White Rose Research Online
eprints@whiterose.ac.uk
https://eprints.whiterose.ac.uk
Universities of Leeds, Sheffield and York

Deposited via The University of Leeds.

White Rose Research Online URL for this paper:
https://eprints.whiterose.ac.uk/id/eprint/109763/

Version: Accepted Version

**Article:**

Gavish, Y, Marsh, CJ, Kuemmerlen, M et al. (2017) Accounting for biotic interactions through alpha-diversity constraints in stacked species distribution models. Methods in Ecology and Evolution, 8 (9). pp. 1092-1102.

https://doi.org/10.1111/2041-210X.12731

© 2017 The Authors. Methods in Ecology and Evolution © 2017 British Ecological Society. This is the peer reviewed version of the following article: Gavish, Y., Marsh, C. J., Kuemmerlen, M., Stoll, S., Haase, P. and Kunin, W. E. (2017), Accounting for biotic interactions through alpha-diversity constraints in stacked species distribution models. Methods Ecol Evol. doi:10.1111/2041-210X.12731, which has been published in final form at https://doi.org/10.1111/2041-210X.12731. This article may be used for non-commercial purposes in accordance with the Wiley Terms and Conditions for Self-Archiving.

**Reuse**
Items deposited in White Rose Research Online are protected by copyright, with all rights reserved unless indicated otherwise. They may be downloaded and/or printed for private study, or other acts as permitted by national copyright laws. The publisher or other rights holders may allow further reproduction and re-use of the full text version. This is indicated by the licence information on the White Rose Research Online record for the item.

**Takedown**
If you consider content in White Rose Research Online to be in breach of UK law, please notify us by emailing eprints@whiterose.ac.uk including the URL of the record and the reason for the withdrawal request.

UNIVERSITY OF LEEDS
University of Sheffield
UNIVERSITY of York

**Title:**

# Accounting for biotic interactions through alpha-diversity constraints in stacked species distribution models

**Authors:**

Gavish, Yoni[1,#] gavishyoni@gmail.com

Marsh, Charles J.[1] C.Marsh@leeds.ac.uk

Kuemmerlen, Mathias[2] mathias.kuemmerlen@senckenberg.de

Stoll, Stefan[2,3] Stefan.Stoll@senckenberg.de

Haase, Peter[2,4] peter.haase@senckenberg.de

Kunin, William E.[1] W.E.Kunin@leeds.ac.uk

**Affiliations:**

[1] School of Biology, Faculty of biological Science, University of Leeds, Leeds LS2 9JT, UK

[2] Senckenberg Research Institute and Natural History Museum Frankfurt, Department of River Ecology and Conservation, Clamecystr. 12, D-63571, Gelnhausen, Germany

[3] Institute for Environmental Sciences, University of Koblenz-Landau, Fortstr. 7, 76829 Landau, Germany.

[4] University of Duisburg-Essen, Faculty of Biology, Essen, Germany

[#] Corresponding Author: Yoni Gavish; gavishyoni@gmail.com 0044-759-999-1988

**Running title:**

The Alpha-adjusted SDM

**Counts:**

| | | |
|---|---|---|
| Total: | 6989 | words |
| Abstract: | 322 | words |
| Key-Words: | 15 | words |
| Main text: | 6312 | words |
| Table + Figure Legends: | 340 | words |

1 table, 5 figures, 35 citations, 6 supporting information

# ABSTRACT

1. Species Distribution Models (SDM) are widely used to predict occupancy patterns at fine resolution over wide extents. However, SDMs generally ignore the effect of biotic interactions and tend to overpredict the number of species that can coexist at a given location and time (hereafter, the alpha-capacity). We developed an extension of SDMs that integrates species-level and community-level modelling to account for the above drivers.
2. The alpha-adjusted SDM takes the Probabilities of Occurrence (PoO) for all species of a community and the site's alpha-capacity and adjusts the PoO, such that: **a.** their sum will equal the alpha-capacity as predicted by probability theory; and **b.** the adjusted PoO are dependent upon the relative suitability of each species for that site. The new method was tested using community data comprising 87 freshwater invertebrate species in an LTER watershed in Germany. We explored the ability of the method to predict alpha and beta-diversity patterns. We further focused on the effect on model performance at the species-level of the error associated with modelling alpha-capacity, of differences in gamma diversity (the size of the community) and of the type of community (random or guild-based).
3. The models that predicted alpha-capacity contained considerable error, and thus adjusting the PoO according to the modelled alpha-capacity resulted with decreased performance at the species level. However, when using the observed alpha-capacity to mimic a good alpha-capacity model, the alpha-adjusted SDMs usually resulted in increased performance. We further found that the alpha-adjusted SDM was better than the original SDM at predicting beta-diversity patterns, especially when using similarity indices that are sensitive to double absences.
4. Using the alpha-adjusted SDM approach may increase the predictive performance at the species and community levels if alpha-capacity can be assessed or modelled with sufficient accuracy, especially in relatively small communities of closely interacting species. With better models to predict alpha-capacity being developed, alpha-adjusted SDM has considerable potential to provide more realistic predictions of species-distribution patterns.

## KEYWORDS

SDM, Beta Diversity, Gamma Diversity, Competition, Co-Existence, Spatial Ecology, Alpha-Capacity, Macroecological Models, Random-Forest, Freshwater Environment

## INTRODUCTION

Species Distribution Models (SDM) are statistical models that relate the species' known presences to environmental data and then extrapolate the identified correlative rules to unsampled locations (Elith & Leathwick 2009) to generate Probability of Occurrence (PoO, the main output of most SDMs) and presence/absence (P/A) maps. SDMs are the main tools available today for predicting species' distributions at fine resolutions over wide spatial extents, and thus have significant potential applications in conservation and management (Guisan *et al.* 2013; Guillera-Arroita *et al.* 2015). SDMs are also used to predict the impact of projected environmental changes on the future distributions of focal species (e.g., Carroll 2010) or biodiversity (e.g., Distler *et al.* 2015). However, SDMs often fail to include other major processes that as well as environmental filtering are also important in determining species distributions. These omissions limit the accuracy of SDM outputs. Among a multitude of omitted processes, accounting for the effects of biotic interactions has been identified as one of the main challenges (Elith & Leathwick 2009; Wiens *et al.* 2009; Guisan & Rahbek 2011; D'Amen *et al.* 2015b; Thuiller *et al.* 2015).

Indeed, considerable effort has been devoted to incorporating the effect of biotic interactions into SDMs, as evident by two recent reviews (Kissling *et al.* 2012; Wisz *et al.* 2013) and additional progress made subsequently (e.g., Harris 2015). In some cases, significant *a priori* knowledge of the nature of the biotic interaction can allow for an almost process-based integration into the SDMs. For example, Heikkinen *et al.* (2007) included the distribution of woodpeckers as additional predictors of owl distributions since owls nest in abandoned woodpeckers nesting cavities. However, in most cases the nature of important biotic interactions are not known *a priori* in sufficient detail. In such cases, Wisz *et al.* (2013) suggested using surrogate variables to reflect gradients and spatial patterns in the distributions of biotic interactions. For example, a group of species competing for a set of limiting resources are likely to create spatial patterns of species richness that reflect the intensity of biotic interactions. A correlative alpha-diversity model (i.e. a model relating known local species richness to environmental variables and extrapolating the learned rules to the entire extent) may be used as a surrogate for the intensity of biotic interactions.

The integration of species-level and community-level modelling lies at the basis of the Spatially Explicit Species Assemblage Modelling framework (SESAM, Guisan & Rahbek 2011), which emerged from repeated observations that stacked P/A maps from SDMs tend to overestimate alpha diversity (Newbold *et al.* 2009; Guisan & Rahbek 2011;  but see: Calabrese *et al.* 2014; D'Amen, Pradervand & Guisan 2015; Distler *et al.* 2015). The SESAM framework suggests three main reasons for this overprediction. First, SDMs tend to ignore dispersal limitations and thus predict species to occur in locations they have not been able to colonize. Second, due to limiting resources, each location can only support a finite number of individuals, and thus a finite number of species even in the absence of biotic interactions (Hubbell 2001). Third, species tend to sustain populations in locations where they can cooccur with other species (Soberón 2007; Peterson & Soberon 2012), and SDMs ignore these assembly rules. In addition, local species richness is affected by the size of the regional species pool (Cornell & Lawton 1992). The dynamic interaction between the above processes, as well as the inability of species to occur in certain locations due to physiological limitations, suggests a dynamic upper constraint on the number of species that will be observed at a given site, which we term the '*alpha-capacity*'.

