], Juliano Sarmento Cabral[4,7], Adam B. Smith[10], Jens-Christian Svenning[9], Hanna Tuomisto[1], Patrick Weigelt[4,11] and Karsten Wesche[12,7]

[1]*Department of Biology, University of Turku, Turku, Finland,* [2]*Institute of Systematic Botany, University of Zurich, Zurich, Switzerland,* [3]*Department of Computational Landscape Ecology, Helmholtz Centre for Environmental Research – UFZ, Leipzig, Germany,* [4]*Biodiversity, Macroecology and Conservation Biogeography Group, University of Göttingen, Göttingen, Germany,* [5]*Department of Community Ecology, Helmholtz Centre for Environmental Research – UFZ, Halle, Germany,* [6]*Geobotany and Botanical Garden, Martin-Luther-University Halle-Wittenberg, Halle, Germany,* [7]*German Centre for Integrative Biodiversity Research (iDiv) Halle-Jena-Leipzig, Leipzig, Germany,* [8]*TUM, Chair for Terrestrial Ecology, Freising-Weihenstephan, Germany,* [9]*Section for Ecoinformatics and Biodiversity, Department of Bioscience, Aarhus University, Aarhus, Denmark,* [10]*Center for Conservation and Sustainable Development, Missouri Botanical Garden, Saint Louis, USA,* [11]*Systemic Conservation Biology, University of Göttingen, Göttingen, Germany,* [12]*Senckenberg Museum of Natural History Görlitz, Görlitz, Germany*

*Correspondence: Dirk Nikolaus Karger, Institute of Systematic Botany, University of Zurich, Zollikerstrasse 107, 8008 Zurich, Switzerland.
E-mail: dirk.karger@systbot.uzh.ch

ABSTRACT

Aim To provide a mechanistic and probabilistic framework for defining the species pool based on species-specific probabilities of dispersal, environmental suitability and biotic interactions within a specific temporal extent, and to show how probabilistic species pools can help disentangle the geographical structure of different community assembly processes.

Innovation Probabilistic species pools provide an improved species pool definition based on probabilities in conjunction with the associated species list, which explicitly recognize the indeterminate nature of species pool membership for a given focal unit of interest and better capture real-world complexity. Probabilistic species pools provide a quantitative assessment of how dispersal, environmental or biotic factors influence estimates of species pool composition and size for a given temporal extent.

Conclusions Based on one simulated and two empirical examples we demonstrate that probabilistic species pools allow us to disentangle the geographical variation in dispersal, environmental and biotic assembly processes for species assemblages in focal units. We also show that probabilistic species pools are fully compatible with traditional definitions of species pools and are applicable over a wide range of spatial and temporal extents. Additionally they are robust to missing data and provide a quantified and transparent approach to estimating the size and composition of species pools in a mechanistic way, providing a valuable tool for studies from community ecology to macroecology.

Keywords

Biodiversity, community assembly, community composition, gamma diversity, local species pool, probabilities, regional species pool, spatial scale, species distribution modelling, species richness.

INTRODUCTION

The concept of a species pool is commonly used in theoretical and applied studies in ecology, biogeography and conservation biology (Ricklefs, 1987; Cornell & Harrison, 2014). A species pool is the set of species that could potentially colonize and establish within a community (hereafter focal unit) (Zobel, 1997; Zobel *et al.*, 1998; Pärtel *et al.*, 2011; Lessard *et al.*, 2012a).

DOI: 10.1111/geb.12422
http://wileyonlinelibrary.com/journal/geb

Species pools have proven especially useful for testing whether the composition of communities differs from random expectation (Connor & Simberloff, 1978), estimating the influence of regional richness on local species richness (Ricklefs, 1987), testing for species saturation (Srivastava, 1999; Shurin & Srivastava, 2005) and understanding how different factors and processes of community assembly are linked to one another (Harrison & Cornell, 2008; Myers & Harms, 2009; Chase & Myers, 2011; Götzenberger *et al.*, 2012). More recently, ecologically explicit definitions of species pools have been suggested as a promising way to quantify the influence of evolutionary and historical processes on local community assembly (Algar *et al.*, 2011; Lessard *et al.*, 2012a; Carstensen *et al.*, 2013).

The term 'species pool' was first introduced in island biogeography, where it was defined as those species that can disperse to an island irrespective of whether they all survive (MacArthur & Wilson, 1967). Thereafter, it found application in non-island-like settings (Ricklefs, 1987; Cornell & Lawton, 1992; Ricklefs & Schluter, 1994; Zobel, 1997; Zobel *et al.*, 1998). Since its introduction, the species pool concept has been used and defined in several ways – from the somewhat simplistic use of regional richness as a species pool (Herben, 2005; Kluge *et al.*, 2006; Knop *et al.*, 2008) to the use of overlapping geographical ranges (dispersion fields; Graves & Rahbek, 2005; Borregaard & Rahbek, 2010; Lessard *et al.*, 2012b; Carstensen *et al.*, 2013). The latter definition proved a major step forward and allowed ecologists to weigh the probability that a species would be included in a community's species pool and to estimate its geographical extent (Graves & Rahbek, 2005; Lessard *et al.*, 2016). Since then, temporal extent has been recognized as an important factor (dispersal or survival over time) that can lead to changes in species pool size (Munguia, 2004; Starzomski *et al.*, 2008; Canning-Clode *et al.*, 2009; White & Hurlbert, 2010; Cornell & Harrison, 2014) but has remained notoriously difficult to integrate into species pools definitions.

SPECIES POOLS FROM A PROBABILISTIC VIEWPOINT

Species pools have nearly always been acted upon by hierarchical 'filters' that restrict membership in the focal community (Zobel, 1997; Zobel *et al.*, 1998). Traditionally, these filters have been conceptualized to act in a binary fashion, determining if species are part of the species pool [1] or not [0] (Pärtel *et al.*, 1996; Algar *et al.*, 2011; González-Caro *et al.*, 2012; Belmaker & Jetz, 2013; Ronk *et al.*, 2015). There is, however, no simple way to decide on the level of constraint or the threshold used to define species pools (Lessard *et al.*, 2016), with the additional problem that binary thresholds or constraints can severely bias the resulting assemblage structure and are therefore generally undesirable (Cavender-Bares *et al.*, 2006; Kraft *et al.*, 2007; Kissling *et al.*, 2012; Eiserhardt *et al.*, 2013; Lessard *et al.*, 2016). To avoid this problem, an approach which enables the examination of several species pool definitions over the entire range of possible levels of constraint or thresholds has recently been proposed (Lessard *et al.*, 2016).

If species pools are thought of as the result of combining species membership probabilities, however, it becomes immediately apparent that a species pool does not necessarily need to include binary thresholds or constraints at all. Any information that is available about the species, especially geographical ranges, dispersal abilities and habitat preferences (Lessard *et al.*, 2012a, 2016), can be used to estimate probabilities of species pool membership across the full range [0, 1]. This allows us to move beyond the need to introduce distinct occurrence thresholds. Here we formulate this framework by specifically addressing environmental, biotic and dispersal-related filters in a probabilistic fashion. Each of these filters (or hereafter factors, x) can be thought of as a set of probabilities associated with each species (Zobel *et al.*, 1998). If the pool is filtered by a single process then its size is simply the sum of all probabilities for all species:

$$\text{species pool size } {}^{i}\Psi = \sum_{s=1}^{S} P_s \quad (1)$$

where P_s is the probability of species s and S is the total number of species.

When an additional filter is applied (e.g. habitat suitability in addition to dispersal), the species pool becomes smaller: not all species that can disperse to a focal unit will be able to grow under the environmental conditions that prevail there. The size of the species pool can hence be obtained by multiplying, for each species, the probabilities associated with each of the n applied filters (assuming they act independently), and then summing over the species:

$$\text{species pool size } {}^{i}\Psi = \sum_{s=1}^{S} \prod_{x=1}^{n} P_{xs} \quad (2)$$

where x designates the particular factor and n is the number of factors. Notably, equations 1 and 2 still hold under a threshold concept, even if P values are binary rather than continuous. Hence, a species pool can simply be defined as a function of probabilities of a species' occurrence in the focal unit given the unit's environmental and biotic conditions, geographical location and the time frame of interest (Lessard *et al.*, 2012a). The resulting set of values $\{P_1, P_2, P_3, \ldots, P_S\}$ that quantify for each of S species its probability, given a specific factor, can then be taken as the probabilistic species pool (Ψ).

