OIKOS

Research

Evaluating ecological uniqueness over broad spatial extents using species distribution modelling

Gabriel Dansereau, Pierre Legendre and Timothée Poisot

G. Dansereau (https://orcid.org/0000-0002-2212-3584) ✉ (gabriel.dansereau@umontreal.ca), P. Legendre (https://orcid.org/0000-0002-3838-3305) and T. Poisot (https://orcid.org/0000-0002-0735-5184), Dépt de Sciences Biologiques, Univ. de Montréal, Montréal, Canada and Québec Centre for Biodiversity Science, Montréal, Canada.

Oikos
2022: e09063
doi: 10.1111/oik.09063
Subject Editor: Thorsten Wiegand
Editor-in-Chief: Dries Bonte
Accepted 6 February 2022

Local contributions to beta diversity (LCBD) can be used to identify sites with high ecological uniqueness and exceptional species composition within a region of interest. Yet, these indices are typically used on local or regional scales with relatively few sites, as they require information on complete community compositions difficult to acquire on larger scales. Here, we investigated how LCBD indices can be predicted over broad spatial extents using species distribution modelling and examined the effect of scale changes on beta diversity quantification. We used Bayesian additive regression trees (BARTs) to predict warbler species distributions in North America based on observations recorded in the eBird database. We then calculated LCBD indices for observed and predicted data and compared the site-wise difference using direct comparison, a spatial association test and generalized linear regression. We also examined the relationship between LCBD values and species richness in different regions and at various spatial extents. Our results showed that species distribution models provided uniqueness estimates highly correlated with observed data. The form and variance of the LCBD–richness relationship varied according to the region and the total extent size. The relationship was also affected by the proportion of rare species in the communities. Therefore, sites identified as unique over broad spatial extents may vary according to regional characteristics. These results show that species distribution modelling can be used to predict ecological uniqueness over broad spatial extents, which could help identify beta diversity hotspots and important targets for conservation purposes in unsampled locations.

Keywords: beta diversity, broad spatial scale, eBird, ecological uniqueness, local contributions to beta diversity, species distribution modelling

Introduction

Beta diversity, defined as the variation in species composition among sites in a geographic region of interest (Legendre et al. 2005), is an essential measure to describe the organization of biodiversity through space. Total beta diversity within a community can be partitioned into local contributions to beta diversity (LCBD) (Legendre

NORDIC SOCIETY OIKOS
www.oikosjournal.org

and De Cáceres 2013), which allow the identification of sites with exceptional species composition, hence unique biodiversity and potential conservation value. Sites with unique community composition often differ from those with high species richness, possibly as they harbour rare species or help maintain beta diversity (Heino et al. 2017, da Silva et al. 2018, Landeiro et al. 2018). Hence, focusing on uniqueness can prove helpful as a complementary approach to species richness (Heino and Grönroos 2017, da Silva et al. 2018, Dubois et al. 2020, Yao et al. 2021). However, the use of LCBD indices is currently limited in two ways. First, LBCD indices are typically used on data collected over local or regional scales with relatively few sites, for example, on fish communities at intervals along a river or stream (Legendre and De Cáceres 2013). Second, LCBD calculation methods require complete information on community composition; thus, they are inappropriate for partially sampled sites (e.g. where data for some species are missing or uncertain) and cannot directly provide assessments for unsampled ones. Accordingly, this method is of limited use to identify areas with exceptional biodiversity in regions with sparse sampling. However, predictive approaches offer an opportunity to overcome such limitations, as computational methods often uncover novel ecological insights from existing data (Poisot et al. 2019), including in lesser-known locations and on larger spatial scales.

Species distribution models (SDMs) (Guisan and Thuiller 2005) can bring a new perspective to LCBD studies by filling in gaps in community composition data to perform analyses on broader scales. Single-species SDMs aim at predicting the distribution of a species in unsampled locations based on information (such as environmental data) from sampled locations with reported occurrences. Many approaches allow going from single-species SDMs to a whole community on which to evaluate community-level metrics, yet their relevance has not been explicitly evaluated for ecological uniqueness and LCBD indices. The most straightforward approach is stacked distribution models (S-SDMs) (Ferrier and Guisan 2006, Guisan and Rahbek 2011). Single-species SDMs are first performed separately, then combined to form a community prediction on which community-level analyses can be applied. S-SDMs tend to overestimate species richness (Dubuis et al. 2011, D'Amen et al. 2015, Zurell et al. 2020), which could result from thresholding the probabilities into presence–absence data before stacking the species distributions (Calabrese et al. 2014). Summing the occurrence probabilities without applying a threshold is an alternative (Calabrese et al. 2014), but it may limit some analyses as it does not return species identities for every site (Zurell et al. 2020), as is required with LCBD calculations. In comparison, joint species distribution models (JSDMs) (Pollock et al. 2014) try to improve predictions by incorporating species co-occurrence or shared environmental responses into the models. However, these models do not always improve community-level predictions compared to S-SDMs (Zurell et al. 2020). Spatially explicit species assemblage modelling (SESAM) (Guisan and Rahbek 2011), hierarchical modelling of species communities (HMSC) (Ovaskainen et al. 2017) and Bayesian networks (BN) (Staniczenko et al. 2017) are other alternatives that could yield better community predictions than S-SDMs. On the other hand, they add methodological and computational overload, impeding their use for broad spatial extents. Moreover, their relevance for community prediction is often validated against extensive work on species richness. By comparison, ecological uniqueness and LCBD indices have rarely been used in predictive frameworks. Therefore, S-SDMs may prove an appropriate first step to establish some prediction baselines.

Combining LCBD indices with a predictive approach through SDMs will allow measuring uniqueness over broader spatial extents, across continuous landscapes and on a higher number of sites than what has previously been studied. LCBD scores have typically been used at local or regional scales with relatively few sites (up to 60 sites on extents covering 10–400 km, Legendre and De Cáceres 2013, da Silva and Hernández 2014, Heino and Grönroos 2017, Heino et al. 2017). Some studies did use the measure over broader, near-continental extents (Yang et al. 2015, Poisot et al. 2017, Taranu et al. 2020), but the total number of sites in these studies were relatively small (maximum 51 sites). Recent studies also investigated LCBD and beta diversity on sites distributed in contiguous grids or as pixels, hence uniform sampling intervals and no spatial gaps, but these did not cover large extents and a high number of sites (up to 1250 sites and 6 km^2, Tan et al. 2017, 2019, Legendre and Condit 2019, D'Antraccoli et al. 2020). Two recent studies have, however, adopted promising predictive approaches on regional extents. First, Niskanen et al. (2017) predicted LCBD values of plant communities (and three other diversity measures) on a continuous scale and a high number of sites (> 25 000) using boosted regression trees (BRTs). However, they modelled the diversity measures directly after calculating them on a smaller number of sampled sites. Second, Vasconcelos et al. (2018) used ecological niche models (ENMs) to predict anurans ecological niches according to actual and forecasted environmental conditions, then calculated the LCBD values on the predictions to identify biodiversity hotspots. Using this approach, predicted LCBD values are calculated in a way closer to the original formulation. This development of predictive techniques is exciting, especially as it could be pushed a step further to continental extents, a higher number of sites and more species occurrences using SDMs and massive data sources. Still, it should be accompanied by an investigation of the determinant of ecological uniqueness in such conditions.

Measuring ecological uniqueness from LCBD indices over broad spatial extents and spatially continuous data also raises the question of which sites will be identified as exceptional and for what reason. The method intends that sites stand out and receive a high LCBD score whenever they display an exceptional community composition, be it a unique assemblage of species with high conservation value or a community richer or poorer than others in the region (Legendre and De Cáceres 2013). Both the original study and many of the later empirical ones have shown a negative relationship

between LCBD scores and species richness (Legendre and De Cáceres 2013, da Silva and Hernández 2014, Heino and Grönroos 2017, Heino et al. 2017), although other studies observed both negative and positive relationships at different sites (Kong et al. 2017) or quadrats (Yao et al. 2021). Some studies showed that the direction of the relationship is related to the percentage of rare species in the community (da Silva et al. 2018, Yao et al. 2021). However, beta diversity and species rarity are both concepts that depend on scale. For instance, total beta diversity increases with spatial extent (Barton et al. 2013) and varies because of higher environmental heterogeneity and sampling of different local species pools (Heino et al. 2015). Therefore, the LCBD-richness relationship and the effect of rare species on LCBD values should be investigated over broad spatial extents, as they might not be constant across scales.

