The potential of ecoregional range maps for boosting taxonomic coverage in large-scale ecology and conservation

Stefan Pinkert[1,2,3], Yanina Sica[1,2], Kevin Winner[1,2], and Walter Jetz[1,2,4]

[1]Yale University, Department of Ecology and Evolutionary Biology, New Haven, CT, US
[2]Center for Biodiversity and Global Change, Yale University, New Haven, CT, US
[3]University of Marburg, Department of Conservation Ecology, Marburg, DE
[4]EO Wilson Foundation, Biodiversity Science, Durham, NC, US

December 20, 2022

INTRODUCTION

Information about species' geographical distribution is central to many ecological and evolutionary questions and underpins effective conservation decision-making (Meyer *et al.*, 2015; Jetz *et al.*, 2019; Oliver *et al.*, 2021; Jetz *et al.,* 2022). Ideally, distributional data for a species is contiguous in space and time and covers its entire global distribution, at a scale reasonable to inform conservation action and research (Jetz *et al.*, 2019). Expert range maps (ExpRMs) arguably come closest to this standard (Rondinini *et al.*, 2006). These maps of aggregated knowledge and field experience about a species range have been the most frequently used type of distributional data in global-scale biogeographical analyses, biodiversity research, and area-based conservation (Hurlbert & Jetz, 2007; Herkt *et al.*, 2017; Jung *et al.*, 2021). Nevertheless, their limited taxonomic scope critically hampers conclusions about the vast majority of species on our planet (Dauby *et al.*, 2017).

The importance of ExpRMs in conservation stems from their use in assessing crucial measures of a species' threat status by the International Union for the Conservation of Nature [IUCN Red List criterion B; see Schatz (2002)]. The IUCN is the largest provider of ExpRMs, currently holding ranges for approximately 115,000 species (iucn.org retrieved March 11, 2022). Given the lack of monitoring data to assess species range size (i.e., B1 - extent of occurrence) and population density (B2 - area of occupancy), these measures are often estimated using ExpRMs. In addition, changes in these measures over time are used to further track population declines and range contractions. Within IUCN specialists groups, experts draw simplified polygons around occurrence records and then refine and/or extend these polygons based on ecological knowledge of the species in concert with habitat layers (Hawkins *et al.*, 2008; *IUCN Standards and Petitions Subcommittee, 2022*). Typically, many experts are involved in the generation and evaluation of range maps ensuring that IUCN ExpRMs generally provide a reliably high-quality representation of species' distributions. Other sources of ExpRMs, especially for plant and invertebrate taxa, include monographies of taxa as well as regional and global field guides (e.g., Scott, 1997; Glassberg, 2017).

Over the last few years, an increasing effort has been put into mobilizing range maps from literature sources under metadata standards to not only make them publicly available but also for clarifying decisions of the production process (see birdlife.org or *mol.org* digitized expert ranges; Marsh *et al.*, 2022). However, due in part to the immense work necessary to produce or digitize each ExpRM, their availability is often limited to more popular or well-studied taxa. Currently, comprehensive and high quality ExpRMs are available for a large proportion of vertebrate species (*IUCN.org*), whereas they are available only for a few selected plants and invertebrates subgroups and typically limited in their geographical extent (e.g. mol.org/patterns).

Species occurrence data, particularly from museum collections and citizen science efforts, have grown rapidly in recent decades. The Global Biodiversity Information Facility hosts occurrence records for 1,723,634 animal and plant species (*gbif.org* retrieved March 11, 2022). Approaches integrating ExpRMs with increasingly complete, spatially explicit, and readily available occurrence data promise unique advances for incorporating a significant proportion of all species on Earth into large-scale assessments on the status and trend of biodiversity. The IUCN has adopted two alternative integrative approaches to address limitations of transparency and reproducibility of ExpRMs. First, hydro basin layers are used to infer species' ranges from intersections with observation- and literature-based occurrence records. Although hydro basin-based ranges are limited to species affiliated with lotic (running) waters, this approach vastly improved the availability of baseline distributional data for the assessment of species' threat status of crabs, crayfishes, shrimps, and Odonata (*IUCN.org*). Secondly, simple, non-parametric occurrence-based estimates such as Minimum Convex Hulls (MCVs) have been proposed as range estimates to calculate the extent of occurrence and the area of occupancy of species (Dauby *et al.*, 2017; see also *ala.org.au*). However, both alternatives do not resolve the internal structure of species' ranges that result from barriers to dispersal, geological differences, and ecological gradients and are therefore likely to significantly overestimate the true species range in many cases. As a result, these range surrogates should be more sensitive (i.e., cover more suitable habitat or potential presences) but less precise (i.e., have a lower occupancy of suitable habitat) than ExpRMs, at least for data-rich species (see Figure 1).

We propose the use of terrestrial ecoregions to develop alternative ExpRM surrogates, what we denominate Ecoregional Range Maps (EcoRMs). Ecoregions define the natural extent of areas with similar environmental conditions and distinct ecological communities. Freshwater, marine, and terrestrial ecoregions are established baseline layers used in conservation efforts by the World Wildlife Fund and The Nature Conservancy as well as in assessments of the progress of conservation strategies (e.g., Sayre *et al.*, 2014; Dinerstein *et al.*, 2020). Being based on broad geological and ecological zonation, ecoregions imply a high surrogacy value for species distributions of a broad spectrum of organisms but their congruence with single species distributions and biodiversity patterns has thus far not been evaluated.

Here, we statistically compare the sensitivity and precision of MCVs and EcoRMs based on predicted absence-presence information from ExpRMs and SDMs at the species-level as well as congruence in the resulting species richness patterns. We generally expect MCVs to cover areas with a high proportion of true presences (i.e., to have a higher precision), but EcoRMs to cover a higher overall number of species' true presences (i.e., to have a higher sensitivity). In addition, we investigate the spatial dependence as well as the relationships of sensitivities and precision with the number of underlying occurrence records. With these evaluations, we aim to inform applications about the potential of non-parametric, readily applicable, updateable, and occurrence-based alternatives to ExpRMs and SDMs for boosting the integration of data-poor species into both conservation and ecological research.

METHODS

Cleaning of occurrence records

Occurrence data for all 792 butterfly species of Canada and the US (Pinkert *et al.*, 2022) was downloaded from the Global Biodiversity Information Facility, querying accepted names and synonyms (gbif.org, retrieved on July 7, 2021). Subsequently, species names were harmonized using the most recent taxonomy of butterflies (Pinkert *et al.*, 2022). We removed presence records with date or coordinate issues (based on GBIF flags) resulting in 8,129,916 total records. In addition, 138,791 records were removed because they were located near country centroids, natural history institutions, or GBIF headquarters using methods of the R-package *CoordinateCleaner* (Zizka *et al.*, 2021) with default distance parameters. A total of 6,215,405 of the occurrences were located outside of the study domain (-48° to -175° longitude and 10° to 83° latitude) or in seas. In addition, 571,604 spatio-temporal duplicates and 1,072 records with a minimal interpoint distance of 500 km were removed. Finally, using a recent country checklist from Pinkert et al. (2022), 377 records were removed because they were more than 1000 km away from the borders of species' checklist countries, resulting in a final tally of 1,202,667 records for 792 species. We applied an additional filtering

step for the data used to model species distributions to reduce memory requirements and limit the run time of the modeling procedure. Specifically, only for species with more than 5,000 records, the occurrence data was successively thinned by subsampling to one point per grid cell with grids of increasing grain size (1 km, 2 km, ..., 32 km) until fewer than 5,000 records remained. Note that, although fairly common practice in species distribution modeling, we did not remove records older than 1970 to facilitate comparability with the literature-based ExpRMs. To assess the overall data availability for butterfly species, all global butterfly occurrences records were cleaned using the above-mentioned filters except for limiting the extent and species set (order Lepidoptera excluding moth families). 17% of the 19,327 accepted butterfly species have fewer than 5 records (hereafter 'extremely data-poor species') and 43% have fewer than 100 occurrence records (hereafter 'data-poor species'; all other species were considered 'data-rich').

Range map types

For comparisons of the performance, we used five range map types: Expert range maps, Minimum Convex Hulls, original and simplified ecoregional range maps, and species distribution models.

