

# ECOGRAPHY

## Research

### A complete inventory of North American butterfly occurrence data: narrowing data gaps, but increasing bias

Vaughn Shirey, Michael W. Belitz, Vijay Barve and Robert Guralnick

V. Shirey (<https://orcid.org/0000-0002-3589-9699>) ✉ ([vms55@georgetown.edu](mailto:vms55@georgetown.edu)), Dept of Biology, Georgetown Univ., Washington, DC, USA. – M. W. Belitz (<https://orcid.org/0000-0002-8162-5998>), V. Barve (<https://orcid.org/0000-0002-4852-2567>) and R. Guralnick (<https://orcid.org/0000-0001-6682-1504>), Florida Museum of Natural History, Univ. of Florida, Gainesville, FL, USA.

#### Ecography

44: 1–11, 2021

doi: 10.1111/ecog.05396

Subject Editor: Joaquin Hortal  
Editor-in-Chief: Miguel Araújo  
Accepted 23 December 2020



Aggregate biodiversity data from museum specimens and community observations have promise for macroscale ecological analyses. Despite this, many groups are under-sampled, and sampling is not homogeneous across space. Here we used butterflies, the best documented group of insects, to examine inventory completeness across North America. We separated digitally accessible butterfly records into those from natural history collections and burgeoning community science observations to determine if these data sources have differential spatio-taxonomic biases. When we combined all data, we found startling under-sampling in regions with the most dramatic trajectories of climate change and across biomes. We also used multiple methods with each supporting the hypothesis that community science observations are filling more gaps in sampling but are more biased towards areas with the highest human footprint. Finally, we found that both types of occurrences have familial-level taxonomic completeness biases, in contrast to the hypothesis of less taxonomic bias in natural history collections data. These results suggest that higher inventory completeness, driven by rapid growth of community science observations, is partially offset by higher spatio-taxonomic biases. We use the findings here to provide recommendations on how to alleviate some of these gaps in the context of prioritizing global change research.

Keywords: butterflies, community science, GBIF, global change, North America, sampling bias

#### Introduction

The mobilization of openly and freely available digital data from natural history collections has increased the ability for researchers to access information about species distribution and abundance in a given time and place. In recent years, these data have been augmented by community science programs which facilitate collection of biodiversity observations and digital vouchers from a network of volunteers. Aggregated occurrences from both natural history collections and community science programs have been used to answer broad questions in ecology, including assessing extinction risks for understudied groups (Carlson et al. 2017, Seppälä et al. 2018) and modeling species response to environmental change (Eskildsen et al. 2015, Soroye et al. 2020).



[www.ecography.org](http://www.ecography.org)

© 2021 The Authors. Ecography published by John Wiley & Sons Ltd on behalf of Nordic Society Oikos  
This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

Despite the utility of these occurrence records, many taxa are still under-sampled (Troudet et al. 2017) and prevalent biases in the spatiotemporal distribution of these data are noteworthy (Beck et al. 2013, Meyer et al. 2015). These biases imply that inventory completeness (how many species have been recorded versus how many are expected to occur) is also uneven across time and space. Given the urgency to understand ecological responses to many global change processes, knowing where sampling has and has not occurred to a sufficient degree is critical for both prioritizing effort to close information gaps and choosing extents and scales for macroecological analyses. The enormous growth of community science reporting for some groups promises to rapidly close inventory gaps, but it remains unclear how specimens from natural history collections and community science data may have their own biases. For instance, community science volunteers may stay closer to developed areas to sample biodiversity than collectors who may be more attentive to collecting in under-sampled regions. This may lead to larger under-sampling by community scientists in remote regions, including regions at high latitude, which are projected to experience the most dramatic effects of climate change. Under-sampling in the Arctic and other sparsely populated regions compromises the ability to assess how climate has impacted communities over time.

Butterflies (Lepidoptera: Papilionoidea) are a diverse group of organisms that are relatively less sampled compared to vertebrate fauna (Troudet et al. 2017), which have been the focus on previous inventory completeness assessments (Meyer et al. 2015). Although much less diverse than moths, butterflies are a charismatic group that attracts the attention of many professionals and hobbyists. Additionally, butterflies have been widely used to detect signals of global change (Parmesan et al. 1999, Eskildsen et al. 2015). Given the value of butterflies as an indicator group, we aim to test how well sampled North America is for butterflies using natural history collections and community science data, as gaps in openly accessible biodiversity data limit efforts to address ecological, evolutionary and conservation questions. More specifically, we utilize estimates of distributions from field guides to establish a baseline richness value at multiple, coarse scales (Jetz et al. 2012). We then compare that value to richness derived from occurrence records from the Global Biodiversity Information Facility (GBIF), Integrated Digitized Biocollections (iDigBio) and eButterfly.

We distinguished occurrence records into those from natural history collections and from community science-based observations and examined temporal trends in the number of records and completeness for each. We then tested the hypothesis that both types of occurrences were biased to areas of higher human population density, but that those biases were particularly severe for community science records. We also examine if there are differences across butterfly families among these record types, hypothesizing that records from natural history collections are less likely to show taxonomic bias towards selected families. To provide further context for these results, we ask how biomes and climate regimes are

sampled differently. Finally, we discuss potential strategies to mitigate under-sampling across the continent in the future.

## Material and methods

Occurrence records from North America (Canada, Mexico and the United States) were obtained from GBIF (GBIF 2020), iDigBio (Supporting information) (iDigBio 2020) and eButterfly (Prudic et al. 2017) from 1950 through 2019. Range maps of species found in the United States and Canada were digitized from the Kaufman Field Guide to Butterflies of North America (Brock and Kaufman 2006). For species found in Mexico, range maps were digitized from 'A Swift Guide to Butterflies of Mexico and Central America' (Glassberg 2018) as part of the ButterflyNet project, which are digitally available for visualization on Map of Life (Jetz et al. 2012). These maps only include known source population locations and do not include distributions of strays. All range maps from the two sources were merged into a single shapefile consisting of many spatial polygons, which were clipped to only terrestrial areas within North America. These range maps were then intersected with continent-wide equal area grids at 100, 200 and 400 km resolution. A species was considered to occupy a 100 km cell if its range passed within two km of the grid centroid and was considered to occupy a cell at coarser resolutions if its range intersected the grid cell irrespective of distance to the cell centroid. We used this cut-off for the 100 km cell range to increase the precision of a species being an actual occupant of the grid cell. Taxonomic names across the fishnet grids and occurrence data were harmonized to a single taxonomic list using R package taxotools (Barve 2020) and the minority of names that could not be resolved manually after the process were discarded from the analysis (8.1% of names, many of which were genus only determinations). We analyzed only occurrence records that fell within the boundaries of their species' range map but recorded how many records fell outside of these boundaries over time to assess any potential temporal degradation of range maps. In order to address issues with identification quality in community science records, we conducted an expert verification of five taxa from iNaturalist images in our dataset. Each iNaturalist image of an adult butterfly was scored based on how strongly the expert agreed with the original identification. We also examined georeferenced locations for occurrence records, discarding records that did not fall within the terrestrial boundaries of North America and resolving unreported geographic coordinates using a previously constructed gazetteer for occurrence data. The gazetteer was based on downloading all insect records in iDigBio (idigbio.org) and generating a list of unique localities with existing georeferences. Next, we further lumped localities that differed only in capitalization and punctuation. We then checked non-georeferenced butterfly records for matches against this list. In cases with more than one georeference for a matching locality, we chose a best-match based on whether the record had a reported uncertainty and based on precision. This approach

