Supplementary Material for:

Measuring metrics: what diversity indicators are most appropriate for different forms of data bias?

Huijie Qiao, Michael C. Orr, & Alice C. Hughes*

Detailed supplements have been developed to accompany all analysis and provide greater insights and definitions for the methods. Each section of supplemental methods and supplemental results is aligned with that in the main text. Below we detail all background information, and detail where it aligns with the main text to enhance the usability and aid interpretation and replicability.

Table of Contents

Measuring metrics: what diversity indicators are most appropriate for different forms of data bias? 1

Supplemental text 2

I- Supplemental methods 2

S1. Species checklist per country (in relation to BirdLife data)- First step 2

S2. Assessing biases in species occurrence data in relation to roads and urban areas for each country -Second step 3

S3. Mapping BirdLife –based richness patterns- Third step 3

S4. eBird data and species richness maps- Forth step 3

S5. Simulating virtual species ranges- Fifth step 4

S6. Applying the diversity metrics to the different scenarios- Sixth step 10

S7. Mapping diversity indices- Seventh step 10

II-Definitions and functions of each index assessed 12

III-Supplemental results 15

Mapping diversity with different metrics 15

Dealing with small volumes of data 15

Mapping diversity 16

Regional drivers of diversity based on distribution modelling 16

Dealing with small volumes of data (retaining sampled and unsampled cells) 17

IV-Supplemental Result Figures 19

1). Parameterising bias 19

a). Spatial completeness figures 19

b). Spatial biases figures 26

2). Understanding how indices overcome biases 39
c). Mapping diversity with different metrics figures and table 39
d). Dealing with small volumes of data figures. 42
3). Mapping and Modelling diversity 46
e). Mapping diversity figures. 46
4). Supplemental Analysis 51
f). Regional assessments and driver analysis figures. 51
g). Dealing with small volumes of data (retaining sampled and unsampled cells) figures. 55

Supplemental text

I- Supplemental methods

Overview

We performed all the analyses in R v4.2 (R Core Team, 2022) and ArcGIS 10.8. More details about the function implementations in R script can be found on < https://github.com/qiaohj/ES_50>. The overall analysis frameworks are shown in Figure 1 of the main text.

S1. Species checklist per country (in relation to BirdLife data)- First step

The first step of methods is to understand what biases existed in real observation data, this is referenced in the first two paragraphs of the methods. For the BirdLife–based checklist (Richness shown in Figure S1), we assessed the correspondence between each administrative boundary and the range map of species. To focus solely on extant species, we retained the categories 'Extant' and 'Probably Extant' in the 'PRESENCE' attribute, as well as 'Native' in the 'ORIGIN' attribute, while removing the 'Seasonal Occurrence Uncertain' in the 'SEASONALITY' attribute from the BirdLife range maps. The definitions for 'PRESENCE, ORIGIN and SEASONALITY' distribution codes can be found via https://nc.iucnredlist.org/redlist/resources/files/1539614211–Mapping_attribute_codes_v1.16_2018.pdf which is consistent with codes also used by Birdlife.

For the eBird–based checklist, we calculated the overlap between all filtered eBird records and administrative boundaries (considering all eBird data) and compiled the species checklist based on eBird records falling within the specified administrative boundaries. Subsequently, we determined the completeness of each region by calculating the ratio of species present in both eBird and BirdLife checklists to the total number of species in the BirdLife–based checklist. Completeness was expressed as the percentage of species recorded

in BirdLife and also found in eBird data for each country, state, and ecoregion. To visualize the spatial coverage and the degree of completeness, we generated corresponding maps. The flowchart in Figure 1 shows all the functions and the relationship between the functions that were used to understand and simulate biases in data, virtual landscapes based on BirdLife maps were resampled using the simulated biases, and these data provided to each index to assess how well original diversity patterns could be replicated.

S2. Assessing biases in species occurrence data in relation to roads and urban areas for each country -Second step

Biases needed to be determined in relation to certain landscape features. This is noted in the main text in the first two paragraphs of the methods section, in addition to in later text where road observation data is noted. To assess data biases in response to accessibility and population, we created 2km and 5km buffers for the road layer (GLOBIO, 2022) and a dataset of building layer (Europa, 2022). We then determined the relationships between eBird and the buffered road and building layers generated above to calculate the percentages of the occurrences falling in 2km/5km buffer of road and building. These metrics were subsequently employed in our ensuing simulations.