Adopting the concept of a probabilistic species pool gives us the flexibility to retain information on probabilities underlying estimates of species occurrences and partition the independent effects of the different factors on the pool. We can estimate to what degree species are actually relevant for different focal units, and even provide statistical estimates of uncertainty derived from the probabilities employed to build the delineation of the species pool.

In the following, we provide case studies to demonstrate how: (1) species-specific probabilities of pool membership can be estimated, and whether they translate into actual events; (2) how probabilistic species pools can be used to investigate the geographical distribution of community assembly processes; and (3) how probabilistic pools compare with results of binary

Table 1 Scale and factors of the estimated species pools included in the calculation for the serpentine grasslands, Ranunculaceae and simulated data.

	Serpentine grassland	Ranunculaceae	Simulations
Scale:			
Data extent	64 m^2	389,220 km^2	225 cells
	All species present ($S = 28$)	All Ranunculaceae species present ($S = 52$)	400 simulated species
Focal unit grain	0.5 m × 0.5 m	10′ longitude × 6′ latitude (arc-minutes; *c.* 130 km^2)	One cell
Temporal extent	2 years	5,000 years	10,000 years (time steps)
Temporal grain	1 year	1 year	1 year (last time step)
Factors:			
Dispersal (D)	x	x	x
Environment (E)	x	x	x
Biotic (B)	–	–	x

Table 2 Abbreviations of different factors used to calculate the various pool probabilities (P) and related probabilistic species pools (Ψ).

Abbreviation	Description	Applied to dataset:
D	Dispersal factors	
DD	Dispersal based on geographical distances	S, R, Sim
DDT	Dispersal based on geographical distances weighted by species traits	S, R
DR	Dispersal based on resistance distances over environmentally suitable surfaces (suitable dispersal pathways)	R
DRT	Dispersal based on resistance distances over environmentally suitable surfaces (suitable dispersal pathways) additionally weighted by species traits	R
E	Environmental factors	
ER	Environmental suitability based on niche modelling (dataset-scale species distribution models)	S, R, Sim
EB	Environmental suitability based on Beals' smoothing	S, R
B	Biotic factors	Sim
	×, denotes combinations of factor groups (dispersal, environmental, biotic)	S, R, Sim

Dataset abbreviations: S, serpentine grasslands; R, Ranunculaceae; Sim, simulations.

pools. We use one simulated and two empirical datasets, chosen because of their relative completeness, data extents and representation of different spatial and temporal scales that researchers might frequently encounter when delineating species pools (Table 1).

Estimating and validating probabilities – serpentine grasslands

Species-specific probabilities are the basic measurement unit of probabilistic species pools. They need to be calculated based on mechanistic assumptions about dispersal, environmental suitability and biotic interactions. However, it often remains unclear how such calculated probabilities are actually translated into real events. To assess if probabilities based on dispersal, environmental factors and current distributions of species are able to actually predict future distributions we use a dataset that includes local community composition recorded in two consecutive years in serpentine grassland largely dominated by annuals in California, USA. We estimate species-specific probabilities in the first year and compare these with actual changes in species composition in the following year.

The serpentine grassland data set originated from a repeated survey of vascular plants in an 8 m × 8 m plot at the University of California's McLaughlin Natural Reserve in the Coast Range of California, USA (Green *et al.*, 2003; Smith *et al.*, 2013). The plot was divided into 256 cells of size 0.5 m × 0.5 m, where the abundance of each species was recorded in 2005 and again in 2006. In any year *c.* 75% of species in the plot were annuals and many perennial seedlings did not survive for more than a year, meaning that the community could change state rapidly across years. The plot contained *c.* 60,000 individuals in 2005 and *c.* 43,000 in 2006.

Dispersal-related P

For estimating distance-based dispersal P_{DD} for each cell (see Table 2 for an explanation of the abbreviations used), we assumed that each species had a probability of reaching a cell (n) based on its presence in a total of N occupied cells located distance d_n from each occupied cell:

$$P_{DD,n} = 1 - \prod_{n=1}^{N} (1 - e^{-kd_n}) \quad (3)$$

(Bischoff, 2005) with N being the total number of cells occupied by the species, and k being a rate constant representing the dispersal ability of species over a given distance and time. We assigned k on the basis of traits related to dispersal (height and seed mass). We assumed that plant height is the primary determinant of dispersal distance, and that, within a height class, larger seed masses reduce dispersal (Thomson *et al.*, 2011). As direct measurements of plant height were not available we classified species into four height categories (prostrate, subshrubs, shrubs, trees). This score was incremented by 1 if the species was an annual to take into account that annuals have a higher probability of producing seeds at the end of their growing season, as they depend on this for survival. We further multiplied the score by log(seed mass), rescaled to [−0.5,0.5], assuming that species with larger seeds do not disperse as far as species with small seeds (Muller-Landau *et al.*, 2008). This acknowledges that seed mass and height do not usually explain more than 50% of the variation in dispersal distances (Muller-Landau *et al.*, 2008). Additionally we assumed a scenario in which dispersal traits are not known, and set all species dispersal values to 0.45, 0.75 or 1.5 m year^{-1} (see Appendix S1 in Supporting Information), which are in the range of known dispersal values for wind-dispersed species within 1 year (Andersen, 1993).

Environmental-related P

We used several methods to calculate probabilities related to environmental factors (P_E) (Table 2). P_{EB}, was calculated by Beals' smoothing (Beals, 1984; De Cáceres & Legendre, 2008). Beals' smoothing is a multivariate transformation specially designed for species presence/absence community data containing noise and/or a lot of zeros, which replaces the observed values of the target species by predictions of occurrence on the basis of its co-occurrences with the remaining species in the overall dataset (De Cáceres & Legendre, 2008). The index provides the probability of occurrence of a target species at a site given its co-occurrences with other species (Beals, 1984).

For P_{ER}, we used environmental covariates to characterize environmental suitability for each cell and species. For the serpentine grasslands data, we focused on soil characteristics since other factors such as climate, exposure and elevation can be seen as constant at this small scale. We used soil core volume (an index of rockiness), pH, concentration of Ca, K, Mg, Mn, Na and Ni, and the ratio Ca/Mg, all of which are known to be linked to plant species occurrences (Janssens *et al.*, 1998). Euclidean distances between the focal cell and all other cells were calculated in multivariate environmental space defined by these factors. The environmental suitability of the cell was then defined as the ranked distance of the closest cell occupied by the species minus 1 divided by the total number of cells ($n = 256$). This yielded a value in the range [0, 1] for each species for each cell.

We validated our estimates of different values of P by using P calculated from 2005 distributions to predict new colonization in cells in 2006. For this test, we calculated the area under the receiver operating characteristic curve (AUC) to evaluate the performance of P from 2005 in predicting species occurrences in 2006 (Fig. 1). We also used these methods to evaluate the reverse prediction, asking which cells a species will colonize, rather than which species a cell will gain.

We assessed sensitivity against data incompleteness by randomly removing cells across 100 iterations and calculating the species pool each time (removing 16, 32, 48, . . ., 240 cells). We then calculated the correlation between combined ($P_{DD\times EB}$) probabilities between the complete data set and the rarefied set.