Correlative alpha-diversity models usually provide better predictions of alpha diversity than stacked SDMs, despite their tendency to slightly overpredict when the observed diversity is low and to underpredict when the diversity is high (Newbold *et al.* 2009; Guisan & Rahbek 2011). Thus, Guisan and Rahbek (2011) suggested that stacked SDMs should be constrained by the site's alpha-capacity (as predicted by the alpha-diversity model) and that assembly rules should be used to allocate the available slots. Currently, There are two main approaches to exploring the SESAM framework with respect to assembly rules: (i) ranking the species according to their PoO and selecting the top number of species that equals the predicted alpha-capacity as presences and the rest as absences (D'Amen *et al.* 2015a; D'Amen, Pradervand & Guisan 2015, hereafter, the 'top-alpha' approach); or (ii) adjusting the PoO of all species in all sites, such that the sum of the adjusted PoO in each site will be as close as possible to the predicted alpha-capacity (Calabrese *et al.* 2014, hereafter, the ‘summed-PoO’ approach). The adjusted PoO can produce P/A maps by selecting a threshold, above which sites are considered presences and below which sites are considered absences (Liu *et al.* 2005).

Both approaches have pros and cons. For example, under the top-alpha approach the PoO remains unchanged, such that species’ PoO are not affected by the PoOs of other species or by the alpha-capacity. Therefore performance indices at the species level that are based on PoO, such as the area under the ROC curve (AUC, Fielding & Bell 1997), will not be changed, despite considerable changes to the predicted P/A maps (and P/A-based performance indices). In fact, this approach may be considered as a thresholding method, yet instead of finding a single threshold per species, it finds a single threshold per site. On the other hand, Calabrese *et al.* (2014) used probability theory to show that the expected diversity of stacked SDMs in a site should equal the sum all species’ PoOs. Thus, adjusting the PoO according to a predicted alpha-capacity, as done by the summed-PoO approach, is better rooted in theory than the top-alpha approach. In addition, the AUC of the original and adjusted PoO will likely differ from one another.

Furthermore, the top-alpha approach is sensitive to the relative PoO of species in a site, while the summed-PoO approach is not. For example, imagine a species that has the same PoO in two sites, and the sites have identical alpha-capacity, but in the first site the PoO is greater than all other species and in the second site it has the lowest PoO. The summed-PoO approach will return identical adjusted PoO regardless of the species’ PoO rank, and therefore their P/A status will be the same. In contrast, the top-alpha approach will tag the species as present in the first site (where it is the highest ranked species) but absent in the second (where it is the lowest). In other words, since SDMs aim to identify which sites are best for a given species, they tend to ignore the second complimentary question: which species are best for a given site. The summed-PoO approach does not account for this second question while the top-alpha approach does.

In this manuscript we describe a new method to integrate alpha-diversity models and stacked PoO, which we call '*alpha-adjusted SDMs*'. The method aims to simultaneously satisfy two criteria. First, similar to the summed-PoO approach, the model should return adjusted PoO, based on probability theory, whose sum equals the alpha-capacity. Second, similar to the top-alpha approach, the model should be sensitive to relative PoO between species, i.e., it should account for both the best-sites-for-a-species and the best-species-for-a-site questions. We hypothesise that the model will perform well for small communities of highly interactive species for which alpha-capacity can be predicted with sufficient accuracy. While applying the method to a community comprising 87 stream macroinvertebrate species from an LTER watershed in Central Germany, we aim to answer the following questions:

**1.** Can the alpha-adjusted SDM predict alpha-diversity patterns?
**2.** Can the alpha-adjusted SDM predict beta-diversity patterns?
**3.** What affects the performance of the alpha-adjusted SDM at the species level?
    **3.1.** The error associated with modelling alpha-capacity?
    **3.2.** The gamma diversity (size of the species pool)?
    **3.3.** The type of community (random species or guild-based)?

## MATHERIALS AND METHODS

### *The Alpha-Adjusted SDM*

The alpha-adjusted SDM makes dual usage of the presence/absence of a set of $S$ species in $J$ sampled sites, alongside a set of predictors describing the conditions in each site. First, the sum over all species of the presence-absence table for each site is used to train an alpha-diversity model, which aims to predict the alpha diversity in each sampled and/or unsampled site ($\alpha_j$). The predicted alpha-diversity values are then rounded to the nearest integer in the range [0, S]. Second, we estimate the PoO of each species in each site ($P_{s,j}$) by training an SDM, and assume that 1- $P_{s,j}$ is the probability of absence. Note, that the assumption that the value returned by SDMs and its complement to 1 can be treated as probabilities of presence and absence (respectively) is fundamental to the alpha-adjusted SDMs approach. Thus, it is important to apply this method only in cases where the absences in the data can be considered true absences and where detectability issues are dealt with (Guillera-Arroita *et al.* 2015).

For any given α value, there are $T_\alpha = \binom{S}{\alpha}$ ways to select exactly α species from the S species. If the PoO values from the SDM are viewed as the probabilities of success in independent, yet different, Bernoulli trials, then the probability of any realization can be estimated using the Poisson-binomial distribution (Calabrese *et al.* 2014). We denote by $I_{s,j,\alpha,t}$ a binary variable that takes the value of 1 (or 0) if species *s* occurs in (or is absent from) site *j*, in realization *t* (out of $T_\alpha$) in which exactly $\alpha$ species occur. Thus, the probability of this realization is given by:

$$R_{j,\alpha,t} = \prod_{s=1}^{S}\left[P_{s,j} \cdot I_{s,j,\alpha,t} + \left(1 - P_{s,j}\right) \cdot \left(1 - I_{s,j,\alpha,t}\right)\right] \qquad \text{eqn1}$$

When eqn1 is summed over all $T_\alpha$ realizations, it will return the probability of the site to contain exactly α species, based on the SDMs of all species:

$$R_{j,\alpha} = \sum_{t=1}^{T}\left[\prod_{s=1}^{S}\left[P_{s,j} \cdot I_{s,j,\alpha,t} + \left(1 - P_{s,j}\right) \cdot \left(1 - I_{s,j,\alpha,t}\right)\right]\right] \qquad \text{eqn2}$$

Summing eqn2 over all alpha-diversity levels in the range [0, S] covers the entire probability space and returns a value of one. In addition, we can use the ratio eqn1/eqn2 to estimate the conditional probability of each realization $K_{j,,t}$ if the alpha diversity is known:

$$K_{j,\alpha,t} = \frac{R_{j,\alpha,t}}{R_{j,\alpha}} = \frac{\prod_{s=1}^{S}\left[P_{s,j} \cdot I_{s,j,\alpha,t} + \left(1-P_{s,j}\right) \cdot \left(1-I_{s,j,\alpha,t}\right)\right]}{\sum_{t=1}^{T}\left[\prod_{s=1}^{S}\left[P_{s,j} \cdot I_{s,j,\alpha,t} + \left(1-P_{s,j}\right) \cdot \left(1-I_{s,j,\alpha,t}\right)\right]\right]} \qquad \text{eqn3}$$

Finally, the alpha-adjusted probabilities (*adjP*$_{s,j,\alpha}$) for species *s* in site *j* for alpha values of $\alpha$ can be estimated by summing the conditional probability of all realizations in which the species is present:

$$adjP_{s,j,\alpha} = \sum_{t=1}^{T}\left[I_{s,j,\alpha,t} \cdot K_{j,\alpha,t}\right] = \sum_{t=1}^{T}\left[I_{s,j,\alpha,t} \cdot \left[\frac{\prod_{s=1}^{S}\left[P_{s,j} \cdot I_{s,j,\alpha,t} + (1-P_{s,j}) \cdot (1-I_{s,j,\alpha,t})\right]}{\sum_{t=1}^{T}\left[\prod_{s=1}^{S}\left[P_{s,j} \cdot I_{s,j,\alpha,t} + (1-P_{s,j}) \cdot (1-I_{s,j,\alpha,t})\right]\right]}\right]\right] \qquad \text{eqn4}$$