S3. Mapping BirdLife –based richness patterns- Third step

Following determining bias in Ebird point based data in steps 1-2 we then needed to map the patterns of richness in Birdlife data as a basis for simulating biases under controlled scenarios. This step is noted in the main text in the third paragraphs of the main text methods. We obtained distributional data for extant birds from BirdLife International via http://datazone.birdlife.org/ (accessed date: Jan. 10th 2022). We initially transformed BirdLife's polygon range maps for each species into raster format in 1km resolution equal–area projection (Mollweide) for global coverage using the 'gdalUtilities' (McInerney & Kempeneers, 2015) package in R, then we stacked the raster files and computed the species count within each cell to get the BirdLife–based species richness map at 1km resolution (Figure S1a).

To further investigate the impact of different spatial resolutions, we rasterized the range maps to 2km, 5km, 10km, 20km, 50km and 100km resolutions and repeated the steps above to create the richness maps in multiple resolutions.

S4. eBird data and species richness maps- Forth step

eBird data also needed to be cleaned for later analysis, this section details the cleaning of eBird data prior to further analysis. This is detailed in the forth and fifth paragraphs of the methods section.

For eBird–based species richness, we downloaded the full eBird database via https://download.ebird.org/ebd/prepackaged/ebd_relDec–2021.tar (accessed date: Jan, 25th 2022) and extracted observational records from 2010 to 2019. To enhance the data quality of the eBird dataset and reduce erroneous records, we used a realm consistency approach to eliminate occurrences that fell outside of their expected realms with the following steps. Initially, we established species–realm relationships (SRR) by determining the relationships between the BirdLife range map for each species and realm layer (downloaded via

https://ecoregions.appspot.com/, accessed date: Jan, 27th 2022). Then, we removed all the records from the eBird dataset which were out of the BirdLife–derived realm(s).

This approach was specifically devised to remove alien species, pets/captive animals, inaccurate identifications and erroneous occurrences, which may otherwise give a false impression of species richness. 611,520,112 records of 10,359 species of bird (around 88% of known species) were left after all the filtering processes. With the refined eBird records, we re–projected them to Mollweide and counted the species richness (Figure S1b) and number of observations (Figure S1c) at the same seven resolutions as BirdLife–based species richness maps.

S5. Simulating virtual species ranges- Fifth step

The fifth step is complex and involves a number of sub-components where we replicate the patterns of bias observed in actual eBird observational data from the Birdlife mapped data for a number of plots. This is primarily noted in the fifth paragraph of main text methods.

Simulations of bias included varying sample intensity (volume and point distribution), accounting for rarity. The below methods were used to subsample the ten plots of Birdlife data with select biases, which could later be provided to the various diversity indices to assess their ability to overcome each form of bias, and recreate the full Birdlife dataset diversity patterns. Supplemental Figures have also been provided to clearly illustrate the different elements of the methods here.