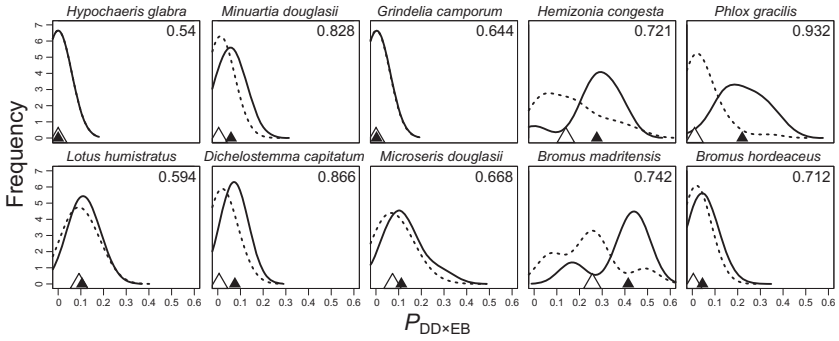

Figure 1 Predicting occurrences for the 10 most abundant species of a serpentine grassland in California. Frequency indicates the distribution (out of 256 cells) of probability values in colonized (black lines) or non-colonized (dotted lines) cells for predicting new colonization and continued absences between the first and second year of the census. In most cases, sites in which a species occurred in the second year (black lines) have higher probability values of presence (the curve is shifted more to the right) in the first year than sites in which species did not occur in the second year (dotted lines). Triangles show the median probability for non-colonized (open triangles) versus colonized probabilities (black triangles). Area under the receiver operating characteristic curve values (AUC, top right corner) are given for each species (with those closer to one being better predictions). Comparison of AUC values indicates that the probabilistic species pool was able to predict colonization (or lack thereof) across years, especially for species with greater discrepancies between P distributions of colonized and uncolonized cells. These results combine dispersal and environment probabilities, using uniform dispersal abilities of $k = 0.45$, exponential dispersal kernels and Beals' smoothing ($P_{DD\times EB;}$ see Table 2 for an explanation of the abbreviations).

Correlations were calculated in two ways: per species across all cells and per cell across all species.

Results and discussion

The median AUC and percentages of sites with better-than-random predictions were: (1) for P_{DD}, 0.875 and 98% and for P_{DDT} 0.863 and 92% (for dispersal only); (2) for P_{ER}, 0.646 and 76% and for P_{EB}, 0.681 and 96% (for environment only). For combined dispersal and environmental probabilities: $P_{DD\times ER}$, 0.68 and 76%; $P_{DD\times EB}$, 0.876 and 99%; $P_{DDT\times ER}$, 0.857 and 92%; $P_{DDT\times EB}$, 0.863 and 92%.

The reverse predictions (predicting the cells that are colonized by a species) also performed well. Estimating *P* using just dispersal factors resulted in a median AUC and proportion of sites with better-than-random predictions (AUC > 0.5) of 0.699 and 80% for P_{DD} and 0.661 and 76% for P_{DDT}; and using environmental factors 0.601 and 80% for P_{EB} and 0.686 and 93% for P_{ER}. Combined dispersal and environmental factors performed well, with values of 0.703 and 84% for $P_{DD\times EB}$, 0.833 and 88% for $P_{DD\times ER}$, 0.654 and 84% for $P_{DDT\times EB}$ and 0.669 and 80% for $P_{DDT\times ER}$. P_{DD} performed better when $k = 0.45$ m year^{-1}, or $k = 0.75$ m year^{-1}, but the differences were only very small. For values outside this range AUC values and percentages were lower than P_{DDT} (see Appendix S2). AUC values of P_{DD} and P_{DDT} values might also be additionally influenced by seed dormancy (dispersal in time), as some occurrences in 2006 might be related to dispersal events before 2005, though other experiments near this site demonstrate fairly low seed bank persistence (A.B.S., unpublished data).

Our results confirm that calculation of *P* can represent the probability that a species will colonize a particular location (Fig. 1). The high AUC values of P_{DDT} provide evidence that the dispersal model performs well when using actual dispersal traits. Although models with $k = 0.45$ m year^{-1} and k = 0.75 m year^{-1} performed slightly better than P_{DDT} (only by a difference of 0.01 in AUC), these values of *k* are arbitrary and usually not known *a priori*. Slightly different assumptions (e.g. $k > 1.5$ m year^{-1}) drastically altered the results.

The results also confirm the hierarchical structure of filters acting on species pools (Zobel *et al.*, 1998), with combined factors ($P_{DD\times EB}$, $P_{DD\times ER}$, $P_{DDT\times EB}$, $P_{DDT\times ER}$) producing better predictions than dispersal alone (P_{DD}, P_{DDT}). Predictions using *P* based on just dispersal were also higher than those based on just environmental factors (P_{ER}, P_{EB}), which seems reasonable given that environmental variation is low at small spatial scales making dispersal a key factor for community assembly.

Values of *P* calculated with missing data and with complete data were fairly well correlated within species across plots, even after discarding up to 94% of the data (Appendix S3a). Values for cells across species were nearly insensitive to rarefaction – abundant species maintained high values and rare species maintained low values (Appendix S3b). This example shows that *P* values are fairly insensitive to data restrictions and give robust estimates of pool membership even when sampling is incomplete.

Inferring geographical patterns of community assembly processes – Ranunculaceae assemblages

Among the more recent applications of species pools are proposals to disentangle the drivers of community composition (Pärtel *et al.*, 2011; de Bello *et al.*, 2012; Lessard *et al.*, 2012a; Cornell & Harrison, 2014). Using a dataset of Ranunculaceae assemblages in Germany, we tested different approaches for estimating the composition and size of probabilistic species pools on a regional spatial scale. We employed different environmental and dispersal filters and compared them with realized species richness in a focal unit. We assessed how inferences about the strength and spatial distribution of different factors on focal units can be drawn from probabilistic species pools.

The German FLORKART data report presence/absence of vascular plant species on a lattice with a spatial resolution of 10′ longitude × 6′ latitude (*c.* 130 km^2) in Germany (Kühn *et al.*, 2006; Manceur & Kühn, 2014). We used the time period 1950–90 and excluded from the analysis grid cells near country borders with an area of less than 117 km^2 located within Germany, resulting in a total of 2994 cells. We chose the family Ranunculaceae because it has a relatively stable taxonomy, is reasonably diverse in Germany (52 species) and has a range of life-forms (annual, perennial, woody lianas) as well as different dispersal and pollination syndromes. We explored a range of possibilities to estimate the species pool (see Table 2) to assess how different processes, parameterizations and definitions affect species pool size and composition.

Dispersal-related P

For the Ranunculaceae, P_{DD} was calculated using the same equation as for the serpentine grassland species, setting *k* to 0.002 distance units per time step (this translates into *c.* 20 m year^{-1}), as exact values of dispersal over time are unknown. We additionally calculated the probability of pool membership by resistance distance over environmentally suitable surfaces P_{DR} for every species based on environmental niche models and the environmental conditions in each grid cell (for detailed methodologies see Pompe *et al.*, 2008; see also McRae, 2006; McRae *et al.*, 2008) using the R package *gdistance*. This takes into account that species can have different dispersal probabilities based on the habitat they need to cross (Janin *et al.*, 2009; Eiserhardt *et al.*, 2013; Weigelt & Kreft, 2013). We selected three dispersal modes potentially capable of covering long distances (wind dispersal, animal dispersal, water dispersal) based on the Dispersal Diaspore Database (Hintze *et al.*, 2013; http://www.seed-dispersal.info accessed on 3 December 2013) and ranked them relative to the overall German flora to include dispersal traits P_{DRT}. Whenever a species was in the upper 50th percentile for one of the three dispersal syndromes, it was considered a long-distance disperser and received $1/k = 1.5$, which was set to $1/k = 0.5$ in other cases. This approach has proved to be valid in a similar study (Ozinga *et al.*, 2009).