Interestingly, eqn4 can also be approximated using methods that do not require listing of all potential realizations (supporting information S1), which may become computationally difficult when alpha and/or gamma are high. Eqn4 (or eqn7 from S1) can be used to adjust the PoO of all species in any given site using the alpha-capacity from the alpha-diversity model or any other alpha value (e.g., the observed alpha diversity). Summing the adjusted probabilities over all species will return the alpha-capacity. The alpha-adjusted SDM will return an adjusted PoO of 0 for all species when the alpha-capacity is 0. When the alpha-capacity equals the size of the species pool (S), the adjusted probabilities will equal 1 for all species. For all other alpha values, in any given site the original ranking of species according to their PoO will be kept in the alpha-adjusted PoO. However, the relative ranking of sites for a given species will change. Table 1 provides a simple example of four species in two sites, differing in PoO rankings and values. Within each site, he ranking is kept in the alpha-adjusted probabilities (as long as alpha is not 0 or S). However, if the alpha diversity of the two sites is identical, then for species A, the adjusted PoO will be higher for site 1 than in site 2, despite the lower values in the original SDM. In other words, although site 1 is not the best site for species A, species A is the best species for site 1, and the alpha-adjusted SDM accounts for that.

### *Study site and sampling scheme*

To test this model framework, we used stream community data from 185 sites in the LTER site Rhine-Main-Observatory that covers the entire watershed of the Kinzig River (1058 km$^2$) in Central Germany. The number of visits per site in the past 15 years ranged between 1 and 20 (131, 30, and 20 sites were sampled once, twice, or three times, respectively). Sampling at all sites followed the protocol described in Haase *et al.* (2004): At each sampling site benthic invertebrates were sampled using a multi-habitat sampling approach, which reflected the proportion of the microhabitat types that were present with at least 5% cover. At each site, 20 sample units were taken, each 25 x 25 cm in size, resulting in a total of 1.25 m$^2$ of river bottom being sampled. The sample material was sorted and identified in the lab to the species level.

We restricted our analysis to 87 species, which were assigned *a priori* to five feeding guilds based on the *freshwaterecology.info* database (Schmidt-Kloiber & Hering 2015; see Table S1). The first feeding guild included 35 generalist species from five different orders, the second was composed of 24 predators from eight orders; the third had 14 grazers from five orders; the fourth included eight gatherers from three orders; and the fifth feeding guild had six filter feeders from five orders.

### *Species distribution models*

We choose random-forest (Breiman 2001) as the SDM algorithm since it is one of the most accurate algorithms available today, it performs better than other algorithms when predicting alpha and beta-diversity patterns from stacked SDMs (Benito, Cayuela & Albuquerque 2013) and because it does not require external validation datasets due to its internal randomization procedure. We used all sites in which a species was observed as presences. Sites sampled at least three times where a given species was not observed were incorporated as absences, reducing the possibility of false absences in our training data. The explanatory variables included two climatic, three land-cover, two topographic, three geological and two hydrological predictors (Table S2). Some predictors were processed to optimally describe the freshwater ecosystem, such as calculating the relative proportion of land-use and geology classes in the upper sub-catchment (Kuemmerlen *et al.* 2014; Kuemmerlen *et al.* 2016). We used the '*randomForest*' package in R (R Core Team 2016) to fit the models, keeping the default settings, with the number of trees increased to 10,000.

### *Alpha-diversity models*

For each community we first estimated the observed alpha as the mean species richness observed at a site per visit. We used the mean-per-visit instead of the cumulative number of species over all visits because the sites differ in number of sampling events and temporal turnover may arise due to competitive interactions (overestimating the alpha-capacity). We modelled mean alpha diversity against the same predictors as for the SDMs using regression-tree based random-forest.

### *Fitting the Alpha-adjusted SDMs*

For each community, we fitted the alpha-adjusted SDM by applying eqn4 for each site, using the PoO of all species from the original SDM and the alpha diversity (rounded to the nearest integer). When fitting the alpha-adjusted SDM, the number of potential realizations increases considerably with alpha and gamma, making it infeasible to process all potential realizations in a reasonable time. If the number of realizations exceeded 50,000, we randomly selected 50,000 realizations. An alternative method, based on eqn7 (S1), is computationally more efficient and accurate but unstable and failed to return meaningful results in some cases. Supporting information S2 contains a simplified annotated workflow for fitting SDMs, fitting alpha-diversity models and adjusting the PoO, (for both eqn4 and eqn7) using data on the eight gatherers species (Supporting information S3). It also maps the original and adjusted PoO with the river network as background (Supporting information S4).

### *Question 1- Predicted alpha-diversity patterns*

For each community, we explored five different methods that predict alpha-diversity patterns and compared them to the observed mean alpha diversity. First was modelling alpha directly against environmental variables, as described above. Second, following the suggestion of Calabrese *et al.* (2014), we sum the PoO values of the SDMs across all species for each site. Third, for each species we converted the PoO values from the original SDMs to P/A data by applying a threshold that maximizes the True Skills Statistics (TSS, sensitivity+specificity-1; Allouche, Tsoar & Kadmon 2006) and then stacked the predicted P/A data for each site. The fourth and fifth predictions of alpha diversity are from the alpha-adjusted SDM, which we applied twice (see below) on each community, once using the observed alpha when adjusting and once using the modelled alpha (from the alpha-diversity model). In both cases, we converted the adjusted PoO to P/A data using the maximal TSS threshold and then stacked the P/A data for each site. Note that for the alpha-adjusted SDMs we did not stack the PoO, as stacking them will return the exact alpha values used for adjusting.

### *Question 2- Predicted beta-diversity patterns*

We explored the ability of the stacked SDMs and the alpha-adjusted SDMs to predict assemblage composition using pairwise similarity indices. For each community, we took the maximal TSS threshold P/A maps of the original SDMs and the two alpha-adjusted SDMs (using the observed and modelled alpha). Then, for each site we calculated the pairwise similarity between each of the three predicted local assemblages and the observed assemblage (species detected during sampling). For the observed assemblage, we treated in this analysis all sites in which a species was not observed as absences, regardless of the number of times the site was sampled.

We used two pairwise similarity indices, one that does not account for double absences and one that does. The first was Sørenson similarity index: $2 \cdot A / (2 \cdot A + B + C)$, with A being the number of species shared by the two assemblages, and B and C the number of species found only in the first or second assemblage. The second measure was $(2 \cdot A + 2 \cdot D) / (2 \cdot A + B + C + 2 \cdot D)$, with D being the number of species not occurring in either of the two assemblages (Eq. 7.3, pg. 255 in Legendre & Legendre 1998). As this function converges to Sørenson's index when $D=0$, we term it the 'extended-Sørenson' in this publication. We used the extended-Sørenson index since the Sørenson index cannot be calculated when both assemblages are empty. Although this rarely happens in the SDM generated assemblages, it may occur in the alpha-adjusted assemblages whenever the alpha-capacity is set to 0. Thus, the extended-Sørenson allows such sites to contribute to the overall similarity value whereas they would be excluded using the traditional Sørenson measure. Furthermore, the alpha-adjusted assemblages are restricted by the alpha-capacity and thus are likely to predict considerably fewer species than the SDM assemblages, which may translate to large differences in D.

*Question 3- Species-level performance and the 230 modelled communities*

Since SDMs and alpha-adjusted SDMs produce comparable PoO values for each species in each site, we followed a similar procedure when assessing their performance against the training data. We used the '*SDMtools*' package in R to calculate two performance indices: AUC and the maximal TSS (the highest TSS of P/A maps generated using 1000 equally spaced thresholds between 0 and 1).

