Global Ecology and Biogeography, (Global Ecol. Biogeogr.) (2014) **23**, 99–112

A Journal of Macroecology

Global Ecology and Biogeography

RESEARCH PAPER

Stacking species distribution models and adjusting bias by linking them to macroecological models

Justin M. Calabrese[1]*, Grégoire Certain[2], Casper Kraan[3] and Carsten F. Dormann[4]

[1]*Conservation Ecology Center, Smithsonian Conservation Biology Institute, National Zoological Park, 1500 Remount Rd., Front Royal, VA 22630, USA,* [2]*Institute of Marine Research, 9019 Tromsø, Norway,* [3]*National Institute of Water and Atmospheric Research, Hamilton 3216, New Zealand,* [4]*Biometry and Environmental System Analysis, University of Freiburg, 79104 Freiburg, Germany*

*Correspondence: Justin M. Calabrese, Conservation Ecology Center, Smithsonian Conservation Biology Institute, National Zoological Park, 1500 Remount Road, Front Royal, VA 22630, USA.
E-mail: calabresej@si.edu

ABSTRACT

Aim Species distribution models (SDMs) are common tools in biogeography and conservation ecology. It has been repeatedly claimed that aggregated (stacked) SDMs (S-SDMs) will overestimate species richness. One recently suggested solution to this problem is to use macroecological models of species richness to constrain S-SDMs. Here, we examine current practice in the development of S-SDMs to identify methodological problems, provide tools to overcome these issues, and quantify the performance of correctly stacked S-SDMs alongside macroecological models.

Locations Barents Sea, Europe and Dutch Wadden Sea.

Methods We present formal mathematical arguments demonstrating how S-SDMs should and should not be stacked. We then compare the performance of macroecological models and correctly stacked S-SDMs on the same data to determine if the former can be used to constrain the latter. Next, we develop a maximum-likelihood approach to adjusting S-SDMs and discuss how it could potentially be used in combination with macroecological models. Finally, we use this tool to quantify how S-SDMs deviate from observed richness in four very different case studies.

Results We demonstrate that stacking methods based on thresholding site-level occurrence probabilities will almost always be biased, and that these biases will tend toward systematic overprediction of richness. Next, we show that correctly stacked S-SDMs perform very similarly to macroecological models in that they both have a tendency to overpredict richness in species-poor sites and underpredict it in species-rich sites.

Main conclusions Our results suggest that the perception that S-SDMs consistently overpredict richness is driven largely by incorrect stacking methods. With these biases removed, S-SDMs perform similarly to macroecological models, suggesting that combining the two model classes will not offer much improvement. However, if situations where coupling S-SDMs and macroecological models would be beneficial are subsequently identified, the tools we develop would facilitate such a synthesis.

Keywords

Boosted regression trees, Kumaraswamy distribution, macroecological models, maximum likelihood, poisson binomial distribution, richness regression models, species richness, stacked species distribution models.

DOI: 10.1111/geb.12102
http://wileyonlinelibrary.com/journal/geb

INTRODUCTION

Species distribution models (SDMs) have become standard tools in biogeography and conservation biology, both for understanding the factors that affect species' geographical ranges and for predicting the response of species to global change (Scott *et al.*, 2002; Franklin, 2009; Peterson *et al.*, 2011). The increasing availability of large-scale, multispecies data sets, coupled with advances in modelling techniques and software, has resulted in the development of SDMs for all or many constituent species of some communities (Thuiller, 2003; Elith *et al.*, 2006). When community-wide SDM coverage exists, it is natural to attempt to combine the species-level models into a descriptor of community-level properties such as species richness. The appeal of this approach is that it may facilitate an easy assessment of site-level biodiversity from knowledge of a handful of readily measurable – or perhaps already available – covariates. The process by which SDMs are combined into community-level models is often referred to as 'stacking' (e.g. Ferrier & Guisan, 2006), so these models of species richness are also called stacked species distribution models (S-SDMs, e.g. Mateo *et al.*, 2012). Despite their potential, current evidence suggests that individual SDMs do not aggregate well into an unbiased description of species richness (Guisan & Rahbek, 2011). Specifically, S-SDMs are thought to systematically overpredict site-level richness (Guisan & Rahbek, 2011; Hortal *et al.*, 2012).

To remedy this situation, Guisan & Rahbek (2011) proposed an integrated framework, SESAM, that starts with site-level richness and species composition predictions from S-SDMs, and then progressively refines these predictions, by applying macroecological models (MEMs), dispersal filters and ecological assembly rules. They suggest that in addition to delivering better predictions of richness than current S-SDMs, this approach would also yield improved prediction of species composition via the collection of refined site-level SDM occurrence probabilities. The core assumptions upon which SESAM rests are that: (1) S-SDMs consistently overpredict richness relative to MEMs; and (2) the reason for this discrepancy is that S-SDMs lack biotic filters such as dispersal limitation and ecological assembly rules (Guisan & Rahbek, 2011; Hortal *et al.*, 2012). Although SESAM represents a bold attempt at synthesis and integration, we feel that its foundational assumptions need to be examined before it can be accepted, and several technical hurdles must be overcome before it could be implemented.

First, before pursuing biotic explanations for the frequently observed discrepancy between S-SDMs and MEMs, which might require significant new research effort, simple statistical artifacts introduced by the way S-SDMs are built should be ruled out. A key candidate for such an artifactual cause is the currently common practice of applying thresholds to SDM-predicted occurrence probabilities to produce binary presence/absence predictions. These presence/absence predictions are then summed for each site to 'stack' the S-SDM. We are aware of no formal, theoretical justification for this practice, and the large array of different ad hoc thresholding schemes currently proliferating in the literature (Table 1) suggests there is confusion around this issue. Furthermore, the few studies that have compared S-SDMs stacked both with and without thresholds show dramatic differences between the two approaches (Aranda & Lobo, 2011; Dubuis *et al.*, 2011). A deeper exploration of the process of stacking individual SDMs into an S-SDM is therefore clearly warranted.

Second, the assumption that S-SDMs overpredict richness relative MEMs needs to be evaluated. MEMs have contributed substantially to our understanding of large-scale ecology and biodiversity, and have been valuable research tools (Brown, 1995; Gaston, 2000; Gaston & Blackburn, 2000; Hawkins & Diniz-Filho, 2004; Kerr *et al.*, 2007; Algar *et al.*, 2009). We might therefore expect MEMs to predict site-level richness well, and probably better and more consistently than S-SDMs, which have substantial problems dealing with rare species (Graham & Hijmans, 2006; Pineda & Lobo, 2009; Guisan & Rahbek, 2011). Although this assumption is a core premise of the SESAM framework, we are not aware of any systematic studies of how accurately MEMs and S-SDMs predict observed species richness on the same data sets. Given that MEMs and S-SDMs tend to use the same predictor variables, it could be that S-SDMs, when correctly stacked, perform similarly to MEMs.

Third, we explore how correctly-stacked S-SDMs deviate from observed richness. Understanding how SDM occurrence probabilities need to be adjusted to bring S-SDM predictions in line with observed richness should allows us to examine the specific ways in which correctly stacked S-SDMs are deficient, and should also suggest how they may be improved in the future. Such adjustments would also be necessary to use MEMs to constrain S-SDMs, as suggested by Guisan & Rahbek (2011), should further evidence suggest that such an approach is warranted, but the methods to do this are currently lacking.

Here, we bring together four diverse data sets to explore how S-SDMs perform relative to MEMs. First, we present the relevant probability theory to establish the distribution of a site-level richness prediction under an S-SDM and the correct (and exact) way to stack individual SDMs into such a prediction. We review the S-SDM literature to document the different stacking approaches that have been used, most of which are based on thresholding site-level occurrence probabilities, and then present a mathematical argument to demonstrate why thresholding will generally be incorrect. Next, we examine the accuracy of the species richness predictions of both correctly stacked S-SDMs and MEMs side-by-side on our four data sets. Finally, we develop a maximum-likelihood approach to adjusting S-SDM occurrence probabilities given some estimate or prediction of site-level species richness, and then apply this tool to our data sets to quantify the ways in which correctly-stacked S-SDMs deviate from observed site-level richness.

Case studies

The four different data sets we use to illustrate our points throughout the paper were chosen to represent contrasting properties, in an attempt to maximize the generality of any pattern we might detect.

Table 1 Non-exhaustive chronology of studies stacking SDMs to yield richness estimates (to July 2012). (For explanation of goodness-of-classification measures, see Fielding & Bell, 1997).