was a prototype for future efforts to speed georeferencing, with more formal approaches to be described in later work.

Inventory completeness was calculated as the ratio of species observed in occurrence data within a grid cell to the number of overlapping range maps within that grid cell. In some cases, due to range map exclusion along coastlines and because we only included species present in the fishnet if it occurred within 2 km of the grid centroid, this ratio was slightly higher than 1.0 and was thus floored to 1.0. We filtered the occurrence data by the basis of the record, year and taxonomy attributes to examine how specimen-only (listed as preserved specimen or material sample from the 'basisOfRecord' field in Darwin Core), community observation-only (listed as human observations from the 'basisOfRecord' field in Darwin Core), time period and the taxon-rank of family (which are monophyletic, Espeland et al. 2018) impacted completeness scores. Machine observations were a small fraction of these data and were not included in the analysis.

Overall, average inventory completeness was assessed using a t-test. We then tested differences in average completeness among families using an ANOVA on the combined, specimen and community observation datasets, and differences in the number of cells complete at or over 50% using a Chi-square test for families between specimen and observation based datasets. Post-hoc testing was conducted with Bonferroni correction in the case of Chi-square (Beasley and Schumacker 1995). We followed t-test, ANOVA, Chi-square and other linear modelling approaches as outlined in Experimental design and data analysis for biologists (Quinn and Keough 2002).

We also assembled spatial data including velocity of climate under the representative concentration pathways (RCP) 4.5 and RCP 8.5 forecasts into the 2050s and 2080s (AdaptWest 2015); human footprint, representing areas where there are built environments, roads or converted land (Venter et al. 2016); protected regions (Dept of Forestry and Natural Resources, Clemson Univ. for CEC 2010); and biomes as designated by the World Wildlife Fund (WWF) (Olson et al. 2001). RCP forecasts examine the greenhouse gas concentration of the atmosphere given future scenarios in which RCP 4.5 represents a 'intermediate' scenario and RCP 8.5 represents a continuous rise in greenhouse gas emissions (Pachauri et al. 2014). For human footprint and climate velocity, we calculated average values, and for protected areas, the percent coverage of those areas, within each 100 km grid cell. For biome type, we determined the proportion of each biome within each 100 km grid to calculate a weighted mean completeness value for each biome. We used these resampled values alongside the completion scores to identify drivers of inventory completeness and under-sampled regions described in more detail below.

### **Human footprint, protected areas and population density**

For potential drivers of completeness, we considered human footprint and protected areas to each represent places where

humans may be actively reporting butterfly occurrences, and specified separate linear models for the combined, museum specimen and community observation datasets as (i.e. inventory completeness ~ human footprint + protected region cell coverage). We also ran these as two univariate models using either human footprint or protected areas as predictors. Model selection was then performed using AIC as the selection criterion to determine the top model (Burnham et al. 2011). We compared selected model goodness of fit for natural history versus community science in order to assess the differential impact these factors may have on datasets with potentially different underlying observation strategies. Additionally, we sought to examine if completeness bias towards regions of high human population density is becoming stronger over time. To do this, we utilized human population density rasters from 1950 to 2010 (Fang and Jawitz 2018), partitioned the original occurrence data by decade and ran a fixed-effects, three way interaction model to predict inventory completeness by population density, decade and basis of record (i.e. inventory completeness ~ human population density × decade × basis of record). Unfortunately, these historical population raster data were only available for the United States, so these analyses are spatially restricted.

### **Sampling in projected novel climate regimes and biomes**

Finally, we examined the inventory completeness within the cells alongside climate velocities from each RCP scenario and the inventory completeness across the WWF biomes found in North America. We removed from our analysis biomes for which the number of 100 × 100 km cells was less than 10. This included just montane forests. We calculated the weighted mean completeness of each biome by considering the proportion of each cell occupied by the biome raster. All data preparation and analysis was performed in R ver. 3.6.3 'Holding the Windsock' (<[www.r-project.org](http://www.r-project.org)>) using the packages tidyverse (Wickham 2019), sp (Pebesma and Bivand 2005), sf (Pebesma 2018), raster (Hijmans 2020), data.table (Dowle and Srinivasan 2019), mapdata (Brownrigg 2018), maptools (Bivand and Lewin-Koh 2020), gridExtra (Auguie 2017), stringr (Wickham et al. 2019), rgdal (Bivand et al. 2020), ggforce (Pedersen 2019), exactextractr (Bastion 2020), sjPlot (Lüdecke 2020), cowplot (Wilke 2019) and scales (Wickham and Seidel 2020). The scripts utilized here are available from a public GitHub repository at <<https://github.com/vmshirey/butterflySampling>>. They are also available with our generated datasets via a Zenodo archive (<<https://doi.org/10.5281/zenodo.4354947>>).

## **Results**

We obtained approximately 2.8 million records from our aggregate GBIF, iDigBio and eButterfly datasets. Overall, 91.2% of occurrence records across the entire dataset fell within range map delineations for their respective species.

This has changed little over time with an average annual in-range percentage of 88.6% from 1950 to 2019 and a recent increase within the last decade of sampling to 91.4% of observations being in-range. In addition, our expert analysis of iNaturalist identifications showed no pattern of consistent misidentification for our select taxa (Supporting information). From 1950 to 2019, the ratio of cells sampled biyearly at 80% completeness by museum specimen data to those by community observations alone decreased dramatically, especially in the last decade of sampling with community-based completion becoming more prevalent as the number of community observations increases (Fig. 1).

### Human footprint, protected areas and population density

In all cases, the best performing model to predict inventory completeness according to AIC included human footprint alone without the percentage of protected natural areas (Table 1). For museum records, the variance explained by the model was low ( $R^2=0.09$ ) compared to the composite dataset ( $R^2=0.25$ ) and the community science dataset ( $R^2=0.29$ ).