As noted above, to tease apart the effect of the error associated with modelling alpha (question 3.1), in each community we fitted the alpha-adjusted SDM twice—once using the observed alpha and once using the modelled alpha. We refer to the performance at the species level using the observed and modelled alpha-adjusted SDMs as $\alpha_{obs}Adj_{TSS}$ and $\alpha_{mod}Adj_{TSS}$, respectively, and the unadjusted SDM performance as $SDM_{TSS}$ (with similar annotation for AUC). For each community we focused on several comparisons. To explore the overall performance of the alpha-adjusted SDM we compared for each species the $\alpha_{mod}Adj_{TSS}$ to $SDM_{TSS}$. To explore the potential of the alpha-adjusted SDM to increase the performance at the species level if alpha can be modelled with no error, we compared the $\alpha_{obs}Adj_{TSS}$ to the $SDM_{TSS}$. Finally, to focus on the effect of the error associated with the alpha-diversity model we compared the $\alpha_{obs}Adj_{TSS}$ to the $\alpha_{mod}Adj_{TSS}$.

To explore the effect of gamma diversity on the performance of the alpha-adjusted SDM (question 3.2) we created communities from the same guild, but with different gamma diversity. We first fitted the alpha-adjusted SDM to each guild, using all of the species. We then created random communities of each guild with only a subset of the species, by randomly selecting a pre-defined number of species form the entire species pool of the guild. For the generalist species, in addition to the full community of 35 species, we created 15 random communities of size 24, 14, 8 and 6 species (for a total of 61 communities). For the predators, in addition to the full 24 species community we created 15 random communities of size 14, 8 and 6 (totalling 46 communities). We repeated this also for the grazers (1 community of 14 species, 15 of 8 species and 15 of 6 species) and gatherers (1 community of 8 and 15 of 6 species). We did not explore communities smaller than 6, so only a single community of filterers was modelled.

We have focused on species from the same guild since the strength of biotic interaction within a guild is expected to be higher than between guilds. To explore if indeed the type of community affect the performance of the alpha-adjusted SDM (question 3.3), we repeated the analysis for communities of the same gamma diversities (35, 24, 14, 8, and 6), in which the species were randomly selected from the entire species pool of 87 species. We repeated this procedure 15 times for each value of gamma diversity. If the strength of biotic interaction is stronger within guilds and the alpha-adjusted SDM accounts for this interaction, we would expect the performance of the alpha-adjusted SDM in the guild-based communities to be better than in random communities of the same size. In total, we fitted the alpha-adjusted SDM to 230 communities.

## RESULTS

Overall, the SDMs had mean TSS and AUC scores of 0.639 (±0.019 SE) and 0.814 (±0.013) respectively, indicating good model performance. The highest mean TSS and AUC were observed for grazers, followed by generalists, gatherers and predators, while the filterers received the lowest values (Table S1).

### *Question 1- Predicted alpha-diversity patterns*

The predicted alpha diversity according to the five methods is summarized in Fig. 1. The alpha-diversity models tended to overestimate diversity when observed diversity was low and underestimate diversity when observed diversity was high (Alpha Model, Fig. 1). Both stacking the PoO of the SDMs (PoO-SDM) and their thresholded P/A maps (PA-SDM) tended to overpredict alpha-diversity (Fig. 1). Interestingly, stacking the P/A maps of the alpha-adjusted PoO using modelled alpha (PA-Adj-Mod) also overestimated alpha diversity, whereas there was high correlation to observed alpha when the adjustment was based on the observed alpha (PA-Adj-Obs). In general, in most communities, the total number of occupancies (summed over all species across all sites) was higher using the PA-SDM than using PA-Adj-Obs and PA-Adj-Mod (Figure S1).

### *Question 2- Predicted beta-diversity patterns*

The predictions of assemblage composition from the alpha-adjusted assemblage where slightly better than those made of the SDM assemblage when using the Sørenson index (Fig. 2). However, when using the extended-Sørenson index, which accounts for double absences, the alpha-adjusted assemblages were considerably more similar to the observed assemblages than the SDM assemblages. This pattern was evident both when using the observed and modelled alpha diversity in the alpha-adjusted SDM.

### *Question 3- Species-level performance*

We found mixed results for the performance of the alpha-adjusted SDM at the species-level (question 3). The error associated with modelling alpha diversity had a considerable negative effect on the ability of the alpha-adjusted SDM to predict species distribution patterns (question 3.1). Thus, for most species in most communities, $\alpha_{obs}Adj_{TSS}$ was considerably higher than $\alpha_{mod}Adj_{TSS}$ (Fig. 3). Similar results were observed for AUC. In most cases, we observed $\alpha_{obs}Adj_{TSS}$ to be higher than $SDM_{TSS}$, while $\alpha_{mod}Adj_{TSS}$ was mostly lower than $SDM_{TSS}$ (Fig. 4). More specifically, $\alpha_{obs}Adj_{TSS}$ was larger than $SDM_{TSS}$ in 100%, 90.8%, 76.8%, 58.6%, 58.3% of the filterers, gatherers, grazers, predators and generalist combinations of species and communities, respectively. Similar values were observed when focusing only on 6 species communities (100%, 91.1%, 76.6%, 64.4% and 68.8%, respectively).

The ability of the alpha-adjusted SDM to enhance the performance at the species-level decreased with increasing gamma diversity when the observed alpha was used (question 3.2, Fig. 4). The increase in mean and median TSS with decrease of gamma diversity was mainly observed in the generalist and predator guilds, and was not observed in the gatherers and grazers guilds (Fig. 5). The same pattern was also observed for random communities (Fig. 5). Interestingly, model performance seemed to increase with increasing gamma diversity when the modelled alpha was used (Fig. 4).

The alpha-adjusted SDM increased species' TSS and AUC more when the species were included in functionally-defined communities (question 3.3). This was mostly evident when comparing communities from the filterers and gatherers guilds to random communities (Fig. 4, top row). Similar patterns, albeit less obvious, were observed for small communities of predators and generalists. As mentioned above, this relation is reversed when the modelled alpha was used, where random communities exhibited higher mean TSS and AUC than guild-based communities (Fig. 4, bottom row).

## DISCUSSION

Here, we introduce a new approach, alpha-adjusted SDMs, to incorporate surrogates for biotic interactions as well as alpha-capacity constraints into correlative SDMs for community assemblages. Our aim was to develop a method that combines the advantages of the two other published approaches (Calabrese *et al.* 2014; D'Amen *et al.* 2015a; D'Amen, Pradervand & Guisan 2015). More specifically, we aimed for a method that *(a)* returns PoOs whose sum equals the alpha-capacity, *(b)* incorporates both the logic of the best-sites-per-species and best-species-per-site predictions. The alpha-adjusted SDM satisfies both these criteria. For our case-study, we found that the alpha-adjusted SDM may increase the accuracy by which we predict alpha-diversity patterns (Fig. 1), assemblage composition (Fig. 2) and species-distribution patterns (Figs. 3, 4, 5), if we can model alpha with sufficient accuracy and if we can focus on small communities of closely interacting species.

However, for most communities we were unable to model alpha diversity with high accuracy (question 1, Fig. 1). Thus, the alpha-adjusted SDM (constrained by the modelled alpha) was outperformed by the original SDMs for most species in most communities (question 3.1, Figs. 3, 4). Despite this error, we still produced more reliable predictions of assemblage composition, especially when using the extended Sørenson index that accounts for dual absences (question 2, Fig. 2). This result suggests that the alpha-adjusted SDM may be better at predicting absences, even when it cannot accurately identify presences. Indeed, the alpha-adjusted SDM predicted lower total occupancies than the original SDMs (Fig. S1, note that although alpha-capacity is used to constrain the PoO, it has no direct effect on the PA map and the prevalence of each species).

Our results, when using the observed alpha as constraints, are even more encouraging. In most cases, the alpha-adjusted SDM returned higher TSS than the original SDM. However, the alpha-adjusted SDM performance differed between guilds, with better performance for filterer, gatherers and grazers than for predators and generalist (Fig. 3). One potential explanation is that guilds differed in gamma diversity (question 3.2). Although quantitative data on competitive interaction coefficients of species rich communities is scarce, empirical data on foodwebs suggest a skewed distribution of interaction strengths with very few strong interactions and a large number of weak interactions (Wootton & Emmerson 2005). The same pattern was found when experimentally manipulating 428 pairs of  freshwater green algae species (Fritschie *et al.* 2014). Thus, when applying the alpha-adjusted SDM on large communities (as we have done for the predators and generalists), we may be adjusting a focal species’ PoO to account for the potential presence of a large number of species with which it interacts only very weakly. Similarly, we observed (Figs. 4, 5) a decrease in model performance with increasing gamma diversity within guilds.

