

1

Université de Montréal

2

Spatially continuous identification of beta diversity hotspots using species distribution models

3

4

By

5

Gabriel Dansereau

6

20147609

7

Département de sciences biologiques

8

Faculté des arts et des sciences

9

Advisory Committee Meeting

10

November 29, 2019

¹¹ **Contents**

¹²	Abstract	3
¹³	Introduction	4
¹⁴	Methods	6
¹⁵	1. Data Collection	6
¹⁶	2. Data Manipulation	7
¹⁷	3. SDM – The BIOCLIM Method	7
¹⁸	4. LCBD Calculation	8
¹⁹	5. Prediction Validity	9
²⁰	6. Alternative methods	10
²¹	7. Climate Change Scenarios and Temporal Beta Diversity	11
²²	Preliminary Results	11
²³	References	14
²⁴	Appendix	17

²⁵ **List of Tables**

²⁶	1 Structure of the Warblers data in the eBird checklists for the countries used in the analyses	17
²⁷	2 Description of the WorldClim 2 climate variables used in the analyses	18

²⁹ **List of Figures**

³⁰	1 Distribution of a single species, the Yellow Warbler (<i>Setophaga petechia</i>), based on the raw occurrence data (fig. 1a) and on the probabilistic SDM predictions from the BIOCLIM model (fig. 1b). Purple spots in fig. 1a represent sites where the species was observed. fig. 1b present the probabilities of occurrence as a gradient ranging from 0.0 (species absent) to 1.0 (species present).	19
³⁵	2 Distribution of species richness in North America, defined as the number of Warblers species per site. The raw occurrence observations from eBird (fig. 2a) and the SDM predictions from the BIOCLIM model (fig. 2b) were both transformed into presence-absence data per species before calculating richness.	20

39	3	Distribution of the LCBD values in North America, calculated from the vari-	
40		ance of the community matrix Y and scaled to the maximum value observed.	
41		The Hellinger transformation was applied on the raw occurrence data (fig. 3a)	
42		before calculating the LCBD indices. SDM predictions (fig. 3b) were con-	
43		verted into presence-absence data, but no transformation was applied before	
44		calculating the LCBD indices.	21
45	4	Relationship between the species richness and the LCBD value of the each	
46		site for raw occurrence data (blue) and SDM predictions (orange). Species	
47		richness was calculated as the number of species in a site (α), divided by the	
48		total number of species (γ). LCBD values were scaled to the maximum value	
49		observed. Hellinger transformation was applied on the raw occurrence data	
50		before calculating LCBD indices.	22

51 **Abstract**

52 Beta diversity is an essential measure to describe the organization of biodiversity in space.
53 The calculation of local contributions to beta diversity (LCBD), specifically, allows for the
54 identification of sites with exceptional diversity within a region of interest, which is useful
55 for both community ecology and conservation purposes. However, beta diversity implies a
56 comparison among the sites of a given region, thus, its use is restricted to sites with known
57 species composition, and to discontinuous spatial scales. We therefore propose a method
58 to calculate LCBD indices on continuous scales for a whole region of interest, including
59 unsampled sites. First, species distributions can be predicted on continuous scales using
60 species distribution models (SDM). These models, such as the BIOCLIM method, use the
61 environmental conditions at sampled sites to predict the presence or absence of each species at
62 unsampled locations. Second, LCBD statistics can then be computed on the SDM predictions.
63 We show that it is therefore possible to identify beta diversity hotspots on spatially continuous
64 and extended scales. Our results confirm that LCBD values are related to species richness,
65 and that species-poor sites contribute most to beta diversity.

66 **Introduction**

67 Beta diversity, defined as the variation in species composition among sites in a geographic
68 region of interest (Legendre, Borcard, and Peres-Neto 2005), is an essential measure to
69 describe the organization of biodiversity in space. Total beta diversity within a community
70 can be partitioned into local contributions to beta diversity (LCBD) (Legendre and De Cáceres
71 2013), which allows for the identification of sites with exceptional species composition,
72 hence exceptional biodiversity. Such a method is useful for both community ecology and
73 conservation biology, as it highlights sites that are most important for their research or
74 conservation values. However, LCBD calculation methods require complete information
75 on community composition, such as a community composition matrix Y , thus they are
76 inappropriate for partially sampled or unsampled sites. To our knowledge, these methods
77 have mostly been applied on community data from sampled sites, hence on discontinuous
78 spatial scales, e.g. at intervals along a river stream (Legendre and De Cáceres 2013). This
79 raises the following questions: 1) could LCBD indices be extended to continuous spatial
80 scales, and 2) could this provide novel ecological insights in poorly sampled regions? We
81 aim to answer these questions by combining the LCBD calculation methods with predictive
82 biogeography approaches, and suggest that this would allow for the identification of hotspots
83 with high conservation value in poorly sampled regions.

84 Species distribution models (SDMs) already allow to make predictions on continuous spatial
85 scales, and these predictions could therefore be used to calculate LCBD indices. SDMs, also
86 known as bioclimatic envelope models (Araújo and Peterson 2012), aim to predict species
87 presence or absence based on previous observations of occurrence, and the environmental
88 conditions at which these were made (Poisot et al. 2019). Examples of uses include climate
89 change impact and invasion risk assessment, reserve selection and design, and discovery
90 of new populations (Araújo and Peterson 2012). This way, they generate novel ecological
91 insights for unsampled or lesser-known locations (Poisot et al. 2019), an approach yet to
92 be applied to the LCBD framework. We believe that a predictive approach such as this one
93 would bring a new perspective to biodiversity study and community ecology. By using SDMs,
94 we would be able to expand community information already available, and thus work on a
95 much larger community matrices than in typical LCBD studies, which might highlight new

96 diversity hotspots.

97 Climate and biodiversity data on extended spatial scales are increasingly available online.
98 For instance, the Worldclim 2.0 database (Fick and Hijmans 2017) provides interpolated
99 climate data for global land areas at very high spatial resolution, and the eBird platform
100 (Sullivan et al. 2009) provides a rapidly growing, citizen-contributed database of worldwide
101 bird observations. Both of these are commonly used in SDMs, and offer relevant information
102 on extended spatial scales. Therefore, we believe that these datasets could be used to predict
103 community composition and calculate LCBD indices on continuous spatial scales, and that
104 the result would be representative of the true community structure.