Our three-way interaction model to examine spatial bias in inventory completeness demonstrated that the relationship between human population density and inventory completeness is becoming stronger in recent decades for community observations, but that the opposite is true for specimen records (Fig. 2, Supporting information).

### Geographic and taxonomic inventory completeness

Inventory completeness was spatially heterogeneous across scales with noticeable geographic gaps in the far north, mid-west and northern Mexico as illustrated in Fig. 3. Mean specimen and observation-based completeness was significantly different according to our t-test ( $-13.27$ , 2919 DF,  $p < 0.0001$ ), with observations having a higher average completion ratio ( $0.40 \pm 0.007$  SE to  $0.27 \pm 0.006$  SE). Sampling was also irregular across families, especially within the Lycaenidae. To illustrate this better, in the composite dataset, differences among completeness across families were significant according to ANOVA ( $F_{(4, 7267)} = 51.49$ ,  $p < 0.0001$ ) (Supporting information) and ANOVA also supported significant differences across families for the specimen based ( $F_{(4, 5368)} = 86.44$ ,  $p < 0.0001$ ) (Supporting information) and observation based ( $F_{(4, 6325)} = 44.72$ ,  $p < 0.0001$ ) (Supporting information) datasets (post-hoc test results in Supporting information). Chi-square tests to assess differences in the number of  $100 \times 100$  km cells completed at 50% (half of the species expected being reported) or more between specimens and observations revealed there is a significant association with family-level completeness and basis of record ( $\chi^2 = 31.04$ , 4 DF,  $p < 0.0001$ ) (Supporting information). Post hoc comparisons revealed that this association was significant for Nymphalidae and Pieridae with observations having more cells at 50% or more complete in these families ( $p < 0.01$ ) (Supporting information).

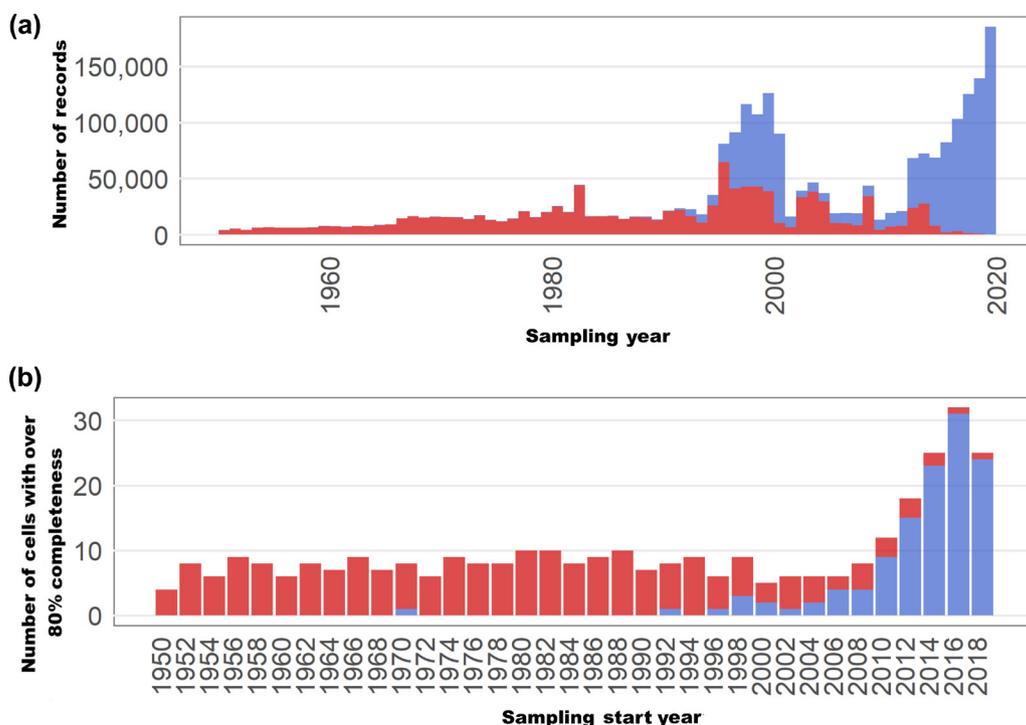


Figure 1. (a) The number of museum specimens and community observation-based occurrence records over time, stacked by year. (b) Number of cells at 100 km resolution that are over 80% complete and that meet that threshold by museum or community observation data alone biyearly. Red = museum specimens; blue = community observations.

Table 1. Coefficient estimates of each multiple regression model for the full record set, specimen-only record set and community observation record set predicting inventory completeness from human footprint and protected areas (e.g. inventory completeness ~ human footprint+protected areas). Delta-AIC values indicate the difference between the multiple regression and simple regression model which included only human footprint as a predictor variable. In all cases, models excluding protected area percentage outperformed the simple regression according to AIC.

	Estimate	95% CI	SE	t	p-value	R <sup>2</sup>	Delta-AIC
All records						0.25	1327.03
Intercept	0.354	0.337–0.370	0.0083	42.63	< 0.0001		
Human footprint	0.027	0.025–0.029	0.0011	23.92	< 0.0001		
Specimens						0.09	1160.43
Intercept	0.194	0.176–0.211	0.0089	21.57	< 0.0001		
Human footprint	0.014	0.011–0.016	0.0012	12.14	< 0.0001		
Observations						0.29	1144.99
Intercept	0.238	0.221–0.256	0.0088	26.96	< 0.0001		
Human footprint	0.028	0.026–0.030	0.0011	24.79	< 0.0001		

### Sampling in projected novel climate regimes and biomes

Sampling was most incomplete in regions experiencing high climate velocity under all four examined RCP scenarios, as illustrated in Fig. 4. In addition, sampling across biomes at the 100 × 100 km resolution was irregular, with some biomes being sampled on average more completely than others as illustrated in Fig. 5. Only the Mediterranean woodland/scrub biome demonstrated over 80% inventory completeness on average with notable under-sampling occurring in deserts, tropical and boreal/arctic regions. Moderate inventory completeness (between 50% and 80% completeness on

average) was demonstrated within most mid-latitude temperate regions.