Expert range maps (ExpRMs)

We compiled ExpRMs for 792 butterflies species of North America from Scott (1997) and Glassberg (2019). Data were georeferenced in shapefile format, quality controlled, taxonomically harmonized using the most recent taxonomy of butterflies (Pinkert *et al.*, 2022), and spatially merged using the R-package *rgdal* (Bivand *et al.*, 2018).

Minimum Convex Hulls (MCVs)

We used the cleaned occurrences and calculated MCVs in the R-package *adehabitatHR* under default settings (Calenge & Fortmann-Roe, 2021). Given potentially strongly unreliable estimates arising from too few occurrence records, we excluded species with fewer than five records, resulting in a total of 662 species with MCVs. We did not use alpha-convex hulls because this approach requires careful tuning of a hyperparameter, α, for each species.

Ecoregional range maps (EcoRMs) - original and simplified

We intersected the cleaned occurrences with a standard ecoregion delineation to produce EcoRMs. For this study, we used the 846 terrestrial ecoregions from (Dinerstein *et al.*, 2017; downloaded at *oneearth.org*) as the most frequently used and globally consistent ecoregion definition. We generated two alternative sets of EcoRMs: 'original' EcoRMs, where ecoregional polygons intersecting the occurrence records were used, and 'simplified' EcoRMs where the outline of the ecoregion was smoothed. For data-rich species we removed ecoregions with only 1 or 2 records (those with [?] 2 records/ecoregion for species with 1000+ total records and those with only 1 record/ecoregion for species with 100+ records). For simplified EcoRMs, we smoothed the outline of the shape depending on whether 10 or less, 100 or less, 1000 or less, or more than 1000 records were available. For the first two cases, we buffered the occurrence records and masked them with a smoothed ecoregion outline (point buffer = 3°/1°; range buffer = 0.25°/0.50°; smooth = 25/50, respectively). For the last two cases, we buffered the selected polygons, filled holes of less than 50 km^2 size and more strongly smoothed the outline (range buffer = 0.5°/0.5°; smooth = 50/100, respectively; for details see protocol and code in data appendix). We acknowledge that further refinements to these maps are possible, e.g. through the consideration of elevational ranges or measures of spatial distance (e.g. Huang *et al.*, 2021, Palacio *et al.*, 2021). However, we herein focused on a simple and readily applicable approach to provide a solution most useful for poorly documented species and that avoids circularity with environmental niche models (e.g., SDMs).

Species distribution models (SDMs)

We generated maximum entropy SDMs (Phillips *et al.*, 2006) using the cleaned occurrences and 11 selected environmental variables with functions of the R-package *dismo* (Hijmans *et al.*, 2021). Only 644 species had both enough cleaned occurrence records to be modeled (i.e., >5 records) and ExpRMs. Ten covariates

were used to produce the SDMs for each species, including climate, topological, and productivity variables. Five climate variables describing annual and seasonality trends were selected from 19 biologically relevant variables (Bio1, Bio4, Bio10, Bio12, Bio15; CHELSA v2 current condition records; Karger *et al.*, 2017, 2018). The average elevation and the coefficient of elevation variation were retrieved from Amatulli et al. (2018). Annual EVI (Enhanced Vegetation Index), Winter EVI, and Summer EVI were retrieved from Tuanmu and Jetz (2015). Standard deviation of interannual variation in MODIS-based cloud cover was taken from Wilson and Jetz (2016). All variables were cropped to the extent of the study area and resampled to a 1 km resolution using bilinear interpolation, if necessary. The modeling domain was set to a buffer of ±5° longitude/latitude around the cleaned occurrence records. MaxEnt models were fitted using 1000 randomly sampled background points and default settings. Models were evaluated on a held-out test set consisting of 20% of the original presences and sampled pseudo-absences. For each species we projected the habitat suitability at 1 km resolution using the final model and considered all cells with a predicted suitability above the 95% quantile of the suitability values extracted from the underlying occurrences records. SDMs with an exceptionally low AUC (<0.5; 20 species) were excluded from further analyses. Most species with an AUC lower than 0.7 (13 species) were data-poor (see also Proosdij *et al.*, 2015) and we kept them for the analysis of the relationships of performance measures with data availability.

Performance assessment

We used both ExpRMs and SDMs as expectations for validation of each of the other range map types (including comparing ExpRMs and SDMs with one another). We took this approach due to the lack of reliable absence data for most species at the continental scale and reflecting an interest in using EcoRMs and other surrogates as data- and computationally efficient alternatives to these two species range estimates. We resampled all five range map types for 624 butterfly species to grids with an approximate grain size of 25 km, 50 km, 100 km, and 200 km using the R-packages *raster*, *sp* , and *sf* (Hijmans *et al.*, 2022; Pebesma *et al.*, 2022b,a). For species in southern parts of the study region MCVs, EcoRMs, and SDMs typically extended into adjacent and potentially suitable regions (see Figure 2). Therefore, in the analyses of all species, range maps were masked using country polygons for Mexico, the US, and Canada (data from *gadm.org*) as our expert range maps are limited to this extent. Following Sofaer et al. (2019), at the species-level we calculated the sensitivity and precision in relation to ExpRMs and SDMs as detailed in Figure 1. We chose to use precision over specificity because specificity is unreliable when using pseudoabsences and thus precision is commonly preferred in presence-only SDM analyses and similarly when building surrogates from presence-only data (Elith & Leathwick, 2009). At the assemblage-level, aggregated distribution data was used to investigate the congruence of species richness patterns among data types with Spearman rank correlations and for mapping richness contrast.

All calculations of sensitivity and precision were repeated for range data aggregated to grain sizes of 25 km, 50 km, 100 km, and 200 km to investigate the scale-dependence of our results. These analyses showed that sensitivity consistently decreased, and precision consistently increased with increasing grain size (Table S1). However, the gain in precision was markedly stronger and the loss of sensitivity markedly lower for a grain size of 100km. We therefore focused the discussion and analyses of congruence of species richness patterns on this grain.

RESULTS

Example species

We exemplified comparisons across range map types for *Typhedanus undulatus,* a small-ranging butterfly species found from the southernmost parts of the US over Mexico to Central America (Figure 2). We find that when validated against the ExpRM, the SDM for *T. undulatus* had the highest sensitivity and precision. The MCV outperformed the simplified and the original EcoRM in terms of sensitivity. The simplified EcoRM outperformed the MCV and the original EcoRM in terms of precision. When validated against the SDM, the simplified EcoRM for *T. undulatus* had the highest sensitivity and mean performance (mean sensitivity and precision). The simplified ExpRM outperformed other approaches in terms of precision. The mean

performance was similarly high for ExpRM and MCV and lowest for the original EcoRM. All approaches based on occurrence records highlighted areas beyond the ExpRM, including the Yucatán peninsula and Guatemala. The occurrence of *T. undulatus* in Florida and the Caribbean was only supported by the SDM, showcasing the need for integrating reasonable range offsets into species distribution modelling.

All species

Extending this assessment to all 624 species, we found that simplified EcoRMs consistently had a greater sensitivity than original EcoRMs, MCVs, and SDMs (0.96, 0.82, 0.79, 0.72, respectively; Figure 3a). SDMs had a greater precision than MCVs, original and simplified EcoRMs (0.69, 0.63, 0.52, and 0.46 respectively). The median of species' mean performance was similar for simplified EcoRMs, MCVs, SDMs, and original EcoRMs (0.71, 0.71, 0.70, 0.67).

Using SDMs as proxies of the true distribution of species, simplified EcoRMs had a greater sensitivity than MCVs, original EcoRMs, and ExpRMs (0.86, 0.68, 0.64, 0.44; Figure 3b). ExpRMs had a greater precision than MCVs, original and simplified EcoRMs (0.73, 0.68, 0.55, 0.51). The mean performance was similarly high for simplified EcoRMs and MCVs, and lowest for both SDMs and original EcoRMs (0.69, 0.68, 0.59, 0.59).

Sample and grain size dependence

The sensitivity and precision of the different types of range maps were generally positively correlated with the number of available occurrence records (Figure 4). Using ExpRMs as proxies of the true distribution of species, the mean performance of SDMs for species with less than 100 occurrence records (i.e., data-poor species) was greater compared to that of simplified EcoRMs, MCVs, and original EcoRMs. Using SDMs as proxies of the true distribution of species, the mean performance was greater for simplified EcoRMs for species with less than 100 occurrence records compared to MCVs, original EcoRMs, and ExpRMs. Separate analyses for range maps aggregated to grain sizes of 25 km, 50 km, 100 km and 200 km revealed that the ranking of mean performance of ranges was generally consistent across spatial scales (Figure 5, Table S1).