Box 1:
Key terms and concepts

Binary species pool: a species pool delineated by including or excluding a species from the species pool. Binary species pools can be seen as special cases of probabilistic species pools where each species' probability of membership equals either 0 or 1.
Data extent: the geographical or taxonomic extent from which probabilistic species pools are derived. Unless complete datasets are used, limited data extents can act as binary constraints on species pools (constraining dispersal or biotic interactions).
Focal unit: the geographical or environmental space, a habitat, or community for which one wants to know the species pool (e.g. a grid cell in the empirical examples).
Probabilistic species pool (Ψ): a set of values $\{P_1, P_2, P_3, \ldots, P_S\}$ that quantify for each of S species its probability of being able to occur in a particular focal unit. This can reflect probabilities of, for example, performance or arrival.
Probability of pool membership (P): the probability that a given species disperses to, establishes, or survives in a focal unit. Probabilities can be estimated by using an a priori known response of a species to a given factor x. They can depend on, for example, survival for a minimum period of time, survival up to a certain ontogenetic stage (seedling, juvenile) or the maintenance of a viable population. The probability of pool membership gives the relative contribution of a species to the species pool.
Relative probabilities: used for estimates of P when data limitations do not allow estimation of absolute probabilities. If relative probabilities are used then the pool size ($^i\Psi$) serves as an index of focal unit richness.
Species pool size ($^i\Psi$): $\sum_{s=1}^{S}\prod_{x=1}^{n}P_{xs}$ where s is species, S is the number of species, x is the factor affecting pool membership, n is the number of factors and P is probability of pool membership. A measure of the expected number of species present in a focal unit (if true probabilities are used, this can be interpreted as number of species, with relative probabilities only as an index thereof).

Environmental-related P

For the Ranunculaceae assemblages, we used available suitability maps with a probability range [0, 1] based on climatic, soil and land-use parameters on a European scale (Pompe *et al.*, 2008). We used the EUSOILS soil characteristics (Panagos, 2006) and 38 bioclimatic variables at a $10' \times 10'$ (arcmin) resolution (Pompe *et al.*, 2008) and included four land-use classes at the $10' \times 10'$ grid resolution from PELCOM (Mücher *et al.*, 2001): forest, grassland, cropland and urban landscape (Mucher *et al.*, 2000). We applied a principal components analysis on the bioclimatic variables and a correspondence analysis (CA) on the soil data to avoid multicollinearity. Six principal components (explained variance 93%) and six CA axes (explained variance 56%) were subsequently used as environmental niche model predictors (Pompe *et al.*, 2008). Probability surfaces for P_{ER} were then calculated using generalized linear models.

Calculating different Ψ

We calculated a set of probabilistic species pools as well as a binary species pools using the entire set of focal units (here grid cells; see Appendix S4 for example results). Further analyses for the Ranunculaceae data, however, concentrated on a core region, defined by being at least 100 km away from the closest grid cell outside Germany, resulting in 1472 grid cells where we assumed that edge effects were not important. The data extent was set to the area of Germany, ignoring seed sources outside the study areas, although these may be important. Given that our aim was to compare effects of different pool estimators, we regard any remaining artefacts as limited. For reference, a binary species pool was delineated using a 'moving window' approach, with a fixed distance threshold of 100 km around a focal unit to include or exclude species in the species pool.

An important emergent property of the probabilistic species pool is the ability to calculate the size of the site-specific species pool (the species pool size index, $^i\Psi$; Box 1) from equation 2. $^i\Psi$ represents the expected number of species present in a focal unit based on the factors used to delineate the species pool. We calculated correlations between $^i\Psi$ across grid cells estimated using different definitions and combinations of dispersal and environmental pool.

Results and discussion

Values based on dispersal ($^i\Psi_D$) were intercorrelated and were also correlated with actual richness in a 100-km radius (all $r > 0.78$; Fig. 2, Appendix S5). Values of $^i\Psi_E$ were only moderately correlated with actual richness in a 100-km radius ($r = 0.56$), but were more strongly correlated with cell richness (especially for Beals' smoothing at $r = 0.87$ between $^i\Psi_{EB}$ and cell richness). $^i\Psi$ values based on environmental factors thus more closely reflect the cell richness than richness in a 100-km radius, confirming that richness in a 100-km radius is only a rough approximation of cell richness. Richness in a 100-km radius may reflect dispersal processes as it is strongly correlated with $^i\Psi_D$ ($r = 0.75$ for Ψ_{DR} ~ richness in a 100-km radius; $r = 0.83$ for Ψ_{DD} ~ richness in a 100-km radius).

$^i\Psi$ values based on Beals' smoothing were only moderately correlated with those from niche modelling (e.g. $r = 0.69$ between $^i\Psi_{EB}$ and $^i\Psi_{ER}$), indicating again that methodological choices have important effects. The limited number of Ranunculaceae species in Germany poses a methodological constraint, as Beals' smoothing calculates probabilities purely based on co-occurrence of species (De Cáceres & Legendre, 2008). Beals' smoothing may also implicitly capture several factors that are not environmental, i.e. biotic factors or even congruent history-linked dispersal-generated patterns (Normand *et al.*,

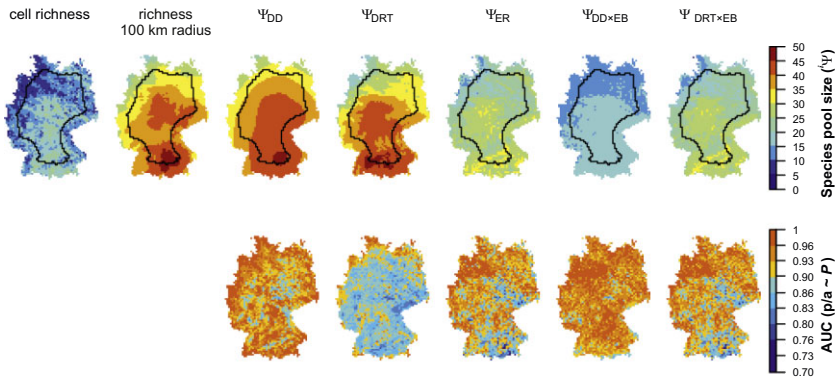

Figure 2 Inferring geographical patterns of community assembly processes. Upper row: Observed species richness at the resolution of single grid cells and within a moving window with a radius of 100 km around a focal cell for Ranunculaceae in Germany compared with species pool size index values (${}^{i}\Psi$) calculated using different dispersal and environmental factors. Lower row: area under the receiver operator curve (AUC) values showing the spatial distribution of predictive power of species pools for presence or absence (p/a) of species compared with their probability of pool membership (*P*). Comparing the spatial distribution of AUC across pool definitions allows inferences about the geographically varying nature of assembly processes. ${}^{i}\Psi$ and AUC were calculated separately for each 10′ longitude × 6′ latitude grid cell. Only the cells enclosed by the black line were used as focal cells in correlative analyses to avoid effects of truncated data availability near the country borders. Ψ subscripts: DD, dispersal based on geographical distances; DRT, suitable dispersal pathways; ER, environmental suitability based on niche modelling; EB, environmental suitability based on Beals' smoothing; × denotes combinations of factor groups. Correlations among all possible combinations of species pool delineations are shown in Appendix S6.

2011). If Beals' smoothing is used to calculate probabilistic species pools, the similarities to and differences from environmental niche modelling should be kept in mind.

To compare how different approaches estimate the composition of Ψ and the respective *P* we used pair-wise Euclidean distances of *P* for each species within the respective species pool of the Ranunculaceae data. Distance matrices of *P* were compared for all combinations of the applied dispersal and environmental pool definitions using pair-wise Mantel correlations. To investigate the spatial distribution of how well *P* was delineated in different ways in congruence with realized communities, we calculated the AUC for each grid cell separately.

Mantel correlations among matrices of species pool composition based on *P* were on average weaker than correlations among ${}^{i}\Psi$ values (Appendix S5). Species pool delineations accounting for both dispersal and environment were only moderately correlated with a simple pool corresponding to richness in a 100-km radius (all $r \leq 0.6$), highlighting the discrepancy between the classical binary species pool and the probabilistic species pool Ψ. It should be noted, however, that the latter was derived by summing relative probabilities. Hence, comparison between ${}^{i}\Psi$ (an index) should therefore be considered with caution. Correlations, however, are not affected by the error in absolute values, and show that the probabilistic species pool differs from actual diversity in any given reference area. They also show that considering dispersal, environmental suitability or the combined effects of both has an important effect on the composition of the pool.