105 The predictive approach we suggest would be especially useful in poorly sampled regions, or
106 in regions with only sparse sampling. While it does not replace a full sampling within the
107 community, predictions and exploratory analyses do provide relevant ecological insights that
108 could be used in different ways. For instance, our method could help identify unsampled
109 sites with potential conservation value which should be targeted as soon as possible in future
110 studies. We believe that the method could also be combined with IPCC climate change
111 scenarios, which provide projections for climate variables, in a way that would allow us to
112 model beta diversity changes with climate change and to identify the sites where the changes
113 in the community will be most important. Once again, this would prove very relevant in
114 an informative approach, suggesting sites to prioritize for future conservation and more
115 structured research.

116 In this document, we cover in more details the methods that we suggest for this M.Sc. research
117 project. The preparation part of the project, including data collection and manipulation, has
118 already been done. A workflow for the analyses, including code implementation, has been
119 defined as well. We also detail preliminary analyses and results intended as proof-of-concept
120 for the approach, which, of course, needs to be refined. Finally, we discuss methods that we
121 intend to use in future analyses, and whose feasibility is not as clearly stated.

122 **Methods**

123 **1. Data Collection**

124 We decided to focus our analyses on bird species and collected the data available on eBird for
125 the Warblers family. The complete database contains nearly 600 million observations. We
126 chose to focus specifically on the Warblers family, as it is a diverse group, popular among
127 birders, with over 30 million observations. Global citizen-contributed databases often present
128 additional challenges compared to conventional datasets due to their lack of structure, as well
129 as spatial and taxonomic biases (Johnston et al. 2019). For instance, there was a clear bias in
130 our data towards the United States, where there were many more observations and sampling
131 events (tbl. 1). However, eBird offers two advantages over other large scale datasets (Johnston
132 et al. 2019): 1) the data is structured as checklists and users can explicitly specify their
133 observations as “complete checklists” when all detected species were reported, which allows
134 to infer information on species absences, and 2) the dataset is semi-structured and checklists
135 are associated with metadata describing sampling effort, such as duration of search, distance
136 travelled and number of observers, which can be used as controls in the analyses. Hence,
137 model performance can be improved by inferring absences and subsampling checklists, while
138 spatial bias can be compensated by including effort covariates in the model (Johnston et al.
139 2019). Therefore, we believe the dataset can be appropriately used to achieve our objective of
140 expanding measures of exceptional biodiversity through space.

141 We collected the data available in the WorldClim 2 database (Fick and Hijmans 2017) for
142 North America, to which we decided to restrict our analyses. The WorldClim data consists
143 of spatially interpolated monthly climate data for global areas, available for resolutions
144 from 10 arc-minutes to 30 arc-seconds (around 18 km² and 1 km² at the equator). Since the
145 release of the first version of the database in 2005 (Hijmans et al. 2005), it became the most
146 common source of climate data for SDM studies (Booth et al. 2014). The variables we used
147 were different measures of temperature and precipitation (tbl. 2), which very high global
148 cross-validation coefficients (> 0.99 and 0.86 respectively) (Fick and Hijmans 2017). We chose
149 to use the coarser 10 arc-minutes resolution in our preliminary analyses, as we believed it
150 was sufficient for proof of concept of our method. However, Hijmans et al. (2005) showed
151 high within-grid cell variation in the 10 arc-minutes data, and therefore recommended the

152 use of the finer resolution, which hid less of the variation known to the model. Given this, we
153 might reconsider the resolution to use in our final analyses.

154 We chose to restrict our analyses to North America given the high amount of data available in
155 eBird. We believed it represented a suitable scale for our models, large enough to cover a lot
156 of variation in environmental variables and community structure, as well as phenomena
157 such as species migration. We also expected such extent of the spatial scale to cover for
158 imprecision in estimated species ranges.

159 **2. Data Manipulation**

160 WorldClim variables and eBird occurrence data are provided in different formats, so they
161 required some manipulations to be combined together. WorldClim variables are provided in
162 a 2-dimensional grid format, useful for large scale analyses and visualization, where each
163 cell or pixel has a size corresponding to the resolution of 10 arc-minutes. Each of the 19
164 variables forms a different grid. On the other hand, eBird records are occurrence-based, so
165 each entry in the dataset corresponds to an observation of a single species at a given time
166 and location. These entries can easily be matched to the 2D grid format of the WorldClim
167 variables through their spatial coordinates, which we found more useful for large scale
168 analyses and visualization. Hence, for each species, we matched all occurrences in eBird to
169 the grid format of the WorldClim variables, and then created a presence-absence community
170 matrix Y , taking all the grid cells as sites. At the 10 arc-minutes resolution, we obtained
171 39 024 sites with occurrences and 62 species in total. All data manipulations and further
172 analyses were realized in *Julia v1.2.0* (Bezanson et al. 2017), with the basic structure built
173 around the soon-to-be-released `SimpleSDMLayers.jl` package.¹

174 **3. SDM – The BIOCLIM Method**

175 We predicted species distributions using the BIOCLIM method (Nix 1986), a climate-envelope
176 model, considered a classic in the field. This method simply relates a species' distribution
177 to the ranges of bioclimatic variables at known locations (Booth et al. 2014). It has long
178 been outperformed by other methods (Elith et al. 2006), but it is still commonly used for
179 its simplistic approach and ease of understanding, as well as its simple relation to niche

¹<https://github.com/EcoJulia/SimpleSDMLayers.jl>

theory (Booth et al. 2014; Hijmans et al. 2017). It is also primarily designed for presence-only data. Despite that, we chose this method for our preliminary analyses as it was easier to implement and because we believe it to be sufficient for proof-of-concept. We discuss possible alternatives in the “Alternative Methods” section below.

The BIOCLIM method defines species potential ranges as a multidimensional environmental hypervolume bounded by the minimum and maximum values for all occurrences (Franklin 2010). For each species, we established the percentile distribution of each environmental variable at the known locations of occurrence (Hijmans et al. 2017). All sites were then compared to those percentile distributions and given a score per variable according to their ranking between 0.0 (1st percentile) and 1.0 (100th percentile). The median or 50th percentile was considered the most suitable value of the variable, and values larger than 0.5 were subtracted from 1. Therefore, both tails were considered the same. The minimum percentile score across all environmental variables was then selected as the predicted value for each site. Values were multiplied by 2 and could therefore be interpreted as probabilities of species occurrence (Hijmans et al. 2017). Predictions of 1 should be rare by definition, as they require a perfectly median site on all variables, and values of 0 should be frequent, since they occur whenever an environmental value is outside the range of the observed ones (Hijmans et al. 2017).