Study	Method*	Organism group	Comments†
Skov & Borchsenius, 1997	Eqn 2	Ecuadorian palms	
Guisan *et al.*, 1999	*T* κ	Nevada mountain plants	S-SDM compared to CCA, not richness
Erasmus *et al.*, 2002	*T* CCR	Various taxa, South Africa	Birds, mammals, reptiles, butterflies
Ferrier *et al.*, 2002	Eqn 2	various NZ taxa	
Lehmann *et al.*, 2002	Eqn 2	NZ ferns	
Loiselle *et al.*, 2003	*T* (0.5, 0.85, 0.95)	Brazilian birds	
Peppler-Lisbach & Schröder, 2004	*T* κ	Grassland plants	Community composition, not richness
Graham & Hijmans, 2006	*T* κ	Amphibians and reptiles, California	MaxEnt; 159 species
Loiselle *et al.*, 2007	*T* (≈ 0.01)	Plant richness, Bolivia/Ecuador	MaxEnt; one threshold for all species
Algar *et al.*, 2009	*T* LS	Canadian butterflies	MaxEnt; S-SDM & MEM extremely similar
Pineda & Lobo, 2009	*T* (21 different)	Mexican amphibians	MaxEnt; same threshold for all species
Raes *et al.*, 2009	*T* (0.1)	Plant richness, Borneo	MaxEnt; same threshold for all species
Randin *et al.*, 2009	*T* CCR	Swiss alpine plants	Biomod
Barbet-Massin *et al.*, 2010	*T* SES	Iberian/African birds	Biomod
Buisson *et al.*, 2010	*T* TSS	French stream fish	Biomod
Aranda & Lobo, 2011	Eqn 2, *T* (various)	Plants, Tenerife, Spain	MaxEnt; same threshold for all species
Dubuis *et al.*, 2011	Eqn 2, *T* TSS, *B*	Swiss plants	Huge different between Eqn 2 and *T*
Fitzpatrick *et al.*, 2011	*T* TSS	North American ants	
Kaschner *et al.*, 2011	*T* (various)	World marine mammals	Same threshold for all species
Mateo *et al.*, 2012	*T* (min. commission)	Two Andean plant groups	Ensemble; max. commission error 0.05
Rondinini *et al.*, 2011	*T* (3 unknown values)	Global mammals	
Pineda & Lobo, 2012	*T* (21 different)	Mexican amphibians	MaxEnt; same threshold for all species
Schmidt-Lebuhn *et al.*, 2012	Eqn 2	Australian Asteraceae	MaxEnt; suitability interpreted as probability

*Methods: Eqn 2, summing individual probabilities; *T*, thresholding then summing; *B*, drawing repeatedly from a Bernoulli distribution. Thresholds used: κ, maximizing kappa; CCR, maximizing correct classification rate; TSS, maximizing true skill statistic (sensitivity + specificity −1); SES, threshold for which sensitivity equals specificity; LS, threshold determined by the lowest probability under which a species has been observed. Numeric values refer to the threshold(s) used.

†Biomod (Thuiller, 2003; Thuiller *et al.*, 2009) has built-in facilities to average models and may automatically threshold predicted occurrence probabilities; MaxEnt (Phillips *et al.*, 2006) refers to a common modelling approach that does not yield probabilities, but a relative index of habitat suitability ranging from 0 to 100 (Elith *et al.*, 2010). This index is still often converted into presence/absence data and stacked in the same way as probabilities.

The 'Wadden Sea macrozoobenthos' data (Kraan *et al.*, 2007, 2009, 2010) were collected in 2005 and represent 2549 sites in which the abundance of 24 organisms (polychaetes, bivalves and crustaceans) was sampled to the lowest taxonomic level. These sites are arranged in a fixed grid with 250 m intervals between sites, encompassing 225 km^2 of soft-sediment flats which are exposed during low tide. Two environmental predictors (median grain size and elevation relative to the Dutch ordnance datum) were shown to be consistent and good predictors for these species (Kraan *et al.*, 2013).

'EU forest trees' is a gridded (20 km × 20 km) version of the distribution of forest trees in the member states of the European Union (Köble & Seufert, 2001). This data set of 111 species represents a satellite-based classification trained on national forest inventories, and is restricted to forested area, so does not represent the full range of environmental conditions under which these species may occur outside forests (e.g. in parks and gardens). As predictor variables, we used temperature seasonality (standard deviation × 100), maximum temperature of the warmest month, mean temperature of the coldest quarter, annual precipitation, precipitation of the driest quarter, and precipitation of the warmest quarter (these are WorldClim's Bioclim variables 4, 5, 11, 12, 17, 18; Hijmans *et al.*, 2005). Additionally, we used growing degree days (http://www.sage.wisc.edu/atlas/maps.php?datasetid=31&includerelatedlinks=1&dataset=31) and water balance (http://edit.csic.es/Climate.html). These variables were not problematically collinear ($|r| < 0.7$; Dormann *et al.*, 2013).

The 'Barents Sea trawls' data set was collected each summer (August–September) from 2004 to 2008, following a regular sampling scheme of stations separated by 30–40 km (Jakobsen & Ozhigin, 2011; Johannesen *et al.*, 2012). The standard towing time was 15 min at 3 knots, equivalent to a towing distance of 0.75 nautical miles. Data were collected for 81 taxonomic groups of fishes (mainly species) and 29 taxonomic groups of invertebrates (Johannesen *et al.*, 2012). The community is restricted to the 79 species that were frequent enough for an SDM to be implemented. We used 25 environmental predictors including: bottom depth; bottom depth gradient; bottom temperature in August–October averaged over the last 30 years; year bottom temperature anomaly; year bottom temperature gradient; surface temperature in August–October averaged over the last

Table 2 Case study characteristics, sorted by increasing mean site-level richness, $S_j^{(obs)}$.

Data set	Sample size	Mean $S_j^{(obs)}$ (min, max)	Reference
Wadden Sea benthos	2549	4.0 (1, 11)	Kraan *et al.* (2010)
EU forest trees	13038	4.9 (1, 21)	Köble & Seufert (2001)
Barents Sea trawls	2457	16.4 (1, 33)	Johannesen *et al.* (2012)
EU mammals	3036	44.2 (1, 73)	Dormann *et al.* (2010)

30 years; year surface temperature anomaly; year surface temperature gradient; bottom salinity in August–October averaged over the last 30 years; year bottom salinity anomaly; year bottom salinity gradient; surface salinity in August–October averaged over the last 30 years; year surface salinity anomaly; year surface salinity gradient; gradient of potential energy deficit; mixed layer depth; gradient of mixed layer depth; surface concentration of chlorophyll *a* (CHLa); gradient of surface concentration of CHLa; number of ice-covered months in the last 10 months before the summer; median velocity of the bottom current taken in January–March; horizontal component of the bottom current vector averaged over January–March; gradient of horizontal component of the bottom current vector averaged over January–March, vertical component of the bottom current vector averaged over January–March, and gradient of vertical component of the bottom current vector averaged over January–March. None of these predictors were problematically collinear ($|r| < 0.7$; Dormann *et al.*, 2013).

'EU mammals' is a 50 × 50 km gridded version of the European Mammal Assessment (Temple & Terry, 2007). It comprises 140 species in 3036 grid cells. We used 13 uncorrelated environmental predictors, of which five were climatic (growing degree days, annual precipitation, summer precipitation, temperature seasonality and residuals of absolute minimum temperature), six were related to land cover (proportion of crop, grassland, mosaic habitat, shrubland, urban and forest) and two were topographic (residuals of mean elevation, residuals of slope) (see Dormann *et al.*, 2010, for details).

All data were analysed as presence/absence data using boosted regression trees (BRTs; Elith *et al.*, 2008). We used a tree complexity of three, a learning rate of 0.005 and 5-fold cross-validation. For common species, we increased the learning rate to 0.025, aiming for 2000 to 5000 trees per species. In a previous analysis (Dormann *et al.*, 2010), these settings worked well, yielding BRT models with high discriminatory power. We did not investigate the species-specific models in any detail, as we used these models merely as an approach to generate expected probabilities of occurrence under different environmental conditions. To verify that our results were not driven by the BRTs overfitting the data, in Appendix S1 we used GLMs (with quadratic terms and first-order interactions) combined with a conservative approach to model selection to yield more parsimonious SDMs, and then performed the same set of analyses that we describe below on these GLM-based SDMs.

Species richness was analysed in the same way, separately assuming Poisson or normal distributions and using the one that better fitted the data (based on a lower BIC). Observed species richness $S_j^{(obs)}$ was computed as the sum of species recorded for site j and analysed in the MEM also with BRTs, trying both Poisson and normal distributions. In each case, the residual diagnostics indicated the normal model to be slightly better.

Summary statistics for the data sets can be found in Table 2. The EU mammals data are typical of range-map analyses. The EU forest trees data are similar, but much more restricted in their scope, as they only include occurrences in forests. This could introduce a strong bias, because we are thus modelling forest use, rather than a physiological niche. Both marine data sets are based on 'point' samples (even if the points are half-hour trawls), i.e. data with high accuracy for the sampled site. This kind of data is inherently much better than range maps or atlas data, because environmental conditions can be quantified without the loss of subscale variability than is inevitable with grid-based analyses (see Rocchini *et al.*, 2011, and Beck *et al.*, 2012, for recent discussions of data quality issues). However, some predictors used in the analysis (Barents Sea: bottom temperature, speed of currents; Wadden Sea: elevation) are based on models of ocean currents (Barents Sea) or interpolated from a lower-resolution sampling scheme (Wadden Sea) and may thus be of lower quality. Furthermore, these single-visit point samples are very likely to have detected only a proportion of the local community actually present, thus consistently underestimating true local richness.