Here, we examined whether species distribution models (SDMs) can be combined with local contributions to beta diversity (LCBD) to assess ecological uniqueness over broader spatial extents. We also investigated the effect of scale changes on beta diversity quantification. We first predicted species distributions on continental scales using extended occurrence data from eBird and Bayesian additive regression trees (BARTs). We then quantified uniqueness with the LCBD measure for both predicted and observed data. Next, we examined the site-wise difference using direct comparison, a spatial autocorrelation test and generalized linear regression. We then investigated the relationship between uniqueness and species richness for different regions and scales and according to the proportion of rare species.

Methods

Occurrence data

We used occurrence data from eBird (Sullivan et al. 2009) downloaded through the eBird Basic Data set from June 2019 (eBird Basic Dataset 2019). We restricted our analyses to the New World warbler family (Parulidae) in North America (Canada, United States, Mexico). eBird is a semi-structured citizen science data set, meaning that observations are reported as checklists of species detected in an observation run (Johnston et al. 2021). Observers can explicitly specify that their checklist contains all species they could detect and identify during a sampling event, in which case it is labelled as a 'complete checklist.' Using complete checklists instead of regular ones allows researchers to infer non-detections in locations where detection efforts occurred, which offers performance gains in species distribution models (Johnston et al. 2021). Therefore, we selected the data from the complete checklists only. Our final data set comprised 62 warbler species and 22 974 330 observations from 9 103 750 checklists. Warblers are a diverse group with many species, are popular among birders given their charismatic aspect, and are widely distributed in various habitats across North America.

Environmental data

Our environmental data consisted of climatic data from WorldClim ver. 2.1 (Fick and Hijmans 2017) and land cover data from the Copernicus Global Land Service (Buchhorn et al. 2019). We restricted these data to a spatial extent comprised between latitudes 20°N to 75°N and between longitudes 145°W to 50°W. First, the WorldClim data consist of spatially interpolated monthly climate data for global land areas. We used the standard BIOCLIM variables (Booth et al. 2014) from WorldClim 2.1, which represent annual trends, ranges and extremes of temperature and precipitation, but selected only 8 out of the 19 ones to avoid redundancy (bio1, bio2, bio5, bio6, bio12, bio13, bio14, bio15). We downloaded the data at a resolution of 10 arcminutes (around 18 km at the equator), the coarsest resolution available, which should mitigate potential imprecision in the eBird data regarding the extent of the sampled areas in each observation checklist. Moreover, some studies have argued that coarser resolutions lead to less overestimation of species richness and better identification of bird biodiversity hotspots given the patchiness of observation data (Hurlbert and Jetz 2007). We acknowledge that using an arcminutes-based resolution means that the surface area of our pixels will not be equal depending on the latitude.

Second, the Copernicus data are a set of variables representing ten land cover classes (e.g. crops, trees, urban areas) and measured as a percentage of land cover. The data are only available at a finer resolution of 100 m. We coarsened them to the same ten arcminute resolution as the WorldClim data by averaging the pixels' cover fraction values. We removed two variables (moss and snow) from our predictive models as their cover fraction was 0% on all sites with warbler observations.

Species distribution models

We converted the occurrence data to a presence–absence format compatible with community analyses. We considered every pixel from our ten arcminutes environmental layers as a site and then verified, for each species, if there was a single observation in every site. Finally, we recorded the outcome as a binary value: present (1) if a species was ever recorded in a site and absent (0) if it was not. Complete checklists help ensure that these zeros represent non-detections, rather than the species not being reported; hence we considered them as absence data, similar to Johnston et al. (2021), although we recognize that there exists a doubt on whether these truly represent non-detections.

We predicted species distribution data on continuous scales from our presence–absence data using Bayesian additive regression trees (BARTs) (Chipman et al. 2010), a classification and regression trees method recently suggested for species distribution modelling (Carlson 2020). BARTs are based on an ensemble of trees, similarly to boosted regression trees and random forest, but follow a sum-of-trees model and a Bayesian framework. Trees are first constrained as weak learners by priors regarding structure and nodes, then

updated through an iterative Bayesian backfitting Markov Chain Monte Carlo (MCMC) algorithm which ultimately generates a posterior distribution of predicted classification probabilities (Chipman et al. 2010, Carlson 2020). In the context of species distribution modelling, BARTs showed high performance when compared to other predictive algorithms (Konowalik and Nosol 2021, Tytar and Baidashnikov 2021). We first performed BARTs separately for all species and estimated the probability of occurrence in all the sites of our region of interest using the posterior median. We then converted the results to a binary outcome according to the threshold that maximized the true skill statistic (TSS) for each species, as suggested by Carlson (2020).

Quantification of ecological uniqueness

We used the method of Legendre and De Cáceres (2013) to quantify compositional uniqueness from overall beta diversity for both the observed and predicted data. First, we assembled the presence–absence data by site to form two site-by-species community matrices, one from observed data, called Y (39 024 sites by 62 species) and one from predicted data, called $\hat{Y}$ (99 382 sites by 62 species). Next, we measured species richness per site as the number of species present. Finally, we removed the sites without any species from the predicted matrix $\hat{Y}$, for a new total of 85 526 sites (this was unnecessary for the observed matrix Y). We then applied the Hellinger transformation to both matrices in order to compute beta diversity from the community composition data (Legendre and De Cáceres 2013). We measured total beta diversity as the variance of each community matrix and calculated the local contributions to beta diversity (LCBD), which quantify how much a specific site (a row in each matrix) contributes to the overall variance in the community (Legendre and De Cáceres 2013). High LCBD values indicate a unique community composition, while low values indicate a more common species set. We note that our LCBD values, which add up to 1 because the values are divided by the total sum-of-squares of the data matrix, were very low given the high number of sites in both Y and $\hat{Y}$. However, the relative difference between the scores in one set matters more than the absolute value to differentiate their uniqueness.

Comparison of observed and predicted values

We performed three verification to compare the richness and uniqueness estimates obtained from our predicted distributions to those obtained with the eBird occurrence data. First, we performed a direct comparison by subtracting the richness and LCBD estimates obtained from Y (the observed data) from the estimates obtained from $\hat{Y}$ (the predicted data). To do so, we used the richness estimates as-is but modified the LCBD values to achieve a non-biased comparison, given that the values were initially calculated on sets of different lengths. Therefore, we recomputed the LCBD scores only for the sites for which we had occurrences in both Y and $\hat{Y}$, which mostly corresponded to the sites in Y, minus a few sites where the SDMs predicted no species occurrence. We then plotted the richness and LCBD differences to examine their spatial distributions. Second, we performed the modified t test from Clifford et al. (1989) to assess the correlation between the observed and predicted estimates and test for spatial association. We performed the test separately for the richness and the LCBD estimates. Third, we performed generalized linear models between the observed and predicted estimates and plotted the deviance residuals to examine their spatial distribution. We used a negative binomial regression with a log link function for the richness estimates and a beta regression with a logit link function for the LCBD values, similar to Heino and Grönroos (2017) and Yao et al. (2021).

Investigation of regional and scaling variation

To investigate possible regional and scaling effects, we recalculated LCBD values on various subregions at different locations and scales. First, we selected two subregions of equivalent size (20.0 longitude degrees by 10.0 latitude degrees) with contrasting richness profiles and corresponding to different ecoregions to verify if the relationship between species richness and LCBD values was similar. The first subregion was in the northeast (latitude 40°N to 50°N, longitude 80°W to 60°W), was mostly species-rich (for both the observed and predicted data), and corresponded to the Eastern Temperate Forests level I ecoregion (Commission for Environmental Cooperation 1997). The second subregion was in the southwest (latitude 30°N to 40°N, longitude 120°W to 100°W), was mostly species-poor, and covered Mediterranean California, North American Deserts, Temperate Sierras and Southern Semi-Arid Highlands ecoregions (Commission for Environmental Cooperation 1997). Second, we recalculated the LCBD indices at three different extents, starting with a focus on the northeast subregion and progressively extending the extent to encompass the southwest subregion. We did these two verifications with both the observed and predicted data but only illustrate the results with the predicted data as both were qualitatively similar.