The final step was to convert the probabilities into presence-absence data, so they could be compared with the raw occurrence data. We transformed the probabilities into zeros and ones by converting all values greater than zero to one. Although it might tend to overestimate species ranges, such a transformation is common in SDMs and can be accounted for during result validation with specific methods (Franklin 2010). We also considered applying a threshold determined by sensitivity analysis, but we haven’t done it yet. In any case, converting into presence-absence data allowed easier calculation of the richness and beta diversity metric.

4. LCBD Calculation

We calculated the LCBD statistics through the total variance of the matrix Y for both the raw data and SDM predictions. Legendre and De Cáceres (2013) showed that LCBD coefficients

209 can be calculated directly through the total variance of matrix Y , or through a matrix of
210 dissimilarities among sampling units. We chose the first approach as it also allows to compute
211 species contributions to beta diversity (SCBD), which could also prove useful for conservation
212 purposes, but we did not investigate these for now. Before computing the LCBD statistics,
213 the presence-absence matrix Y had to be transformed in an appropriate way (Legendre and
214 De Cáceres 2013). We chose to apply the Hellinger transformation to the raw data and no
215 transformation on the SDM predictions for now, although we did not investigate these in
216 detail. The most appropriate transformation still needs to be determined, especially for
217 the SDM predictions. We then computed a matrix S of squared deviations from column
218 means and summed all the values of S to obtain the total sum of squares (SS) of the species
219 composition data (Legendre and De Cáceres 2013). LCBD coefficients are then computed
220 as $LCBD_i = SS_i/SS_{Total}$, where SS_i is the sum of squares of a sampling unit i . Finally, since
221 our matrix Y is very large, the LCBD coefficients are very small, so we scaled them to the
222 maximum value observed.

223 5. Prediction Validity

224 The exact way of testing the validity of the predictions remains to be determined, and will
225 also depend on the exact methods used to make the SDM predictions. A key element to note
226 is that both SDM predictions and LCBD values will have to be validated, hence they might
227 require different methods. Metrics that measure the accuracy of categorical or probabilistic
228 predictions in SDMs are well documented, and take various forms. Some require absence
229 data to test against, and can be used on probabilistic predictions directly (area-under-curve,
230 AUC) or after a conversion of the predictions to binary presence-absence using a given
231 threshold (Kappa index, measuring the difference between observed and chance agreement
232 in a confusion matrix) (Franklin 2010). Other methods are appropriate for presence-only
233 data, such as the Boyce Index. In any case, measuring prediction error is only one part
234 of the validation. Finding appropriate data for evaluation is also critical (Franklin 2010),
235 especially since we aim to describe community structure. Separating the data into training
236 and testing datasets, with 70% and 30% of the observations for instance, is an approach
237 common in machine learning methods. However, all of the available observations might
238 be needed in some cases (Franklin 2010). An interesting approach, suggested by Elith et al.

239 (2006) for SDMs, would be to find independent, well-structured presence-absence datasets
240 for validation, on which both SDM predictions and beta diversity metrics could be tested.
241 This approach has the advantage that the testing data is truly independent of the training
242 one, hence it could be used with certain tests of significance. Although it might not cover the
243 entire extent of the predictions in a single test, this method would bring a closer comparison
244 to the way LCBD metrics are used in most studies. Therefore, it would provide interesting
245 perspectives if combined with other, full-extent validation methods.

246 **6. Alternative methods**

247 Many methods generally outperform BIOCLIM for the predictions, as shown by Elith et
248 al. (2006). In our case, better predictions will come by two different means: 1) approaches
249 that are better than BIOCLIM to model the relationship between species presence-absence
250 (or even abundance) and environmental variables, and 2) approaches that account for other
251 drivers of species distributions, such as ecological interactions and species migration. Machine
252 learning methods, especially, would be interesting alternatives to consider. MAXENT (Phillips,
253 Anderson, and Schapire 2006), another presence-only method, has come to be one of the most
254 widely used methods in SDM studies, often with WorldClim variables (Booth et al. 2014).
255 Similarly, Random Forests are simple to put in place, take into account both presence and
256 absence data, allow for quantification of the variables importance in explaining variation, and
257 offer intrinsic testing metrics (Franklin 2010). However, while those methods might return
258 more accurate predictions, they do not implicitly model other drivers of species distribution,
259 among which species interactions and functional niche. Integrating those factors might
260 prove more difficult given our dataset and our focus on Warblers species, as no appropriate
261 information on their interaction is available. Joint species distribution models (JSDMs) might
262 be an interesting way to encompass those, as they attempt to model species co-occurrence,
263 rather than the distribution of single species (Pollock et al. 2014). Also, a different taxonomic
264 group and dataset with more details on interactions could simply be used. On the other hand,
265 a method that could be applied to any taxonomic group, especially those well represented
266 in large citizen-contributed datasets, would be most useful for research and conservation
267 purposes.

268 **7. Climate Change Scenarios and Temporal Beta Diversity**

269 We aim to apply our method to environmental conditions from IPCC climate change scenarios.
270 First, community compositions after climate change could be modelled on continuous scales
271 through SDMs. Second, we could identify the sites where the community has changed in the
272 most exceptional ways. This identification can be done by looking at the variation in LCBD
273 values, but also through the use of temporal beta diversity indices (TBI) (Legendre 2019).
274 TBI indices allow to study changes in community composition through time from repeated
275 surveys at given sites. Whereas LCBD values essentially measure the contribution to beta
276 diversity of one site compared to all others, TBI measure changes in community composition
277 site-wise between two surveys. Moreover, TBI indices can be decomposed into species losses
278 and gains, and can be tested for significance using a permutation test (Legendre 2019). An
279 approach similar to that of Legendre and Condit (2019) would be interesting to follow in
280 our case. First, they computed LCBD indices and compared the location of the sites with
281 exceptional compositions between two surveys 30 years apart. The comparison showed that
282 important changes seemed to have occurred in a specific swamp region. Then, they used TBI
283 indices to confirm the sites with significant changes, decompose these changes into losses
284 and gains, and identify the species that had changed the most. An approach such as this one
285 could be highly informative with our data, although the permutation tests and corrections
286 required might cause some problems given the number of sites in our study.