Distribution theory for stacking SDMs

Stacked SDM predictions are built from the $J \times K$ matrix of SDM-predicted occurrence probabilities, $\mathbf{P}$, where J and K are the number of sites and species in the data set, respectively. We use the term 'site' here loosely to mean a location corresponding to a model prediction, which may range from a single point where sampling was conducted to the much coarser grid cells of range-map analyses. The matrix $\mathbf{P}$ has elements $p_{j,k}$, each of which is the occurrence probability of the k-th species at the j-th site. Let the $1 \times K$ vector, $\mathbf{p}_j$, represent the occurrence probabilities of each species at site j. The occurrence of the k-th species at the j-th site is a Bernoulli trial (like a coin toss) with probability of success $p_{j,k}$. As in all other studies that sum the individual SDM occurrence probabilities or presence/absence predictions, we assume that all K Bernoulli trials at site j are independent. The species richness at the j-th site is then the sum of K independent but non-identical Bernoulli trials with probability of success vector $\mathbf{p}_j$.

Given these considerations, the site-level species richness prediction, S_j, follows a Poisson binomial (sometimes also called Poisson's binomial) distribution with probability mass function (Wang, 1993; Fernandez & Williams, 2010)

$$\Pr(S_j|\mathbf{p}_j)=\frac{1}{K+1}\sum_{n=0}^{K}\left(e^{-i2\pi nS_j/(K+1)}\prod_{k=1}^{K}\left[p_{j,k}e^{\frac{i2\pi n}{K+1}}+(1-p_{j,k})\right]\right), \quad (1)$$

where $i=\sqrt{-1}$ is the imaginary unit. The expected value (mean) and variance of S-SDM predictions for the j-th site under the Poisson binomial distribution are

$$\mathrm{E}(S_j)=\sum_{k=1}^{K}p_{j,k} \quad (2)$$

$$\mathrm{Var}(S_j)=\sum_{k=1}^{K}(1-p_{j,k})\,p_{j,k}. \quad (3)$$

Equations 2 & 3 are both exact and provide a formal theoretical basis for stacking S-SDMs. Given the $\mathbf{p}_j$, the expected value of the Poisson binomial distribution, Eqn 2, is the best predictor of site-level species richness. This means that the proper way to aggregate individual SDMs into an S-SDM given the above assumptions is simply to sum the site-level occurrence probabilities. Equation 3 is the exact variance of the site-level richness prediction assuming the $\mathbf{p}_j$ are fixed, known quantities. In reality, the $\mathbf{p}_j$ are estimated with uncertainty, and ignoring this uncertainty would result in misleadingly narrow confidence intervals on S_j. Error propagation techniques (e.g. Chapter 5 in Clark, 2007) could be used in combination with Eqn 3 to fully account for uncertainty in the site-level richness prediction.

Why thresholds should not be used to stack SDMs

In practice, many studies have not applied Eqn 2, but have instead employed various thresholding schemes to convert site-level occurrence probabilities into binary presence/absence predictions, which are then summed (Table 1). The arguments in the previous section provide a formal justification, grounded in probability theory, for stacking via Eqn 2. However, despite the commonness of stacking via thresholded occurrence probabilities (Table 1), we were unable to find a similar theoretical justification for this practice in the literature. Although thresholding methods that account for differences in prevalence among species have been shown to perform somewhat better (Liu *et al.*, 2005), Table 1 illustrates that a single global threshold across species differing vastly in prevalence has even been used. Our goal in this section is to show that thresholding schemes, even if they account for species-specific prevalences, will generally yield incorrect results when the aim is to construct an S-SDM from a set of SDMs.

Under a thresholding scheme, the presence or absence of the k-th species at the j-th site is given by the indicator

$$I(p_{j,k},T_{j,k})=\begin{cases}1 & \text{if } p_{j,k}>T_{j,k}\\ 0 & \text{if } p_{j,k}\le T_{j,k}\end{cases},$$

where the threshold, T, is, for now, both site-specific and species-specific. Thresholded richness at the j-th site is then calculated as

$$S_j^{(\mathrm{tsh})}=\sum_{k=1}^{K}I(p_{j,k},T_{j,k}), \quad (4)$$

where we have used a superscript in parentheses as a label on S_j. The values of $p_{j,k}$ are estimated with uncertainty and can thus be described by probability distributions. Under repeated sampling, the average probability of success (occurrence) of the k-th species at site j over N independent realizations of $p_{j,k}$ is $\bar{p}_{j,k}=\frac{1}{N}\sum_{l=1}^{N}p_{j,k,l}$, where $p_{j,k,l}$ denotes the l-th realization of $p_{j,k}$.

We define the exact threshold for species k at site j such that, as N becomes large, the proportion of thresholded presences is $\bar{p}_{j,k}$. This threshold is exact in the sense that, if all species-by-site combinations had such a threshold, the average result of summing the thresholded presence values for each site j would converge to Eqn 2 as N approaches infinity. The exact threshold is defined as

$$T_{j,k}=Q_{j,k}(\bar{p}_{j,k}), \quad (5)$$

where $Q_{j,k}(y)$ represents the y-th quantile of the probability distribution of $p_{j,k}$. This thresholding scheme is exact only on average over a large number of realizations. Furthermore, this scheme requires $J\times K$ unique thresholds, making it unreasonable in practice.

We now consider the most commonly used thresholding approach (prevalence-based thresholding), where each species has a single threshold, T_k, across all sites. Let $\mathbf{p}_k$ be the vector of occurrence probabilities of the k-th species across all J sites. The values of $\mathbf{p}_k$ are described by a probability distribution with support on [0, 1], which we will refer to as the 'parent distribution'. We further assume that the J sites can be ordered according to their suitability for the k-th species, and that the occurrence probabilities of the k-th species tend to increase with site suitability.

A single random realization of $\mathbf{p}_k$ is generated by drawing J independent values from the parent distribution and ordering them from minimum to maximum value. The minimum is assigned to the least suitable site for species k, the second smallest is assigned to site with the second lowest suitability, and so on until the sample maximum is assigned to the most suitable site. Each of these ordered occurrence probabilities is an order statistic and, under repeated sampling, is described by a probability distribution related to the parent distribution (Casella & Berger, 2002).

For convenience, we label the sites according to their suitability for the k-th species, such that site $j=1$ is the least suitable and $j=J$ is the most suitable. In general, the j-th order statistic of a random sample of size J drawn from a parent distribution with parameter vector θ has expected value $m_j(J,\theta)$, and quantile function $Q_j(y;J,\theta)$. A key property of order statistics is that the functional forms of the mean and quantile functions depend on j, and thus, in our setup, are different at each site. In other words,

the first moments m_1 and m_2, for example, do not just have different numerical values but are different functions. The same is true of the quantile functions. We provide a specific example of this property in Appendix S2. From Eqn 5, a single, species-specific threshold for all sites would be exact only if

$$T_k = Q_j(m_j(J, \theta_k); J, \theta_k) \text{ for all } j. \quad (6)$$

Notice that for species k, J and θ_k will be the same for all sites. However, because each site represents a different order statistic, the functional forms of both m_j and Q_j change at each site j. The particular way in which the functional forms of the mean and quantile function change across the different order statistics (sites) is determined entirely by the parent distribution. However, as all of these functions are very likely to be non-linear in θ_k and J, the conditions under which Eqn 6 would hold across all sites seem to be exceptionally narrow. We are not aware of any probability distribution with support on [0, 1] that would satisfy Eqn 6, and we suspect these conditions will never be met in practice. In Appendix S2, we assume the parent distribution is a Kumaraswamy distribution (an alternative to the beta distribution, which is more mathematically tractable for this analysis; Jones, 2009), and then prove by counterexample that Eqn 6 cannot generally be satisfied, even in the simplified case of only $J = 2$ sites.

We therefore conclude that thresholding schemes will lead, quite generally, to biased results relative to the exact calculation given by Eqn 2. Only the case of a unique threshold for each site-by-species combination guarantees agreement with Eqn 2, and even then only asymptotically. Thus, thresholding approaches suffer from the dual limitation that there is no clear justification for using them to construct S-SDMs and that, in doing so, one is almost certain to obtain an incorrect result.

To date, two studies have compared S-SDMs stacked via thresholding with those stacked via Eqn 2. Both Aranda & Lobo (2011) and Dubuis *et al.* (2011) found dramatic differences between summing probabilities (or, in the case of Aranda & Lobo, 2011, relative suitabilities) and threshold-derived presence/absences. In Appendix S3, we add to this body of results. Specifically, we compare S-SDM predictions with and without thresholding for our four empirical examples. We use the prevalence-based thresholding scheme advocated by Liu *et al.* (2005). In all four cases, thresholding led to the overprediction of richness relative to Eqn 2. In three out of four cases, the bias introduced by thresholding was severe. Our arguments above and in Appendix S2 demonstrate why it is not surprising that such differences exist, but the two case studies of Aranda & Lobo (2011) and Dubuis *et al.* (2011), coupled with our four examples (Appendix S3), provide clear evidence that the direction of the biases will tend towards overprediction, and that the magnitude of these discrepancies will often be very large. In other words, the distinction between stacking according to Eqn 2 and stacking via thresholding is not merely one of fine detail, but can instead result in a qualitatively different relationship between S-SDM predictions and observed richness.