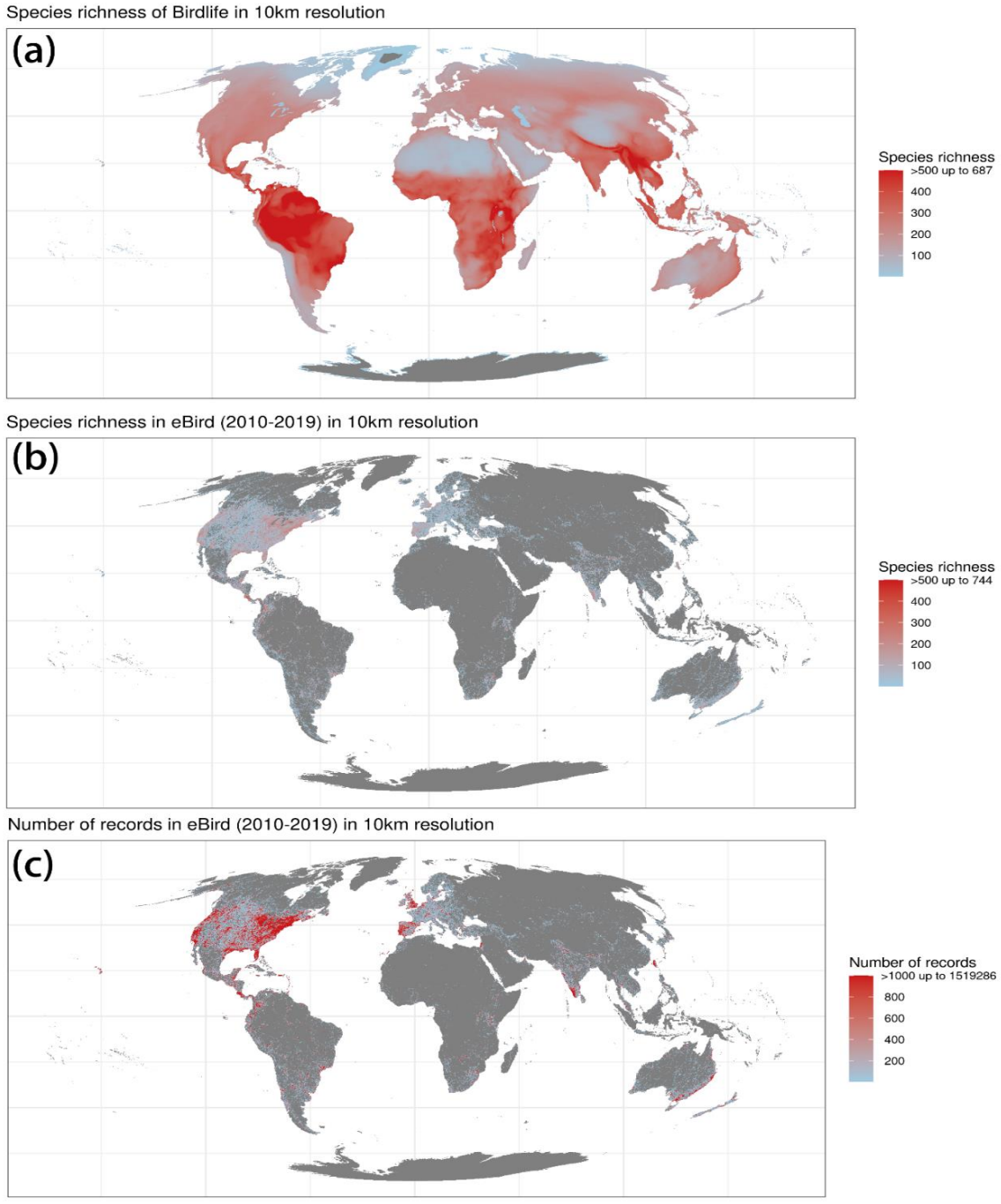


Figure S1. The species richness and number of observations from the different data sources. **(a)** BirdLife range map–based species richness, **(b)** eBird based species richness and **(c)** number of observations from eBird records at a 100km resolution. This figure is associated with the methods for the background to subsampling of plots, along with Methods S5.

A) Diversity plots. To simulate the different patterns of species diversity, we generated ten random 500km×500km boxes (called "virtual lands" as they are simulated based on real areas, Figure 1a) and cropped the 1km–resolution BirdLife–based species richness maps by the 10 virtual lands (Figure 1b). Then we designed two types of observation behavior on the 10 virtual lands, which are random and road–based observing behaviors with three sampling strengths. For the random observation behavior, we assumed that we sampled 100/500/1000 times in each virtual land randomly (Figure S4a, c, and e). For road–based observing

behavior, we assumed that sampling efforts were highly correlated with the distance to roads, using the GRIP global roads database (Meijer et al., 2018). We chose 5,027 random boxes on the road map (to simulate a road–based observation gradient from 0% to 100% of samples within observed proximities to the road, Figure S2), and cropped the road map with the boxes as virtual sampled areas (Figure S3). Then we created 100, 500, and 1000 observing events on each virtual sampled area, with 55% on roads, 33% within 2km of roads, 11% within 5km buffer of roads, and 1% outside of road buffers (Figures S4b, d, and f, Figure S5b).

B) Accounting for rarity. We also took into consideration that different species may have varying probabilities of being observed. The observability of species was determined in relation to their IUCN threat status. The species in Least Concern (LC) group had a 100% probability to be observed if its distribution was included in any of observing behaviours above. The observable probability decreased to 40% in Near Threatened (NT) group, 20% in Vulnerable (VU), and for 10% Critically Endangered (CR) and Endangered (EN); this was based on probable population density and observation probability.