287 The possibility of using climate change scenarios in the SDMs also needs to be assessed. We
288 did not try to download nor find the appropriate data for now. However, interpolated climate
289 change variables are sometimes different than the ones in WorldClim. Therefore, the SDM
290 models to use and the resulting predictions might have to be different too, and potentially
291 less reliable. Nonetheless, we believe it will be possible to do some kind of time analysis
292 linking beta diversity, climate change and species distribution modelling, and that it could
293 return highly informative results for conservation purposes.

294 **Preliminary Results**

295 Our preliminary results consisted of comparisons between the raw occurrence data and the
296 SDM predictions for the four following elements: single-species distribution (fig. 1), species

richness (fig. 2), LCBD coefficients (fig. 3), as well as the relationship between the species richness and LCBD coefficients (fig. 4). Two main results emerged from them: 1) the models provided seemingly valid and relevant results for poorly sampled regions, both expected species-poor and species-rich ones, and 2) there was an association between species richness and LCBD coefficients partially confirming the relationship shown in other studies.

First, the example of the Yellow Warbler (*Setophaga petechia*), one of the most observed species, showed that the single-species models predicted a broad distribution covering poorly sampled areas, with notable patches of absence across the continent (fig. 1). Likewise, species richness, defined as the number of species present per site, showed a clear latitude gradient, with the poorest sites to the North and the richest to the South (fig. 2). A form of altitude gradient could also be observed, with the Rockies and other mountains well delimited by their lower values. In both cases, the results make intuitive sense and highlight the models ability to predict species presence despite poor or no sampling. Mexico, for example, has much sparser sampling and fewer observations, but the models predict Yellow Warblers presence in most areas nonetheless, as well as higher species richness than on the highly sampled Atlantic Coast, which make sense for a more southern location. We believe these to be valid insights on poorly sampled locations, although we reckon that intuitive reasoning is not a proof of anything, and that the model might be wrong in important ways. In any case, it highlights the need for an appropriate method of validation, as well as a thoughtful consideration of other factors such as species migration.

Second, our preliminary LCBD results seemed to confirm the association between species richness and LCBD coefficients, while also being valid and insightful. They were however harder to interpret given the use of the Hellinger transformation for the raw occurrence data only. Raw occurrence data showed a negative relationship between species richness and LCBD coefficients (fig. 4), as observed previously by Heino and Grönroos (2017), with no clear geographic pattern (fig. 3a). If anything, the highest values seemed to be at the borders of the most sampled regions, where are located the sites in which fewer species were observed (fig. 2a). On the other hand, SDM predictions showed a quadratic form, with the LCBD coefficients re-increasing after some threshold (fig. 4). The geographic pattern is also clearer, with the highest values to the northern and southern extremes (fig. 3b). Since

327 LCBD indices should highlight the most exceptional species composition, both species poor
328 or species rich, this result is no necessarily surprising. In a way, extremes and intermediate
329 values might be represented more evenly on the continuous scale of a SDM than in the raw
330 data. Although raw occurrence data results concurred with those of Heino and Grönroos
331 (2017), the different result with the SDM projections might show the importance of going
332 beyond occurrence data when using large but spatially biased databases such as eBird. Once
333 again, our results highlight the need for well-thought method and an investigation of the
334 appropriate transformation to use on the data.

335 Finally, one disappointing aspect of our method is that the result failed to identify patterns
336 on finer scales. The trends shown by the SDMs for both the species richness and LCBD
337 coefficients were large-scale, latitude-related patterns. Except for mountains, few exceptional
338 sites are actually shown in the middle of the landscape. While it might have been unrealistic
339 to expect such results from a coarse analysis like ours, it would be useful for conservation
340 purposes to be able the identify precise sites within smaller regions. This might be achieved
341 by using a finer resolution, which we should probably reconsider in light of these results, or by
342 using a different technique, such as training the models and predicting species distributions
343 on large scales, but computing and scaling LCBD values on finer local ones, which might
344 highlight regional differences in a new way.

345 **References**

- 346 Araújo, Miguel B., and A. Townsend Peterson. 2012. "Uses and Misuses of Bioclimatic
347 Envelope Modeling." *Ecology* 93 (7): 1527–39. <https://doi.org/10.1890/11-1930.1>.
- 348 Bezanson, Jeff, Alan Edelman, Stefan Karpinski, and Viral B. Shah. 2017. "Julia: A Fresh
349 Approach to Numerical Computing." *SIAM Review* 59 (1): 65–98. <https://doi.org/10.1137/141000671>.
- 351 Booth, Trevor H., Henry A. Nix, John R. Busby, and Michael F. Hutchinson. 2014. "BIOCLIM:
352 The First Species Distribution Modelling Package, Its Early Applications and Relevance to
353 Most Current MaxEnt Studies." *Diversity and Distributions* 20 (1): 1–9. <https://doi.org/10.1111/ddi.12144>.
- 355 Elith, Jane, Catherine H. Graham, Robert P. Anderson, Miroslav Dudík, Simon Ferrier, Antoine
356 Guisan, Robert J. Hijmans, et al. 2006. "Novel Methods Improve Prediction of Species'
357 Distributions from Occurrence Data." *Ecography* 29 (2): 129–51. <https://doi.org/10.1111/j.2006.0906-7590.04596.x>.
- 359 Fick, Stephen E., and Robert J. Hijmans. 2017. "WorldClim 2: New 1-Km Spatial Resolution
360 Climate Surfaces for Global Land Areas." *International Journal of Climatology* 37 (12):
361 4302–15. <https://doi.org/10.1002/joc.5086>.
- 362 Franklin, Janet. 2010. "Moving Beyond Static Species Distribution Models in Support of
363 Conservation Biogeography: Moving Beyond Static Species Distribution Models." *Diversity
364 and Distributions* 16 (3): 321–30. <https://doi.org/10.1111/j.1472-4642.2010.00641.x>.
- 365 Heino, Jani, and Mira Grönroos. 2017. "Exploring Species and Site Contributions to Beta
366 Diversity in Stream Insect Assemblages." *Oecologia* 183 (1): 151–60. <https://doi.org/10.1007/s00442-016-3754-7>.
- 368 Hijmans, Robert J., Susan E. Cameron, Juan L. Parra, Peter G. Jones, and Andy Jarvis. 2005.
369 "Very High Resolution Interpolated Climate Surfaces for Global Land Areas." *International
370 Journal of Climatology* 25 (15): 1965–78. <https://doi.org/10.1002/joc.1276>.
- 371 Hijmans, Robert J., Steven Phillips, John Leathwick, and Jane Elith. 2017. *Dismo: Species
372 Distribution Modeling.* <https://CRAN.R-project.org/package=dismo>.