Are macroecological models less biased than S-SDMs?

In the previous sections, we have provided a formal basis from which to build S-SDMs, showed that thresholding will generally yield incorrect results, and illustrated with our cases studies (and two studies from the literature) that thresholding introduces an often pronounced bias in the direction of overprediction. With these issues out of the way, we now examine how correctly stacked S-SDMs compare to MEMs. So far, only two studies, Algar *et al.* (2009) and Dubuis *et al.* (2011), have performed such a comparison, and both show extremely similar patterns for S-SDMs stacked according to Eqn 2 and MEMs. In Algar *et al.* (2009), S-SDMs and MEMs both fit the observed richness very closely: the results of both approaches are similar to each other and to the observed richness. Although both approaches are very similar to each other in the study of Dubuis *et al.* (2011), they both overpredict species richness in species-poor sites, and underestimate in species-rich sites (their Fig. 1a,c), in contrast to the assertion of Guisan & Rahbek (2011) that S-SDMs consistently overestimate richness.

Based on our four data sets (Table 2; see Appendix S4 for visualization), we can tentatively conclude that correctly stacked S-SDMs are no worse than MEMs (mean R^2 across the case studies of 0.71 vs. 0.75; see Table 3); that $\mathbf{S}^{(\text{S–SDM})}$ and $\mathbf{S}^{(\text{MEM})}$ are highly correlated ($r = 0.945$); and that both S-SDMs and MEMs fairly consistently underpredict richness at species-rich sites and overpredict at species-poor sites (i.e. that calibration slopes are < 1 and intercepts > 0; Table 3).

Obviously, six data sets (the four described herein, plus those of Algar *et al.*, 2009, and Dubuis *et al.*, 2011) are not sufficient to generalize these findings. The high consistency does suggest, however, that both correctly stacked S-SDMs and MEMs tend to exhibit the same biases. Combining these results with those of the previous sections on thresholding strongly suggests that the oft-cited overprediction of richness exhibited by S-SDMs is primarily a statistical artifact introduced by threshold-based stacking methods, and is not a result of these models ignoring dispersal filters or species interactions.

A maximum-likelihood approach to adjusting S-SDM predictions

We now develop a general method for adjusting S-SDM richness predictions by modifying the $p_{j,k}$ across a data set as a function of site-level species richness, S_j. When an MEM that accurately predicts richness is available, we can set $S_j = S_j^{(\text{MEM})}$, which is the MEM-predicted richness at site j. Our approach then could be used to facilitate the synthesis of S-SDMs and MEMs envisioned by Guisan & Rahbek (2011). The resulting adjusted occurrence probabilities, $\mathbf{p}_j^{\star}$, should then yield more accurate information about the composition of the community, which the MEM alone could not provide.

When we wish to quantify the nature and magnitude of the discrepancy between a raw S-SDM and observed data, we can set $S_j = S_j^{(\text{obs})}$, the observed species richness at the j-th site. The

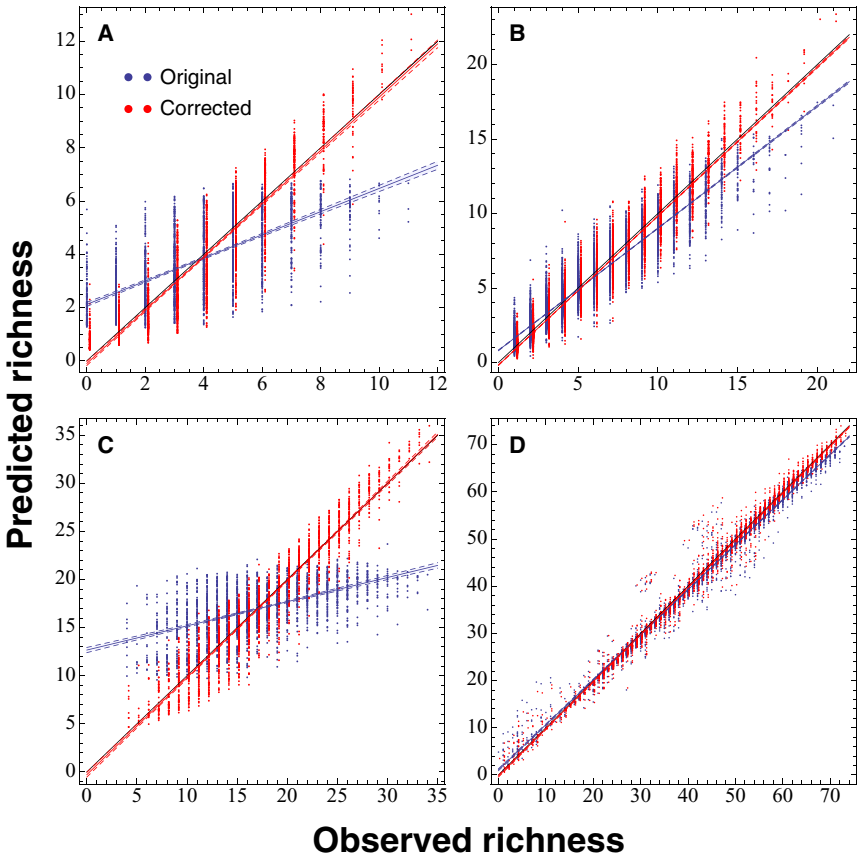

Figure 1 Predicted richness from the original (blue) and adjusted (red) S-SDMs versus observed richness, for the Wadden Sea macrobenthos (a), European forest trees (b), Barents Sea trawls (c) and European mammals (D) data sets. All four examples responded similarly to adjustment, but required adjustments of different strengths. Regression lines (solid) and 95% confidence bands (dashed) are provided to summarize the relationship between the original and adjusted occurrence probabilities and observed richness. Note that the regression line for the adjusted probabilities (red) falls very close to the 1:1 line (black) in all four cases.

Table 3 Calibration regressions (ordinary least squares) of S-SDMs and MEMs on observed species richness. Perfect fits would have intercepts of 0 and slopes of 1. Notice high correlation between both approaches (last column).

	S-SDM			MEM			S-SDM–MEM
	Intercept	slope	R^2	Intercept	slope	R^2	Correlation
Wadden Sea benthos	2.119	0.435	0.46	2.003	0.466	0.49	0.989
EU forest trees	0.816	0.820	0.90	0.665	0.159	0.79	0.934
Barents Sea trawls	8.700	0.472	0.61	6.598	0.602	0.72	0.928
EU mammals	3.636	0.894	0.87	0.454	0.990	1.00	0.930

adjustment is then summarized by the values of the adjustment parameters (see below), and this summary will facilitate comparisons of the performance of S-SDM predictions across data sets and taxa. Such comparisons should help to reveal consistent patterns in the way that S-SDM predictions go wrong, and might suggest concrete ways in which S-SDMs could be modified to improve their performance.

By definition, the occurrence probabilities must satisfy $0 \leq p_{j,k} \leq 1$ for all j and k. We wish to apply an adjustment to the $\mathbf{p}_j$ that depends on S_j, which may be $S_j^{(\mathrm{MEM})}$ or $S_j^{(\mathrm{obs})}$ depending on the context. A simple way to do this, while still respecting the above-mentioned constraint, is to first logit-transform the $\mathbf{p}_j$, apply an additive adjustment that depends on S_j to the logit-transformed values, and then inverse-logit-transform the adjusted values back to the probability scale. Let

$$q_{j,k} = \mathrm{logit}(p_{j,k}) = \log\left(\frac{p_{j,k}}{1-p_{j,k}}\right); \tag{7}$$

the adjusted values can then be defined as

$$q^{\star}_{j,k} = q_{j,k} + aS_j + b, \tag{8}$$

where a and b are the adjustment parameters. The adjusted occurrence probabilities are then

$$p^{\star}_{j,k} = \mathrm{logit}^{-1}(q^{\star}_{j,k}) = \frac{1}{1+e^{-q^{\star}_{j,k}}}. \tag{9}$$

For a given data set, the values of the adjustment parameters a and b can be estimated via maximum likelihood, by calculating

the probability of S_j for each site given the modified $\mathbf{p}_j^\star$ characterizing that site, which depend on the original $\mathbf{p}_j$ and the adjustment parameters. These probabilities are given by the Poisson binomial distribution (Eqn 1). Let $\mathbf{S} = \{S_1, S_2, \ldots, S_J\}$ denote the vector of species richnesses (either predicted or observed) for the J sites within a data set. The log-likelihood function is then

$$L(a, b|\mathbf{S}) = \sum_{j=1}^{J}\sum_{k=1}^{K}\log(\Pr(S_j|p_{j,k}^\star)), \quad (10)$$

where the dependence on the parameters a and b enters through the adjusted occurrence probabilities, $p_{j,k}^\star$. Equation 10 can be maximized numerically to yield the maximum-likelihood estimates $\hat{a}$ and $\hat{b}$. We implemented the estimation procedure in R (R Development Core Team, 2012) using the package POIBIN (Hong, 2011), which provides an efficient implementation of Eqn 1.