However, even if we control for differences in gamma diversity, by standardising community size at 6 species, we still observed differences in relative performance between guilds (Fig. 3). We attribute this to the effect of the strength of biotic interactions (question 3.3). The performance using alpha-adjusted SDMs was higher for single-guild communities than randomly assembled communities (Figs. 4, 5). We assumed that species from the same feeding guild are more likely to compete directly with one another over limiting resources. This assumption is likely valid for the filterers, grazers and gatherers, which directly compete for the same resource, but to a lesser extent for generalist and predators. The generalists often feed opportunistically avoiding competition by exploiting various spatio-temporally abundant resources, while the predators interact primarily with their prey, to which they are intrinsically linked through co-evolved hunting and defence strategies, downgrading the importance of intra-guild interactions. In addition to lower mean performance, the two guilds also showed high variance in model performance when gamma diversity was small (Fig. 4), which may arise from randomly selecting pairs of strongly interacting species in some random communities and failing to do so in others.

In general terms, there are currently two main frameworks aiming to provide more accurate predictions of distribution patterns at the species or the community levels. Both frameworks try to account for the main processes that affect species distribution patterns. The first approach aims to develop dynamic process-based models that explicitly incorporate important processes such as dispersal and biotic interaction into the modelling framework (see examples in Wisz *et al.* 2013). However, process-based models need parametrization which is currently unachievable for the majority of species and systems. Unfortunately, this is unlikely to change in the near future.

Thus we are left with the second framework that does not try to account for the processes directly, but instead uses interacting-correlative models as surrogates. Under this second framework, processes are accounted for by incorporating and integrating information from multiple correlative models in an ecologically meaningful way (Guisan & Rahbek 2011; Boulangeat, Gravel & Thuiller 2012), and much theoretical consideration has been invested in selecting the different ways by which to integrate the constituting models. However, if we are to create successful ‘process-based, correlative frameworks’, we must define *a priori* which criteria should be satisfied by the hierarchical correlative chain of models. Here, we developed one such method that unifies the main advantages of other applications of the SESAM framework (Calabrese *et al.* 2014; D'Amen *et al.* 2015a; D'Amen, Pradervand & Guisan 2015). However, the ability of our framework (and other such frameworks) to provide better predictions lies in the trade-off between incorporating multiple processes and multiple sources of errors. Expressed differently: the addition of an additional correlative model to a chain of models is most successful if the error associated with the new model is low, thus improving the overall modelling framework. In our case, the alpha-diversity model introduced a significant level of error, superseding any potential advantage of accounting for alpha-capacity constraints. However, we also found very encouraging results when using the observed alpha values to inform our models. Thus we believe that developing new and improved methods for correlative alpha-diversity models is one of the most important challenges for predictive ecology in the next few years. If further developed, the alpha-adjusted SDM approach could allow for the prediction of species and community level distribution patterns over wide extents and at fine resolutions with the level of accuracy required for effective conservation and management.

## ACKNOWLEDGEMENTS

This work was financed by the EU BON project (www.eubon.eu) that is a 7th Framework Programme funded by the European Union under Contract No. 308454. We thank the associate editor, one anonymous reviewer and D. Rocchini for their constructive reviews.

## DATA ACCESSIBILITY

All environmental and community data used in this manuscript is stored in the ‘Ecobase’ database of the LTER site Rhine-Main-Observatory. See also LTER database DEIMS, UUID: 9f9ba137-342d-4813-ae58-a60911c3abc1.

## REFERENCES

Allouche, O., Tsoar, A. & Kadmon, R. (2006) Assessing the accuracy of species distribution models: prevalence, kappa and the true skill statistic (TSS). *Journal of Applied Ecology,* **43,** 1223-1232.

Benito, B.M., Cayuela, L. & Albuquerque, F.S. (2013) The impact of modelling choices in the predictive performance of richness maps derived from species-distribution models: guidelines to build better diversity models. *Methods in Ecology and Evolution,* **4,** 327-335.

Boulangeat, I., Gravel, D. & Thuiller, W. (2012) Accounting for dispersal and biotic interactions to disentangle the drivers of species distributions and their abundances. *Ecology letters,* **15,** 584-593.

Breiman, L. (2001) Random forests. *Machine Learning,* **45,** 5-32.

Calabrese, J.M., Certain, G., Kraan, C. & Dormann, C.F. (2014) Stacking species distribution models and adjusting bias by linking them to macroecological models. *Global Ecology and Biogeography,* **23,** 99-112.

Carroll, C. (2010) Role of climatic niche models in focal-species-based conservation planning: Assessing potential effects of climate change on Northern Spotted Owl in the Pacific Northwest, USA. *Biological Conservation,* **143,** 1432-1437.

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.

D'Amen, M., Dubuis, A., Fernandes, R.F., Pottier, J., Pellissier, L. & Guisan, A. (2015a) Using species richness and functional traits predictions to constrain assemblage predictions from stacked species distribution models. *Journal of Biogeography,* **42,** 1255-1266.

D'Amen, M., Pradervand, J.-N. & Guisan, A. (2015) Predicting richness and composition in mountain insect communities at high resolution: a new test of the SESAM framework. *Global Ecology and Biogeography,* **24,** 1443-1453.

D'Amen, M., Rahbek, C., Zimmermann, N.E. & Guisan, A. (2015b) Spatial predictions at the community level: from current approaches to future frameworks. *Biological Reviews***,** n/a-n/a.

Distler, T., Schuetz, J.G., Velásquez-Tibatá, J. & Langham, G.M. (2015) Stacked species distribution models and macroecological models provide congruent projections of avian species richness under climate change. *Journal of Biogeography,* **42,** 976-988.

Elith, J. & Leathwick, J.R. (2009) Species Distribution Models: Ecological Explanation and Prediction Across Space and Time. *Ecology, Evolution, and Systematics,* **40,** 677-697.

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.

Fritschie, K.J., Cardinale, B.J., Alexandrou, M.A. & Oakley, T.H. (2014) Evolutionary history and the strength of species interactions: testing the phylogenetic limiting similarity hypothesis. *Ecology,* **95,** 1407-1417.

Guillera-Arroita, G., Lahoz-Monfort, J.J., Elith, J., Gordon, A., Kujala, H., Lentini, P.E., McCarthy, M.A., Tingley, R. & Wintle, B.A. (2015) Is my species distribution model fit for purpose? Matching data and models to applications. *Global Ecology and Biogeography,* **24,** 276-292.

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., Tingley, R., Baumgartner, J.B., Naujokaitis-Lewis, I., Sutcliffe, P.R., Tulloch, A.I.T., Regan, T.J., Brotons, L., McDonald-Madden, E., Mantyka-Pringle, C., Martin, T.G., Rhodes, J.R., Maggini, R., Setterfield, S.A., Elith, J., Schwartz, M.W., Wintle, B.A., Broennimann, O., Austin, M., Ferrier, S., Kearney, M.R., Possingham, H.P. & Buckley, Y.M. (2013) Predicting species distributions for conservation decisions. *Ecology letters,* **16,** 1424-1435.

Haase, P., Lohse, S., Pauls, S., Schindehütte, K., Sundermann, A., Rolauffs, P. & Hering, D. (2004) Assessing streams in Germany with benthic invertebrates: development of a practical standardised protocol for macroinvertebrate sampling and sorting. *Limnologica - Ecology and Management of Inland Waters,* **34,** 349-365.

Harris, D.J. (2015) Generating realistic assemblages with a joint species distribution model. *Methods in Ecology and Evolution,* **6,** 465-473.

Heikkinen, R.K., Luoto, M., Virkkala, R., Pearson, R.G. & Körber, J.-H. (2007) Biotic interactions improve prediction of boreal bird distributions at macro-scales. *Global Ecology and Biogeography,* **16,** 754-763.

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

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.

Kuemmerlen, M., Schmalz, B., Guse, B., Cai, Q., Fohrer, N. & Jaehnig, S.C. (2014) Integrating catchment properties in small scale species distribution models of stream macroinvertebrates. *Ecological Modelling,* **277,** 77-86.