Proportion of rare species

We investigated the effect of the proportion of rare species in the community on the direction of the relationship between species richness and LCBD values in our northeast and southwest subregions. Following De Cáceres et al. (2012) and Yao et al. (2021), we classified species as rare when they occurred in less than 40% of the sites in each subregion. We calculated the proportion of rare species for every site. We then grouped the sites for both subregions depending on whether they were part of an ascending or a descending portion in the LCBD–richness relationship. Given that the relationship sometimes displays a curvilinear form with a positive quadratic term (Heino and Grönroos 2017, Tan et al. 2019), we separated the ascending and descending portions based on the species richness at the site with the lowest LCBD value (using the median richness if there were multiple sites). This

value corresponds to the inflection point of the relationships. For example, the lowest LCBD value was 7.045×10^{-05} in the northeast subregion and the corresponding richness was 23. All the sites with more than 23 species were assigned to the ascending portion, and all the sites with 23 species or fewer were assigned to the descending portion. In the southwest subregion, the lowest LCBD value and its corresponding richness were 5.438×10^{-05} and 12, respectively. We then mapped the ascending and descending groups to view their spatial distribution. We also examined the distribution of the rare species proportions in both groups using a kernel density estimation plot. Similar to our previous verification, we performed this analysis with both observed and predicted data but once again only illustrate the results with the predicted data as both were qualitatively similar.

Software

We used Julia ver. 1.6.1 (Bezanson et al. 2017) for most of the project and R ver. 4.1.0 (<www.r-project.org>) for some specific steps. We used the Julia package SimpleSDMLayers.jl (Dansereau and Poisot 2021) as the basic framework for our analyses, to download the WorldClim 2.1 data, and to map our results through the package's integration of Plots.jl. We also used StatsPlots.jl to produce the kernel density estimation plots in our rare species analysis. We computed the LCBD indices with our own function implemented in Julia, whose results were verified by comparison to the *beta.div* function from the package adespatial (Dray et al. 2021) in R. We used the R packages auk (Strimas-Mackey et al. 2018) to extract and manipulate eBird data, embarcadero (Carlson 2020) to perform the BART models, vegan (Oksanen et al. 2019) to apply the Hellinger transformations, and SpatialPack (Vallejos et al. 2020) to perform the modified t test (with the function *modified.ttest*) from Clifford et al. (1989). We used MASS (Venables and Ripley 2002) and betareg (Cribari-Neto and Zeileis 2010) to perform the negative binomial and beta regressions, respectively. We also used GDAL (GDAL/OGR Contributors 2021) to coarsen the Copernicus land cover data. All the scripts required to reproduce the analyses are archived on Zenodo (<https://doi.org/10.5281/zenodo.6024392>).

Results

Species distribution models generate relevant community predictions

Species richness from observation data (Fig. 1a) was higher on the east coast and lower on the west coast, with many unsampled patches in the north, south and central west. Richness results from SDM data (Fig. 1b) displayed higher richness on the east coast and sites with few or no species up north and in the central west. There was no clear latitudinal gradient in richness but rather an east–west one. Landmarks such as the Rocky Mountains and croplands in the central west (which should be species-poor habitats) were notably visible on the maps, separating the east and west. LCBD scores from observation data (Fig. 1c) were low on the east coast and higher on the border of sampled sites in the central west. They were also higher in the north and in the south where observations were sparser. Results from SDM predictions were qualitatively similar (Fig. 1d), with lower LCBD values in the east and more unique sites in the central west, Central Mexico and some northern regions. There was no clear latitudinal gradient, and the east–west contrast, while present, was less clear than on the richness maps.

The modified t test of Clifford et al. (1989) showed a high correlation between the observed and predicted estimates of richness and uniqueness, as well as a statistically significant spatial association between the values. For species richness, the correlation coefficient was 0.777, the F-statistic was 20.007 and the p-value was 6.093×10^{-04}. For LCBD scores, the correlation coefficient was 0.518, the F-statistic was 40.083 and the p-value was 5.528×10^{-09}.

The difference between the observed and predicted estimates (predicted richness – observed richness and predicted LCBD – observed LCBD) showed opposite geographic distributions for species richness and ecological uniqueness (Fig. 2). Predicted richness estimates were higher than observed estimates on the east coast, on the west coast and in Mexico but were lower than observed estimates in the central west (Fig. 2a). Predicted LCBD estimates, on the other hand, were lower than observed estimates on the east coast and higher in the central west (Fig. 2b). Regression residuals showed similar geographic distributions to their corresponding difference values (Fig. 3).

Uniqueness displays regional variation as two distinct profiles

The relationship between LCBD values and species richness displayed contrasting profiles in species-rich and species-poor regions (Fig. 4). In the species-rich northeast region, LCBD scores displayed a mostly decreasing relationship with species richness, with a slightly curvilinear form and increase of values for very rich sites. The sites with the highest LCBD values were the species-poor sites while the species-rich sites displayed scores. The southwest subarea showed a different relationship with a sharper initial decline and a larger increase as richness reached 20 species. The sites with the highest LCBD values were the poorest in terms of species richness, as in the northeast region, but the species-rich sites were proportionally more unique in the southwest region. Total beta diversity was higher in the southwest subregion (0.417) than in the northeast (0.179), indicating higher compositional differences between the sites.

Uniqueness depends on the scale on which it is measured

The LCBD-richness relationship showed important variation when scaling up and changing the region's extent (Fig. 5). For

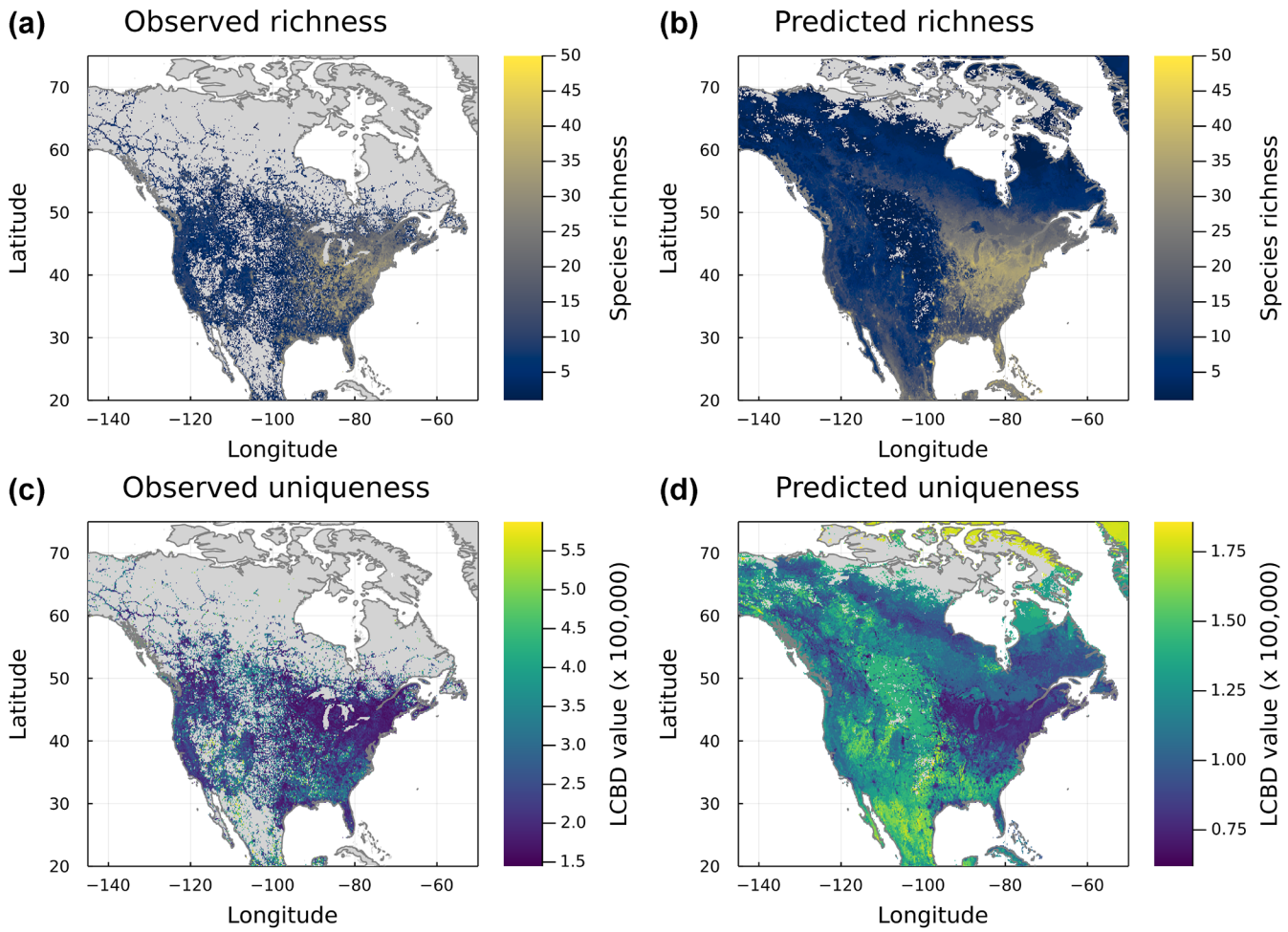

Figure 1. Comparison of species richness and LCBD scores from observed and predicted warbler occurrences in North America. Values were calculated for sites representing ten arcminute pixels. We measured species richness after converting the occurrence data from eBird (a) and the SDM predictions from our single-species BART models (b) to a presence–absence format per species. We applied the Hellinger transformation to the presence–absence data, then calculated the LCBD values from the variance of the community matrices separately for the occurrence data (c) and the SDM predictions (d). Areas in light grey (not on the colour scale) represent mainland sites with environmental data but without any warbler species present.

smaller extents, starting with a species-rich region, the relationship is well defined, mostly decreasing but notably curvilinear (with a lesser increase for richness values higher than the median). However, as the extent increases and progressively reaches species-poor regions, the relationship broadens, displays more variance and loses its curvilinear aspect while keeping a decreasing form. Total beta diversity was higher when increasing the spatial extent, going from 0.121 to 0.284 and 0.687.