Assemblage-level comparisons

The species richness patterns based on MCVs, as well as simplified and original EcoRMs were very similar (Figure 6). The species richness pattern based on ExpRMs was most congruent with that based on MCVs (Spearman's rho = 0.91). Spatial comparisons against ExpRMs highlighted potential overestimations of species richness based on MCVs and EcoRMs in the southern central US and northern central Mexico. The species richness pattern based on SDMs was most congruent with that based on simplified EcoRMs (Spearman's rho = 0.94), with potential overestimations of the former mainly in coastal areas of south-west US and Florida. ExpRMs and SDMs, both of which were used for calculating the performance measures interchangeably, were themselves highly congruent (0.89) but yielded different species richness patterns in south-west US, Florida, and northern Mexico. At a grain size of 100 km, range size estimates of MCVs, original EcoRMs and simplified EcoRMs were similarly congruent with those based on ExpRMs (r = 0.88, 0.88, 0.88; all p-values < 0.001) and SDMs (r = 0.94, 0.94, 0.94; Figure S2).

DISCUSSION

Our results show that Minimum Convex Hulls (MCVs) and ecoregional range maps (EcoRMs) perform similarly well in describing species distributions based on expert range maps (ExpRMs) or species distribution models (SDMs). EcoRMs consistently showed greater sensitivity and MCVs greater precision. The mean sensitivity and precision of both range estimates was similar across all species, but EcoRMs performed better for data-poor species. The species richness pattern based on ExpRMs was most congruent with that from MCVs, whereas the species richness pattern based on SDMs was most congruent with that from simplified EcoRMs. Our results suggest that EcoRMs hold the promise to provide accurate and broadly available baseline range information for species, particularly data-poor ones, across a broad spectrum of taxa.

MCVs as range surrogates

For data-poor species and in the face of strong geographical biases in occurrence records, ExpRMs are likely the most accurate surrogate of species ranges. However, given the immense workload to produce ExpRMs, the limited number of experts available to inform them, and the sheer number of species on Earth, there is an increasing interest in automating and facilitating their production. Specifically, several recent studies suggested the use of MCVs or similar approaches as surrogates for species' ranges (Dauby *et al.*, 2017; Huang *et al.*, 2021; Palacio *et al.*, 2021). Nevertheless, to our knowledge, their performance has not been evaluated before. The herein presented analyses of the species-level sensitivities and precision as well as the concordance of MCVs with ExpRMs and SDMs provide the first empirical evaluation of this putative range surrogate. Our results show that MCVs are generally appropriate surrogates of species distribution and particularly useful for delimiting the most suitable part of a species range as indicated by their high precision. However, MCVs are particularly susceptible to sampling bias and may merge large parts of disjunct ranges resulting in a low sensitivity for detecting true occurrences (here either presences from ExpRMs or SDMs). In addition, we demonstrate that they are generally less appropriate for data-poor species (i.e., those with <100 records), which highlights an important yet rarely considered limitation of MCVs. Given that the minimal data requirements of MCVs are similar to those of SDMs, which outperform the former, our results emphasize that MCVs provide a suboptimal surrogate for the vast majority of species (Figure 3).

Ecoregional ranges

Ecoregions represent generalized expert knowledge that is commonly used for assemblage-level analyses, but their appropriateness as a substitute of other range map types such as SDMs or ExpRMs as well as their potential surrogacy for species level-distributions remains unevaluated. We assessed their surrogacy for species ranges and found that ecoregional ranges provide reliable estimates of species ranges as well as diversity patterns. Their additional technical advantages (e.g., reproducibility and actuality) suggest a high potential for EcoRMs to provide reasonable range estimates for data-poor species as well as for improving the quality of other range estimates for well-documented species. Ecoregional information is currently mainly used for biodiversity and conservation research at large geographical and taxonomic scales (e.g., Fritz*et al.*, 2009; Dinerstein *et al.*, 2017). Here we show that EcoRMs yield diversity patterns comparable to those based on commonly used surrogates of species' ranges, including SDMs and ExpRMs (Figure 5). Our results thereby, for the first time, confirm the appropriateness of using ecoregion-based diversity patterns in large-scale studies and underline the general importance of the geoecological classification on which they are based (Olson *et al.*, 2001). Moreover, the high performance of EcoRMs for species with less than 100 occurrence records highlights their potential for incorporating data-poor species into large-scale analyses of diversity patterns (e.g., hotspot analyses and protected area coverage) to boost taxonomic representation. Acknowledging their limitations at finer scales, EcoRMs should be used for regions and taxa where ExpRMs are outdated or not available, or where primary occurrence data is limited or of poor quality.

We compared the sensitivity and precision of EcoRMs and MCVs using ExpRMs and SDMs as true ranges. Since, EcoRMs represent a limited number of rather coarse regions, whereas MCVs are exclusively based on a narrow definition of spatial proximity, we expected that EcoRMs have a greater sensitivity but a lower precision than MCVs. Our performance evaluations indicate that both original and simplified EcoRMs have a remarkably high sensitivity of detecting species' true distributions (Table S1). The precision of original EcoRMs was consistently lower than for ExpRMs, SDMs or MCVs. For a minor loss in sensitivity, spatial simplifications (i.e., removing very small fragments and smoothing the outline) of EcoRMs resulted in a disproportionate gain in precision. Except for SDMs, the simplified EcoRMs consistently had a greater performance than other range estimate for species with less than 100 records. Analyzing the scale dependence of range estimates, we confirmed that 200 km is the appropriate resolution of ExpRMs (Hurlbert & Jetz, 2007), but that of simplified EcoRMs was twice as fine. Our analyses thereby provide strong support for species-level applications of ecoregional ranges, particularly for data-poor regions and taxa.

Applications

First, we argue that EcoRMs can be used to close an important knowledge gap for data-poor species. Our results suggest that even for very data-poor species (<5 records), EcoRMs provide relatively accurate range

maps. Assessed against the global availability of occurrence records for butterflies, ExpRMs would allow incorporation of 17% of all known species into conservation and biodiversity research that have previously not been accounted for [data from Pinkert et al. (2022)]. Many of these species are rare and therefore a particular focus of conservation action (Lamoreux *et al.*, 2006) because of their high risk of extinction (Courchamp *et al.*, 2006).

Second, similarly to checklists for political or administrative units, assignment of species to ecoregions will facilitate the incorporation of older (less spatially accurate) distributional data from the literature as well as data from inventories at a coarse, yet geoecologically meaningful, grain. This checklist work would be facilitated by tools such as the *ntbox* (Osorio-Olvera *et al.*, 2020), that allow users to overlay all available data, add species and regional information, and modify the synthesis range if needed. Integrative species ranges may, in turn, benefit initiatives such as the NatureServe Canada's EBAR project that aims to archive metadata on the information and decisions underlying species range maps and collect expert reviews for the final range product (natureserve.org). An additional advantage of these range maps that stems from their reproducibility is the possibility to produce them for different periods of time, based on which range shifts, range contractions/extensions could be tracked (Araújo *et al.*, 2002).

Third, EcoRMs could be routinely used as masks for SDMs to better delimitate dispersal barriers. SDMs are uniquely useful to resolve internal structures in data-rich species distributions (Hurlbert & White, 2005; Rondinini *et al.*, 2006; Herkt *et al.*, 2017). The geoecologically delimitation of EcoRMs and their high sensitivity make them ideal masks for SDMs (note the outliers in Florida and the Caribbean in Figure 2). Integrative models combining SDMs with range estimates such as the approach presented by Merow et al. (2017) may directly include EcoRM offsets to better define sampling regions for pseudo-absences and to evaluate the appropriateness of the range offset with different distance decay parameters. Other non-parametric attempts to resolve the internal structure of range maps using presence/absence data, elevational ranges, and measures of spatial proximity (e.g., Huang *et al.*, 2021, Palacio *et al.*, 2021) ultimately face the same limitations — of highly inaccurate and spatially biased occurrence data — as SDMs and even stronger limitations of data availability due to a lack of absence information. The advantages of EcoRMs, particularly for data-poor species, suggest that the integration of ecoregional offsets into SDMs would likely improve models for 26% of all butterfly species, for example [i.e. species with [?]5 but <100 records; data from Pinkert et al. (2022)].