By taking a specific grid cell as a focal unit (cell 4425, Göttingen, Germany, located in the middle of Germany where no edge effects are expected), we estimated the difference in ${}^{i}\Psi$ for a dispersal-based species pool (Ψ_{DD}) and an environment-based species pool (Ψ_{ER}). ${}^{i}\Psi$ values were higher for Ψ_{DD} than for Ψ_{ER}, indicating that environmental filtering imposes a stronger constraint than dispersal filtering in this case (Fig. 3). Again, as continuous variables (*P*) are compared to binary (presence–absence of species in a focal unit) ones, AUC is an appropriate measure of how well Ψ corresponds to realized communities. Ψ_{ER} predicted species occurrences in the focal unit better than Ψ_{DD}, again indicating stronger environmental filtering within the focal unit (Fig. 3).

One of the advantages of the concept of probabilistic species pools is that it provides quantitative confidence in our species pool estimates. To quantify the degree of confidence in ${}^{i}\Psi$, we calculated the probability mass function (PMF) for ${}^{i}\Psi$ which gives the probability of each value of ${}^{i}\Psi$. Additionally we calculated the cumulative density function (CDF) for ${}^{i}\Psi$, which provides the probability of the minimum certain number of species present in the focal unit. Based on these specific probability distributions (Fig. 3), it is apparent that possible values for ${}^{i}\Psi$ are between about 11 and 18 species, with the highest probability of 14 species and a probability of about 1 that fewer than 19 species are part of the species pool.

Spatial variation in the relative values of AUC can be used as an indicator of how the importance of different assembly mechanisms varies geographically (Fig. 2). We found, for example, that environmental filters (Ψ_{EB}) were stronger in the northern part of Germany, while the combination of dispersal and environmental filters ($\Psi_{DD\times EB}$) was stronger in the northern and southern parts of the country, but somewhat weaker in the middle. In contrast, resistance-based dispersal (Ψ_{DRT}) was much less influential relative to other assembly processes. This example demonstrates the utility of comparing observed communities with Ψ in a geographical context.

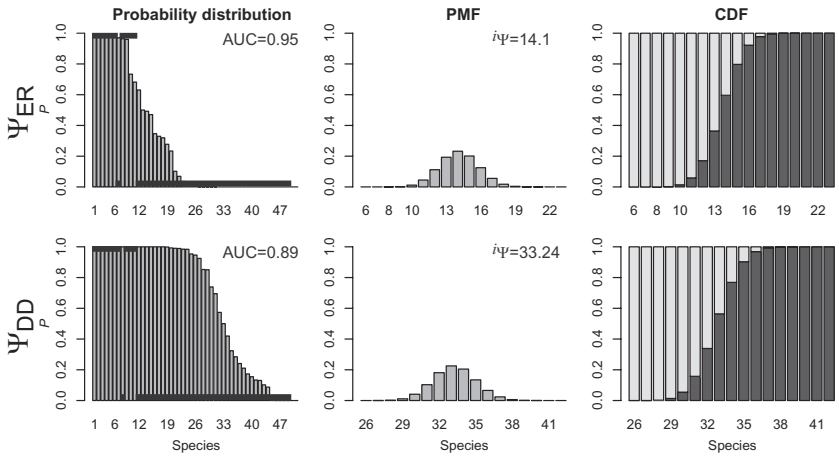

Figure 3 Comparison of species occurrence with community structure for a probabilistic species pool of Ranunculaceae delineated by environmental factors (Ψ_{ER}) and dispersal (Ψ_{DD}) for a timeframe of 5000 years. The focal unit here is a grid cell focused on Göttingen, Germany (130 km^2). Black squares represent observed presences and absences of a species within the focal unit. Area under the receiver operator curve (AUC) is based on the comparison between probabilities in the community structure of the species pool and species occurrences. $^{i}\Psi$ is the species pool index value. The probability mass function (PMF) gives the probability of each value of $^{i}\Psi$. The cumulative density function (CDF) provides the probability of a certain minimum number of species present in the focal unit given a specific factor. The observed richness in the focal unit is nine species. This is less than the mean predictions using different pools, indicating that other processes are shaping this assemblage. Probability distributions of Ψ_{ER} and Ψ_{DD} show declines in probabilities towards the full (global) set of species, indicating that all important species have been included in the species pool.

Comparing probabilistic and binary species pools – simulations

The novel aspect of the probabilistic species pool is the use of species-specific probabilities versus binary thresholds, which impose profound constraints on the inferences in a given analysis (Cavender-Bares *et al*., 2006; Kraft *et al*., 2007; Kissling *et al*., 2012; Eiserhardt *et al*., 2013; Lessard *et al*., 2016). In a simulation framework we compared binary species pools with probabilistic species pools to highlight how a probabilistic approach has a profound impact on estimates of the size and composition of species pools. The use of artificial data allowed us to consider all major components affecting species distributions (see Appendix S6), and we can hence apply dispersal and environmental as well as biotic filters to test if a combination of all relevant factors closely reassembles realized communities.

The simulated data set comes from Cabral & Kreft (2012). The model considered different species pools consisting of a global set of 400 species varying in their environmental niche and traits that control their performance. Species pools were simulated under physiological, dispersal and competition constraints. The spatial extent was 15×15 grid cells, where one grid cell was calibrated to implicitly measure 1 km^2. With the simulated data, all major components affecting species distribution are explicitly considered (see Appendix S5). To apply the probabilistic species pool concept we used the last time step of only one simulated pool.

For the dispersal factor, we overlaid Clark's 2Dt kernels (Clark *et al*., 1999) for each species over a hypothetical landscape. Ψ_D was then given by the sum of all overlaid kernels. For P_E, we used a species-specific suitability for each cell scaled between 0 and 1 based on a factor that resembled temperature. For the biotic factors, we included demographic information which was calculated as the survival probability of each species under metapopulation dynamics (explicit local dynamics and dispersal between populations) without interactions (i.e. species were simulated alone) for each environmental condition. The probability P was then given by number of occupied cells divided by the number of suitable cells, considering all species simultaneously (community simulation). Interspecific competition in this case was present as resource competition and facilitation which happened when a species with overcompensatory dynamics had its metapopulation stabilized by resource competition.

We calculated species pool size indices based on: (1) a probabilistic, and (2) a binary classification assuming a species was present if $P > 0.001$ for dispersal and $P > 0$ for the remaining factors for the last time step of the simulation. These are small values for thresholds, but using small values is conceptually no different from assuming that the regional pool of species comprises the actual species pool of a focal unit (Cavender-Bares *et al*., 2006; Swenson & Enquist, 2009; Kissling *et al*., 2012; Eiserhardt *et al*., 2013).

Results and discussion

$^{i}\Psi$ showed strikingly different values depending on the filter and distribution used (Fig. 4). Overall, the dispersal filter resulted in the largest number of species, particularly when a binary species pool concept was applied. In fact, the highest values were close to the total number of species in the dataset across all cells (400 species).

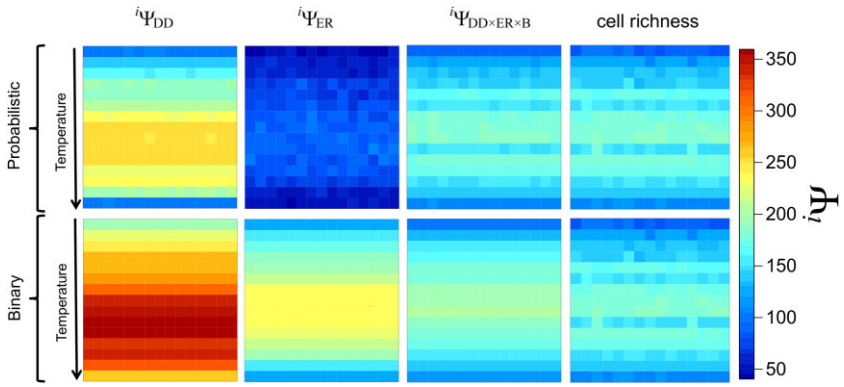

Figure 4 Landscape patterns of probabilistic and binary species pools. Species pool size index values ($^{i}\Psi$) incorporating dispersal ($^{i}\Psi_{DD}$), environment ($^{i}\Psi_{ER}$) and additional biotic factors ($^{i}\Psi_{DD\times ER\times B}$) for simulated data are compared with binary definitions of the species pool over a 15 cell × 5 cell landscape consisting of a single temperature gradient. $^{i}\Psi$ was calculated separately for each cell. Colour (online only) values/shades indicate the species richness within each cell. The top row represents a probabilistic species pool size index, the bottom row the corresponding binary species pool (thresholds set at $P > 0.0001$ for Ψ_{DD}, $P > 0$ for the binary pool). The right-hand column shows the simulated (observed) cell richness.