The features of the $\mathbf{p}_j$, and how adjustment changes them as a function of S_j, can be summarized by fitting beta distributions to them on a site-by-site basis. This can be readily achieved using the method of moments, because the beta distribution has closed-form moment estimators (Chapter 4 in Clark, 2007). The beta distribution is essentially identical to the Kumaraswamy distribution, which we use in Appendix S2, in terms of shape and behaviour, but has contrasting mathematical properties (Jones, 2009). Although the Kumaraswamy distribution is useful for the analyses of Appendix S2, the beta distribution is much more convenient for our purposes here. Comparing the fitted beta distributions of the original and adjusted occurrence probabilities at a site allows one to visualize the effects of adjustment, but, because all of our example data sets have a large number of sites, a site-by-site summary is impractical. Instead, we can look at the average effects of adjustment by averaging the fitted beta-distribution parameters within each data set over sites with the same observed species richness. This approach yields a pair of original and adjusted beta-distribution parameters for each level of species richness in the data set. The two distributions for each richness level can then be plotted against each other to visualize the average effects of the adjustment.

Adjustment will most strongly affect low and high occurrence probabilities, which correspond to rare and common species, respectively. It is therefore useful to visualize the relative effects of adjustment on these two groups of species. Rare species will tend to occur in the lower quantiles of the $\mathbf{p}_j$, and common species in the upper quantiles. We therefore use the 0.05 quantile of $\mathbf{p}_j$ at each site to represent the rare species and the 0.95 quantile to represent the common species. We then average the value of each of these quantiles over all sites within a data set that share a particular richness level. The ratio of these averaged quantile values after adjustment to their values before adjustment ($Q_{\text{adj}}/Q_{\text{org}}$) then indicates the degree of adjustment that was applied. When this ratio is 1, there was no adjustment; when it is above 1, occurrence probabilities were increased by the adjustment; when it is less than 1, the probabilities were decreased. The ratio for the rare species can then be plotted together with that for the common species in order to visualize the degree of adjustment experienced by rare and common species across sites ranging from low richness to high richness.

Table 4 Maximum likelihood estimates (MLEs), standard errors (SEs), z-values and P-values for the correction parameters $\hat{a}$ and $\hat{b}$ across the four data sets.

Data set	Param.	MLE	SE	z-value	Pr(z)
Wadden Sea benthos	$\hat{a}$	0.290	0.006	45.46	$< 10^{-15}$
	$\hat{b}$	−1.248	0.032	−38.74	$< 10^{-15}$
EU forest trees	$\hat{a}$	0.115	0.002	47.83	$< 10^{-15}$
	$\hat{b}$	−0.706	0.018	−38.62	$< 10^{-15}$
Barents Sea trawls	$\hat{a}$	0.101	0.001	79.49	$< 10^{-15}$
	$\hat{b}$	−1.783	0.024	−73.66	$< 10^{-15}$
EU mammals	$\hat{a}$	0.017	$< 10^{-5}$	5319.20	$< 10^{-15}$
	$\hat{b}$	−0.365	$< 10^{-4}$	−5621.20	$< 10^{-15}$

How S-SDM predictions deviate from observed richness

All four example data sets show the same qualitative pattern of discrepancy relative to observed species richness: richness is overestimated at species-poor sites and underestimated at species-rich sites (Fig. 1). All data sets responded well to adjustment, with the mean of the adjusted predictions falling very close to the 1:1 line across the range of observed species richness (Fig. 1). The strength of adjustment required, however, differed among data sets and seems to be related to the maximum single-site species richness in the data set (Table 4). Specifically, the data sets with high site-level richness had low values of the adjustment slope, a, and vice versa (Table 4). The magnitude of the adjustment intercept, however, did not show a clear relationship either with the maximum site-level species richness or with the total number of species in a data set. Both adjustment parameters differed significantly from zero for all four data sets, indicating that adjustment was beneficial in all cases (Table 4).

The distributions of $\mathbf{p}_j$ varied considerably with observed species richness in all data sets (Fig. 2). Species-poor sites featured unimodal $\mathbf{p}_j$ distributions with the mode occurring at 0 (Fig. 2, left column). Species-rich sites, in contrast, had bimodal (U-shaped) distributions, with modes at 0 and 1 (Fig. 2, right column). Intermediate sites were either unimodal or bimodal, with the two data sets with the highest total richness – EU trees and EU mammals – showing the bimodal pattern at intermediate-richness sites (Fig. 2, middle column). Adjustment tended to accentuate the original patterns at the low-richness and high-richness site extremes, but had little effect at intermediate sites.

The patterns revealed by examining the distributions of the $\mathbf{p}_j$ suggest that species tend to have either very low or very high occurrence probabilities. Adjustment should therefore have a strong effect on the tails of the distributions of occurrence probabilities. Focusing on the 0.05 and 0.95 quantiles of the $\mathbf{p}_j$ distributions, adjustment tended to depress the values of both

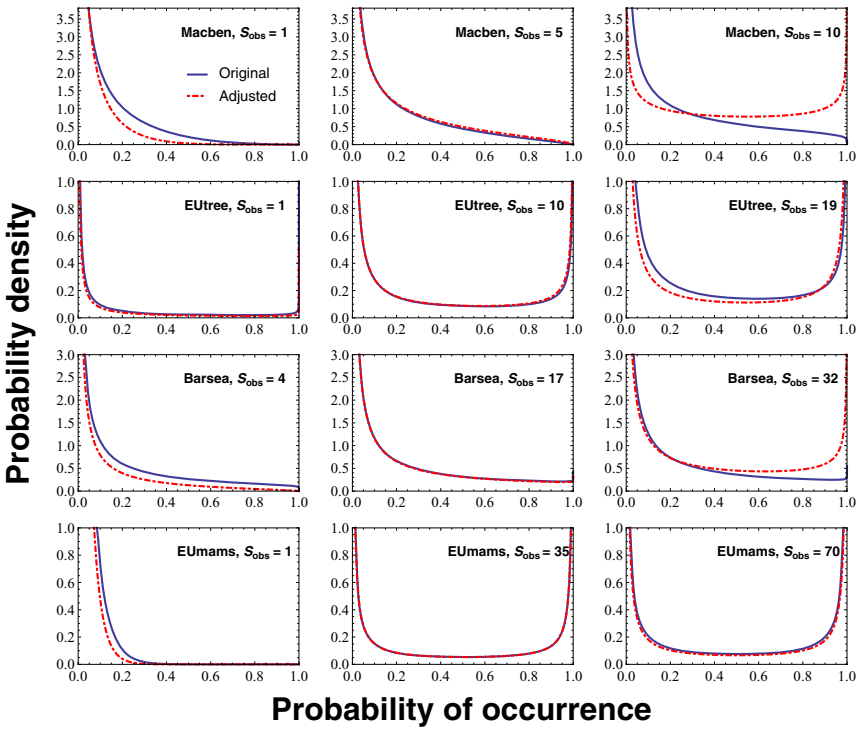

Figure 2 Fitted beta distribution probability density functions for the original (blue) and adjusted (red) $\mathbf{p}_j$ at low (left column), medium (middle column), and high (right column) richness sites for the Wadden Sea macrobenthos (first row), European forest trees (second row), Barents Sea trawls (third row) and European mammals (fourth row) data sets.

common (0.95 quantile) and rare (0.05 quantile) species by a similar amount in the low-richness sites (Fig. 3). In other words, in species-poor sites, the occurrence probabilities of both common and rare species were overpredicted. In the higher-richness sites, adjustment substantially increased the occurrence probabilities of rare species (Fig. 3); the occurrence probabilities of common species also increased, but the magnitude of the increase was small, and tended to level off for sites at the high end of the richness spectrum (Fig. 3). In between, each data set featured a (not necessarily integer) richness value where the adjustment had no effect on either quantile.

DISCUSSION

How (and how not) to stack

From our probability theory argument in the first section, it is clear that the process of stacking SDMs is straightforward and requires only a summation of the per-site occurrence probabilities, $\mathbf{p}_j$. We have also clearly demonstrated that the far more common approach of first converting $\mathbf{p}_j$ values into presence/absence predictions and then summing them will generally be incorrect, and will result in biased predictions of species richness. Additionally, our four example data sets, plus two existing data sets in the literature (Algar *et al.*, 2009; Dubuis *et al.*, 2011), show that thresholding schemes lead to S-SDMs that overpredict richness relative to using Eqn 2, often dramatically. Taken together, our formal argument justifying stacking via Eqn 2, the notable lack of a similar argument justifying thresholding, our analytical proof that thresholding will generally lead to biased results, and six empirical examples showing that thresholding leads to systematic overprediction of site-level richness provide strong evidence that using a thresholding scheme to build an S-SDM is incorrect and will produce biased results. We therefore strongly recommend that this practice be immediately put to rest. It is important to note, however, that our results apply only to using thresholds to stack S-SDMs, and do not imply anything about the use of thresholds for other purposes (e.g. for generating range maps from SDM-predicted occurrence probabilities).