Limitations and extensions

Here, we chose a set of well-documented species from the US and Canada that had sufficient data for species distribution modeling and available expert range maps. North America is, however, classified into rather large ecoregions and it includes relatively weak geographical barriers for dispersal (e.g., Pinkert *et al.*, 2016; Stelbrink *et al.*, 2018). In the light of these limitations, we suggest two future avenues of technical evaluations. Firstly, analyses for tropical regions and widespread species may provide important insights into spatial variation of the performance of EcoRMs and its dependence on the range size of species. The performance of EcoRMs should be even higher in these regions and species, because ecoregions are more fine-scaled in the tropics and advantages such the adequate representation of disjunct ranges and delineation of biogeographical are likely most important in global-scale analyses. Secondly, the framework presented in this study may be applied to different terrestrial taxa as well as freshwater and marine ecoregions to better understand the general surrogacy of ecoregions. In the same vein, we encourage the integration of data on species turn-over and expert knowledge of a broader range of taxa into the continuously developed ecoregional database to further improve and resolve the baseline layers (Olson *et al.*, 2001; Dinerstein *et al.*, 2017).

For demonstration purposes, we in this study, focused on ecoregions developed by Olson et al. (2001) and refined by Dinerstein et al. (2017), both because they are the most frequently used type of regionalization in conservation and biodiversity research and because they exclusively rely on similarities of species communities rather than environmental information. The latter feature of ecoregions is particularly relevant for integrating EcoRMs into SDMs, as a direct dependence on environmental data would introduce circularity for niche

estimation. However, our results also represent a proof of concept for the application of a wider range of regionalizations developed using remote sensing and environmental data, such as global layers of ecological land units (e.g., Sayre *et al.*, 2014) and classifications of mountain regions (e.g., Snethlage *et al.*, 2022), for cases where circularity with niche estimation is not relevant. We acknowledge that, for instance, of ecosystem land units are already available at a resolution 5 times finer than that of ecoregions, they are readily updateable and technically allow for temporally continuous time-series data, whereby they would not only improve the availability of range data, but also data of improved spatial and temporal resolution.

CONCLUSIONS

The main goal of this study was to assess the performance of surrogates for species distributions to facilitate improvements in building a reliable information basis for area-based conservation, threat assessments and biodiversity research. Although ExpRMs are still commonly used in conservation research and SDMs are increasingly used in large-scale biodiversity research, the availability of both range map types is critically limited to well-documented taxa (Jeliazkov *et al.*, 2022). An important yet seldom addressed information gap are the rare and poorly documented species. Any progress in closing this gap will disproportionately reduce regional and taxonomical biases in large-scale analyses. Our findings suggest that EcoRMs are of high potential for both conservation and biodiversity research, including but not limited to applications for evaluating species' threat and diversity. EcoRMs outperformed traditional surrogates for data-poor species and are uniquely applicable to very data-poor species. In addition, the exceptionally high sensitivities of ecoregions promise improvements for well documented species through their integration as masks or offsets in SDMs (e.g., Merow *et al.,* 2017). The broader implications of our findings are that ecoregional information provides a versatile tool with an immense potential to boost the taxonomic coverage in ecology and conservation.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

CODE AVAILABILITY

The code for generating ecoregional range maps is archived and publicly available at mol.org (DOI: https://doi.org/10.48600/711m-x166).

DATA AVAILABILITY STATEMENT

All data supporting our results is archived and publicly available at mol.org (DOI: https://doi.org/10.48600/711m-x166).

REFERENCES

Amatulli, G., Domisch, S., Tuanmu, M.-N., Parmentier, B., Ranipeta, A., Malczyk, J. & Jetz, W. (2018) A suite of global, cross-scale topographic variables for environmental and biodiversity modeling. *Scientific Data*, **5**, 180040.

Araujo, M.B., Williams, P.H. & Fuller, R.J. (2002) Dynamics of extinction and the selection of nature reserves. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, **269**, 1971–1980.

Bivand, R., Keitt, T., Rowlingson, B., Pebesma, E., Sumner, M., Hijmans, R.J. & Rouault, E. (2018) *rgdal: Bindings for the Geospatial Data Abstraction Library*, R-package version 1.2-20.

Brown, J.H., Stevens, G.C. & Kaufman, D.M. (1996) The geography of range size: Size, shape, boundaries, and internal structure. *Annual Review of Ecology and Systematics*, **27**, 597–623.

Calenge, C. & Fortmann-Roe, S. (2021) *adehabitatHR: Home Range Estimation*, R-package version 0.4.19.

Courchamp, F., Angulo, E., Rivalan, P., Hall, R.J., Signoret, L., Bull, L. & Meinard, Y. (2006) Rarity value and species extinction: The anthropogenic allee effect. *PLOS Biology*, **4**, e415.

Dauby, G., Stevart, T., Droissart, V., Cosiaux, A., Deblauwe, V., Simo-Droissart, M. *et al.* (2017) ConR: An R package to assist large-scale multispecies preliminary conservation assessments using distribution data. *Ecology and Evolution*, **7**, 11292–11303.

Dinerstein, E., Joshi, A.R., Vynne, C., Lee, A.T.L., Pharand-Deschenes, F., Franca, M. *et al.* (2020) A "Global Safety Net" to reverse biodiversity loss and stabilize Earth's climate. *Science Advances*, **6**, eabb2824.

Dinerstein, E., Olson, D., Joshi, A., Vynne, C., Burgess, N.D., Wikramanayake, E., Hahn, N. *et al.* (2017) An ecoregion-based approach to protecting half the terrestrial realm. *BioScience*, **67**, 534–545.

Elith, J., & J. R. Leathwick (2009). The contribution of species distribution modelling to conservation prioritization. P. 70–93 in A. Moilanen, K. A. Wilson, & H. Possingham, eds. *Spatial conservation prioritization: quantitative methods* . Oxford University Press, New York, USA; Oxford, UK.

Fritz, S.A., Bininda-Emonds, O.R.P. & Purvis, A. (2009) Geographical variation in predictors of mammalian extinction risk: Big is bad, but only in the tropics. *Ecology Letters*, **12**, 538–549.

Glassberg, J. (2017)*A Swift guide to the butterflies of Mexico and Central America.* 2nd ed., Princeton University Press, NJ, USA.

Herkt, K.M.B., Skidmore, A.K. & Fahr, J. (2017) Macroecological conclusions based on IUCN expert maps: A call for caution. *Global Ecology and Biogeography*, **26**, 930–941.

Hijmans, R.J., van Etten, J., Sumner, M., Cheng, J., Baston, D., Bevan, A. *et al.*(2022) *raster: Geographic Data Analysis and Modeling*, R-package version 3.5-15.

Hijmans, R.J., Phillips, S., Leathwick, J. & Elith, J. (2021) *dismo: Species Distribution Modeling*, R-package version 1.3-5.

Huang, R.M., Medina, W., Brooks, T.M., Butchart, S.H.M., Fitzpatrick, J.W., Hermes, C. *et al.* (2021) Batch-produced, GIS-informed range maps for birds based on provenanced, crowd-sourced data inform conservation assessments. *PLOS ONE*, **16**, e0259299.

Hurlbert, A.H. & Jetz, W. (2007) Species richness, hotspots, and the scale dependence of range maps in ecology and conservation. *Proceedings of the National Academy of Sciences USA*, **104**, 13384–13389.

Hurlbert, A.H. & White, E.P. (2005) Disparity between range map- and survey-based analyses of species richness: patterns, processes and implications.*Ecology Letters*, **8**, 319–327.

Jeliazkov, A., Gavish, Y., Marsh, C.J., Geschke, J., Brummitt, N., Rocchini, D. *et al.*(2022) Sampling and modelling rare species: Conceptual guidelines for the neglected majority. *Global Change Biology*, **28**, 3754–3777.

Jetz, W., McGeoch, M.A., Guralnick, R., Ferrier, S., Beck, J., Costello, M.J., Fernandez, M. *et al.* (2019) Essential biodiversity variables for mapping and monitoring species populations. *Nature Ecology & Evolution*, **3**, 539–551.