- 373 Johnston, A., W. M. Hochachka, M. E. Strimas-Mackey, V. Ruiz Gutierrez, O. J. Robinson,
374 E. T. Miller, T. Auer, S. T. Kelling, and D. Fink. 2019. "Best Practices for Making Reli-
375 able Inferences from Citizen Science Data: Case Study Using eBird to Estimate Species
376 Distributions." *bioRxiv*, March, 574392. <https://doi.org/10.1101/574392>.
- 377 Legendre, Pierre. 2019. "A Temporal Beta-Diversity Index to Identify Sites That Have
378 Changed in Exceptional Ways in Space-Time Surveys." *Ecology and Evolution* 9 (6): 3500–
379 3514. <https://doi.org/10.1002/ece3.4984>.
- 380 Legendre, Pierre, Daniel Borcard, and Pedro R. Peres-Neto. 2005. "Analyzing Beta Diver-
381 sity: Partitioning the Spatial Variation of Community Composition Data." *Ecological
382 Monographs* 75 (4): 435–50. <https://doi.org/10.1890/05-0549>.
- 383 Legendre, Pierre, and Richard Condit. 2019. "Spatial and Temporal Analysis of Beta Diversity
384 in the Barro Colorado Island Forest Dynamics Plot, Panama." *Forest Ecosystems* 6 (1): 7.
385 <https://doi.org/10.1186/s40663-019-0164-4>.
- 386 Legendre, Pierre, and Miquel De Cáceres. 2013. "Beta Diversity as the Variance of Community
387 Data: Dissimilarity Coefficients and Partitioning." *Ecology Letters* 16 (8): 951–63. <https://doi.org/10.1111/ele.12141>.
- 389 Nix, Henry A. 1986. "A Biogeographic Analysis of Australian Elapid Snakes." *Atlas of Elapid
390 Snakes of Australia* 7: 4–15.
- 391 Phillips, Steven J., Robert P. Anderson, and Robert E. Schapire. 2006. "Maximum Entropy
392 Modeling of Species Geographic Distributions." *Ecological Modelling* 190 (3): 231–59.
393 <https://doi.org/10.1016/j.ecolmodel.2005.03.026>.
- 394 Poisot, Timothée, Richard LaBrie, Erin Larson, Anastasia Rahlin, and Benno I. Simmons. 2019.
395 "Data-Based, Synthesis-Driven: Setting the Agenda for Computational Ecology." *Ideas in
396 Ecology and Evolution* 12 (July). <https://doi.org/10.24908/iee.2019.12.2.e>.
- 397 Pollock, Laura J., Reid Tingley, William K. Morris, Nick Golding, Robert B. O'Hara, Kirsten
398 M. Parris, Peter A. Vesk, and Michael A. McCarthy. 2014. "Understanding Co-Occurrence
399 by Modelling Species Simultaneously with a Joint Species Distribution Model (JSDM)." *Methods in Ecology and Evolution* 5 (5): 397–406. <https://doi.org/10.1111/2041-210X.121>

401 80.

402 Sullivan, Brian L., Christopher L. Wood, Marshall J. Iliff, Rick E. Bonney, Daniel Fink, and
403 Steve Kelling. 2009. "eBird: A Citizen-Based Bird Observation Network in the Biological
404 Sciences." *Biological Conservation* 142 (10): 2282–92. <https://doi.org/10.1016/j.biocon.2>
405 009.05.006.

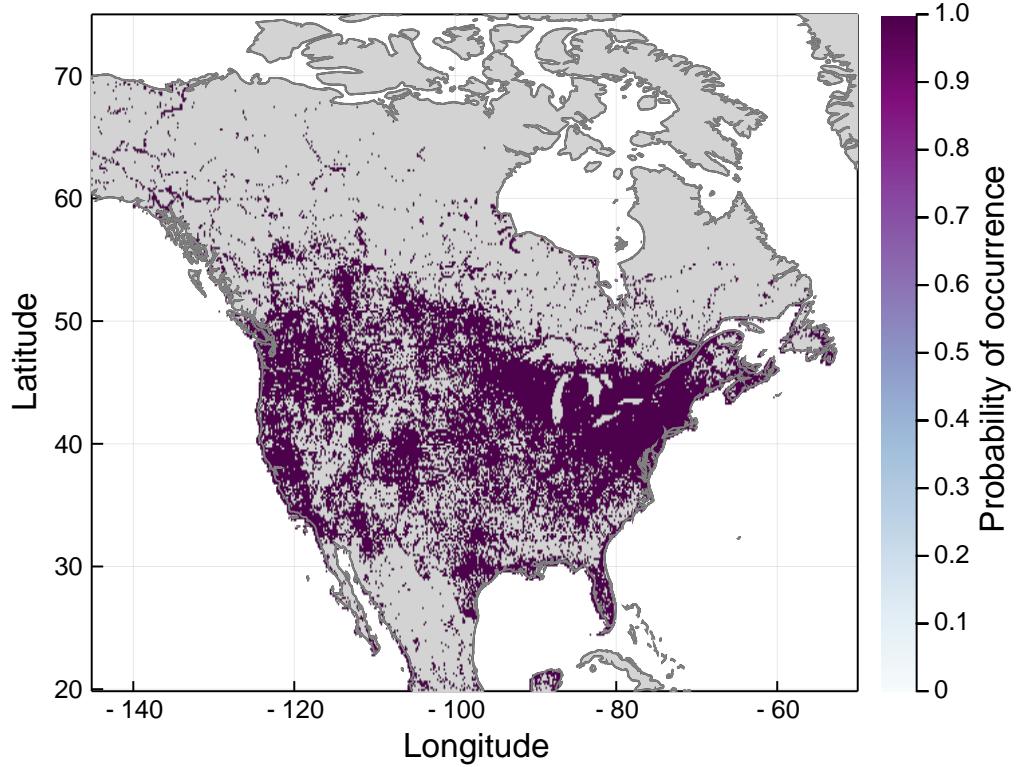
406 **Appendix**

Table 1: Structure of the Warblers data in the eBird checklists for the countries used in the analyses