Researchers working primarily with presence-only data may object that their models do not predict the probability of occurrence but rather an index of habitat suitability (Phillips *et al.*, 2006; Phillips & Dudík, 2008). Although true, the conversion using one or more arbitrary thresholds, as is commonly done (Table 1), does not solve this problem. It implicitly assumes that suitability values are comparable among species, which may or may not be the case. It should only be a matter of time until recent papers on the equivalence in principle of MaxEnt and Poisson point process models (e.g. Renner & Warton, 2013) facilitate a conversion of MaxEnt output to probabilities. Until then, we propose to use our maximum-likelihood approach to adjust MaxEnt-generated suitabilities, i.e. to treat them as if they were probabilities.

Relationships between S-SDMs and MEMs

When stacked via Eqn 2, we find that S-SDMs make richness predictions that are very similar to those of MEMs on the same data. All four of our data sets confirm this pattern, with an average correlation of 0.945 between S-SDMs and MEMs (Table 3). Two other studies in the literature have also noted this similarity. In a

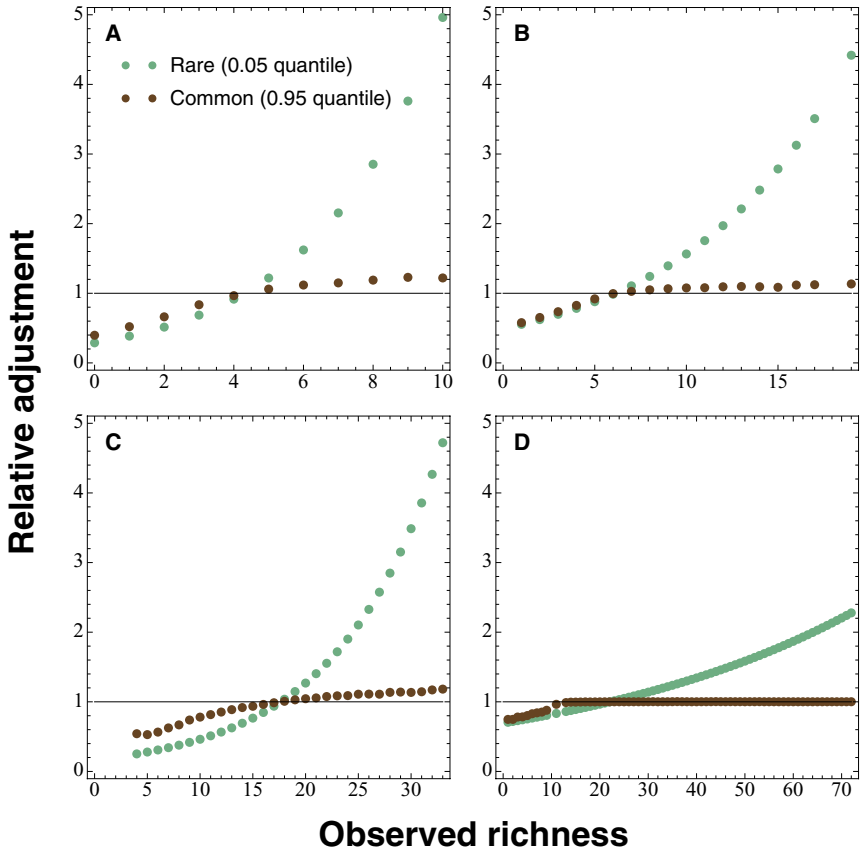

Figure 3 The relative effect of adjustment on rare (green, 0.05 quantile) and common (brown, 0.95 quantile) species as a function of observed species richness. The relative adjustment is the ratio of the adjusted, Q_{adj}, to original, Q_{org}, value of the focal quantile, averaged over all sites sharing the same richness within a given data set. Only richness values observed at ≥ 5 sites within a data set are shown. Relative adjustment values > 1 indicate the occurrence probabilities for the focal species group have been increased by adjustment, values < 1 indicate a decrease, and values = 1 signify no change. A horizontal line for relative adjustment = 1 is shown for reference. The panels are arranged as in Fig. 1. The scale of the *y*-axis is the same in all panels to emphasize the different strengths of adjustment.

study of Canadian butterflies, Algar *et al.* (2009) found very high agreement both between MEMs and S-SDMs and between both methods and observed richness patterns. Similarly, Dubuis *et al.* (2011) report that correctly stacked S-SDMs and MEMs show very similar richness predictions (their Figure 1a,c) and make very similar predictions for the effect of elevation (their Figure 2c,d). These six studies suggest that correctly stacked S-SDMs do not exhibit a systematic tendency to overpredict richness, as is frequently claimed in the literature.

Interestingly, S-SDMs and MEMs for all six data sets (our four; Algar *et al.*, 2009; and Dubuis *et al.*, 2011) show the same tendency to overestimate species richness in species-poor sites, and to underestimate it on species-rich sites (Fig. 1; Appendix S4). Of our examples, the EU mammals data set was the only case where an MEM could potentially be used together with our likelihood-based adjustment method to appreciably improve S-SDM performance. Although the MEM for this example was highly correlated with observed richness ($R^2 = 0.997$), the S-SDM was almost as good ($R^2 = 0.87$). We have yet to encounter an example where the MEM performed very well and a correctly stacked S-SDM performed poorly. More examples will need to appear in the literature before we can conclude whether or not MEMs and S-SDMs can productively be used together in the manner suggested by Guisan & Rahbek (2011), but the available evidence is not encouraging.

Our maximum-likelihood approach to adjusting the $\mathbf{p}_j$ suggests that there is a systematic pattern of deviation that decreases with increasing maximum site-level richness. Furthermore, examining the adjusted $\mathbf{p}_j$ in detail revealed that the rarer species (small $p_{j,k}$) typically required stronger correction, in both species-rich and species-poor sites, to bring S-SDMs into line with observed richness. Although it is tempting to speculate about the origins of these patterns, we urge caution as they are based on only four data sets. However, we have provided the tools to examine patterns of error in a wide range of S-SDMs. If similar patterns are observed across many more examples, these regularities could potentially be used to correct bias in S-SDM predictions.

Based on our examination of stacking methods and our comparison of correctly stacked S-SDMs and MEMs, we found no evidence of systematic differences between the two model types. In other words, the core observation on which SESAM is based – that S-SDMs consistently overpredict richness whereas MEMs do not – appears to be the result of a statistical artifact introduced by using thresholding schemes to produce S-SDMs.

What are the causes of biased predictions?

We suspect that the tendency for both correctly stacked S-SDMs and MEMs to overpredict in species-poor sites and underpredict in species-rich sites may have several causes, which are more likely to be statistical than ecological. First, it is well known that issues of sample size plague the estimation of SDMs for rare species (e.g. Jiménez-Valverde *et al.*, 2009). Also, richness pat-

terns are largely driven by common species (i.e. those with a high prevalence), as shown by Jetz & Rahbek (2002) and more systematically by Lennon *et al.* (2004) and Šizling *et al.* (2009).

A signal of prevalence in model accuracy has been consistently reported, with SDMs for both common and rare species being less reliable than those for species of intermediate prevalence (McPherson *et al.*, 2004; Santika, 2011). It could thus be that both MEM and S-SDM patterns are largely determined by common species, which we may fail to represent properly in our models.

Second, the high consistency between S-SDMs and MEMs suggests a common underlying cause of their biases. The observed bias in S-SDMs may have nothing to do with SDM model accuracy or neglecting species interactions during stacking, but may instead be caused by 'regression dilution' or 'attenuation' (Madansky, 1959; MacMahon *et al.*, 1990; McInerny & Purves, 2011). This refers to a bias in the estimation of the strength of an effect due to unaccounted variability in the predictor (see Frost & Thompson, 2000, for a review). In other words, because environmental predictors may not represent the habitat conditions of the species but rather the average across the whole grid cell, the former are represented by the latter with a large error. This error could lead to underestimation of the strength of the relationship between predictor (say, temperature) and the probability of occurrence. If the true function was, for example, linear with an intercept of 2 and a slope of 2, regression dilution could lead to estimates of 3 and 1.5, respectively. Thus, the intercept would compensate for lower slope estimates. As habitat preferences among species are on average positively correlated (otherwise we would not see any pattern of species richness along environmental gradients), regression dilution will lead to underestimation of occurrence at suitable sites and overestimation at unsuitable sites, which is exactly the pattern we found.

If this hypothesis is correct, analysing species richness directly should yield a very similar discrepancy, because MEMs frequently use the same environmental predictors as S-SDMs. Although speculative at the moment, the issue of subscale variability has been on the list of problems of SDMs for many years (e.g. Rahbek & Graves, 2000, 2001; Vaughan & Ormerod, 2003; Rahbek, 2005; Beck *et al.*, 2012), and we have simply described a particular way in which it may be manifested.