Kuemmerlen, M., Stoll, S., Sundermann, A. & Haase, P. (2016) Long-term monitoring data meet freshwater species distribution models: Lessons from an LTER-site. *Ecological Indicators,* **65,** 122-132.

Legendre, P. & Legendre, L. (1998) *Numerical Ecology, Second English Edition*. Elsevier Science B.V. , Amsterdam, The Netherlands.

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.

Newbold, T., Gilbert, F., Zalat, S., El-Gabbas, A. & Reader, T. (2009) Climate-based models of spatial patterns of species richness in Egypt's butterfly and mammal fauna. *Journal of Biogeography,* **36,** 2085-2095.

Peterson, A.T. & Soberon, J. (2012) Species Distribution Modeling and Ecological Niche Modeling: Getting the Concepts Right. *Natureza & Conservacao,* **10,** 102-107.

R Core Team (2016) R: A language and environment for statistical computing. *R Foundation for Statistical Computing, Vienna, Austria, http://www.R-project.org*.

Schmidt-Kloiber, A. & Hering, D. (2015) www.freshwaterecology.info - An online tool that unifies, standardises and codifies more than 20,000 European freshwater organisms and their ecological preferences. *Ecological Indicators,* **53,** 271-282.

Soberón, J. (2007) Grinnellian and Eltonian niches and geographic distributions of species. *Ecology letters,* **10,** 1115-1123.

Thuiller, W., Pollock, L.J., Gueguen, M. & Muenkemueller, T. (2015) From species distributions to meta-communities. *Ecology letters,* **18,** 1321-1328.

Wiens, J.A., Stralberg, D., Jongsomjit, D., Howell, C.A. & Snyder, M.A. (2009) Niches, models, and climate change: Assessing the assumptions and uncertainties. *Proceedings of the National Academy of Sciences,* **106,** 19729-19736.

Wisz, M.S., Pottier, J., Kissling, W.D., Pellissier, L., Lenoir, J., Damgaard, C.F., Dormann, C.F., Forchhammer, M.C., Grytnes, J.-A., Guisan, A., Heikkinen, R.K., Høye, T.T., Kühn, I., Luoto, M., Maiorano, L., Nilsson, M.-C., Normand, S., Öckinger, E., Schmidt, N.M., Termansen, M., Timmermann, A., Wardle, D.A., Aastrup, P. & Svenning, J.-C. (2013) The role of biotic interactions in shaping distributions and realised assemblages of species: implications for species distribution modelling. *Biological Reviews of the Cambridge Philosophical Society,* **88,** 15-30.

Wootton, J.T. & Emmerson, M. (2005) Measurement of Interaction Strength in Nature. *Annual Review of Ecology, Evolution, and Systematics,* **36,** 419-444.

## SUPPORTING INFORMATION

Table S1 – Species names, acronyms and feeding guilds, along with their prevalence and the performance indices of the original SDM

Table S2 – Predictors used in the SDMs and alpha-diversity models

Supporting information S1 – Poisson-binomial distribution approximations of the alpha-adjusted SDM

Supporting information S2 – R codes for applying the alpha-adjusted SDM analysis (ExampleGatherers.R)

Supporting information S3 – Example data set for the 8 gatherer species (GathererRMO.csv)

Supporting information S4 – The background river network (RiverNetwork.csv)

Figure S1 – The total number of occupancies predicted for each community by the original SDM, and the two alpha-adjusted SDMs (based on max TSS PA maps).

## AUTHORS CONTRIBUTIONS STATEMENT

YG and CJM conceived the ideas and designed the methodology. MK, SS and PH collected the data. YG analysed the data and led the writing of the manuscript. All authors contributed considerably to later drafts and played a crucial role in completing this research.

## TABLES

**Table 1:** An example of the probability of occurrence (PoO) values generated by the alpha-adjusted SDM for 4 species in two sites. Species A has higher PoO values in site 2 than in site 1, but relative to other species, it is ranked lowest in site 2 and highest in site 1. Thus, for similar α values, the adjusted PoO of species A are higher in site 1 than in 2.

| | | SDM | Alpha-adjusted PoO | | | | |
|---|---|---|---|---|---|---|---|
| Site | Species | PoO | α = 0 | α = 1 | α = 2 | α = 3 | α = 4 |
| 1 | A | 0.35 | 0 | 0.41 | 0.70 | 0.89 | 1 |
| | B | 0.20 | 0 | 0.19 | 0.44 | 0.77 | 1 |
| | C | 0.30 | 0 | 0.32 | 0.63 | 0.86 | 1 |
| | D | 0.10 | 0 | 0.08 | 0.23 | 0.48 | 1 |
| 2 | A | 0.45 | 0 | 0.10 | 0.26 | 0.52 | 1 |
| | B | 0.60 | 0 | 0.17 | 0.42 | 0.74 | 1 |
| | C | 0.70 | 0 | 0.27 | 0.58 | 0.83 | 1 |
| | D | 0.80 | 0 | 0.46 | 0.74 | 0.91 | 1 |

## FIGURES

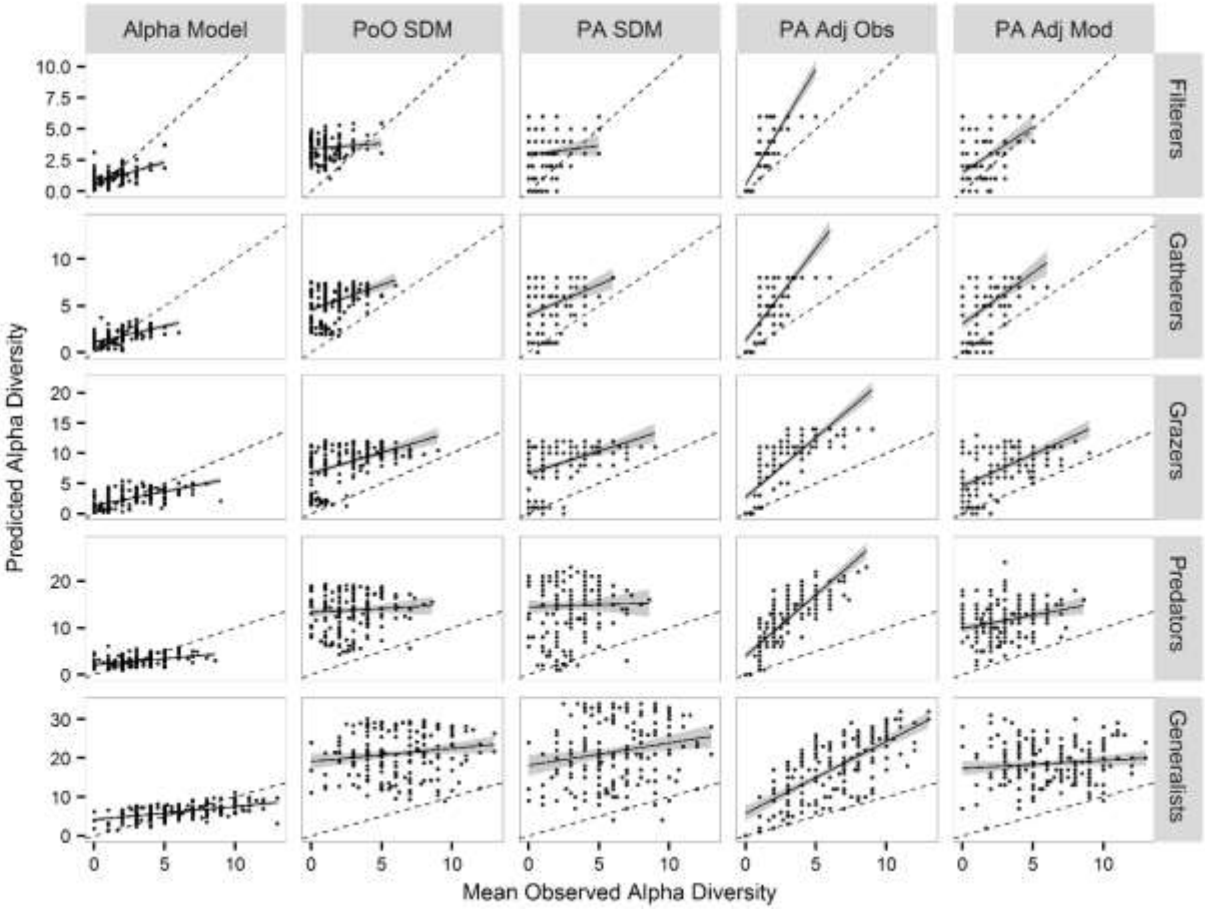

**Figure 1:** The predicted alpha diversity according to five different methods plotted against the mean observed alpha diversity in each site. Result for the full communities of each guild. Dashed line is the unity line. Solid lines are OLS regressions (±S.E.).