Jetz, W., McGowan, J., Rinnan, D.S., Possingham, H.P., Visconti, P., O'Donnell, B. & Londono-Murcia, M.C. (2022) Include biodiversity representation indicators in area-based conservation targets.*Nature Ecology & Evolution*, **6** , 123–126.

Jung, M., Arnell, A., de Lamo, X., Garcia-Rangel, S., Lewis, M., Mark, J. *et al.* (2021) Areas of global importance for conserving terrestrial biodiversity, carbon and water. *Nature Ecology & Evolution*, **5**, 1499–1509.

Karger, D.N., Conrad, O., Bohner, J., Kawohl, T., Kreft, H., Soria-Auza, R.W. *et al.*(2018) Data from: Climatologies at high resolution for the earth's land surface areas - Dryad.

Karger, D.N., Conrad, O., Bohner, J., Kawohl, T., Kreft, H., Soria-Auza, R.W. *et al.*(2017) Climatologies at high resolution for the earth's land surface areas. *Scientific Data*, **4**, 170122.

Lamoreux, J.F., Morrison, J.C., Ricketts, T.H., Olson, D.M., Dinerstein, E., McKnight, M.W. & Shugart, H.H. (2006) Global tests of biodiversity concordance and the importance of endemism. *Nature*, **440**, 212–214.

Marsh, C.J., Sica, Y.V., Burgin, C.J., Dorman, W.A., Anderson, R.C., del Toro Mijares, I., Vigneron, J.G., et al. (2022) Expert range maps of global mammal distributions harmonised to three taxonomic authorities. *Journal of Biogeography*, **49**, 979-992.

Meyer, C., Kreft, H., Guralnick, R. & Jetz, W. (2015) Global priorities for an effective information basis of biodiversity distributions. *Nature Communications*, **6**, 1–8.

Oliver, R.Y., Meyer, C., Ranipeta, A., Winner, K. & Jetz, W. (2021) Global and national trends, gaps, and opportunities in documenting and monitoring species distributions. *PLOS Biology*, **19**, e3001336.

Olson, D.M., Dinerstein, E., Wikramanayake, E.D., Burgess, N.D., Powell, G.V.N., Underwood, E.C., D'amico, 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.

Orr, M.C., Hughes, A.C., Chesters, D., Pickering, J., Zhu, C.-D. & Ascher, J.S. (2021) Global Patterns and Drivers of Bee Distribution. *Current Biology*, **31**, 451-458.e4.

Osorio-Olvera, L., Lira-Noriega, A., Soberon, J., Peterson, A.T., Falconi, M., Contreras-Diaz, R.G. *et al.* (2020) ntbox: An r package with graphical user interface for modelling and evaluating multidimensional ecological niches. *Methods in Ecology and Evolution*, **11**, 1199–1206.

Palacio, R.D., Negret, P.J., Velasquez-Tibata, J. & Jacobson, A.P. (2021) A data-driven geospatial workflow to map species distributions for conservation assessments. *Diversity and Distributions*, **27**, 2559–2570.

Pebesma, E., Bivand, R., Racine, E., Sumner, M., Cook, I., Keitt, T., Lovelace, R., Wickham, H., Ooms, J., Muller, K., Pedersen, T.L., Baston, D. & Dunninton, D. (2022a) *sf: Simple Features for R*, R-package version 1.0-7.

Pebesma, E., Bivand, R., Rowlingson, B., Gomez-Rubio, V., Hijmans, R., Sumner, M. *et al.* (2022b) *sp: Classes and Methods for Spatial Data*, R-package version 1.4-6.

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

Pinkert, S., Barve, V., Guralnick, R. & Jetz, W. (2022) Global geographical and latitudinal variation in butterfly species richness captured through a comprehensive country-level occurrence database. *Global Ecology and Biogeography*, **31**, 830–839.

Proosdij, A.S.J., Sosef, M.S.M., Wieringa, J.J. & Raes, N. (2016) Minimum required number of specimen records to develop accurate species distribution models. *Ecography*, **39** , 542–552.

Rondinini, C., Wilson, K.A., Boitani, L., Grantham, H. & Possingham, H.P. (2006) Tradeoffs of different types of species occurrence data for use in systematic conservation planning. *Ecology Letters*, **9**, 1136–1145.

Saura, S., Bastin, L., Battistella, L., Mandrici, A. & Dubois, G. (2017) Protected areas in the world's ecoregions: How well connected are they? *Ecological Indicators*, **76**, 144–158.

Sayre, R., Karagulle, D., Frye, C., Boucher, T., Wolff, N.H., Breyer, S.*et al.* (2020) An assessment of the representation of ecosystems in global protected areas using new maps of World Climate Regions and World Ecosystems. *Global Ecology and Conservation*, **21** , e00860.

Schatz, G.E. (2002) Taxonomy and Herbaria in Service of Plant Conservation: Lessons from Madagascar's Endemic Families. *Annals of the Missouri Botanical Garden*, **89**, 145–152.

Scott, J.A. (1997) *The Butterflies of North America: A Natural History and Field Guide*, Stanford University Press, CA, USA.

Snethlage, M.A., Geschke, J., Ranipeta, A., Jetz, W., Yoccoz, N.G., Korner, C. *et al.* (2022) A hierarchical inventory of the world's mountains for global comparative mountain science. *Scientific Data*, **9**, 149.

Sofaer, H. R., J. A. Hoeting, & C. S. Jarnevich (2019) The area under the precision-recall curve as a performance metric for rare binary events. *Methods in Ecology and Evolution*, **10** , 565–577.

Tuanmu, M.-N. & Jetz, W. (2015) A global, remote sensing-based characterization of terrestrial habitat heterogeneity for biodiversity and ecosystem modelling. *Global Ecology and Biogeography*, **24**, 1329–1339.

Wilson, A.M. & Jetz, W. (2016) Remotely Sensed High-Resolution Global Cloud Dynamics for Predicting Ecosystem and Biodiversity Distributions. *PLOS Biology*, **14**, e1002415.

Zizka, A., Silvestro, D., Andermann, T., Azevedo, J., Ritter, C.D., Edler, D. *et al.* (2021) CoordinateCleaner: Automated Cleaning of Occurrence Records from Biological Collections, R-package version 2.0-20.

ACKNOWLEDGEMENTS

We thank Kalkidan Fekadu Chefira and John Wilshire for their help with the data integration and visualization in Map of Life. S.P. acknowledges support from the Alexander von Humboldt Foundation. W.J. and S.P. are grateful for support from the E.O. Wilson Biodiversity Foundation and its Half-Earth Project. This research was partly funded by the Gordon and Betty Moore Foundation through Grant GBMF8137 to the E.O. Wilson Biodiversity Foundation to support the work of W.J. and S.P. W.J. acknowledges support from National Science Foundation grant DEB-1541500.

AUTHOR CONTRIBUTIONS

S.P. and W.J. conceived the idea of this study. Y.S. and K.W. helped with cleaning occurrence data and mobilizing range map data as well as contributed to methodological discussions. S.P. processed and analyzed all data as well as wrote the first draft of the manuscript. All authors contributed to revisions of the manuscript.

BIOSKETCH

We are a group of researchers broadly interested in approaches that leverage the complementarity of multiple types of biodiversity data with the aim of improving biodiversity conservation as well as our understanding of the ecological and evolutionary processes that underpin species' distributions.