Uniqueness depends on the proportion of rare species

The proportion of rare species per site differed depending on the classification in the ascending or descending portions of the LCBD–richness relationship (Fig. 6). The proportion of rare species was higher in the sites corresponding to the ascending portions of the relationships (shown in Fig. 4) than in the sites corresponding to the descending portions for both subregions. The classification of the sites in the two portions showed a clear latitudinal gradient in the northeast subregion, while it was distributed in patches in the southwest subregion (Fig. 6).

Discussion

Our results showed a decreasing relationship between species richness and LCBD values on broad spatial extents (Fig. 5c) but also highlighted that the exact form of this relationship varies depending on the region and the spatial extent on which it is measured. Our species-rich northeast subregion (Fig. 4a) showed a decreasing relationship, very similar to previous studies and slightly curvilinear, as described by Heino and Grönroos (2017) and Tan et al. (2019). This result for warbler species is in line with the original study on fish communities

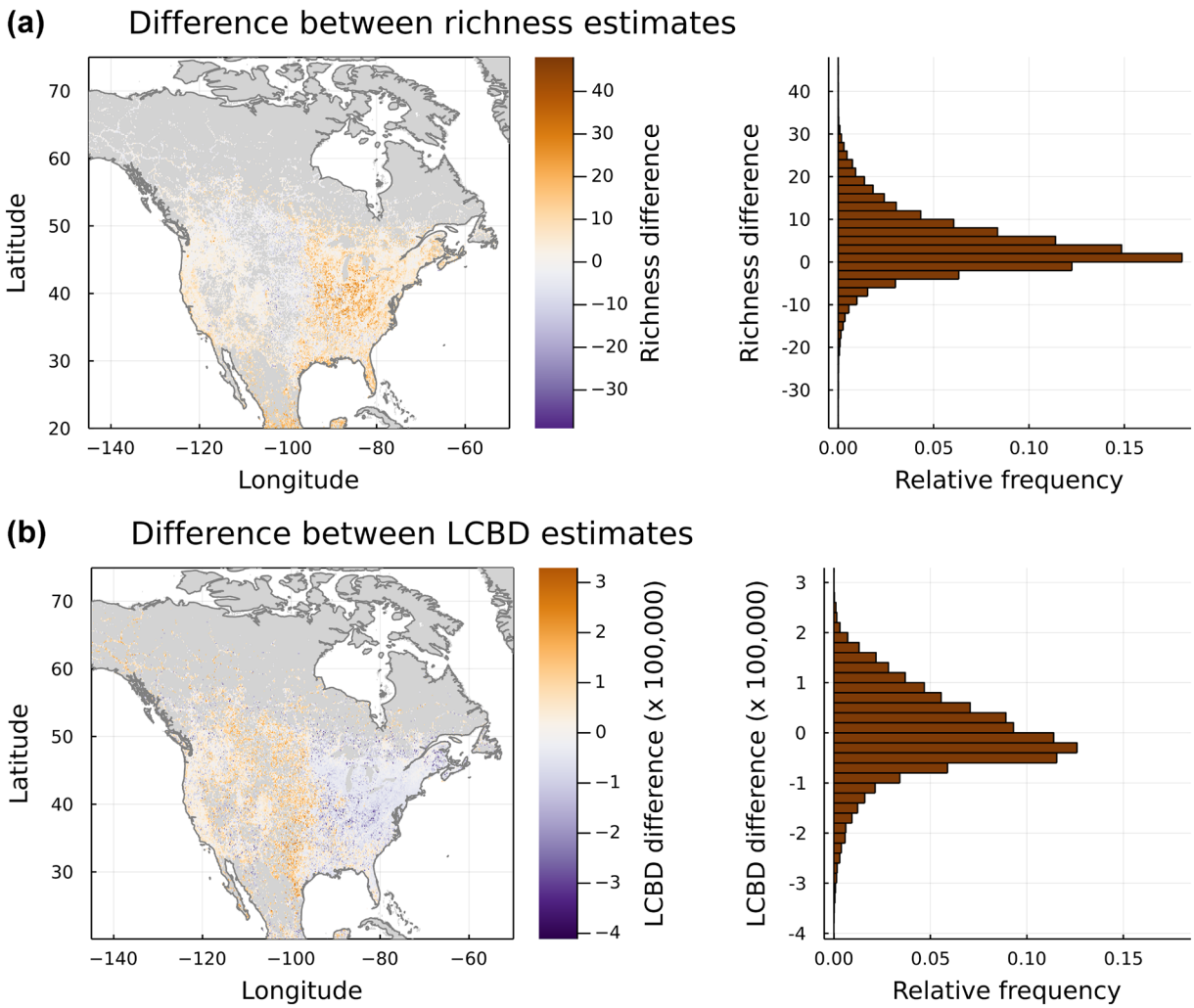

Figure 2. Comparison between observed and predicted estimates of species richness (a) and ecological uniqueness (b). The difference values represent the estimate from the predicted data set minus the estimate from the observed data set. LCBD values were recalculated for the same set of sites with observations in both data sets.

(Legendre and De Cáceres 2013) and with following ones on insect metacommunities (da Silva and Hernández 2014, Heino and Grönroos 2017, Heino et al. 2017), dung beetles (da Silva et al. 2018, 2020), aquatic beetles (Heino and Alahuhta 2019), stream macroinvertebrates (Sor et al. 2018), stream diatoms (Vilmi et al. 2017), multi-trophic pelagic food webs (phytoplankton, zooplankton, fish) (Taranu et al. 2020), temperate forest trees (Tan et al. 2019), mammals (da Silva et al. 2020), wetland birds (de Deus et al. 2020) and various phylogenetic groups (plants, lizards, mites, anurans, mesoinvertebrates) (Landeiro et al. 2018). However, it was originally argued that the negative relationship was not general or obligatory (Legendre and De Cáceres 2013). Different LCBD–richness relationships have also been observed, with both positive and negative relationships for different sites or taxonomic groups in some studies (Kong et al. 2017, Teittinen et al. 2017), as well as a negative relationship with the number of common species but a positive relationship with the number of rare species (Qiao et al. 2015).

Our results further show that the relationship may depend on the region's richness profile, as the relationship was different in our species-poor southwest subregion, with a sharper initial decrease (Fig. 4b). Therefore, the curvilinear form may depend on how pronounced the contrast is between the region's median richness and its richest ecologically feasible sites. The increasing part of the curvilinear form for higher richness values was also more pronounced in our results (Fig. 4a–b, 5c) than in previous studies (Tan et al. 2019), which reinforces the idea that the relationship and its curvilinear form may vary depending on the region.