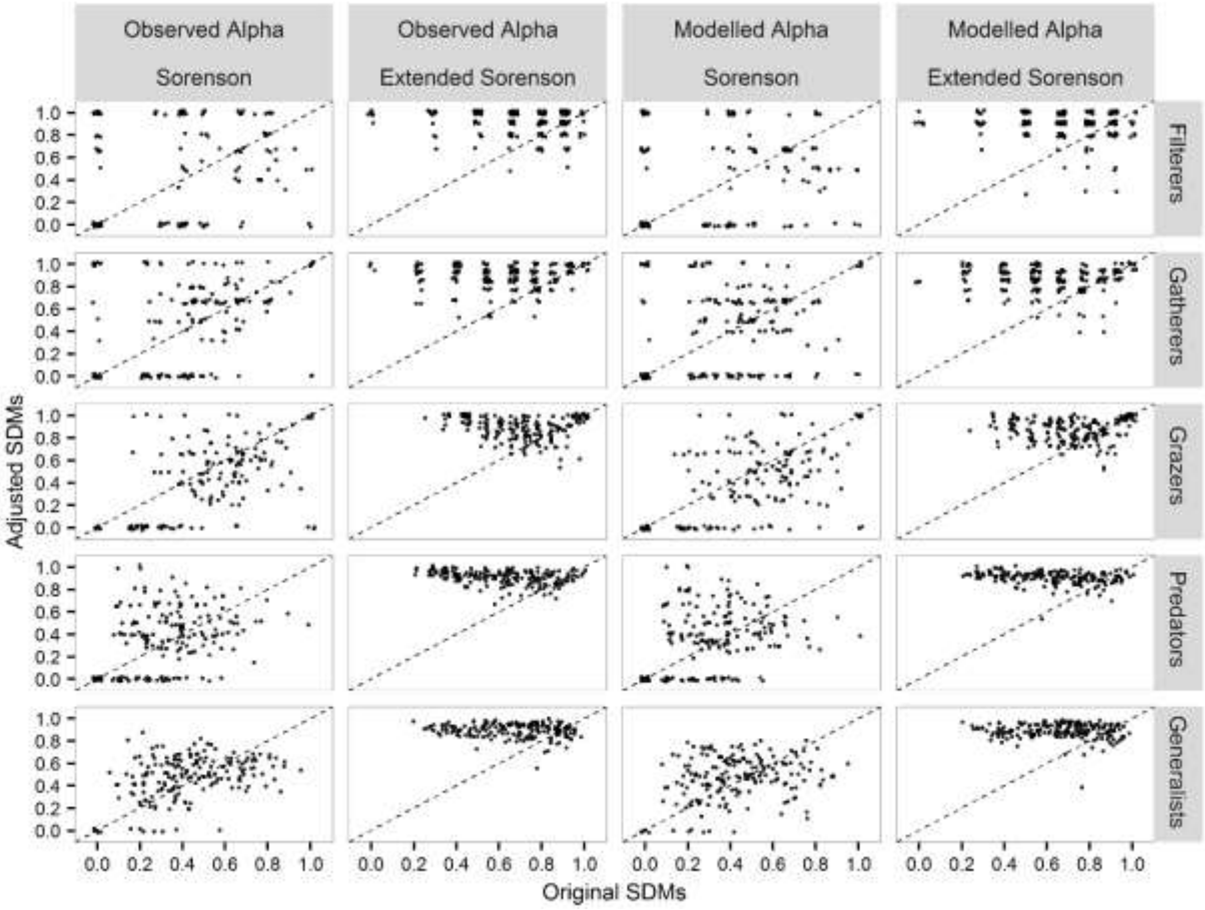

**Figure 2:** The pairwise similarity values between the alpha-adjusted SDM and the observed assemblage, plotted against the pairwise similarity between the original SDM and the observed assemblage. Points above the unity line (dashed) are sites in which the alpha-adjusted SDM outperformed the original SDM. Result for the full communities of each guild.

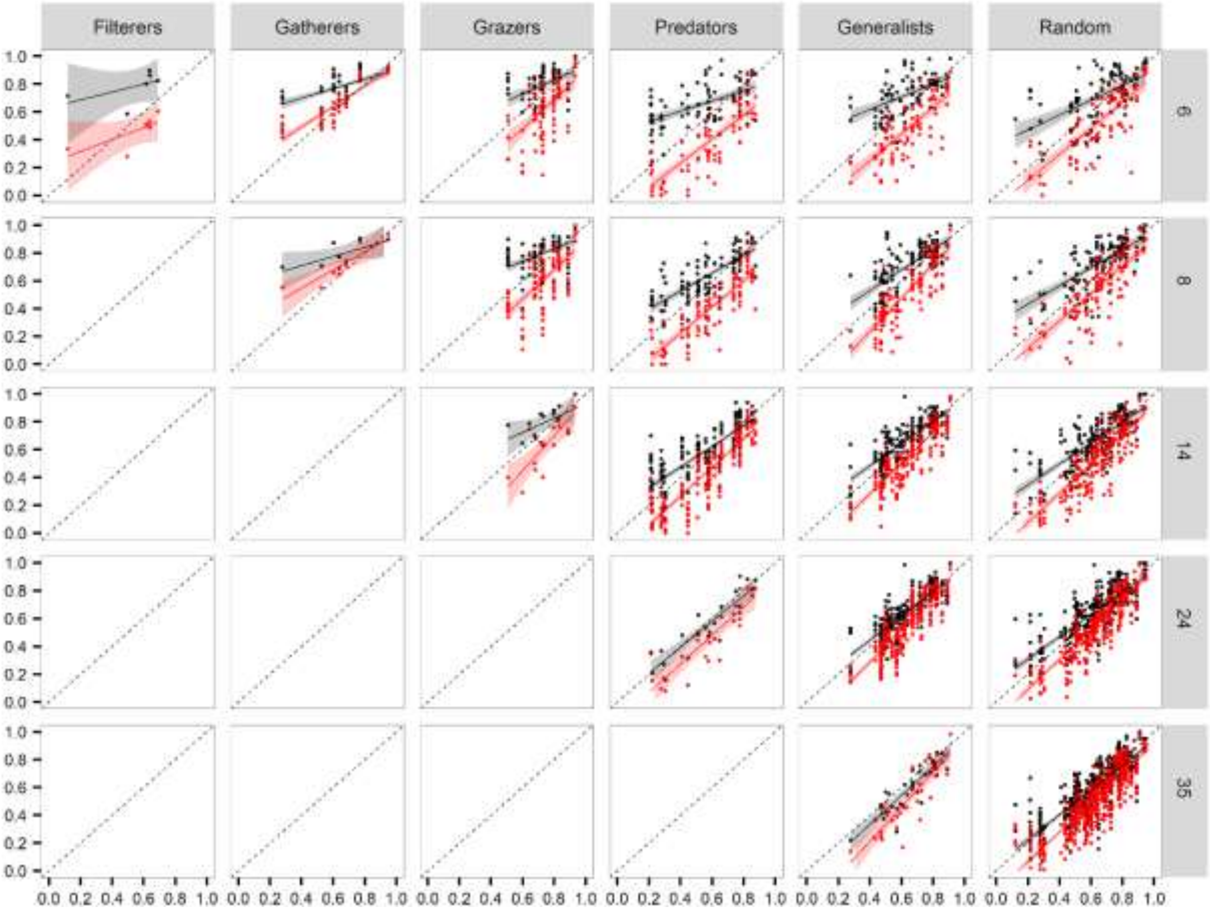

**Figure 3:** The TSS of the adjusted SDM plotted against the TSS of the original SDM. Points above the unity line (dashed) are species in which the alpha-adjusted SDM outperformed the original SDM. The adjusted TSS when using the observed alpha (black) was consistently higher than when using the modelled alpha (red). Each panel is for a combination of guild and community size and contains all 15 communities of the combination. Solid lines are OLS regressions (±S.E.)