### Discussion

Inventory completeness across North America has increased in recent years, driven strongly by the growing number of community observations generated from programs such as iNaturalist, which shares research grade observations with GBIF, and eButterfly (Fig. 1a). The majority of cells with > 80% completeness are now derived from community science data, which has continued to grow each year over the past

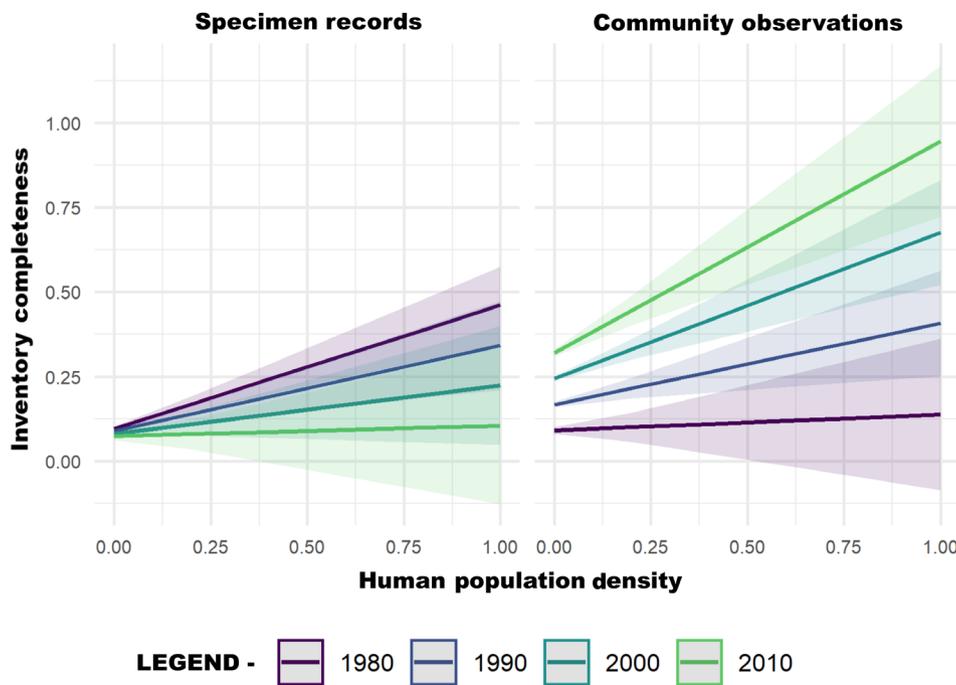


Figure 2. Results of our three-way interaction model for inventory completeness by human population density, decade and basis of record (e.g. inventory completeness ~ human population density × decade × basis of record). Colors indicate the decade of sampling for each record time (specimen records versus community observations). Results indicate that the influence of human population density is becoming stronger for community observations and less strong for specimen records.

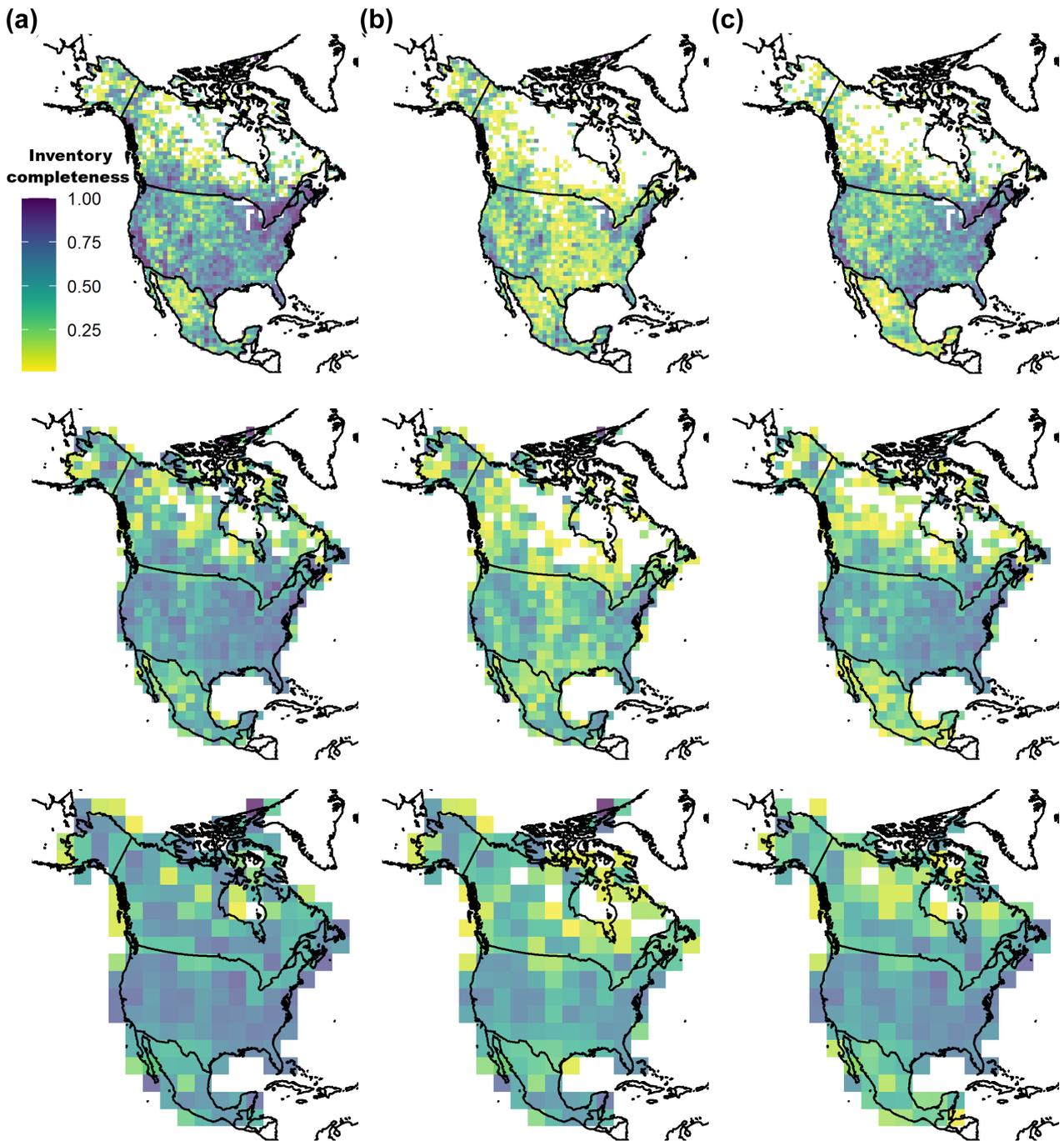


Figure 3. Inventory completeness within cells of varying spatial resolution (100, 200, 400 km) across North America from 1950 to 2019 based on record source (a) all records, (b) specimens and (c) community observations. Projection is North American Albers, Equal Area Conic.