The variation in the LCBD–richness relationship when extending the study extent showed that the uniqueness patterns highlighted are not necessarily the same depending on the scale on which it is used (Fig. 5). The relationship progressively lost its clear definition and curvilinear form as the east and west profiles merged, creating a new distinct profile with more variation and no curvilinear form. Therefore, aggregating too many different sites might possibly mask

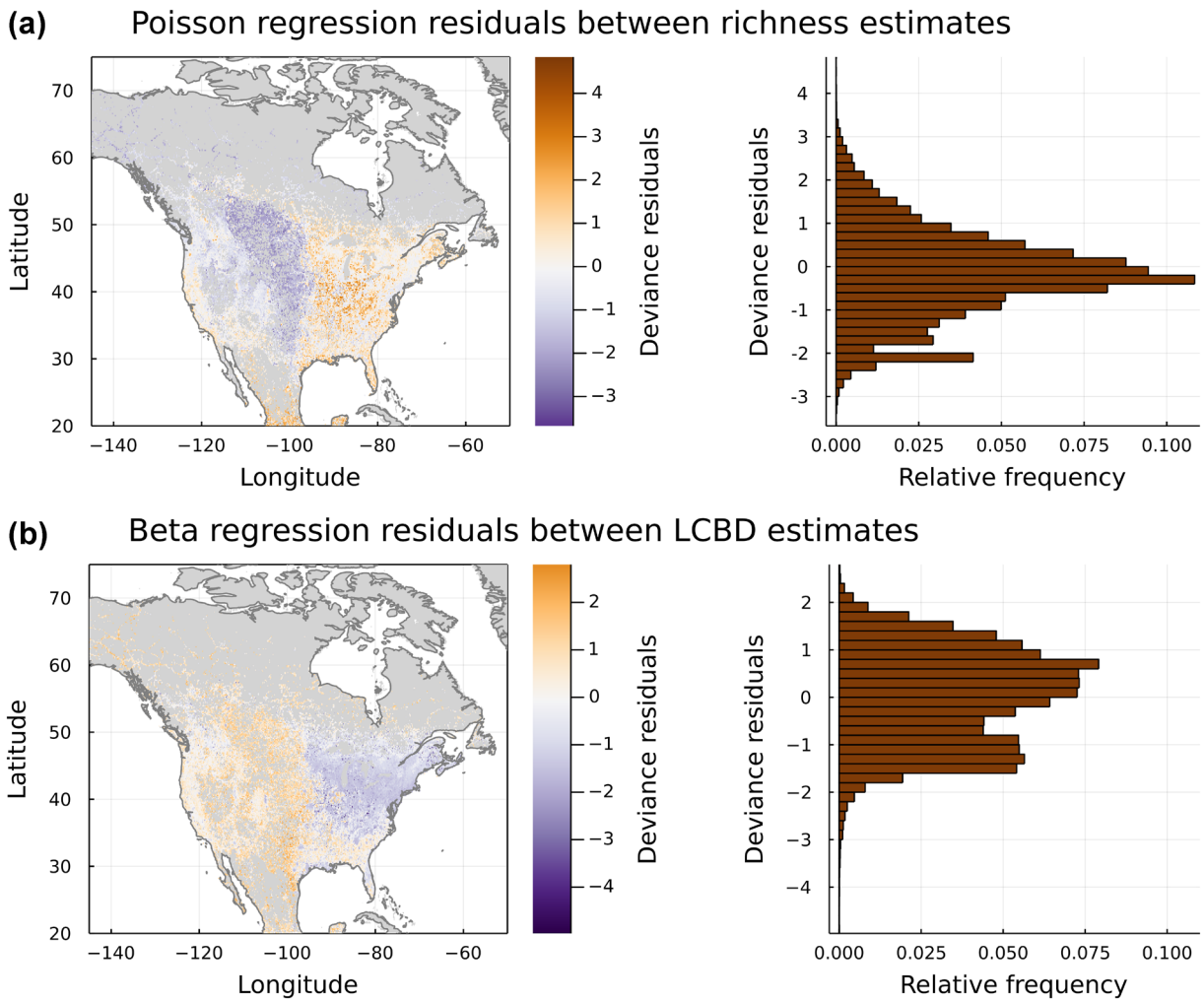

Figure 3. Comparison of the regression residuals between the observed and predicted estimates of species richness (a) and ecological uniqueness (b). The estimate from the predicted data set was used as the dependent variable and the estimate from the observed data set as the independent variable. A negative binomial regression with a log link function was used for species richness, and a beta regression with a logit link function was used for uniqueness. LCBD values were recalculated for the same set of sites with observations in both data sets.

some patterns of uniqueness in species-rich sites. Total beta diversity, on the other hand, showed the variation expected from previous studies, increasing with spatial extent (Fig. 5) (Barton et al. 2013, Heino et al. 2015). Its value was high at the continental scale (0.687) but lower than what has been observed in some studies (e.g. 0.80 on macroinvertebrate communities in the Lower Mekong Basin by Sor et al. 2018).

Our results confirm that the proportion of rare species in the community may affect the direction of the relationship between species richness and ecological uniqueness (Fig. 6). da Silva et al. (2018) suggested that the proportion of rare and common species in the communities determines whether the relationship will be negative, non-significant or positive. Yao et al. (2021) showed an association between the direction of the relationship and the proportion of rare species, with sites with a lower proportion (between 60% and 75% in their case) displaying a negative relationship and sites with a higher proportion (around 85%) showing a positive one. Our results further show that sites associated with a positive relationship within a curvilinear one tended to have a higher rare species proportion (Fig. 6). This also implies that the proportion of rare species was higher in species-rich sites than in species-poor ones in both our northeast and southwest subregions. Further work should attempt to disentangle the effects of the rare species proportion and the region's richness profile.

Our results showed that SDM models provide richness and uniqueness predictions highly correlated to the occurrence data while filling gaps in poorly sampled regions (Fig. 1). The results showed a statistically significant spatial association between predicted and observed estimates despite correcting for autocorrelation using the modified t-test from Clifford et al. (1989). A positive autocorrelation on large distances indicates aggregates or structures repeating through space (Legendre and Fortin 1989). This is consistent with our results, as the distribution of richness and uniqueness values was visibly spatially structured in both our observed and predicted data (Fig. 1). Nonetheless, it is possible that the autocorrelation in the predicted values could represent an artifact

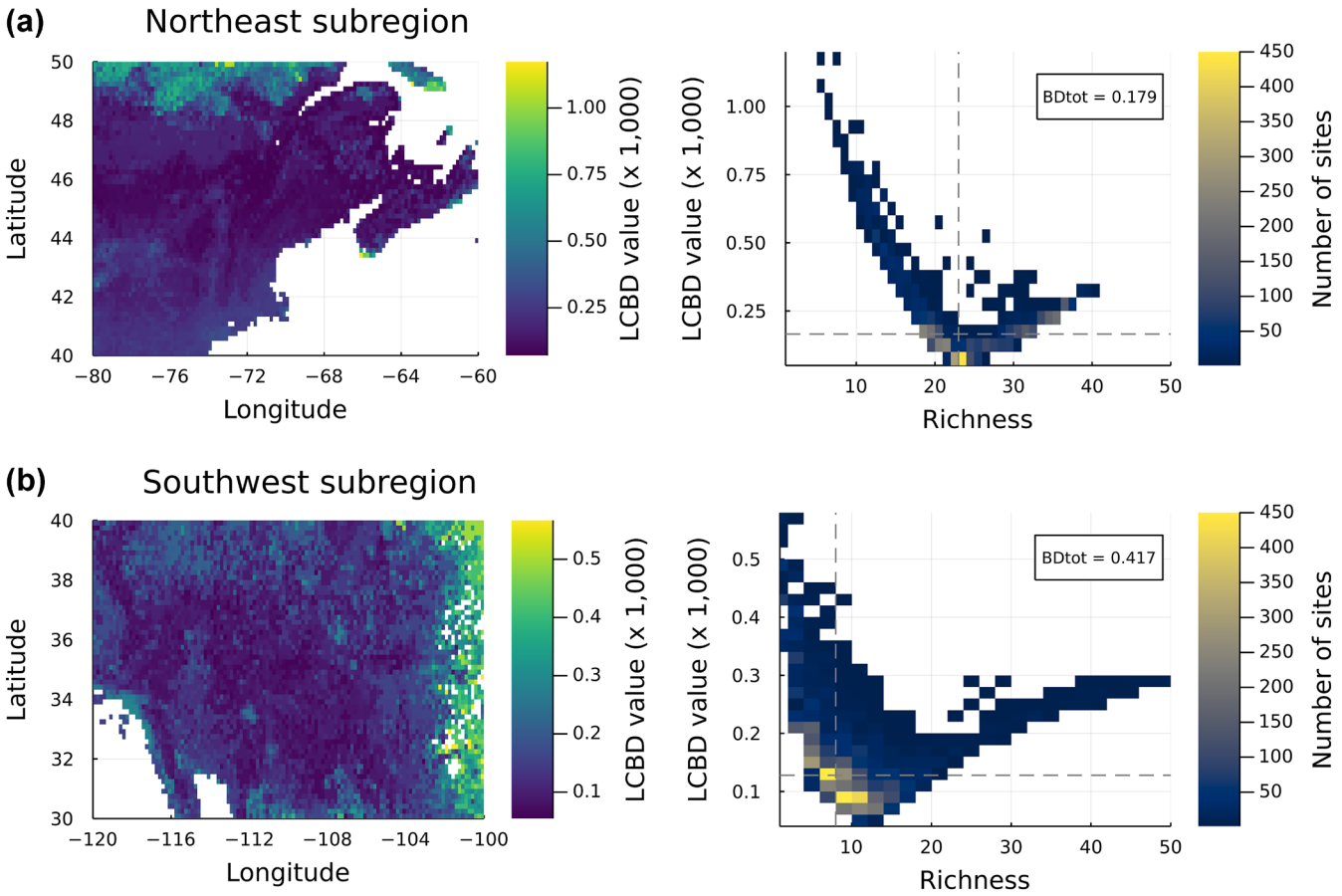

Figure 4. Comparison between a species-rich region (northeast, a) and a species-poor one (southwest, b) based on the SDM predictions for warbler species in North America. The left-side figures represent the LCBD scores for the assembled presence–absence predictions, calculated separately in each region. The colour scales are set to the respective range of LCBD scores to highlight the relative change within each region rather than compare the scores between both regions. The right-side two-dimensional histograms represent the decreasing and slightly curvilinear relationship between LCBD values and species richness. The vertical and horizontal dashed lines respectively represent the median richness and LCBD value in each region, while BDtot represents the total beta diversity.