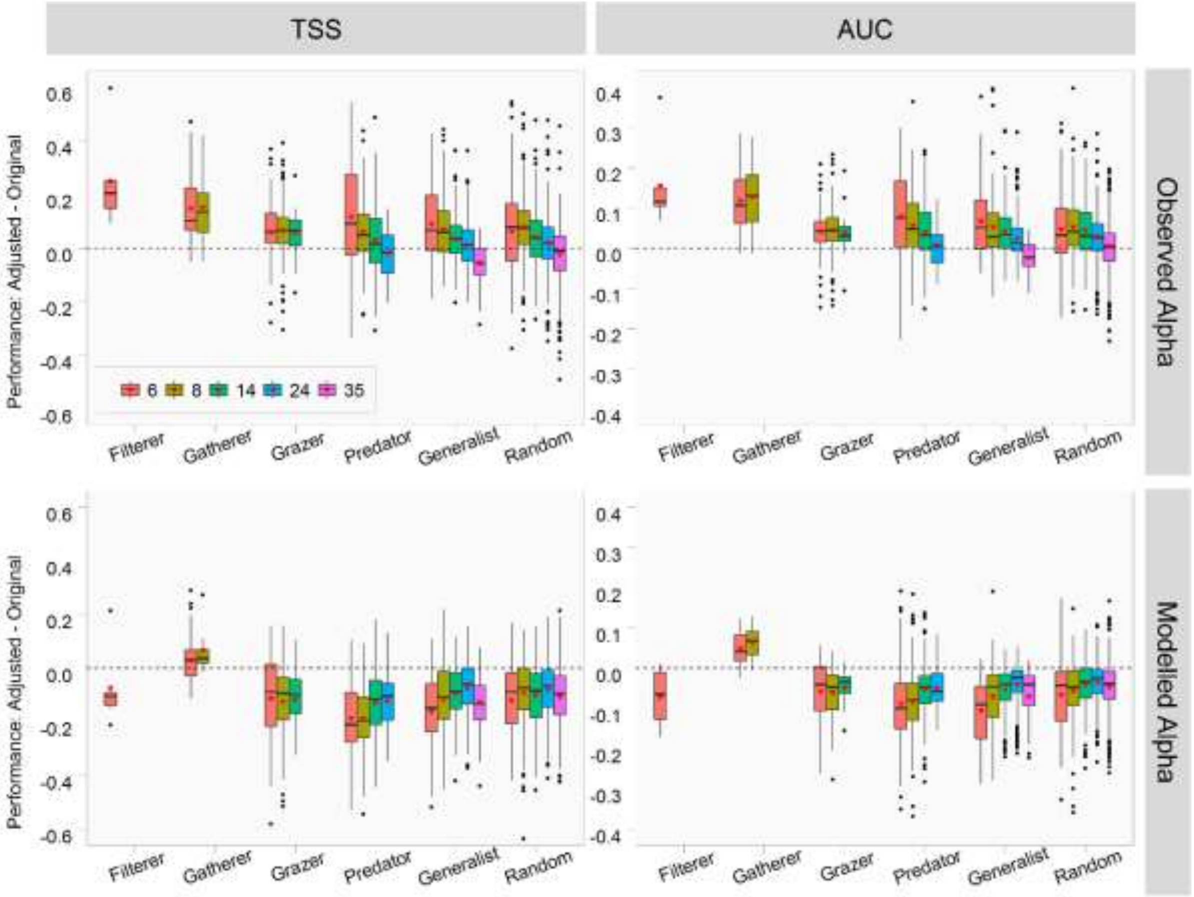

**Figure 4:** The difference between the adjusted and the original TSS (and AUC) for different guilds and gamma diversities, based on the observed and modelled alpha diversity. Values larger than 0 (horizontal line) are cases in which the alpha-adjusted SDM outperformed the original SDM. The means are given as red triangles.

605

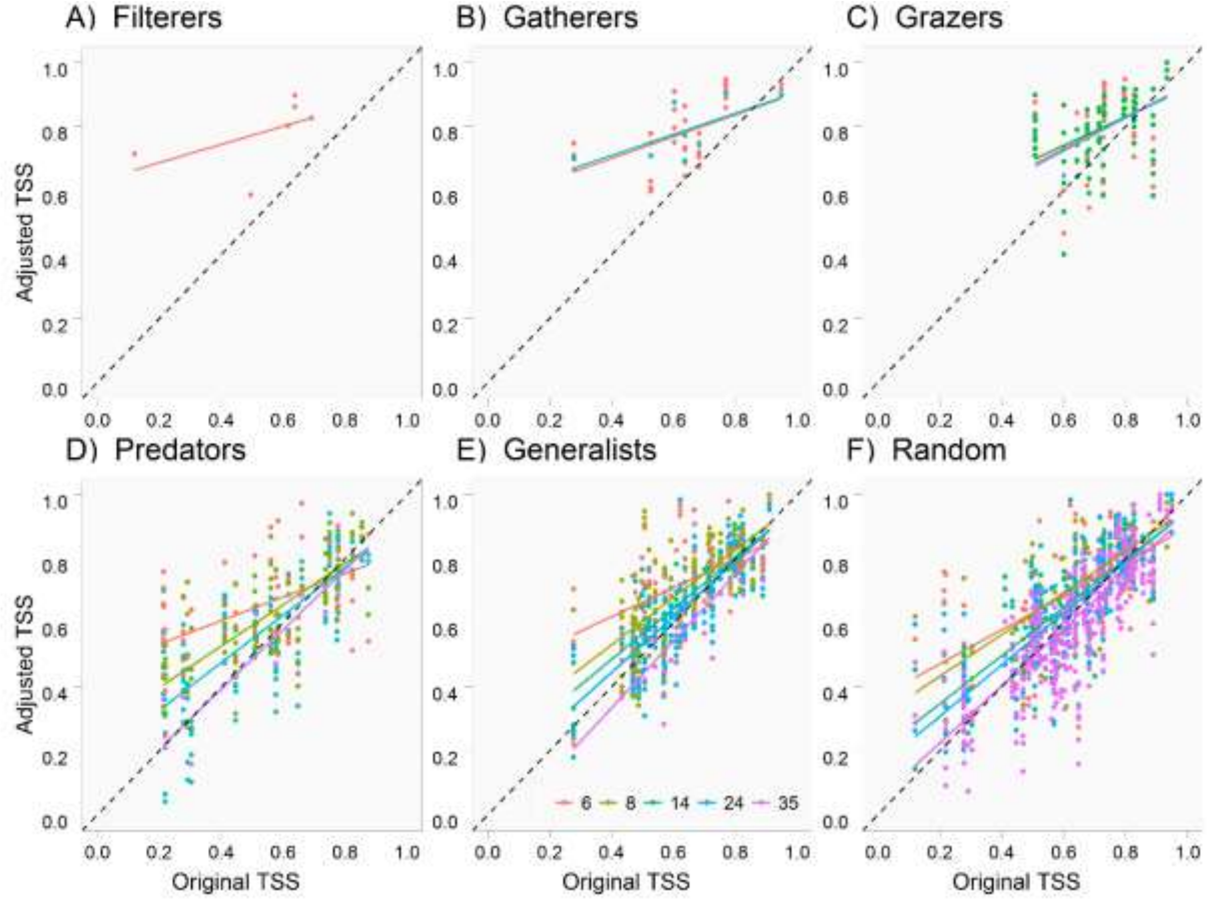

Figure 5: The adjusted vs. original TSS for communities of different guilds (panels) and gamma diversity (colours). Each point in each panel represent a single species and the dashed line is the line of unity. All results are based on the alpha-adjusted SDMS with the observed alpha diversity.