CONCLUSIONS

The use of S-SDMs to predict species richness is still in its infancy and many problems remain to be solved. We have attempted to put the process of stacking S-SDMs on firmer statistical ground by demonstrating the correct way to build S-SDMs. Our results strongly suggest that the use of ad hoc stacking methods based on thresholding introduces a systematic bias in S-SDM richness predictions that can account for the frequently noted discrepancies between S-SDMs and MEMs. These results therefore cast substantial doubt on the core assumptions of the SESAM framework – that S-SDMs systematically overpredict richness because they lack dispersal filters and ecological assembly rules. We recommend that future investigations into the relationship between S-SDMs and MEMs first rule out artifactual causes for differences between these modelling frameworks before invoking biotic mechanisms. We also suggest that more effort be directed at studying the extent to which regression dilution can account for the similar biases observed in MEMs and correctly stacked S-SDMs.

ACKNOWLEDGEMENTS

We are grateful to Renate Köble at the EC Joint Research Centre in Ispra for providing the EU tree data, and to Christian Hof, Holger Kreft and two anonymous reviewers for comments on an earlier version of the manuscript. The benthic monitoring carried out in the Dutch Wadden Sea was initiated, organized and led by Theunis Piersma and Anne Dekinga, and received full support from the Royal Netherlands Institute for Sea Research. A large number of colleagues, students and volunteers contributed to the collection of the field data. G.C. was funded by the BarEcoRe project (NRC contract number 200793/S30), under the Norwegian Research Council NORKLIMA programme. C.K. was supported by the Marsden Fund Council from Government funding, administered by the Royal Society of New Zealand.

REFERENCES

Algar, A.C., Kharouba, H.M., Young, E.R. & Kerr, J.T. (2009) Predicting the future of species diversity: macroecological theory, climate change, and direct tests of alternative forecasting methods. *Ecography*, **32**, 22–33.

Aranda, S.C. & Lobo, J.M. (2011) How well does presence-only-based species distribution modelling predict assemblage diversity? A case study of the Tenerife flora. *Ecography*, **34**, 31–38.

Barbet-Massin, M., Thuiller, W. & Jiguet, F. (2010) How much do we overestimate future local extinction rates when restricting the range of occurrence data in climate suitability models? *Ecography*, **33**, 878–886.

Beck, J., Ballesteros-Mejia, L., Buchmann, C.M., Dengler, J., Fritz, S.A., Gruber, B., Hof, C., Jansen, F., Knapp, S., Kreft, H., Schneider, A.-K., Winter, M. & Dormann, C. (2012) What's on the horizon for macroecology? *Ecography*, **35**, 673–683.

Brown, J.H. (1995) *Macroecology*. University of Chicago Press, Chicago, IL.

Buisson, L., Thuiller, W., Casajus, N., Lek, S. & Grenouillet, G. (2010) Uncertainty in ensemble forecasting of species distribution. *Global Change Biology*, **16**, 1145–1157.

Casella, G. & Berger, R. (2002) *Statistical inference*. Duxbury Press, Belmont, CA.

Clark, J.S. (2007) *Models for ecological data: an introduction*. Princeton University Press, Princeton, NJ.

Dormann, C.F., Gruber, B., Winter, M. & Herrmann, D. (2010) Evolution of climate niches in European mammals? *Biology Letters*, **6**, 229–232.

Dormann, C.F., Elith, J., Bacher, S., Buchmann, C., Carl, G., Carré, G., García Marquéz, J.R., Gruber, B., Lafourcade, B.,

Leitão, P.J., Münkemüller, T., McClean, C., Osborne, P.E., Reineking, B., Schröder, B., Skidmore, A.K., Zurell, D. & Lautenbach, S. (2013) Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. *Ecography*, **36**, 27–46.

Dubuis, A., Pottier, J., Rion, V., Pellissier, L., Theurillat, J. & Guisan, A. (2011) Predicting spatial patterns of plant species richness: a comparison of direct macroecological and species stacking modelling approaches. *Diversity and Distributions*, **17**, 1122–1131.

Elith, J., Graham, C.H., Anderson, R.P. *et al.* (2006) Novel methods improve prediction of species' distributions from occurrence data. *Ecography*, **29**, 129–151.

Elith, J., Leathwick, J.R. & Hastie, T. (2008) A working guide to boosted regression trees. *Journal of Animal Ecology*, **77**, 802–813.

Elith, J., Kearney, M. & Phillips, S. (2010) The art of modelling range-shifting species. *Methods in Ecology and Evolution*, **1**, 330–342.

Erasmus, B.F.N., Van Jaarsveld, A.S., Chown, S.L., Kshatriya, M. & Wessels, K.J. (2002) Vulnerability of South African animal taxa to climate change. *Global Change Biology*, **87**, 679–693.

Fernandez, M. & Williams, S. (2010) Closed-form expression for the Poisson-binomial probability density function. *IEEE Transactions on Aerospace and Electronic Systems*, **46**, 803–817.

Ferrier, S. & Guisan, A. (2006) Spatial modelling of biodiversity at the community level. *Journal of Applied Ecology*, **43**, 393–404.

Ferrier, S., Watson, G., Pearce, J. & Drielsma, M. (2002) Extended statistical approaches to modelling spatial pattern in biodiversity in northeast New South Wales. I. Species-level modelling. *Biodiversity and Conservation*, **11**, 2275–2307.

Fielding, A.H. & Bell, J.F. (1997) A review of methods for the assessment of prediction errors in conservation presence/absence models. *Environmental Conservation*, **24**, 38–49.

Fitzpatrick, M.C., Sanders, N.J., Ferrier, S., Longino, J.T., Weiser, M.D. & Dunn, R. (2011) Forecasting the future of biodiversity: a test of single- and multi-species models for ants in North America. *Ecography*, **34**, 836–847.

Franklin, J. (2009) *Mapping species distributions: spatial inference and prediction*. Cambridge University Press, Cambridge, UK.

Frost, C. & Thompson, S.G. (2000) Correcting for regression dilution bias: comparison of methods for a single predictor variable. *Journal of the Royal Statistical Society A*, **163**, 173–189.

Gaston, K.J. (2000) Global patterns in biodiversity. *Nature*, **405**, 220–227.

Gaston, K.J. & Blackburn, T.M. (2000) *Pattern and process in macroecology*. Blackwell, Malden, MA.

Graham, C.H. & Hijmans, R.J. (2006) A comparison of methods for mapping species ranges and species richness. *Global Ecology and Biogeography*, **15**, 578–587.

Guisan, A. & Rahbek, C. (2011) SESAM – a new framework integrating macroecological and species distribution models for predicting spatio-temporal patterns of species assemblages. *Journal of Biogeography*, **38**, 1433–1444.

Guisan, A., Weiss, S.B. & Weiss, A.D. (1999) GLM versus CCA spatial modeling of plant species distribution. *Plant Ecology*, **143**, 107–122.

Hawkins, B.A. & Diniz-Filho, J.A.F. (2004) 'Latitude' and geographic patterns in species richness. *Ecography*, **27**, 268–272.

Hijmans, R.J., Cameron, S.E., Parra, J.L., Jones, P.G. & Jarvis, A. (2005) Very high resolution interpolated climate surfaces for global land areas. *International Journal of Climatology*, **25**, 1965–1978.

Hong, Y. (2011) *poibin: the Poisson binomial distribution*. Available at: http://cran.r-project.org/web/packages/poibin/index.html (accessed 11 July 2013).

Hortal, J., De Marco, P., Jr, Santos, A.M.C. & Diniz-Filho, J.A.F. (2012) Integrating biogeographical processes and local community assembly. *Journal of Biogeography*, **39**, 627–628.

Jakobsen, T. & Ozhigin, V. (eds) (2011) *The Barents Sea. Ecosystem, resources, management. Half a century of Russian–Norwegian cooperation*. Tapir Academic Press, Trondheim, Norway.

Jetz, W. & Rahbek, C. (2002) Geographic range size and determinants of avian species richness. *Science*, **297**, 1548–1551.

Jiménez-Valverde, A., Lobo, J.M. & Hortal, J. (2009) The effect of prevalence and its interaction with sample size on the reliability of species distribution models. *Community Ecology*, **10**, 196–205.

Johannesen, E., Høines, Å.S., Dolgov, A.V. & Fossheim, M. (2012) Demersal fish assemblages and spatial diversity patterns in the Arctic-Atlantic transition zone in the Barents Sea. *PLoS ONE*, **7**, e34924.

Jones, M.C. (2009) Kumaraswamy's distribution: a beta-type distribution with some tractability advantages. *Statistical Methodology*, **6**, 70–81.

Kaschner, K., Tittensor, D.P., Ready, J., Gerrodette, T. & Worm, B. (2011) Current and future patterns of global marine mammal biodiversity. *PLoS ONE*, **6**, e19653.

Kerr, J.T., Kharouba, H.M. & Currie, D.J. (2007) The macroecological contribution to global change solutions. *Science*, **316**, 1581–1584.

Köble, R. & Seufert, G. (2001) Novel maps for forest tree species in Europe. *Proceedings of the 8th European Symposium on the Physico-Chemical Behaviour of Air Pollutants: 'A Changing Atmosphere!'*, Turin, Italy, 17–20 September 2001.

Kraan, C., Piersma, T., Dekinga, A., Koolhaas, A. & van der Meer, J. (2007) Dredging for edible cockles (*Cerastoderma edule*) on intertidal flats: short-term consequences of fisher patch-choice decisions for target and non-target benthic fauna. *ICES Journal of Marine Science*, **64**, 1735–1742.