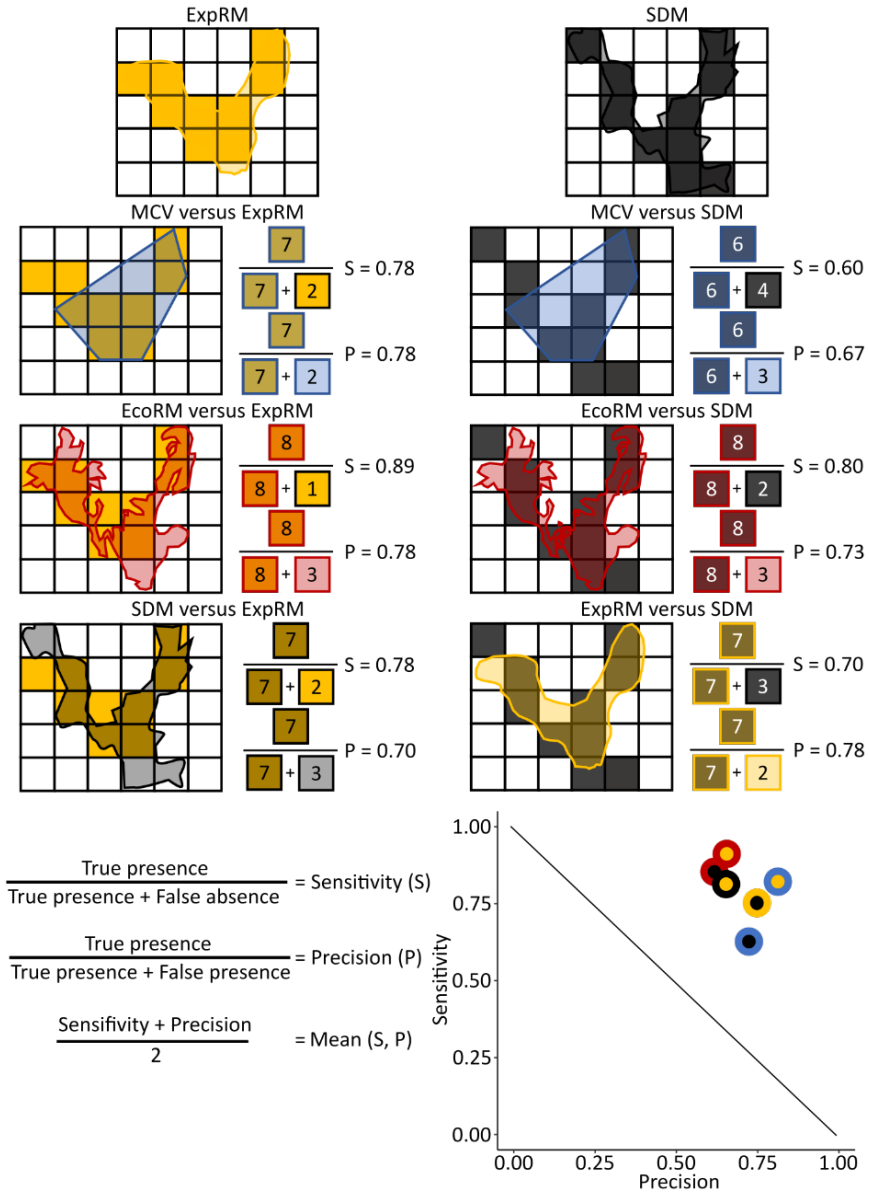

Figure 1 | Performance metrics of different range map types assessed against an expert range map (ExpRM) and a species distribution model (SDM) of a hypothetical species at coarse grain. Orange cells in the left panel (ExpRM) and black cells in the right panel (SDM) indicate the alternative true presences and white cells are true absences. Focal data types are the minimum convex hull (MCV) and ecoregional range map (EcoRM). Sensitivity (true positive rate) is defined as the proportion of presences correctly identified, precision (positive predictive value) is the proportion of positive predictions that are correct, and the overall performance the arithmetic mean of these metrics - all with respect to the 'truth' provided by the validation dataset. Cells are considered a true presence when a range map covers more than 50% their area. The scatterplot in the lower right corner provides a graphical representation of the performance statistics above. Inner circles of points indicate the basis of comparison (orange = ExpRM or black = SDM) and rings the respective range map type. In this example, the EcoRMs has high sensitivity and would be preferrable for applications wishing to minimize false absences while still offering good precision. In turn, MCVs or ExpRMs may be preferred for applications wishing to maximize precision (e.g. species conservation).

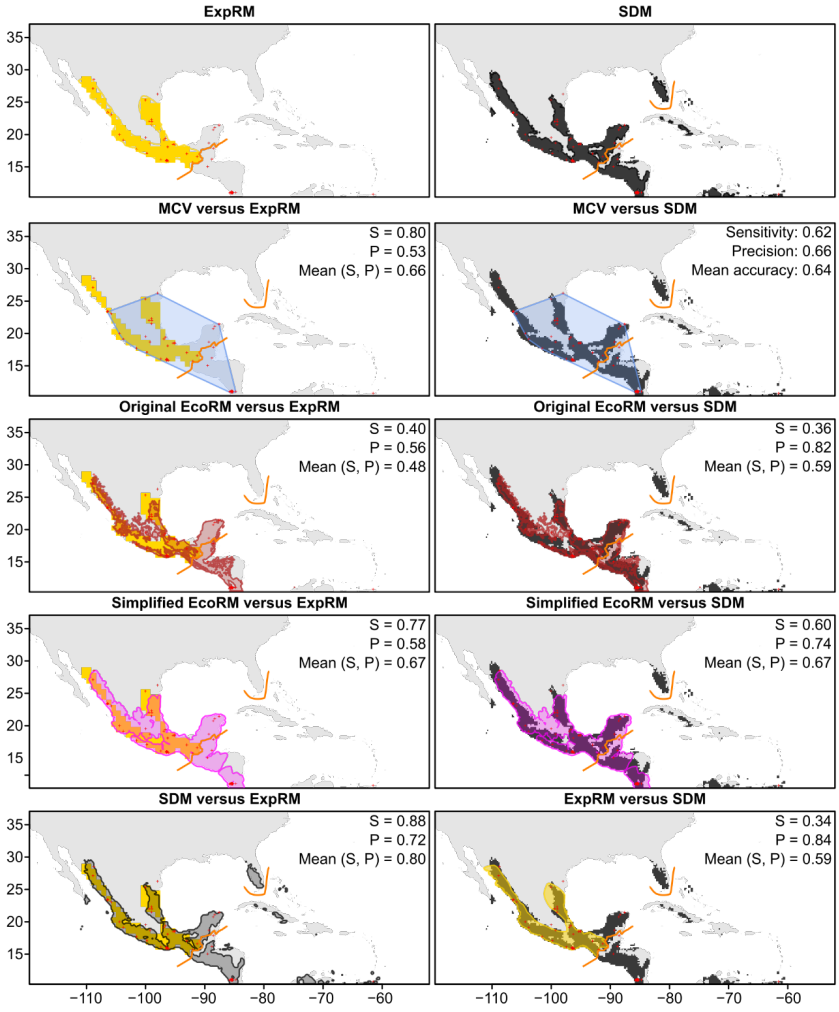

Figure 2 | Range map types of an example species (*Typhedanus undulatus*, 99 records). ExpRMs and SDMs (top row) are used to calculate sensitivity (S), precision (P) and their mean (see Figure 1). All data were resampled to a grid of cells with a grain of approximately 100 km. Note that most range map types extend to central America; for the final analyses all data was cropped to the boundaries of continental North America (Canada, US, Mexico; below orange line) because ExpRMs are limited to this extent.

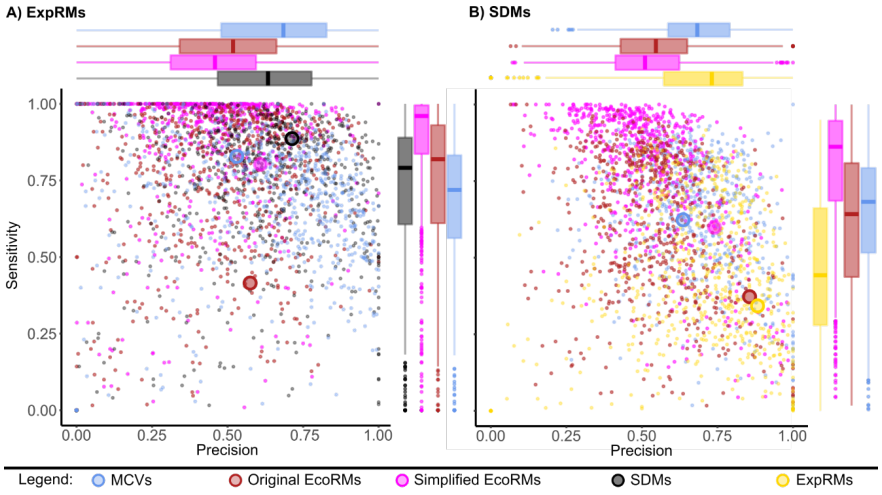

Figure 3 | Scatterplots and boxplots of the sensitivity and precision of the five range map types for 624 North American butterfly species. Performance measures were calculated based on presence/absence information from **A)** ExpRMs and **B)** SDMs. Ranges were analyzed at a grain of approximately 100 km. Larger points highlight performance statistics for *Typhedanus undulatus* shown in Figure 2.