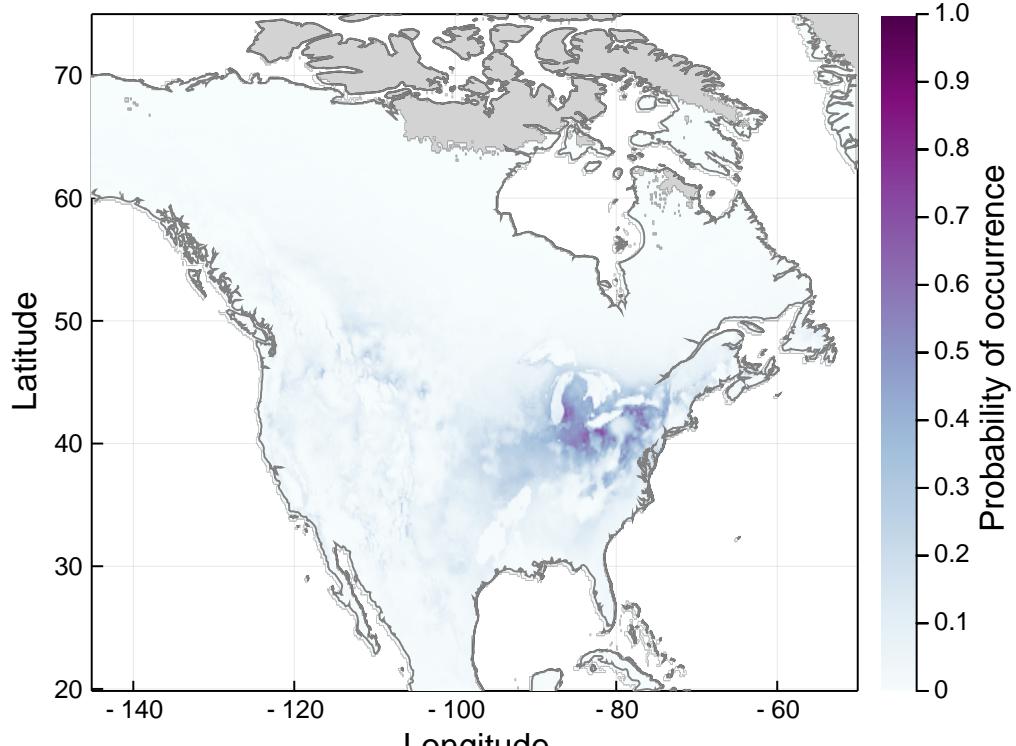
Country	Observations	Checklists	Species	Species per checklist (mean)	Species per checklist (median)	Species per checklist (maximum)
US	19 206 453	7 840 526	56	2.450	2.0	34
CA	3 360 650	1 115 625	45	3.012	2.0	31
MX	407 227	147 599	61	2.759	2.0	21
Total	22 974 330	9 103 750	63	2.523	2.0	34

Table 2: Description of the WorldClim 2 climate variables used in the analyses

Variable	Description
1	Annual Mean Temperature
2	Mean Diurnal Range (Mean of monthly (max temp - min temp))
3	Isothermality (BIO2/BIO7) (* 100)
4	Temperature Seasonality (standard deviation *100)
5	Max Temperature of Warmest Month
6	Min Temperature of Coldest Month
7	Temperature Annual Range (BIO5-BIO6)
8	Mean Temperature of Wettest Quarter
9	Mean Temperature of Driest Quarter
10	Mean Temperature of Warmest Quarter
11	Mean Temperature of Coldest Quarter
12	Annual Precipitation
13	Precipitation of Wettest Month
14	Precipitation of Driest Month
15	Precipitation Seasonality (Coefficient of Variation)
16	Precipitation of Wettest Quarter
17	Precipitation of Driest Quarter
18	Precipitation of Warmest Quarter
19	Precipitation of Coldest Quarter