10 years (Fig. 1b), demonstrating the importance of these data for closing distribution knowledge gaps. A large volume of community science observations may be due to the ease of submission and a rather agnostic sampling protocol. For example, iNaturalist submissions can be completed by simply taking a photograph on a mobile phone. Networks such as eButterfly often appeal more directly to dedicated lepidopterists of various skill levels, and do not require photo vouchers

to publish data, which has the potential to allow for more observations of butterfly species that are difficult to photograph. This is in contrast to specimen-based data in which preparation, curation and digitization are all required steps to publish occurrence data.

Despite this influx of community science data, sampling is still heterogeneous across space and taxonomy (Fig. 3, Supporting information). Regions with low human footprint

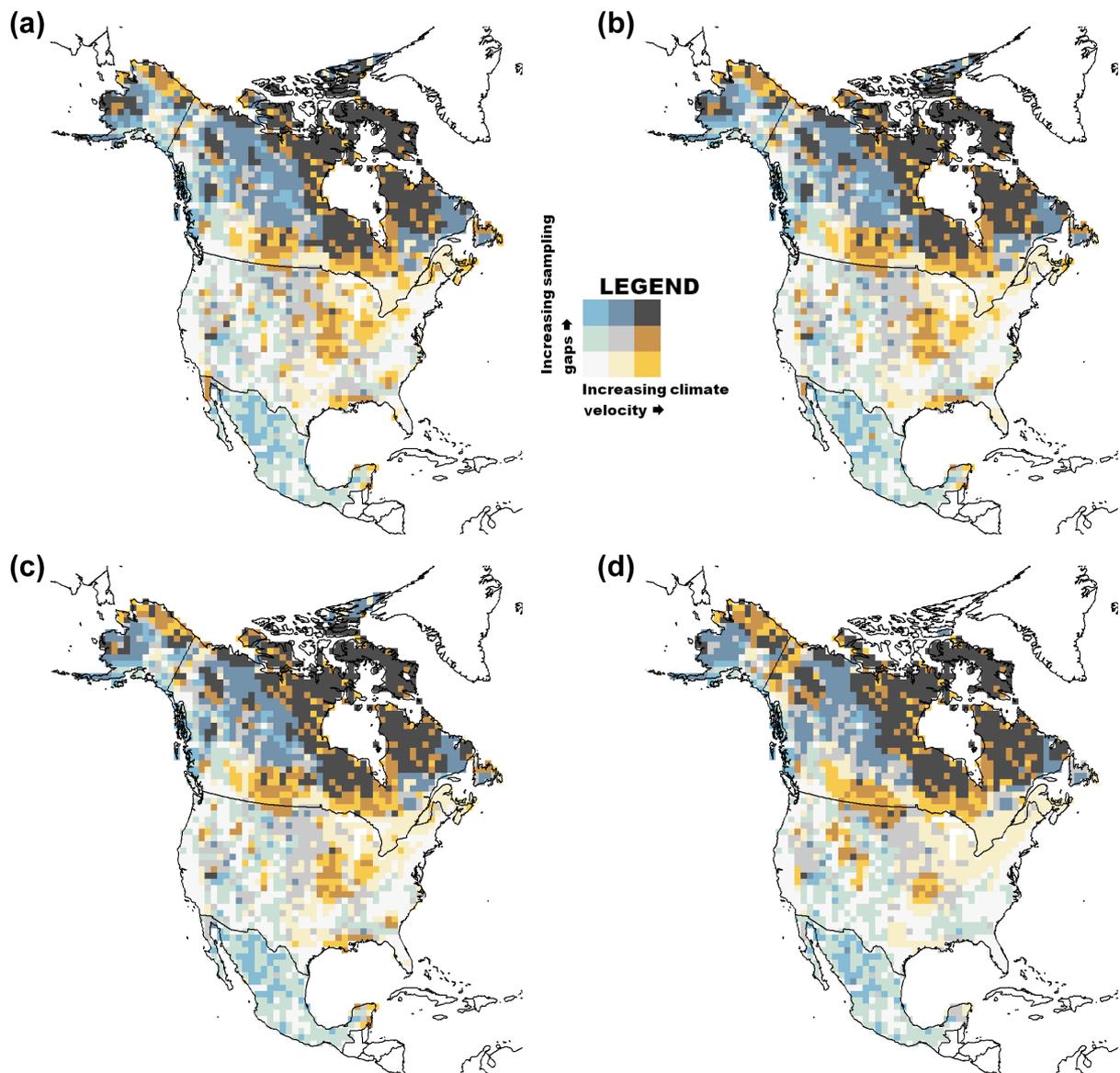


Figure 4. Intensity of sampling gaps versus intensity of climate velocity under four scenarios: (a) RCP 4.5 into the 2050s, (b) RCP 4.5 into the 2080s, (c) RCP 8.5 into the 2050s and (d) RCP 8.5 into the 2080s. Darker colors indicate both high climate velocity and increased sampling gaps. Climate velocity rasters do not extend into northern Nunavut for some forecasts and climate velocity colors are relative to each individual RCP scenario. Projection is North American Albers, Equal Area Conic.

are frequently under-sampled or not sampled at all, and our simple model validates this finding alongside other studies that have examined the relationship between human population densities and record densities (Girardello et al. 2019). A key finding is that these biases towards sampling where human infrastructure is most developed are stronger for community observation data than for specimens (Table 1). Thus, community science observations are not likely to be a panacea for closing inventory gaps. Indeed, our analysis of the relationship between human population density and inventory completeness over time revealed that community science driven completeness is becoming increasingly associated with regions of higher human population densities (Fig. 3). We suspect that the development of agnostic, community science

projects such as iNaturalist have driven this trend as people are probably more likely to be engaged with their immediate surroundings. Surprisingly, the opposite trend was found for museum specimen records, with inventory completeness seeming to be less associated with human population densities over time (although they are still spatially biased towards areas of high human population density). This result may suggest that scientific butterfly collectors are increasingly focusing collecting efforts in areas outside of metropolitan centers.