of the predictive models (capturing the spatial structure from the environmental variables, for example), and might not represent the true autocorrelation expected for the uniqueness estimates. Further work could verify this by quantitatively comparing the autocorrelation and spatial structures in the observed and predicted uniqueness estimates.

Predicted values also tended to underestimate uniqueness in species-rich regions and overestimate it in species-poor ones, with the opposite trend for species richness (Fig. 2, 3). Overprediction of richness using S-SDMs was reported previously (Dubuis et al. 2011, D'Amen et al. 2015, Zurell et al. 2020). No comparable baseline exists for predictions of LCBD values, as our study is the first to compare LCBD estimates from observed and predicted data in such a way. However, some studies showed that LCBD distributions were spatially structured across sampling sites (da Silva et al. 2018). On the other hand, the spatial structure in our results did not exactly concord with the one reported by Heino and Alahuhta (2019), who showed a negative relationship between LCBD values and latitude for diving beetles communities in northern Europe. In comparison, our LCBD scores increased both in the north and south (Fig. 1), hence did not strictly increase with latitude, and also showed a clear east–west gradient. Overall, our distribution results (Fig. 1, 2, 3) also have implications for conservation, as they confirm that species richness and ecological uniqueness measured from LCBD values may conflict and highlight different potential hotspots (Dubois et al. 2020, Yao et al. 2021), thus reinstating the need to protect both with complementary strategies.

Our predictions for regions with sparse sampling are of interest as they allow a quantitative evaluation, however imperfect, for sites where we would otherwise have no information. Our SDMs also offered relevant LCBD predictions using eBird, arguably one of the largest presence–absence data sets available (when using its complete checklist system), showing the measure's potential on such massive data. Together, the potential to generate uniqueness predictions in new locations and through massive data opens new opportunities for LCBD analyses on extended spatial scales and on a broader diversity of taxons. An interesting way forward would be to test these results using more advanced community assembling techniques than S-SDMs. The use of SESAM

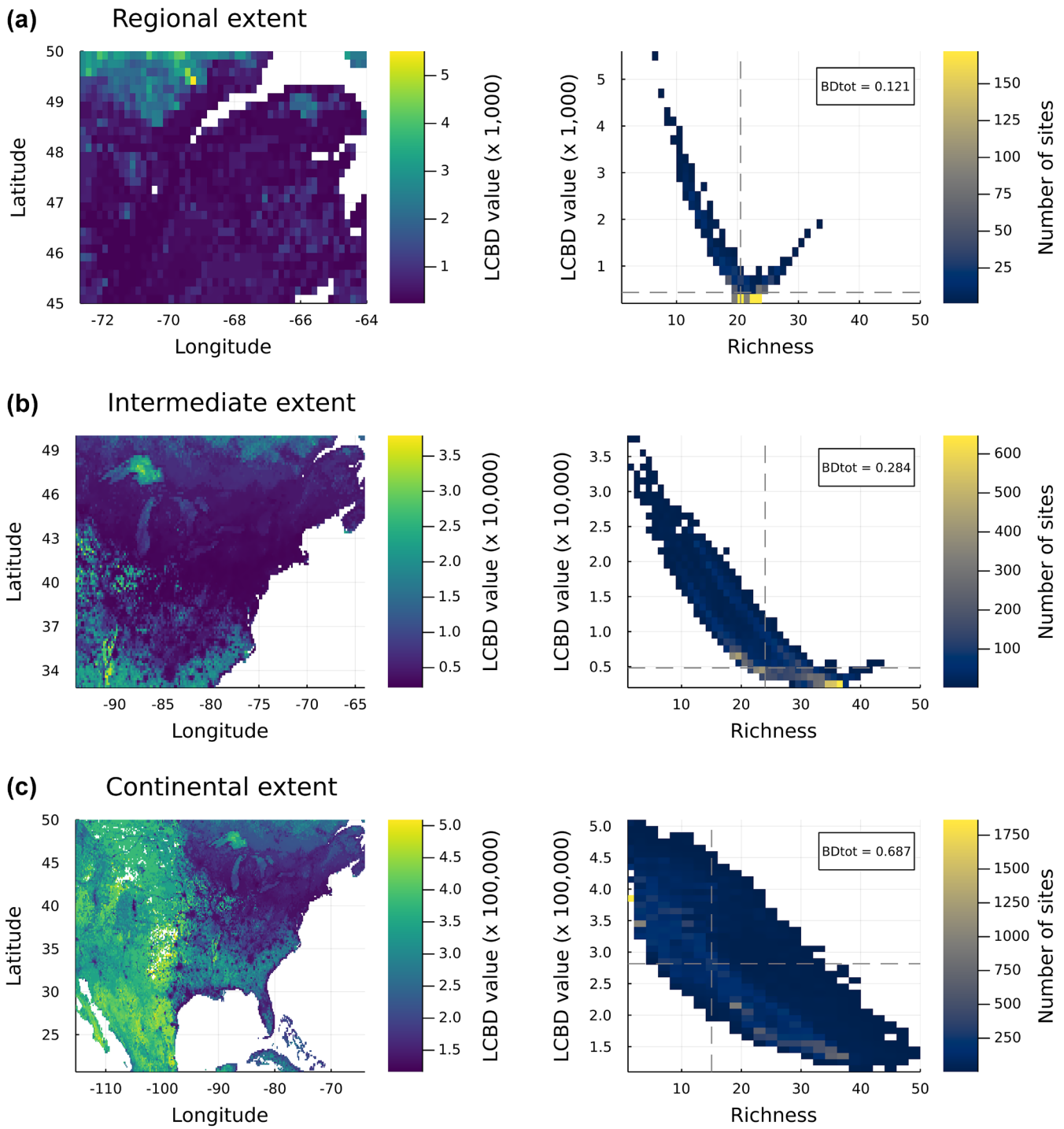

Figure 5. Effect of extent size on the relationship between site richness and LCBD values based on the SDM predictions for warbler species in North America. The relationship progressively broadens and displays more variance when scaling up while total beta diversity increases. The LCBD values were recalculated at each scale based on the sites in this region. The vertical and horizontal dashed lines respectively represent the median richness and LCBD value in each region, while BDtot represents the total beta diversity.

(Guisan and Rahbek 2011) with probabilistic SDMs, probability ranking and species richness predictions as macroecological constraints returns high site-level prediction accuracy (Zurell et al. 2020) and would be compatible with presence–absence LCBD calculations. The use of probabilistic stacks rather than binary ones (Calabrese et al. 2014) could also constitute a novel way to calculate LCBD indices. Both these procedures should reduce the richness deviation we observed, and it would be interesting to verify if this can also be the case with LCBD values. An ensemble of SDM algorithms