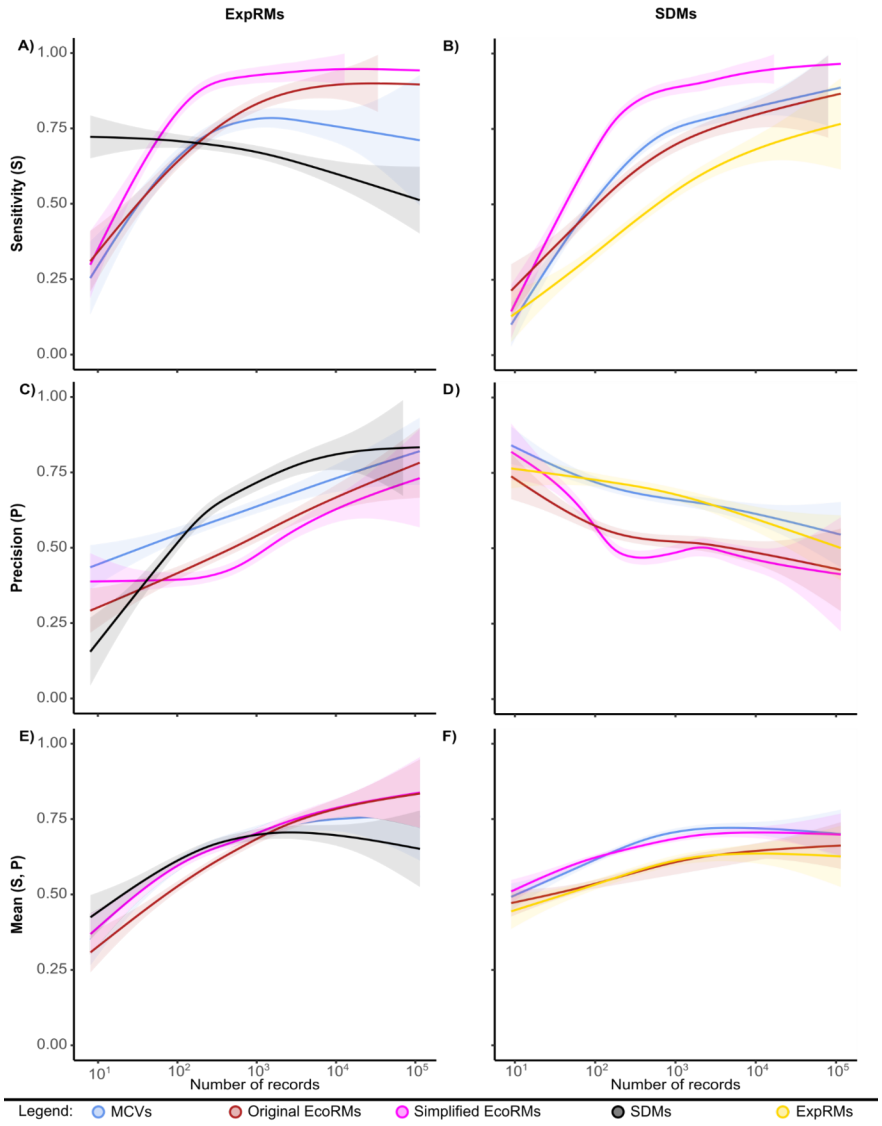

Figure 4 | Sensitivity, precision, and mean performance of the five range map types for 624 North American butterfly species related to the number of cleaned occurrence records available. Performance measures were calculated based on presence/absence information from **A,C,E)** ExpRMs and **B,D,F)** SDMs. Lines are spline-based smoothed regressions across species points (not shown) and semi-transparent areas indicate the 95% confidence interval of these regressions. For calculations and all other information see Figure 3.

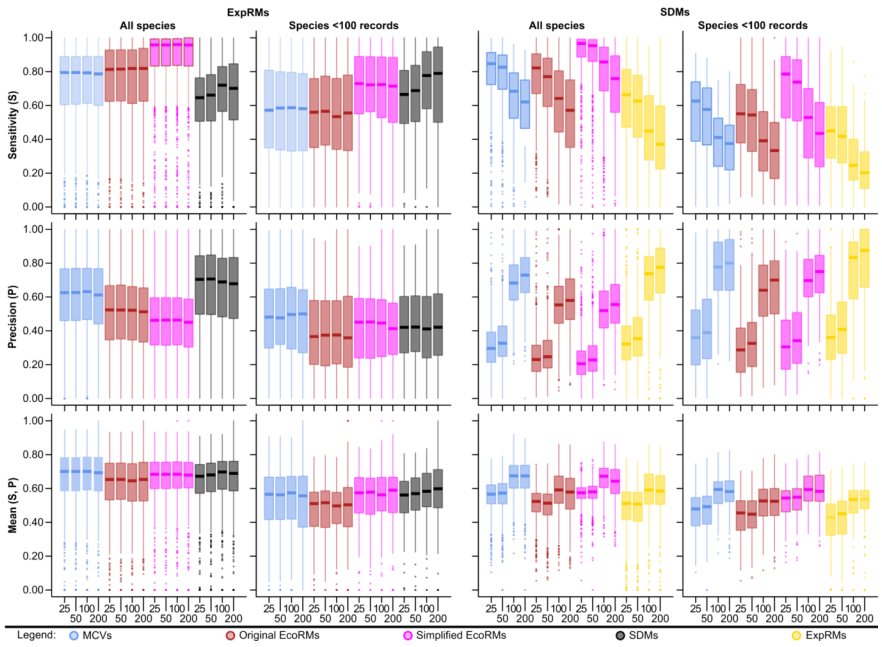

Figure 5 | Scale-dependence of the sensitivity, precision, and mean performance of the five range map types for 624 North American butterfly species. Performance measures were calculated based on presence/absence information from ExpRMs and SDMs at four different grain sizes (x-axis in km) both for all species and for species with fewer than 100 occurrence records (127 species).

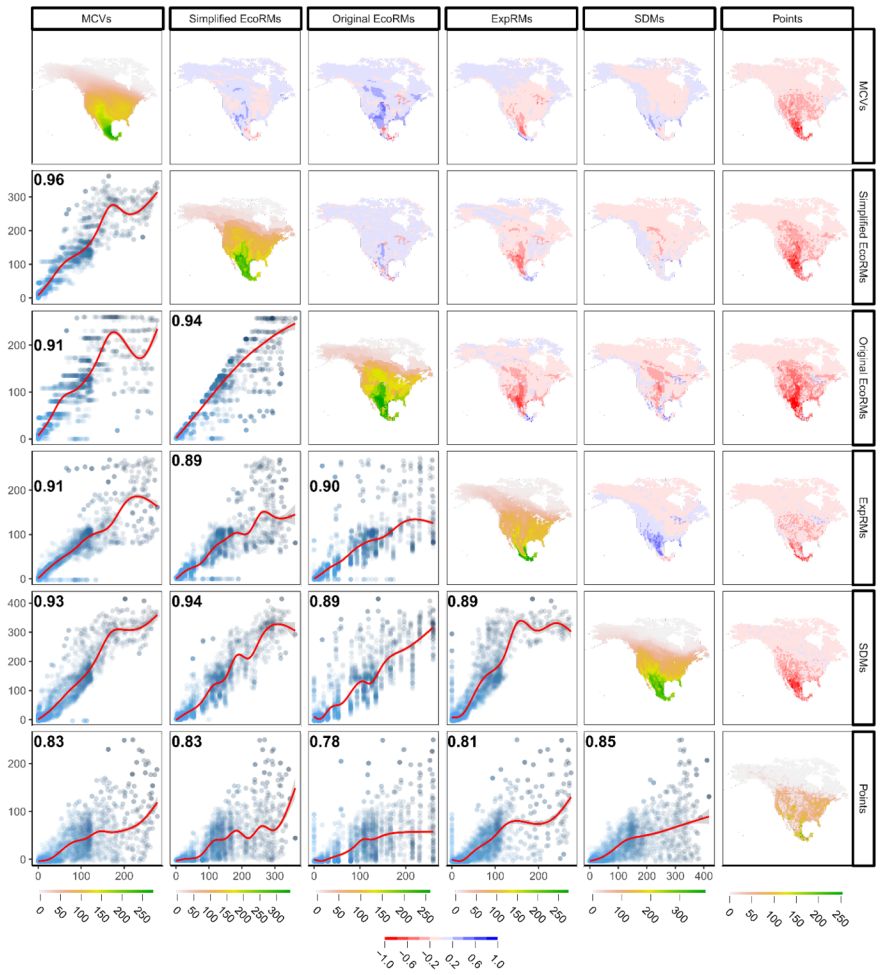

Figure 6 | Comparisons of different types of distributional data for 624 North American butterfly species. Scatterplots in the lower left triangle show the relationships of species richness estimates based on ExpRMs, MCVs, simplified EcoRMs, original EcoRMs, SDMs and the occurrence records based on which the latter four range map types were calculated. Red lines are spline-based smoothed regressions and point color indicates the point density (light blue = low, dark blue = high). Values above scatterplots are Spearman rank coefficients calculated for pairs. Maps in the diagonal show species richness patterns and those in the upper triangle contrasts of scaled species richness patterns of pairs. All data was resampled to a grid of cells with an approximate grain size of 100 km.