While some areas of North America are likely to be inventoried at increasingly finer spatial grain with burgeoning growth of community science data, other areas may remain perniciously under-sampled. This likely continuing butterfly

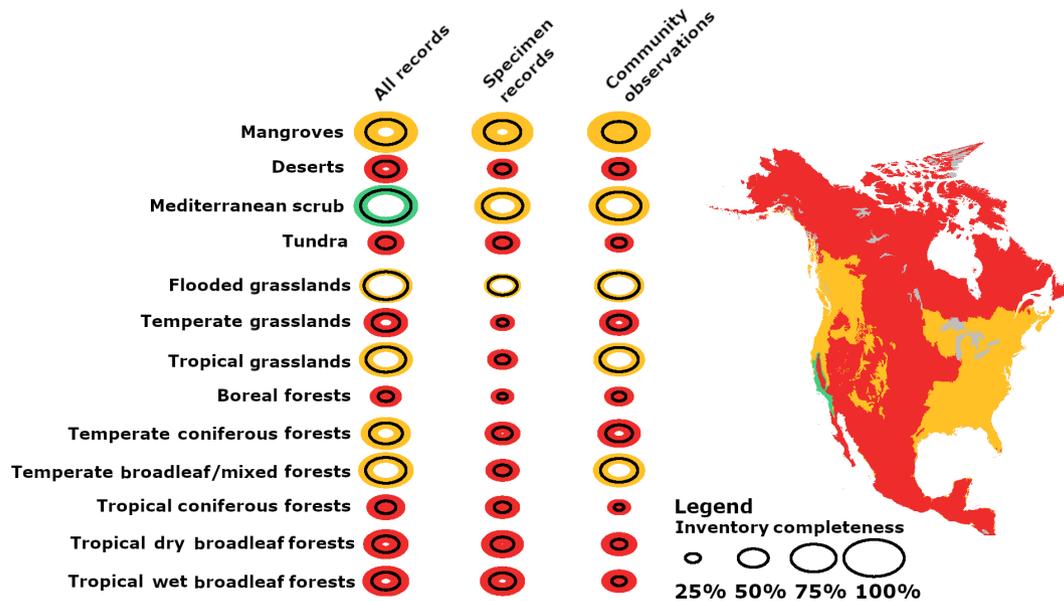


Figure 5. Weighted mean  $\pm$  SD inventory completeness across WWF biomes, (a) composite dataset, (b) community observations only and (c) museum specimens only. Panel (d) displays the biomes utilized without delineation for clarity and includes coloration based on average composite inventory completeness. Red = sampling below 50% average completeness, yellow = sampling average between 50% and 80% completeness, green = sampling average at or above 80% completeness.

inventory knowledge gap in remote regions is thus both particularly challenging and crucial to overcome since these are exactly the areas forecasted to experience the most climatic change. In the polar/subpolar regions of North America, climate velocities are particularly high, yet sampling is woefully incomplete (Fig. 4). Even more concerning is that regardless of RCP scenario or timeline, these northern regions remain consistently at the most extreme climate velocity classification. As well, even some mid and low-latitude biomes are under-sampled, including deserts and many tropical biomes in which butterfly diversity is extremely high (Willig et al. 2003) (Fig. 5). We argue that community science alone is unlikely to solve existing gaps in biodiversity monitoring unless those programs are directed into sparsely populated regions through socially responsible excursions or other research campaigns that consult with local stakeholders and Indigenous communities. These directed and collaborative efforts, requiring partnerships and coordination, will help to provide a critical basis for mapping and ultimately monitoring butterfly diversity to detect changes in the face of shifting climate regimes.

We had anticipated that traits that make butterflies easier to detect, photograph and identify might be different across butterfly families, thus leading to familial-level biases in completeness. We expected these biases to be more acute for community scientists, compared to professional collectors. This is not to say that community scientists are not experts in their own right (as certainly, many are), but that data from community science programs may be biased differently than those from museum collections based on observation and reporting strategies. We demonstrated preliminary evidence for one form of reduced spatial bias in natural

history specimen collecting, which might also suggest sampling across more habitats, potentially reducing taxonomic biases. In the composite dataset, Lycaenidae exhibit lower average completeness with most other groups differing from each other as well (Supporting information), supporting our hypothesis of taxonomic biases in completeness. However, we did not find evidence that natural history specimen collecting led to less taxonomically biased sampling, at least at the familial level. We did, however, find that completeness from community science observations was higher compared to natural history specimen records only for nymphalids and pierids and not for other butterfly families. We had instead anticipated either similar trends across taxonomic groups, or that showy families such as the swallowtail butterflies (i.e. Papilionidae) were more likely to be biased in favor of community science observations given they are generally colorful, large and charismatic. Further exploration using species-level trait data to tease apart these patterns is warranted. In particular, species-level rarity may be particularly important, especially if phylogenetically conserved. Other traits that may be worth examining include habitat and flight preferences (canopy versus understory fliers) that directly relate to ease of human observation. For example, canopy fliers, like many species in Papilionidae, may be difficult for community scientists to photograph, whereas high relative abundance as exhibited by *Pieris rapae*, may make photography and reporting much easier. Species with cryptic coloration such as members of the genus *Oeneis* may be difficult to detect in situ and as fast fliers, also difficult to photograph. Further, species that are hard to identify such as members of the genus *Speyeria* may dissuade community scientists from reporting (Riva et al. 2020).

## Comparisons with prior work and recommendations

In comparison to similar work conducted on other taxa, our results show a consistent pattern of records being generally biased towards human population and infrastructure and with oversampling of rare or targeted species in many cases (Girardello et al. 2019, Haque et al. 2020). With respect to butterflies, our study expands upon prior work done specifically on butterfly inventory completeness (Girardello et al. 2019) by including an independent baseline richness via digitized maps at coarse resolution and by examining the contributions of specimens and community observations. Girardello et al. (2019) utilized species accumulation curves (SACs) which estimate the expected species richness in a given area based on repeated sampling. In this previous study, they utilized the slope of the last 10% of the SAC to estimate saturation in complete sampling. SAC construction relies on an estimate of the size of the species pool based on sampling, which is not repeated in opportunistic data. By establishing a hard boundary on how many species we expected to occur via range map overlap, we removed this estimation process from our approach. One caveat to our use of range maps from published literature is that, in some cases, range maps are partially based on museum specimen data as well as other books and personal experience and thus may be over or under-generalized. However, we believe these results still hold merit by focusing on North America and by including an assessment of inventory completeness in regions with high climate velocity and across biomes, we can better assess which areas are in need of targeted sampling in the future. Specifically, and in contrast to previous work (Girardello et al. 2019), we found a severe lack of sampling in the most northern regions of North America. This urgency to sample the north is further supported by the stark reality that these regions are also experiencing the most drastic impacts of climate change (Manabe and Stouffer 1980, Gauthier et al. 2015). Additionally, at more coarse spatial resolutions (200 and 400 km), our results do diverge from those of Girardello et al. (2019) which illustrate more completeness in high latitude region of North America. This is likely because the species pool in these regions is smaller, which leads to a quicker SAC asymptote with fewer species observations. Different approaches to SAC construction have been shown to produce different estimates in completeness (Pelayo-Villamil et al. 2018), and completeness is also influenced by spatial scale (Lobo et al. 2018). By setting a hard boundary on expected species richness using range maps, our study avoids this potential pitfall. It is worth noting that while we tested and found no evidence for the temporal degradation of our range maps, range maps are ultimately just one source of expert knowledge regarding distributions (Jetz et al. 2012). Despite this, our approach of comparing observed versus expected species richness has revealed similar patterns at the 100 km scale to previous work, and critical regions of under-sampling across the continent.