Using an environmental filter ($^{i}\Psi_{E}$) gave the lowest $^{i}\Psi$ values for the probabilistic species pools, but the second highest values for binary species pools. This difference is remarkable, and shows that differences in binary and probabilistic species pools can be quite substantial depending on the filters involved, which confirms the differences between species pools that emerged in the analysis of Ranunculaceae. Additionally, the simulated dataset revealed that including biotic filters into Ψ led to a close resemblance of $^{i}\Psi$ values with the realized species richness values at the grid cell level (Fig. 4).

The simulated example also shows how the strength of binary constraints in species pool definitions can be directly assessed by comparing it with probabilistic species pools, without the need to examine the entire range of possible thresholds on species pool definitions as recently proposed (Lessard *et al.*, 2016). Binary and probabilistic species pools can vary greatly in $^{i}\Psi$ along specific environmental gradients (temperature in this case). The inclusion of biotic filters only caused a small difference in $^{i}\Psi$ compared with binary values, but the difference is more pronounced for dispersal- and environmental-delineated species pools. Obviously, inference about the strength of processes can vary depending on the choice of a binary or probabilistic species pool concept. This is especially important as dispersal and environmental filters are the most commonly studied ones, no doubt because data for these are more readily available than data on biotic or demographic factors.

PROSPECTS AND LIMITATIONS OF PROBABILISTIC SPECIES POOLS

The perception of processes that shape local community composition and diversity largely depends on the circumscription of the species pool (Kraft *et al.*, 2011; Anacker & Harrison, 2012; Lessard *et al.*, 2012a,b; Karger *et al.*, 2014, 2015) and how this relates to the spatial and temporal scale of sampling (Tuomisto & Ruokolainen, 2012). Incorporating the probabilistic species pool into studies on community composition and diversity can help disentangle the strength of different processes that drive local species assembly and their geographical distribution. The analysis of Ranunculaceae shows how the null expectation generated using a probabilistic species pool can be compared with the observed community composition in the focal unit (Fig. 2). The geographical distribution of AUC values in this example demonstrates that a probabilistic species pool based on dispersal and environment ($\Psi_{DRT\times EB}$) explains species assemblage composition well in some areas but not in others. In such locations, other factors might be operational which are not included in the species pool definition (e.g. biotic interactions).

The value of $^{i}\Psi$ (equation 2) reflects the expected number of species within a focal unit. However, the physical space of the focal unit needs to be able to accommodate all those species (e.g. a focal unit of 1 m^2 might only hold one tree species, even if $^{i}\Psi > 1$). $^{i}\Psi$ might therefore be more appropriate for testing community saturation (Srivastava, 1999; Loreau, 2000).

Probabilistic species pools that are only based on regional abundance and dispersal of species (excluding environmental filters) could be used to test neutral versus deterministic theories (Hubbell, 2001). Probabilistic species pools that also include environmental filters can be used to estimate whether sampling intensity is adequate, as they allow one to quantify the likelihood of the observed number of species (see Fig. 3) within a focal unit (see Manceur & Kühn, 2014, for single-species examples).

Limitations in employing the probabilistic species pool arise mostly from the availability of data, especially on biotic factors and demography but also on dispersal traits. The limited spatial extent of data can impose a binary constraint on a probabilistic pool and are therefore somewhat similar to arbitrary thresholds. Limited data extents are, however, not a conceptual problem like arbitrary thresholds, but rather an operational one. When using a binary pool, a threshold needs to be defined to determine

when a species survives or dispersal is assumed to be sufficient. Problems inherent in this approach can be circumvented by applying a probabilistic approach. Data extents, however, impose an operational constraint which can be overcome by accumulating information on species outside the extent of the original data. Small data extents might be fully sufficient for specific temporal extents, but can lead to the exclusion of important data in others. The data extent of serpentine grassland, for example, seems incredibly small at first; but these species disperse less than about 1 m year^{-1} on average, meaning that the 8 m × 8 m plot captures much of the extent of any one cell. In general, if many species within a given data extent have about zero probability of belonging to a pool, it can be seen as an indicator of insufficient data (e.g. Figs 2 & S3).

The exact dispersal kernels over time for many species are, however, still largely unknown. For the Ranunculaceae, we simply assumed a baseline dispersal of 20 m year^{-1} across 5000 years (P_{DD}), additionally adjusted by species traits (P_{DDT}). Although this gives a temporal reference frame for the species pool, it remains to be verified if it actually reflects future dispersal events in this particular case. Such verifications are possible using repeated sampling, as in the serpentine grassland example. Examples like these highlight how important it is to integrate, or for the time being accurately state, time-related variables (dispersal/time) in species pool definitions, as they give a temporal reference frame for the study (exact values are usually not given in approaches that use regional richness or dispersion fields). Even the most recent advances in delineating species pools (Graves & Rahbek, 2005; Lessard *et al.*, 2016) still fall short in stating exact temporal extents due to the problem of unknown dispersal distances over time. To advance the ability of species pools to inform us of assembly processes, it is therefore necessary to gather exact data on dispersal distances of species and repeatedly sample communities and assemblages.

Approximating probabilities also introduces error of unknown size that can lead to problems if we want to calculate the expected number of species within a focal unit. These errors could quickly increase non-additively when combining species pools. The problem of unknown errors when assigning probabilities raises the question: does replacing one arbitrary decision involved in defining a binary species pool with many arbitrary decisions for defining other kinds of species pools yield a qualitatively better estimate? In our view, an important advantage of the probabilistic species pool is that it requires decisions about pool membership to be explicit. The calculation of probabilities makes the underlying assumptions and their associated uncertainties visible, replicable, challengeable and correctable. This may in itself encourage new research approaches towards understanding species pools and their importance in species assembly.

CONCLUSION

The probabilistic approach to species pools expands former definitions and adds useful properties to the concept. In particular, treating species pools probabilistically allows ecologists to better link focal units with species pools by using different sets of parameters, thereby determining the relative roles of major potential community assembly factors. The probabilistic approach to species pools should therefore be a valuable tool in macroecology, community ecology, climate change research, invasion ecology and ecological restoration.

ACKNOWLEDGEMENTS

D.N.K. acknowledges funding from the Swiss National Science Foundation (SNF 148691). J.S.C. acknowledges financial support from the German Research Foundation (DFG SA-21331) and from the German Initiative of Excellence. H.K. and P.W. acknowledge funding from the German Research Council (DFG) Free Floater Program in the Excellence Initiative at the University of Göttingen and in the scope of the BEFmate project from the Ministry of Science and Culture of Lower Saxony. I.K. and S.P. acknowledges funding from the German Federal Agency of Nature Conservation (project 'Flora and Climate', FKZ 80581001). We thank Cristina Castanha for providing the environmental data for the serpentine dataset and UC McLaughlin Natural Reserve for logistical support, Adriana Ruggiero, Daniel W. Carstensen and two anonymous referees for valuable comments on a former version of the manuscript, and Kith Vargas Karger for language editing. The final manuscript is a joint effort of the working group sREGPOOL and an outcome of a workshop supported by sDiv, the Synthesis Centre for Biodiversity Sciences – a unit of the German Centre for Integrative Biodiversity Research (iDiv) Halle-Jena-Leipzig, funded by the German Research Foundation (FZT 118).

REFERENCES

Algar, A.C., Kerr, J.T. & Currie, D.J. (2011) Quantifying the importance of regional and local filters for community trait structure in tropical and temperate zones. *Ecology*, **92**, 903–914.

Anacker, B.L. & Harrison, S.P. (2012) Historical and ecological controls on phylogenetic diversity in Californian plant communities. *The American Naturalist*, **180**, 257–269.

Andersen, M.C. (1993) Diaspore morphology and seed dispersal in several wind-dispersed Asteraceae. *American Journal of Botany*, **80**, 487–492.