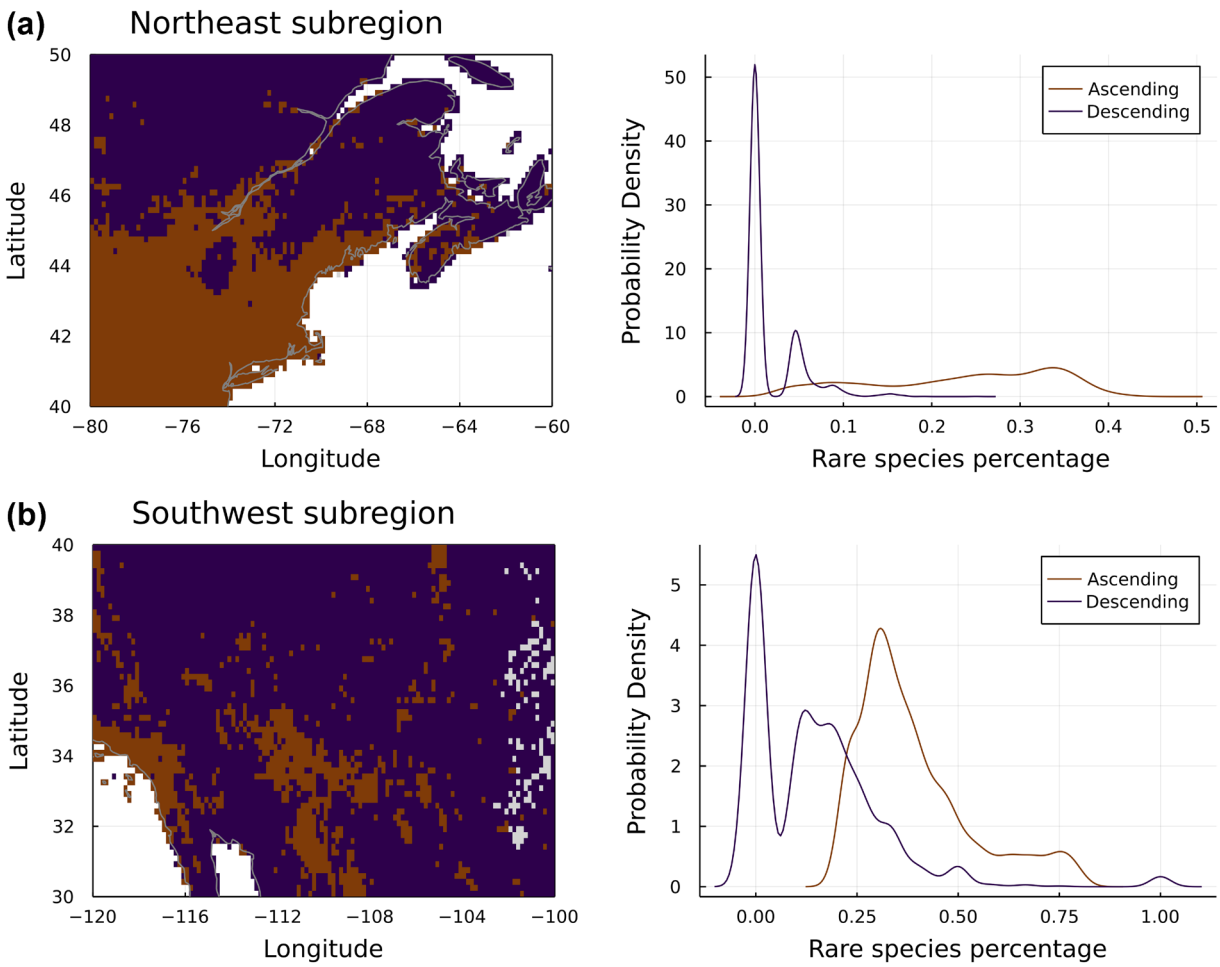

Figure 6. Proportion of rare species in the ascending and descending portions of the LCBD–richness relationship for the northeast (a) and southwest (b) subregions. The left side figures show the geographic distribution of the sites from each group. Sites were assigned to the ascending portion if their species richness was higher than the richness of the site with the lowest LCBD value, which corresponds to the inflection point of the right side figures of Fig. 4, and in the descending portion otherwise. The right side figures represent the kernel density estimation of the proportion of rare species in each group. Values on the y-axis are probability densities scaled so that the area under the curve equals one. Similarly, the area under the curve for a given range of values on the x-axis (proportions of rare species) represents the probability of observing a value in that range. Species were classified as rare when they occurred in fewer than 40% of the sites in the subregion. The proportion of rare species was then calculated for every site.

could also be used to capture a broader range of possible outcomes for the LCBD predictions. However, given the high performance of BARTs in model comparisons (Konowalik and Nosol 2021, Tytar and Baidashnikov 2021) and the large extent we covered, we do not believe the changes to the LCBD gradients would be strong enough to affect our results in a meaningful way.

This study showed how ecological uniqueness can be measured over broad spatial extents, including for regions with sparse sampling, and how scale changes may affect beta diversity quantification. It is the first study to assess the relevance of local contributions to beta diversity calculated on the output of species distribution models. It is also the first to compare the relationship between LCBD values and species richness for different regions and spatial extents. First, our results showed that the negative relationship often observed between species richness and LCBD scores can take different forms depending on the richness profile of the regions on which it is measured. Therefore, species-rich and species-poor regions may display different ways to be unique. Second, the negative relationship was not constant when varying the spatial study extent and may be less clearly defined at broad scales when contrasting regional relationships are present. The broad-scale uniqueness profile might then be completely distinct from the regional profiles constituting it. Finally, species distribution models offer a promising way to generate uniqueness predictions on broad spatial extents and could prove useful to identify beta diversity hotspots in unsampled locations on large spatial scales, which could be important targets for conservation purposes.

Acknowledgements – We acknowledge that this study was conducted on land within the traditional unceded territory of the Saint Lawrence Iroquoian, Anishinabewaki, Mohawk, Huron-Wendat

and Omàmiwininiwak nations. We thank Élise Filotas and Anne-Lise Routier for their helpful comments on this manuscript.
Funding – We received financial support from the Fonds de recherche du Québec – Nature et Technologie (FRQNT, grant no. 275686) and the Computational Biodiversity Science and Services (BIOS2) NSERC CREATE training program.

Author contributions

Gabriel Dansereau: Conceptualization (equal); Formal analysis (lead); Funding acquisition (lead); Methodology (equal); Project administration (supporting); Software (lead); Visualization (lead); Writing – original draft (lead); Writing – review and editing (lead). **Pierre Legendre**: Conceptualization (equal); Methodology (equal); Project administration (supporting); Supervision (supporting); Writing – review and editing (equal). **Timothée Poisot**: Conceptualization (lead); Funding acquisition (supporting); Methodology (equal); Project administration (lead); Resources (lead); Software (equal); Supervision (lead); Writing – original draft (supporting); Writing – review and editing (equal).

Data availability statement

All data used in this work come from publicly accessible data sets. The WorldClim climate data are available at <www.worldclim.org/data/worldclim21.html>. The Copernicus land cover data are archived on Zenodo <https://doi.org/10.5281/zenodo.3243509>. The eBird Basic Dataset is available for download from eBird after completing a data request form at <https://ebird.org/science/use-ebirddata/download-ebird-data-products>. Pre-processed data ready for analysis are available alongsidethe scripts on Zenodo <https://doi.org/10.5281/zenodo.6024392>.