Kraan, C., van Gils, J.A., Spaans, B., Dekinga, A., Bijleveld, A.I., van Roomen, M., Kleefstra, R. & Piersma, T. (2009) Landscape-scale experiment demonstrates that Wadden Sea intertidal flats are used to capacity by molluscivore migrant shorebirds. *Journal of Animal Ecology*, **78**, 1259–1268.

Kraan, C., Aarts, G., van der Meer, J. & Piersma, T. (2010) The role of environmental variables in structuring landscape-scale species distributions in seafloor habitats. *Ecology*, **91**, 1583–1590.

Kraan, C., Aarts, G., Piersma, T. & Dormann, C.F. (2013) Temporal variability of ecological niches: a study on intertidal macrobenthic fauna. *Oikos*, **122**, 754–760.

Lehmann, A., Overton, J.M. & Austin, M.P. (2002) Regression models for spatial prediction: their role for biodiversity and conservation. *Biodiversity and Conservation*, **11**, 2085–2092.

Lennon, J.J., Koleff, P., Greenwood, J.J.D. & Gaston, K.J. (2004) Contribution of rarity and commonness to patterns of species richness. *Ecology Letters*, **7**, 81–87.

Liu, C., Berry, P.M., Dawson, T.P. & Pearson, R.G. (2005) Selecting thresholds of occurrence in the prediction of species distributions. *Ecography*, **28**, 385–393.

Loiselle, B.A., Howell, C.A., Graham, C.H., Goerck, J.M., Brooks, T., Smith, K.G. & Williams, P.H. (2003) Avoiding pitfalls of using species distribution models in conservation planning. *Conservation Biology*, **17**, 1591–1600.

Loiselle, B.A., Jørgensen, P.M., Consiglio, T., Jiménez, I., Blake, J.G., Lohmann, L.G. & Montiel, O.M. (2007) Predicting species distributions from herbarium collections: does climate bias in collection sampling influence model outcomes? *Journal of Biogeography*, **35**, 105–116.

McInerny, G.J. & Purves, D.W. (2011) Fine-scale environmental variation in species distribution modelling: regression dilution, latent variables and neighbourly advice. *Methods in Ecology and Evolution*, **2**, 248–257.

MacMahon, S., Peto, R., Cutler, J., Collins, R., Sorlie, P., Neaton, J., Abbott, R., Godwin, J., Dyer, A. & Stamler, J. (1990) Blood pressure, stroke, and coronary heart disease: part 1, prolonged differences in blood pressure: prospective observational studies corrected for the regression dilution bias. *The Lancet*, **335**, 765–774.

McPherson, J.M., Jetz, W. & Rogers, D.J. (2004) The effects of species' range sizes on the accuracy of distribution models: ecological phenomenon or statistical artefact? *Journal of Applied Ecology*, **41**, 811–823.

Madansky, A. (1959) The fitting of straight lines when both variables are subject to error. *Journal of the American Statistical Association*, **54**, 173–205.

Mateo, R.G., Felicísimo, Á.M., Pottier, J., Guisan, A. & Muñoz, J. (2012) Do stacked species distribution models reflect altitudinal diversity patterns? *PLoS ONE*, **7**, e32586.

Peppler-Lisbach, C. & Schröder, B. (2004) Predicting the species composition of *Nardus stricta* communities by logistic regression modelling. *Journal of Vegetation Science*, **15**, 623–634.

Peterson, A.T., Soberón, J., Pearson, R.G., Anderson, R.P., Martínez-Meyer, E., Nakamura, M. & Araújo, M. (2011) *Ecological niches and geographic distributions*. Monographs in Population Biology 49. Princeton University Press, Princeton, NJ.

Phillips, S.J. & Dudík, M. (2008) Modeling of species distributions with Maxent: new extensions and a comprehensive evaluation. *Ecography*, **31**, 161–175.

Phillips, S.J., Anderson, R.P. & Schapire, R.E. (2006) Maximum entropy modeling of species geographic distributions. *Ecological Modelling*, **190**, 231–259.

Pineda, E. & Lobo, J.M. (2009) Assessing the accuracy of species distribution models to predict amphibian species richness patterns. *Journal of Animal Ecology*, **78**, 182–190.

Pineda, E. & Lobo, J.M. (2012) The performance of range maps and species distribution models representing the geographic variation of species richness at different resolutions. *Global Ecology and Biogeography*, **21**, 935–944.

R Development Core Team (2012) *R: a language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria.

Raes, N., Roos, M.C., Slik, J.W.F., Van Loon, E.E. & ter Steege, H. (2009) Botanical richness and endemicity patterns of Borneo derived from species distribution models. *Ecography*, **32**, 180–192.

Rahbek, C. (2005) The role of spatial scale and the perception of large-scale species-richness patterns. *Ecology Letters*, **8**, 224–239.

Rahbek, C. & Graves, G.R. (2000) Detection of macro-ecological patterns in South American hummingbirds is affected by spatial scale. *Proceedings of the Royal Society B: Biological Sciences*, **267**, 2259–2265.

Rahbek, C. & Graves, G.R. (2001) Multiscale assessment of patterns of avian species richness. *Proceedings of the National Academy of Sciences USA*, **98**, 4534–4539.

Randin, C.F., Engler, R., Normand, S., Zappa, M., Zimmermann, N.E., Pearman, P.B., Vittoz, P., Thuiller, W. & Guisan, A. (2009) Climate change and plant distribution: local models predict high-elevation persistence. *Global Change Biology*, **15**, 1557–1569.

Renner, I.W. & Warton, D.I. (2013) Equivalence of MAXENT and Poisson point process models for species distribution modeling in ecology. *Biometrics*, **69**, 274–281.

Rocchini, D., Hortal, J., Lengyel, S., Lobo, J.M., Jiménez-Valverde, A., Ricotta, C., Bacaro, G. & Chiarucci, A. (2011) Accounting for uncertainty when mapping species distributions: the need for maps of ignorance. *Progress in Physical Geography*, **35**, 211–226.

Rondinini, C., Di Marco, M., Chiozza, F., Santulli, G., Baisero, D., Visconti, P., Hoffmann, M., Schipper, J., Stuart, S.N., Tognelli, M.F., Amori, G., Falcucci, A., Maiorano, L. & Boitani, L. (2011) Global habitat suitability models of terrestrial mammals. *Philosophical Transactions of the Royal Society B: Biological Sciences*, **366**, 2633–2641.

Santika, T. (2011) Assessing the effect of prevalence on the predictive performance of species distribution models using simulated data. *Global Ecology and Biogeography*, **20**, 181–192.

Schmidt-Lebuhn, A.N., Knerr, N.J. & González-Orozco, C.E. (2012) Distorted perception of the spatial distribution of plant diversity through uneven collecting efforts: the example of Asteraceae in Australia. *Journal of Biogeography*, **39**, 2072–2080.

Scott, J.M., Heglund, P.J., Morrison, M.L., Haufler, J.B., Raphael, M.G., Wall, W.A. & Samson, F.B. (2002) *Predicting species occurrences: issues of accuracy and scale*. Island Press, Washington, DC.

Šizling, A.L., Šizlingová, E., Storch, D., Reif, J. & Gaston, K.J. (2009) Rarity, commonness, and the contribution of individual species to species richness patterns. *The American Naturalist*, **174**, 82–93.

Skov, F. & Borchsenius, F. (1997) Predicting plant species distribution patterns using simple climatic parameters: a case study of Ecuadorian palms. *Ecography*, **20**, 347–355.

Temple, H.J. & Terry, A. (2007) *The status and distribution of European mammals*. Office for Official Publications of the European Communities, Luxemburg.

Thuiller, W. (2003) BIOMOD – optimizing predictions of species distributions and projecting potential future shifts under global change. *Global Change Biology*, **9**, 1353–1362.

Thuiller, W., Lafourcade, B., Engler, R. & Araújo, M.B. (2009) BIOMOD – a platform for ensemble forecasting of species distributions. *Ecography*, **32**, 369–373.

Vaughan, I.P. & Ormerod, S.J. (2003) Improving the quality of distribution models for conservation by addressing shortcomings in the field collection of training data. *Conservation Biology*, **17**, 1601–1611.

Wang, Y.H. (1993) On the number of successes in independent trials. *Statistica Sinica*, **3**, 295–312.

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of this article at the publisher's web-site.

Appendix S1 GLM-based results.
Appendix S2 Single, species-specific threshold when the $\mathbf{p}_k$ values follow a Kumaraswamy distribution.
Appendix S3 A comparison of S-SDMs with and without thresholding.
Appendix S4 Additional figures.

BIOSKETCH

Justin Calabrese is a quantitative ecologist at the Smithsonian Conservation Biology Institute. His work mixes ecological theory, statistics and probability models, and empirical data. His research interests range broadly and include studying the drivers of animal movement, exploring how phenology affects individual fitness and population dynamics, and understanding the factors governing parasite/host interactions.

Editor: José Alexandre Diniz-Filho