- SUPPORTING INFORMATION -

The potential of ecoregional range maps for boosting taxonomic coverage in large-scale ecology and conservation

Table S1 | Scale-dependence of performance metrics of different range map types for 624 North American butterfly species. Median sensitivity and precision were calculated across different grain sizes for the three focal range types (MCVs, original and simplified EcoRMs) based on available occurrence records. Values highlighted in bold indicate the data type with the highest sensitivity (S)/ precision (P)/ mean performance (mean sensitivity and precision) per group (i.e., per grain size nested in subset and basis of comparison) and shaded cells highlight the data type with the highest mean performance across grains per subset. For metric definitions see Figure 1.

			ExpRMs	**ExpRMs**	**ExpRMs**	**SDMs**	**SDMs**	**SDMs**
Subset	**Grain**	**Type**	**S**	**P**	**Mean (S, P)**	**S**	**P**	**Mean (S, P)**
All	25	MCVs	0.78	0.63	**0.71**	0.85	0.30	0.58
		Simplified EcoRMs	**0.96**	0.46	**0.71**	**0.97**	0.20	**0.59**
		Original EcoRMs	0.81	0.52	0.67	0.83	0.23	0.53
		ExpRMs				0.67	**0.31**	0.49
		SDMs	0.64	**0.71**	0.67			
	50	MCVs	0.78	0.63	**0.71**	0.83	0.33	0.58
		Simplified EcoRMs	**0.96**	0.46	**0.71**	**0.96**	0.22	**0.59**
		Original EcoRMs	0.81	0.52	0.66	0.78	0.24	0.51
		ExpRMs				0.63	**0.35**	0.49
		SDMs	0.65	**0.71**	0.68			
	100	MCVs	0.78	0.64	**0.71**	0.68	0.69	0.68
		Simplified EcoRMs	**0.96**	0.45	**0.71**	**0.86**	0.51	**0.69**
		Original EcoRMs	0.81	0.51	0.66	0.64	0.55	0.60
		ExpRMs				0.44	**0.73**	0.59
		SDMs	0.71	**0.69**	0.70			
	200	MCVs	0.77	0.60	0.69	0.62	0.73	**0.68**
		Simplified EcoRMs	**0.96**	0.44	**0.70**	**0.77**	0.55	0.66
		Original EcoRMs	0.82	0.51	0.66	0.58	0.57	0.57
		ExpRMs				0.37	**0.78**	0.57
		SDMs	0.70	**0.68**	0.69			
<100	25	MCVs	0.54	**0.49**	0.51	0.63	**0.37**	0.50
		Simplified EcoRMs	**0.75**	0.42	**0.58**	**0.79**	0.31	**0.55**
		Original EcoRMs	0.56	0.36	0.46	0.55	0.31	0.43
		ExpRMs				0.45	0.36	0.41
		SDMs	0.65	0.40	0.53			
	50	MCVs	0.54	**0.49**	0.51	0.58	0.40	0.49
		Simplified EcoRMs	**0.74**	0.41	**0.58**	**0.75**	0.35	**0.55**
		Original EcoRMs	0.58	0.36	0.47	0.55	0.33	0.44
		ExpRMs				0.40	**0.41**	0.41
		SDMs	0.67	0.41	0.54			
	100	MCVs	0.57	**0.50**	0.53	0.42	0.79	0.60
		Simplified EcoRMs	0.75	0.41	0.58	**0.54**	0.70	**0.62**
		Original EcoRMs	0.54	0.37	0.45	0.40	0.65	0.52
		ExpRMs				0.24	**0.83**	0.53
		SDMs	**0.78**	0.40	**0.59**			
	200	MCVs	0.56	0.50	0.53	0.39	0.80	**0.60**
		Simplified EcoRMs	0.73	0.41	0.57	0.44	0.76	**0.60**
		Original EcoRMs	0.56	0.35	0.45	0.35	0.70	0.52
		ExpRMs				0.21	**0.89**	0.55
		SDMs	**0.79**	**0.42**	**0.60**	**0.85**	0.30	0.58

Abbreviations: Subset = threshold of the number of records per species used, this is no threshold (All) or species with fewer than 100 records only (<100; n = 127 species); Gain = grain size in km. MCVs = Minimum convex hulls, EcoRMs = Ecoregional range maps, ExpRMs = Expert range maps, SDMs = Species distribution models.

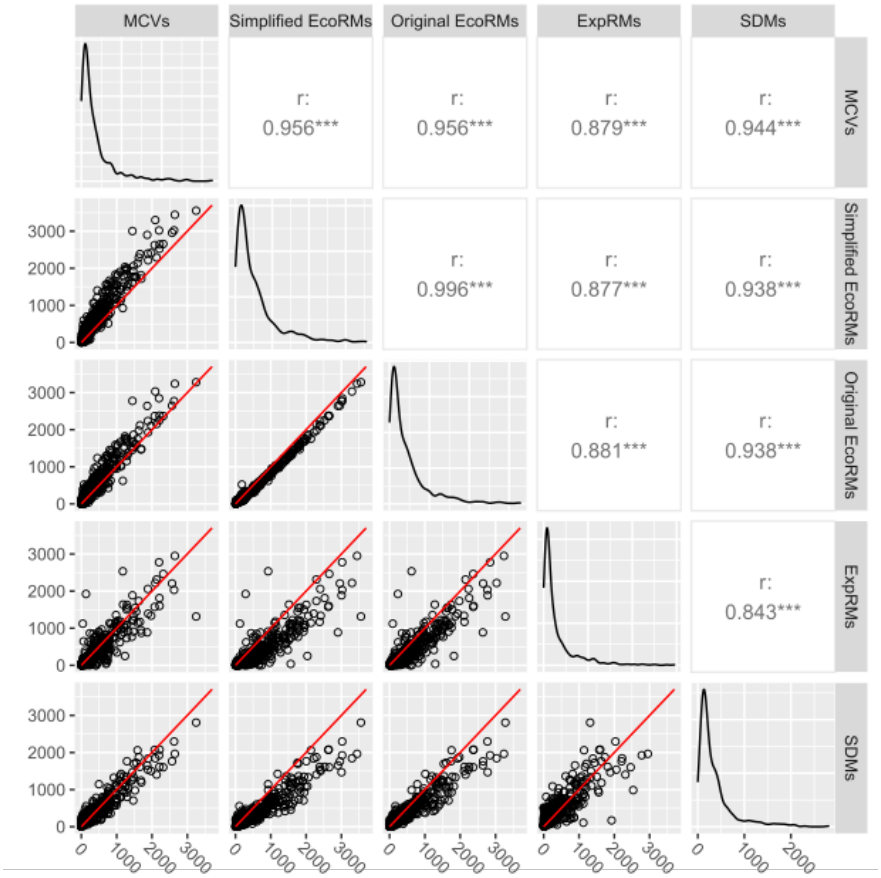

Figure S1 | Comparisons of different types of range maps for 624 North American butterfly species. Scatterplots in the lower left triangle show the relationships of species' range size estimates based on ExpRMs, MCVs, simplified EcoRMs, original EcoRMs and SDMs. 1:1 lines are indicated in red. The diagonal shows histograms of range size estimates based on the different types of range maps. Values in the upper triangle are Pearson correlation coefficients calculated for pairs. All data was resampled to a grid of cells with an approximate grain size of 100 km.