Beals, E.W. (1984) Bray–Curtis ordination: an effective strategy for analysis of multivariate ecological data. *Advances in Ecological Research,* **14**, 1–55.

Belmaker, J. & Jetz, W. (2013) Spatial scaling of functional structure in bird and mammal assemblages. *The American Naturalist*, **181**, 464–478.

Bischoff, A. (2005) Analysis of weed dispersal to predict chances of re-colonisation. *Agriculture, ecosystems & environment*, **106**, 377–387.

Borregaard, M.K. & Rahbek, C. (2010) Dispersion fields, diversity fields and null models: uniting range sizes and species richness. *Ecography*, **33**, 402–407.

Cabral, J.S. & Kreft, H. (2012) Linking ecological niche, community ecology and biogeography: insights from a mechanistic niche model. *Journal of Biogeography*, **39**, 2212–2224.

Canning-Clode, J., Bellou, N., Kaufmann, M.J. & Wahl, M. (2009) Local–regional richness relationship in fouling assemblages – effects of succession. *Basic and Applied Ecology*, **10**, 745–753.

Carstensen, D.W., Lessard, J.-P., Holt, B.G., Krabbe Borregaard, M. & Rahbek, C. (2013) Introducing the biogeographic species pool. *Ecography*, **36**, 1310–1318.

Cavender-Bares, J., Keen, A. & Miles, B. (2006) Phylogenetic structure of Floridian plant communities depends on taxonomic and spatial scale. *Ecology*, **87**, S109–S122.

Chase, J.M. & Myers, J.A. (2011) Disentangling the importance of ecological niches from stochastic processes across scales. *Philosophical Transactions of the Royal Society B: Biological Sciences*, **366**, 2351–2363.

Clark, J.S., Silman, M., Kern, R., Macklin, E. & HilleRisLambers, J. (1999) Seed dispersal near and far: patterns across temperate and tropical forests. *Ecology*, **80**, 1475–1494.

Connor, E.F. & Simberloff, D. (1978) Species number and compositional similarity of the Galapagos flora and avifauna. *Ecological Monographs*, **48**, 219–248.

Cornell, H.V. & Harrison, S.P. (2014) What are species pools and when are they important? *Annual Review of Ecology, Evolution, and Systematics*, **45**, 45–67.

Cornell, H.V. & Lawton, J.H. (1992) Species interactions, local and regional processes, and limits to the richness of ecological communities: a theoretical perspective. *Journal of Animal Ecology*, **61**, 1–12.

de Bello, F., Price, J.N., Münkemüller, T., Liira, J., Zobel, M., Thuiller, W., Gerhold, P., Götzenberger, L., Lavergne, S., Lepš, J., Zoebel, K, & Pärtel, M. (2012) Functional species pool framework to test for biotic effects on community assembly. *Ecology*, **93**, 2263–2273.

De Cáceres, M. & Legendre, P. (2008) Beals smoothing revisited. *Oecologia*, **156**, 657–669.

Eiserhardt, W.L., Svenning, J.-C., Baker, W.J., Couvreur, T.L.P. & Balslev, H. (2013) Dispersal and niche evolution jointly shape the geographic turnover of phylogenetic clades across continents. *Scientific Reports*, **3**, 1164.

González-Caro, S., Parra, J.L., Graham, C.H., McGuire, J.A. & Cadena, C.D. (2012) Sensitivity of metrics of phylogenetic structure to scale, source of data and species pool of hummingbird assemblages along elevational gradients. *PLoS ONE*, **7**, e35472.

Götzenberger, L., de Bello, F., Bråthen, K.A., Davison, J., Dubuis, A., Guisan, A., Lepš, J., Lindborg, R., Moora, M., Pärtel, M., Pellissier, L., Pottier, J., Vittoz, P., Zobel, K. & Zobel, M. (2012) Ecological assembly rules in plant communities – approaches, patterns and prospects. *Biological Reviews*, **87**, 111–127.

Graves, G.R. & Rahbek, C. (2005) Source pool geometry and the assembly of continental avifaunas. *Proceedings of the National Academy of Sciences USA*, **102**, 7871–7876.

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

Harrison, S. & Cornell, H. (2008) Toward a better understanding of the regional causes of local community richness. *Ecology Letters*, **11**, 969–979.

Herben, T. (2005) Species pool size and invasibility of island communities: a null model of sampling effects. *Ecology Letters*, **8**, 909–917.

Hintze, C., Heydel, F., Hoppe, C., Cunze, S., König, A. & Tackenberg, O. (2013) D3: the dispersal and diaspore database – baseline data and statistics on seed dispersal. *Perspectives in Plant Ecology, Evolution and Systematics*, **15**, 180–192.

Hubbell, S.P. (2001) *The unified neutral theory of biodiversity and biogeography*. Princeton University Press, Princeton, NJ.

Janin, A., Léna, J.-P., Ray, N., Delacourt, C., Allemand, P. & Joly, P. (2009) Assessing landscape connectivity with calibrated cost–distance modelling: predicting common toad distribution in a context of spreading agriculture. *Journal of Applied Ecology*, **46**, 833–841.

Janssens, F., Peeters, A., Tallowin, J.R.B., Bakker, J.P., Bekker, R.M., Fillat, F. & Oomes, M.J.M. (1998) Relationship between soil chemical factors and grassland diversity. *Plant and Soil*, **202**, 69–78.

Karger, D.N., Weigelt, P., Amoroso, V.B., Darnaedi, D., Hidayat, A., Kreft, H. & Kessler, M. (2014) Island biogeography from regional to local scales: evidence for a spatially scaled echo pattern of fern diversity in the Southeast Asian archipelago. *Journal of Biogeography*, **41**, 250–260.

Karger, D.N., Tuomisto, H., Amoroso, V.B., Darnaedi, D., Hidayat, A., Abrahamczyk, S., Kluge, J., Lehnert, M. & Kessler, M. (2015) The importance of species pool size for community composition. *Ecography*, **38**, 1243–1253.

Kissling, W.D., Dormann, C.F., Groeneveld, J., Hickler, T., Kühn, I., McInerny, G.J., Montoya, J.M., Römermann, C., Schiffers, K., Schurr, F.M., Singer, A., Svenning, J.-C., Zimmermann, N.E. & O'Hara, R.B. (2012) Towards novel approaches to modelling biotic interactions in multispecies assemblages at large spatial extents. *Journal of Biogeography*, **39**, 2163–2178.

Kluge, J., Kessler, M. & Dunn, R.R. (2006) What drives elevational patterns of diversity? A test of geometric constraints, climate and species pool effects for pteridophytes on an elevational gradient in Costa Rica. *Global Ecology and Biogeography*, **15**, 358–371.

Knop, E., Schmid, B. & Herzog, F. (2008) Impact of regional species pool on grasshopper restoration in hay meadows. *Restoration Ecology*, **16**, 34–38.

Kraft, N.J., Cornwell, W.K., Webb, C.O. & Ackerly, D.D. (2007) Trait evolution, community assembly, and the phylogenetic structure of ecological communities. *The American Naturalist*, **170**, 271–283.

Kraft, N.J.B., Comita, L.S., Chase, J.M., Sanders, N.J., Swenson, N.G., Crist, T.O., Stegen, J.C., Vellend, M., Boyle, B., Anderson,

M.J., Cornell, H.V., Davies, K.F., Freestone, A.L., Inouye, B.D., Harrison, S.P. & Myers, J.A. (2011) Disentangling the drivers of beta diversity along latitudinal and elevational gradients. *Science*, **333**, 1755–1758.

Kühn, I., Bierman, S.M., Durka, W. & Klotz, S. (2006) Relating geographical variation in pollination types to environmental and spatial factors using novel statistical methods. *New Phytologist*, **172**, 127–139.

Lessard, J.-P., Belmaker, J., Myers, J.A., Chase, J.M. & Rahbek, C. (2012a) Inferring local ecological processes amid species pool influences. *Trends in Ecology and Evolution*, **27**, 600–607.