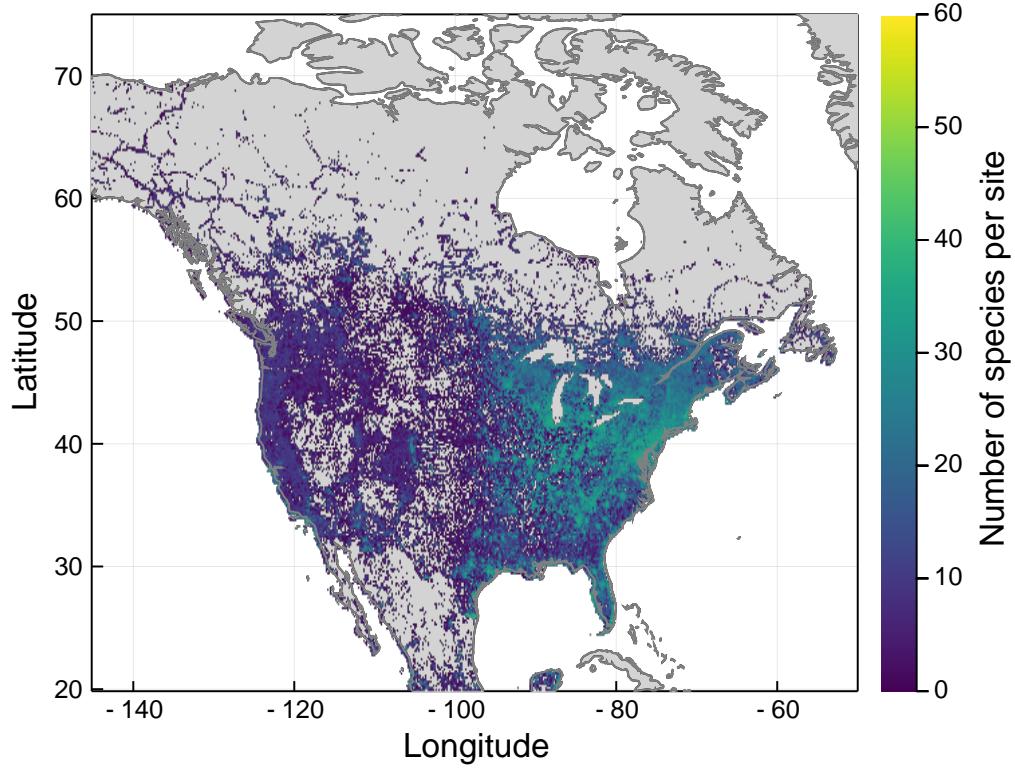


(a) Raw occurrence data

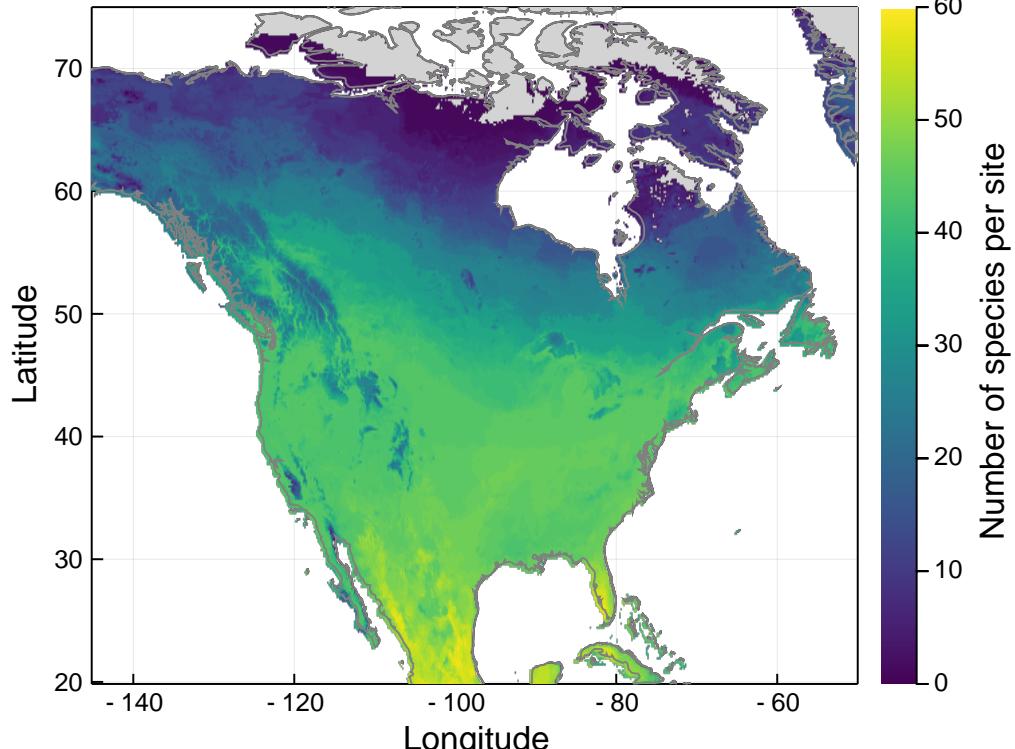


(b) SDM predictions

Figure 1: Distribution of a single species, the Yellow Warbler (*Setophaga petechia*), based on the raw occurrence data (fig. 1a) and on the probabilistic SDM predictions from the BIOCLIM model (fig. 1b). Purple spots in fig. 1a represent sites where the species was observed. fig. 1b present the probabilities of occurrence as a gradient ranging from 0.0 (species absent) to 1.0 (species present).