Overall, several key regions should be prioritized for sampling including: (a) tundra and boreal forest, (b) tropical forests and (c) deserts. Given the relatively low human

population densities of these regions, funding directed towards establishing community science initiatives, and partnerships among organizations with interests in butterfly monitoring, will likely be critical alongside complementing these initiatives with specimen collection and focal digitization of records in these regions. It will be important to target regions with low inventory completeness through an approach that integrates knowledge about species specific environmental requirements to either confirm or deny a species presence in each locality. Species distribution and occupancy models will be a crucial tool for discerning where these unverified occurrences may be found (Fois et al. 2018).

## Conclusions

Butterfly inventory completeness is not uniform across North America. Our research has revealed continued under-sampling in regions facing threats from climate change as well as within specific biomes across the continent. These biases are becoming stronger overtime as community science observations are increasingly being recorded in areas with high human population density. Additionally, family level differences in inventory completeness may be driven by species traits and abundance, leading to disparities in completeness across taxa. In order to mitigate some of these biases, attention should be drawn towards establishing community partnerships of both opportunistic and structured survey systems in under-sampled regions. It is clear that community science provides a strong mechanism for alleviating sampling shortfalls and has potential to provide finer-grained views of butterfly communities, but only if such initiatives are also directed farther from regions with the densest human populations and transport infrastructure. Furthermore, additional curation and digitization of museum specimens will be critical in developing a historical backbone for analyses across time and space. Millions of specimens still remain undigitized in arthropod natural history collections (Cobb et al. 2019), and the continuation of funding for museum staff and biodiversity informatics infrastructure will be critical in mobilizing these data needed for ecological research, especially for some kinds of temporal trend analyses (Soroye et al. 2020). Still, continued digitization in some regions has shown that well-sampled areas remain consistently well-sampled with the addition of new data, rather than contributing to under-sampled regions (Stropp et al. 2016, 2020). However, collections are broadening in their scope of which research communities they serve, and repeat sampling of species in the same area is helpful for certain research questions aside from inventory completeness assessments. Supporting digitization in tandem with concerted efforts to direct community science and collection initiatives towards under-sampled regions specifically will move us towards unlocking the full potential of these opportunistic data in an era of global change.

## Data availability statement

Scripts and author generated data are stored both within a Github repository cited in the text as well as a static Zenodo

archive. Other, third party datasets are cited throughout the text. iDigBio datasets used in this analysis are referenced in the Supporting information. Archived code can be found in both a Zenodo (<<https://doi.org/10.5281/zenodo.4354947>>) and GitHub (<<https://github.com/vmshirey/butterflySampling>>) repository.

*Acknowledgements* – We would like to thank the countless museum staff and community scientists for their tremendous work in digitizing and documenting butterfly records from across the continent. We appreciate Minji Ku, Michelle Duong and the Map of Life informatics team for help digitizing butterfly distributions. Map of Life provides visualizations of range products utilized in this study. We would also like to thank Leslie Ries, Mariana Abarca, Elise Larsen, Naresh Neupane and Grace Jeschke for helpful feedback. Finally, we would like to thank four anonymous reviewers for their feedback which contributed greatly to the strengthening of this research.

*Funding* – VS was supported by Georgetown Univ. and this material is based upon work supported by the National Science Foundation Graduate Research Fellowship under grant number 1937959. MWB was supported by a Univ. of Florida Biodiversity Inst. Fellowship. VB and RG were supported by the ButterflyNet project (DEB-1541500).

*Conflicts of interest* – The authors report no conflicts of interest.

## Author contributions

**Vaughn Shirey:** Conceptualization (equal); Data curation (equal); Formal analysis (lead); Investigation (lead); Visualization (equal); Writing – original draft (lead); Writing – review and editing (equal). **Michael Belitz:** Conceptualization (equal); Data curation (equal); Investigation (equal); Methodology (equal); Writing – review and editing (equal). **Vijay Barve:** Conceptualization (equal); Data curation (equal); Investigation (equal); Methodology (equal); Resources (equal); Writing – review and editing (equal). **Robert Guralnick:** Conceptualization (equal); Data curation (equal); Formal analysis (equal); Funding acquisition (equal); Investigation (equal); Methodology (equal); Project administration (equal); Supervision (equal); Validation (equal); Writing – review and editing (equal).

## References

AdaptWest Project. 2015. Gridded climate velocity data for North America at 1 km resolution. – Available at <[www.adaptwest.databasin.org](http://www.adaptwest.databasin.org)>.

Auguie, B. 2017. GridExtra: miscellaneous functions for ‘grid’ graphics. R package ver. 2.3. – <<https://CRAN.R-project.org/package=gridExtra>>.

Barve, V. 2020. Taxotools: tools to handle taxonomic lists. R package ver. 0.0.43. – <<http://doi.org/10.5281/zenodo.3934939>>.

Baston, D. 2020. Exactextractr: fast extraction from raster datasets using polygons. R package ver. 0.3.0. – <<https://CRAN.R-project.org/package=exactextractr>>.

Beasley, T. M. and Schumacker, R. E. 1995. Multiple regression approach to analyzing contingency tables: post hoc and planned comparison procedures. – J. Exp. Educ. 64: 79–93.

Beck, J. et al. 2013. Online solutions and the Wallacean shortfall: what does GBIF contribute to our knowledge of species’ ranges? – Divers. Distrib. 19: 1043–1050.

Bivand, R. and Lewin-Koh, N. 2020. Maptools: tools for handling spatial objects. R package ver. 1.0-1. – <<https://CRAN.R-project.org/package=maptools>>.

Bivand, R. et al. 2020. Rgdal: bindings for ‘geospatial’ data abstraction library. R package ver. 1.5-10. – <<https://CRAN.R-project.org/package=rgdal>>.

Brock, J. P. and Kaufman, K. 2006. Kaufman field guide to butterflies of North America. – Houghton Mifflin Harcourt.

Brownrigg, R. 2018. Mapdata: extra map databases. R package ver. 2.3.0. – <<https://CRAN.R-project.org/package=mapdata>>.

Burnham, K. P. et al. 2011. AIC model selection and multimodel inference in behavioral ecology: some background, observations and comparisons. – Behav. Ecol. Sociobiol. 65: 23–25.