Lessard, J.-P., Borregaard, M.K., Fordyce, J.A., Rahbek, C., Weiser, M.D., Dunn, R.R. & Sanders, N.J. (2012b) Strong influence of regional species pools on continent-wide structuring of local communities. *Proceedings of the Royal Society B: Biological Sciences*, **279**, 266–274.

Lessard, J.-P., Weinstein, B.G., Borregaard, M.K., Marske, K.A., Martin, D.R., McGuire, J.A., Parra, J.L., Rahbek, C., Graham, C.H., Harrison, A.E.S. & Bronstein, E.J.L. (2016) Process-based species pools reveal the hidden signature of biotic interactions amid the influence of temperature filtering. *The American Naturalist*. Available at: http://www.jstor.org/stable/10.1086/684128.

Loreau, M. (2000) Are communities saturated? On the relationship between α, β and γ diversity. *Ecology Letters*, **3**, 73–76.

MacArthur, R.H. & Wilson, E.O. (1967) *The theory of island biogeography*, Princeton University Press, Princeton NY.

McRae, B.H. (2006) Isolation by resistance. *Evolution*, **60**, 1551–1561.

McRae, B.H., Dickson, B.G., Keitt, T.H. & Shah, V.B. (2008) Using circuit theory to model connectivity in ecology, evolution, and conservation. *Ecology*, **89**, 2712–2724.

Manceur, A.M. & Kühn, I. (2014) Inferring model-based probability of occurrence from preferentially sampled data with uncertain absences using expert knowledge. *Methods in Ecology and Evolution*, **5**, 739–750.

Mucher, C.A., Steinnocher, K.T., Kressler, F.P. & Heunks, C. (2000) Land cover characterization and change detection for environmental monitoring of pan-Europe. *International Journal of Remote Sensing*, **21**, 1159–1181.

Muller-Landau, H.C., Wright, S.J., Calderón, O., Condit, R. & Hubbell, S.P. (2008) Interspecific variation in primary seed dispersal in a tropical forest. *Journal of Ecology*, **96**, 653–667.

Munguia, P. (2004) Successional patterns on pen shell communities at local and regional scales. *Journal of Animal Ecology*, **73**, 64–74.

Mücher, C.A., Champeaux, J.L., Steinnocher, K.T., Griguolo, S., Wester, K., Heunks, C., Winiwater, W., Kressler, F.P., Goutorbe, J.P., ten Brink, B., Katwijk, V.F., van Furberg, O., Perdigao, V. & Nieuwenhuis, G.J.A. (2001) *Development of a consistent methodology to derive land cover information on a European scale from remote sensing for environmental monitoring; the PELCOM report*. Alterra, Wageningen.

Myers, J.A. & Harms, K.E. (2009) Seed arrival, ecological filters, and plant species richness: a meta-analysis. *Ecology Letters*, **12**, 1250–1260.

Normand, S., Ricklefs, R.E., Skov, F., Bladt, J., Tackenberg, O. & Svenning, J.-C. (2011) Postglacial migration supplements climate in determining plant species ranges in Europe. *Proceedings of the Royal Society of London B: Biological Sciences*, **278**, 3644–3653.

Ozinga, W.A., Römermann, C., Bekker, R.M., Prinzing, A., Tamis, W.L.M., Schaminée, J.H.J., Hennekens, S.M., Thompson, K., Poschlod, P., Kleyer, M., Bakker, J.P. & Van Groenendael, J.M. (2009) Dispersal failure contributes to plant losses in NW Europe. *Ecology Letters*, **12**, 66–74.

Panagos, P. (2006) The European soil database. *GEO: Connexion*, **5**, 32–33.

Pärtel, M., Zobel, M., Zobel, K. & van der Maarel, E. (1996) The species pool and its relation to species richness: evidence from Estonian plant communities. *Oikos*, **75**, 111–117.

Pärtel, M., Szava-Kovats, R. & Zobel, M. (2011) Dark diversity: shedding light on absent species. *Trends in Ecology and Evolution*, **26**, 124–128.

Pompe, S., Hanspach, J., Badeck, F., Klotz, S., Thuiller, W. & Kühn, I. (2008) Climate and land use change impacts on plant distributions in Germany. *Biology Letters*, **4**, 564–567.

Ricklefs, R.E. (1987) Community diversity – relative roles of local and regional processes. *Science*, **235**, 167–171.

Ricklefs, R.E. & Schluter, D. (1994) *Species diversity in ecological communities*. University of Chicago Press, Chicago.

Ronk, A., Szava-Kovats, R. & Pärtel, M. (2015) Applying the dark diversity concept to plants at the European scale. *Ecography*, **38**, 1015–1025.

Shurin, J.B. & Srivastava, D.S. (2005) New perspectives on local and regional diversity: beyond saturation. *Metacommunities* (ed. by M. Holyoak, R. Holt and M. Leibold), pp. 399–417. University of Chicago Press, Chicago.

Smith, A.B., Sandel, B., Kraft, N.J.B. & Carey, S. (2013) Characterizing scale-dependent community assembly using the functional-diversity–area relationship. *Ecology*, **94**, 2392–2402.

Srivastava, D.S. (1999) Using local–regional richness plots to test for species saturation: pitfalls and potentials. *Journal of Animal Ecology*, **68**, 1–16.

Starzomski, B.M., Parker, R.L. & Srivastava, D.S. (2008) On the relationship between regional and local species richness: a test of saturation theory. *Ecology*, **89**, 1921–1930.

Swenson, N.G. & Enquist, B.J. (2009) Opposing assembly mechanisms in a Neotropical dry forest: implications for phylogenetic and functional community assembly. *Ecology*, **90**, 2161–2170.

Thomson, F.J., Moles, A.T., Auld, T.D. & Kingsford, R.T. (2011) Seed dispersal distance is more strongly correlated with plant height than with seed mass. *Journal of Ecology*, **99**, 1299–1307.

Tuomisto, H. & Ruokolainen, K. (2012) Comment on 'disentangling the drivers of β diversity along latitudinal and elevational gradients. *Science*, **335**, 1573.

Weigelt, P. & Kreft, H. (2013) Quantifying island isolation – insights from global patterns of insular plant species richness. *Ecography*, **36**, 417–429.

White, E.P. & Hurlbert, A.H. (2010) The combined influence of the local environment and regional enrichment on bird species richness. *The American Naturalist*, **175**, E35–E43.

Zobel, M. (1997) The relative role of species pools in determining plant species richness: an alternative explanation of species coexistence? *Trends in Ecology and Evolution*, **12**, 266–269.

Zobel, M., Maarel, E. & Dupré, C. (1998) Species pool: the concept, its determination and significance for community restoration. *Applied Vegetation Science*, **1**, 55–66.

SUPPORTING INFORMATION

Additional Supporting Information may be found in the online version of this article:

Appendix S1 Examples of species probabilities (P) for five species from the serpentine grasslands dataset.

Appendix S2 Performance (area under the receiver operating characteristic curve, AUC) of different fixed dispersal values (k) that distinguish which cells a species will colonize compared with the AUC of trait-based dispersal values (P_{DDT}) in the serpentine grassland dataset.

Appendix S3 Sensitivity of probability values (P) to reductions in data completeness in the serpentine grasslands dataset.

Appendix S4 Probability distributions and richness of *Clematis recta* based on different methods of approximation.

Appendix S5 Comparison of species pool size index values ($^{a}\Psi$) and species pool membership values (P).

Appendix S6 Importance of different factors for species pool delineations and their causal links.

BIOSKETCH

The idea of probabilistic species pools is a joint and equal effort of the working group sREGPOOL of the German Center for Integrative Biodiversity Research (iDiv) (http://www.idiv-biodiversity.de/de/sdiv/workshops/workshops-2013/sregpool).

D.N.K., I.K., J.-C.S., M.K., H.K. and H.T. developed the theoretical concept. A.F.C., S.P., P.W. and K.W. analysed the Ranunculaceae data, A.B.S. and B.S. performed analysis with the serpentine grassland data, J.S.C. performed the work on the simulated data. D.N.K. wrote the first draft of the manuscript, and all authors contributed substantially to revisions.

Editor: Adriana Ruggiero