References

Barton, P. S. et al. 2013. The spatial scaling of beta diversity. – Global Ecol. Biogeogr. 22: 639–647.
Bezanson, J. et al. 2017. Julia: a fresh approach to numerical computing. – SIAM Rev. 59: 65–98.
Booth, T. H. et al. 2014. BIOCLIM: the first species distribution modelling package, its early applications and relevance to most current MaxEnt studies. – Divers. Distrib. 20: 1–9.
Buchhorn, M. et al. 2019. Copernicus global land service: land cover 100m: Epoch 2015: Globe. – Zenodo, <https://doi.org/10.5281/zenodo.3243509>.
Calabrese, J. M. et al. 2014. Stacking species distribution models and adjusting bias by linking them to macroecological models. – Global Ecol. Biogeogr. 23: 99–112.
Carlson, C. J. 2020. Embarcadero: species distribution modelling with Bayesian additive regression trees in R. – Methods Ecol. Evol. 11: 850–858.
Chipman, H. A. et al. 2010. BART: Bayesian additive regression trees. – Ann. Appl. Stat. 4: 266–298.
Clifford, P. et al. 1989. Assessing the significance of the correlation between two spatial processes. – Biometrics 45: 123–134.
Commission for Environmental Cooperation 1997. Ecological regions of North America. – Commission for Environmental Cooperation, <www3.cec.org/islandora/en/item/1701-ecological-regions-north-america-toward-common-perspective/>.
Cribari-Neto, F. and Zeileis, A. 2010. Beta regression in R. – J. Stat. Softw. 34: 1–24.
D'Amen, M. et al. 2015. Using species richness and functional traits predictions to constrain assemblage predictions from stacked species distribution models. – J. Biogeogr. 42: 1255–1266.
D'Antraccoli, M. et al. 2020. More species, less effort: designing and comparing sampling strategies to draft optimised floristic inventories. – Perspect. Plant Ecol. Evol. Syst. 45: 125547.
da Silva, P. G. and Hernández, M. I. M. 2014. Local and regional effects on community structure of dung beetles in a Mainland–Island scenario. – PLoS One 9: e111883.
da Silva, P. G. et al. 2018. Disentangling the correlates of species and site contributions to beta diversity in dung beetle assemblages. – Divers. Distrib. 24: 1674–1686.
da Silva, P. G. et al. 2020. Can taxonomic and functional metrics explain variation in the ecological uniqueness of ecologically-associated animal groups in a modified rainforest? – Sci. Total Environ. 708: 135171.
Dansereau, G. and Poisot, T. 2021. SimpleSDMLayers.jl and GBIF.jl: a framework for species distribution modeling in Julia. – J. Open Source Softw. 6: 2872.
De Cáceres, M. et al. 2012. The variation of tree beta diversity across a global network of forest plots. – Global Ecol. Biogeogr. 21: 1191–1202.
Deus, F. et al. 2020. Avian beta diversity in a neotropical wetland: the effects of flooding and vegetation structure. – Wetlands 40: 1513–1527.
Dray, S. et al. 2021. Adespatial: multivariate multiscale spatial analysis. – <https://CRAN.R-project.org/package=adespatial>.
Dubois, R. et al. 2020. Ecological uniqueness of plant communities as a conservation criterion in lake-edge wetlands. – Biol. Conserv. 243: 108491.
Dubuis, A. et al. 2011. Predicting spatial patterns of plant species richness: a comparison of direct macroecological and species stacking modelling approaches. – Divers. Distrib. 17: 1122–1131.
eBird Basic Dataset 2019. Version: EBD_relJun-2019. – Cornell Lab of Ornithology.
Ferrier, S. and Guisan, A. 2006. Spatial modelling of biodiversity at the community level. – J. Appl. Ecol. 43: 393–404.
Fick, S. E. and Hijmans, R. J. 2017. WorldClim 2: new 1-km spatial resolution climate surfaces for global land areas. – Int. J. Climatol. 37: 4302–4315.
GDAL/OGR Contributors 2021. GDAL/OGR geospatial data abstraction software library. Manual. – Open Source Geospatial Foundation, <https://gdal.org>.
Guisan, A. and Rahbek, C. 2011. SESAM a new framework integrating macroecological and species distribution models for predicting spatio-temporal patterns of species assemblages. – J. Biogeogr. 38: 1433–1444.
Guisan, A. and Thuiller, W. 2005. Predicting species distribution: offering more than simple habitat models. – Ecol. Lett. 8: 993–1009.
Heino, J. and Alahuhta, J. 2019. Knitting patterns of biodiversity, range size and body size in aquatic beetle faunas: significant relationships but slightly divergent drivers. – Ecol. Entomol. 44: 413–424.
Heino, J. and Grönroos, M. 2017. Exploring species and site contributions to beta diversity in stream insect assemblages. – Oecologia 183: 151–160.

Heino, J. et al. 2015. A comparative analysis reveals weak relationships between ecological factors and beta diversity of stream insect metacommunities at two spatial levels. – Ecol. Evol. 5: 1235–1248.
Heino, J. et al. 2017. Unravelling the correlates of species richness and ecological uniqueness in a metacommunity of urban pond insects. – Ecol. Indic. 73: 422–31.
Hurlbert, A. H. and Jetz, W. 2007. Species richness, hotspots and the scale dependence of range maps in ecology and conservation. – Proc. Natl Acad. Sci. USA 104: 13384–13389.
Johnston, A. et al. 2021. Analytical guidelines to increase the value of community science data: an example using eBird data to estimate species distributions. – Divers. Distrib 27: 1265–1277.
Kong, H. et al. 2017. Spatio-temporal variation of fish taxonomic composition in a south-east asian flood-pulse system. – PLoS One 12: e0174582.
Konowalik, K. and Nosol, A. 2021. Evaluation metrics and validation of presence-only species distribution models based on distributional maps with varying coverage. – Sci. Rep. 11: 1482.
Landeiro, V. L. et al. 2018. Species-poor and low-lying sites are more ecologically unique in a hyperdiverse Amazon region: evidence from multiple taxonomic groups. – Divers. Distrib. 24: 966–977.
Legendre, P. and Condit, R. 2019. Spatial and temporal analysis of beta diversity in the Barro Colorado Island Forest Dynamics Plot, Panama. – For. Ecosyst. 6: 7.
Legendre, P. and De Cáceres, M. 2013. Beta diversity as the variance of community data: dissimilarity coefficients and partitioning. – Ecol. Lett. 16: 951–963.
Legendre, P. and Fortin, M.-J. 1989. Spatial pattern and ecological analysis. – Vegetatio 80: 107–138.
Legendre, P. et al. 2005. Analyzing beta diversity: partitioning the spatial variation of community composition data. – Ecol. Monogr. 75: 435–450.
Niskanen, A. K. J. et al. 2017. Drivers of high-latitude plant diversity hotspots and their congruence. – Biol. Conserv. 212: 288–299.
Oksanen, J. et al. 2019. Vegan: community ecology package. – <https://CRAN.R-project.org/package=vegan>.
Ovaskainen, O. et al. 2017. How to make more out of community data? A conceptual framework and its implementation as models and software. – Ecol. Lett. 20: 561–576.
Poisot, T. et al. 2017. Hosts, parasites and their interactions respond to different climatic variables. – Global Ecol. Biogeogr. 26: 942–951.
Poisot, T. et al. 2019. Data-based, synthesis-driven: setting the agenda for computational ecology. – Ideas Ecol. Evol. 12: 9–21.
Pollock, L. J. et al. 2014. Understanding co-occurrence by modelling species simultaneously with a joint species distribution model (JSDM). – Methods Ecol. Evol. 5: 397–406.
Qiao, X. et al. 2015. Beta diversity determinants in badagongshan, a subtropical forest in central China. – Sci. Rep. 5: 17043.
Sor, R. et al. 2018. Uniqueness of sampling site contributions to the total variance of macroinvertebrate communities in the Lower Mekong Basin. – Ecol. Indic. 84: 425–432.
Staniczenko, P. P. A. et al. 2017. Linking macroecology and community ecology: refining predictions of species distributions using biotic interaction networks. – Ecol. Lett. 20: 693–707.
Strimas-Mackey, M. et al. 2018. Auk: eBird data extraction and processing with AWK. – <https://cornelllabofornithology.github.io/auk/>.
Sullivan, B. L. et al. 2009. eBird: a citizen-based bird observation network in the biological sciences. – Biol. Conserv. 142: 2282–2292.
Tan, L. et al. 2017. How beta diversity and the underlying causes vary with sampling scales in the Changbai Mountain Forests. – Ecol. Evol. 7: 10116–10123.
Tan, L. et al. 2019. Understanding and protecting forest biodiversity in relation to species and local contributions to beta diversity. – Eur. J. For. Res. 138: 1005–1013.
Taranu, Z. E. et al. 2020. Large-scale multi-trophic co-response models and environmental control of pelagic food webs in Québec Lakes. – Oikos 130: 377–395.
Teittinen, A. et al. 2017. Local and geographical factors jointly drive elevational patterns in three microbial groups across subarctic ponds. – Global Ecol. Biogeogr. 26: 973–82.
Tytar, V. and Baidashnikov, O. 2021. Associations between habitat quality and body size in the Carpathian-podolian land snail Vestia turgida: species distribution model selection and assessment of performance. – Zoodiversity 55: 25–40. <http://ojs.akademperiodyka.org.ua/index.php/Zoodiversity/article/view/67>.
Vallejos, R. et al. 2020. Spatial relationships between two georeferenced variables: with applications in r. – Springer. <http://srb2gv.mat.utfsm.cl/>.
Vasconcelos, T. S. et al. 2018. Expected impacts of climate change threaten the anuran diversity in the brazilian hotspots. – Ecol. Evol. 8: 7894–7906.
Venables, W. N. and Ripley, B. D. 2002. Modern applied statistics with S, 4th edn. – Springer. <www.stats.ox.ac.uk/pub/MASS4/>.
Vilmi, A. et al. 2017. Ecological uniqueness of stream and lake diatom communities shows different macroecological patterns. – Divers. Distrib. 23: 1042–1053.
Yang, J. et al. 2015. The compositional similarity of urban forests among the world's cities is scale dependent. – Global Ecol. Biogeogr. 24: 1413–1423.
Yao, J. et al. 2021. Ecological uniqueness of species assemblages and their determinants in forest communities. – Divers. Distrib. 27: 454–462.
Zurell, D. et al. 2020. Testing species assemblage predictions from stacked and joint species distribution models. – J. Biogeogr. 47: 101–113.