Carlson, C. J. et al. 2017. Parasite biodiversity faces extinction and redistribution in a changing climate. – Sci. Adv. 3: e1602422.

Cobb, N. S. et al. 2019. Assessment of North American arthropod collections: prospects and challenges for addressing biodiversity research. – PeerJ 7: e8086.

Dept of Forestry and Natural Resources, Clemson Univ. for the Commission for Environmental Cooperation. 2010. Terrestrial protected areas of North America, 2010. – <<http://www.cec.org/naatlas>>.

Dowle, M. and Srinivasan, A. 2019. Data.table: extension of ‘data.frame’ R package ver. 1.12.8. – <<https://CRAN.R-project.org/package=data.table>>.

Eskildsen, A. et al. 2015. Ecological specialization matters: long-term trends in butterfly species richness and assemblage composition depend on multiple functional traits. – Divers. Distrib. 21: 792–802.

Espeland, M. et al. 2018. A comprehensive and dated phylogenomic analysis of butterflies. – Curr. Biol. 28: 770–778.

Fang, Y. and Jawitz, J. W. 2018. High-resolution reconstruction of United States human population distribution, 1790–2010. – Sci. Data 5: 180067.

Fois, M. et al. 2018. Using species distribution models at local scale to guide the search of poorly known species: review, methodological issues and future directions. – Ecol. Model. 385: 124–132.

Gauthier, S. et al. 2015. Boreal forest health and global change. – Science 349: 819–822.

GBIF.org 2020. GBIF occurrence. – Download <<https://doi.org/10.15468/dl.8v7jwf>>.

Girardello, M. et al. 2019. Gaps in butterfly inventory data: a global analysis. – Biol. Conserv. 236: 289–295.

Glassberg, J. 2018. A swift guide to butterflies of Mexico and Central America. – Princeton Univ. Press.

Haque, M. M. et al. 2020. Taxonomic shortfalls in digitized collections of Australia’s flora. – Biodivers. Conserv. 29: 333–343.

Hijmans, R. J. 2020. Raster: geographic data analysis and modeling. R package ver. 3.1-5. – <<https://CRAN.R-project.org/package=raster>>.

iDigBio 2020. iDigBio occurrence download (records citation file in Supplemental materials).

Jetz, W. et al. 2012. Integrating biodiversity distribution knowledge: toward a global map of life. – Trends Ecol. Evol. 27: 151–159.

- Lobo, J. M. et al. 2018. KnowBR: an application to map geographical variation of survey effort and identify well-surveyed areas from biodiversity databases. – *Ecol. Indic.* 91: 241–248.
- Lüdtke, D. 2020. sjPlot: data visualization for statistics in social science. R package ver. 2.8.4. – <<https://CRAN.R-project.org/package=sjPlot>>.
- Manabe, S. and Stouffer, R. J. 1980. Sensitivity of a global climate model to an increase of CO<sub>2</sub> concentration in the atmosphere. – *J. Geophys. Res. Oceans* 85: 5529–5554.
- Meyer, C. et al. 2015. Global priorities for an effective information basis of biodiversity distributions. – *Nat. Comm.* 6: 8221.
- Olson, D. M. et al. 2001. Terrestrial ecoregions of the world: a new map of life on Earth: a new global map of terrestrial ecoregions provides an innovative tool for conserving biodiversity. – *Bioscience* 51: 933–938.
- Pedersen, T. L. 2019. Ggforce: accelerating ‘ggplot2.’ R package ver. 0.3.1. – <<https://CRAN.R-project.org/package=ggforce>>.
- Pelayo-Villamil, P. et al. 2018. Completeness of national freshwater fish species inventories around the world. – *Biodivers. Conserv.* 27: 3807–3817.
- Parmesan, C. et al. 1999. Poleward shifts in geographical ranges of butterfly species associated with regional warming. – *Nature* 399: 579–583.
- Pachauri, R. K. et al. 2014. Climate change 2014: synthesis report. – In: Contribution of working groups I, II and III to the fifth assessment report of the Intergovernmental Panel on Climate Change. IPCC.
- Pebesma, E. J. and Bivand, R. S. 2005. Classes and methods for spatial data in R. – *R News* 5: 9–13, <<https://cran.r-project.org/doc/Rnews/>>.
- Pebesma, E. 2018. Simple features for R: standardized support for spatial vector data. – *R J.* 10: 439–446.
- Prudic, K. L. et al. 2017. eButterfly: leveraging massive online citizen science for butterfly conservation. – *Insects* 8: 53.
- Quinn, G. P. and Keough, M. J. 2002. Experimental design and data analysis for biologists. – Cambridge Univ. Press.
- Riva, F. et al. 2020. Identification ‘by eye’: integrative character assessment informs regional field identification of greater fritillary butterflies (Nymphalidae: *Speyeria*). – *J. Insect Conserv.* 24: 259–267.
- Seppälä, S. et al. 2018. Species conservation profiles of a random sample of world spiders I: Agelenidae to Filistatidae. – *Biodivers. Data J.* 6: e23555.
- Soroye, P. et al. 2020. Climate change contributes to widespread declines among bumble bees across continents. – *Science* 367: 685–688.
- Stropp, J. et al. 2016. Mapping ignorance: 300 years of collecting flowering plants in Africa. – *Global Ecol. Biogeogr.* 25: 1085–1096.
- Stropp, J. et al. 2020. The ghosts of forests past and future: deforestation and botanical sampling in the Brazilian Amazon. – *Ecography* 43: 979–989.
- Troudet, J. et al. 2017. Taxonomic bias in biodiversity data and societal preferences. – *Sci. Rep.* 7: 9132.
- Venter, O. et al. 2016. Global terrestrial human footprint maps for 1993 and 2009. – *Sci. Data* 3: 160067.
- Wickham, H. 2019. Stringr: simple, consistent wrappers for common string operations. R package ver. 1.4.0. – <<https://CRAN.R-project.org/package=stringr>>.
- Wickham, H. and Seidel, H. 2020. Scales: scale functions for visualization. R package ver. 1.1.1. – <<https://CRAN.R-project.org/package=scales>>.
- Wickham, H. et al. 2019. Welcome to the tidyverse. – *J. Open Source Softw.* 4: 1686.
- Wilke, C. O. 2019. Cowplot: streamlined plot theme and plot annotations for ‘ggplot2.’ R package ver. 1.0.0. – <<https://CRAN.R-project.org/package=cowplot>>.
- Willig, M. R. et al. 2003. Latitudinal gradients of biodiversity: pattern, process, scale and synthesis. – *Annu. Rev. Ecol. Evol. Syst.* 34: 273–309.