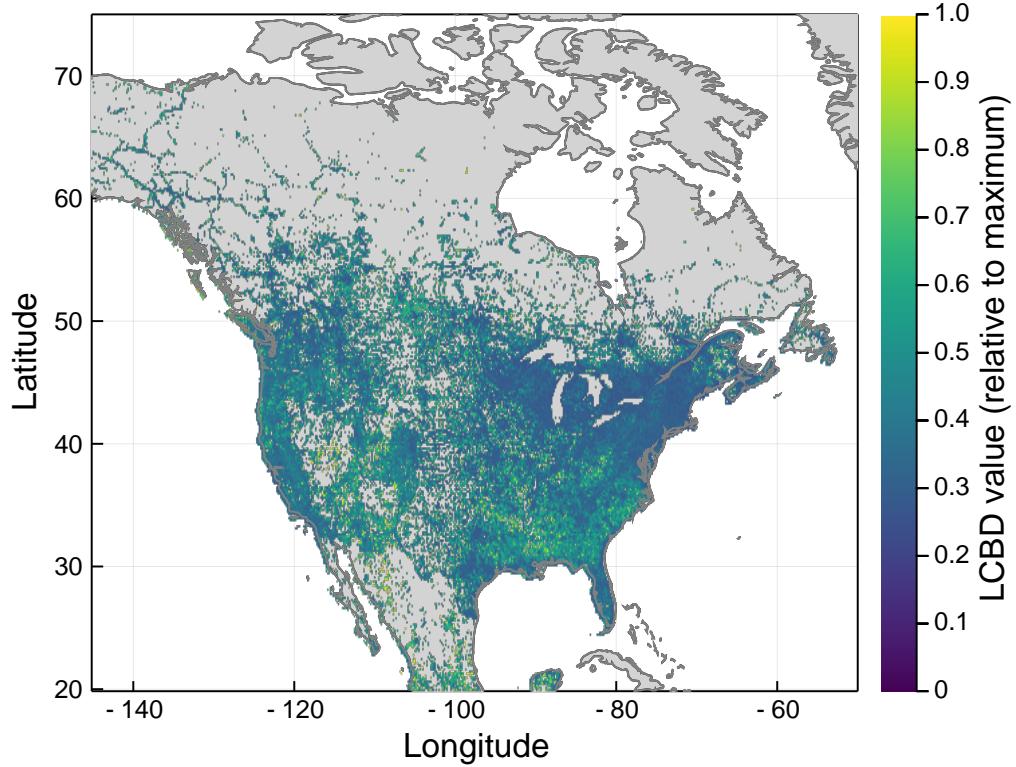


(a) Raw occurrence data

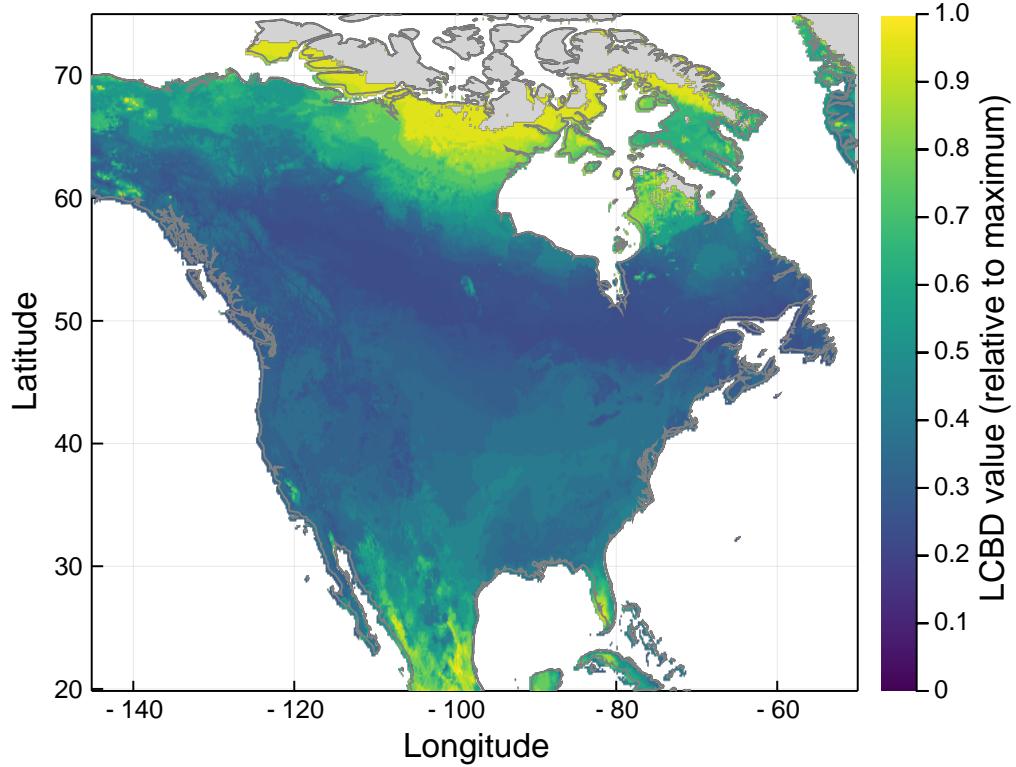


(b) SDM predictions

Figure 2: Distribution of species richness in North America, defined as the number of Warblers species per site. The raw occurrence observations from eBird (fig. 2a) and the SDM predictions from the BIOCLIM model (fig. 2b) were both transformed into presence-absence data per species before calculating richness.



(a) Raw occurrence data (Hellinger transformed)



(b) SDM predictions

Figure 3: Distribution of the LCBD values in North America, calculated from the variance of the community matrix Y and scaled to the maximum value observed. The Hellinger transformation was applied on the raw occurrence data (fig. 3a) before calculating the LCBD indices. SDM predictions (fig. 3b) were converted into presence-absence data, but no transformation was applied before calculating the LCBD indices.

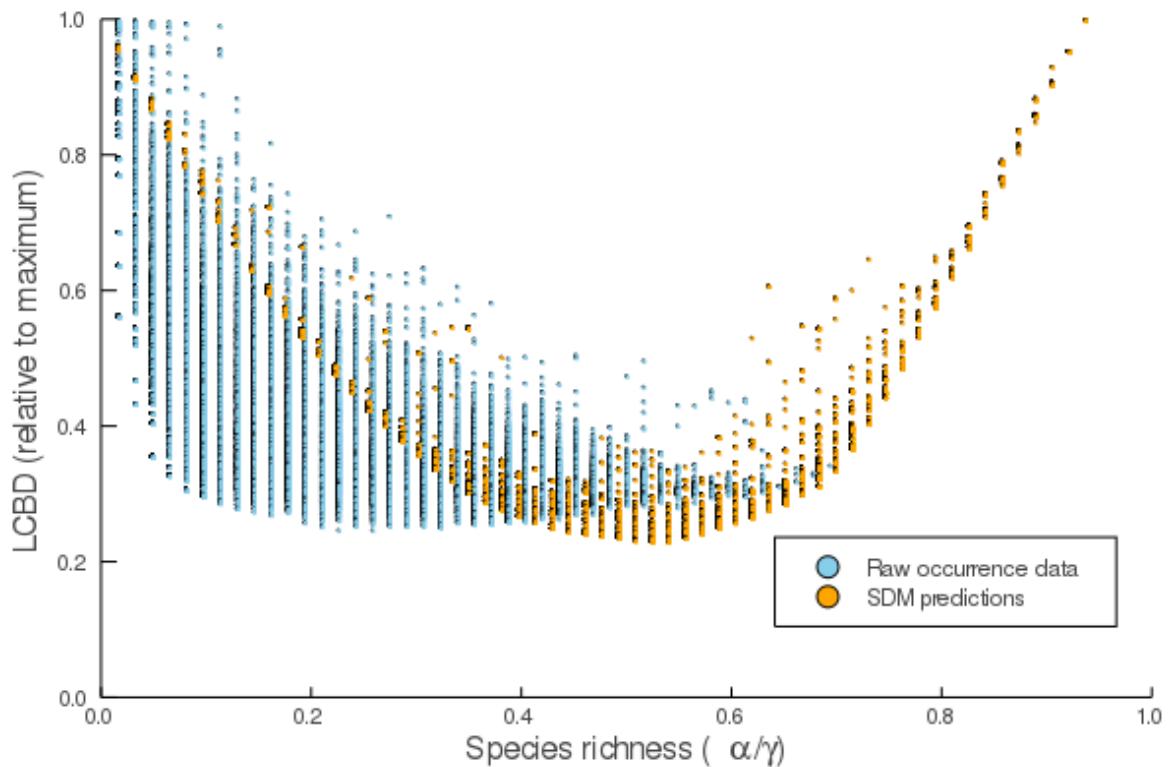


Figure 4: Relationship between the species richness and the LCBD value of the each site for raw occurrence data (blue) and SDM predictions (orange). Species richness was calculated as the number of species in a site (α), divided by the total number of species (γ). LCBD values were scaled to the maximum value observed. Hellinger transformation was applied on the raw occurrence data before calculating LCBD indices.