Following this, we calculated the observed species by considering the species richness map generated by accounting for rarity and the simulated observing behaviours. Obviously, these factors vary by taxa as well as region, thus this is just to account for detectability biases and the probability of more threatened species having smaller ranges or lower abundance.

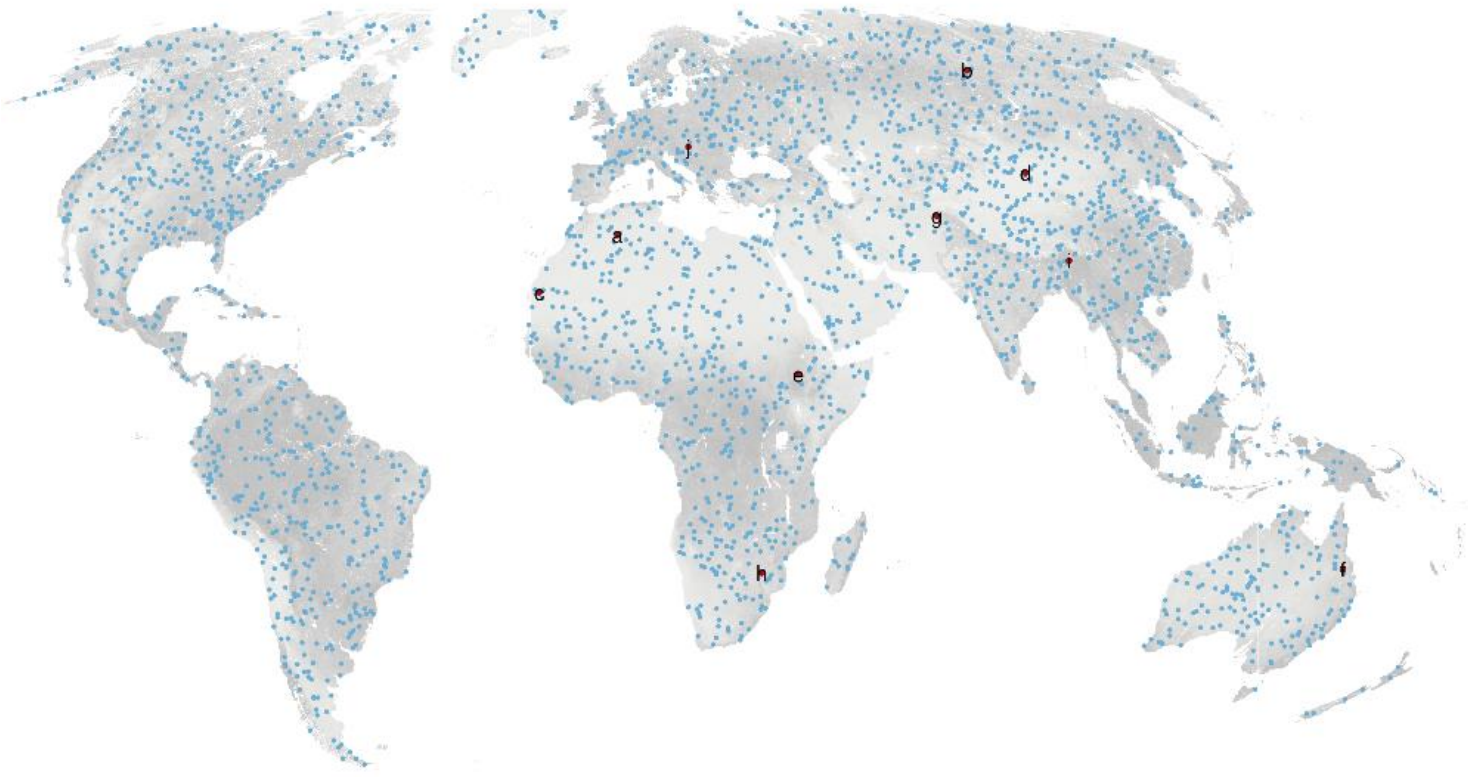


Figure S2. Creating the virtual lands based on the real–world road map. The blue points on the map represent a total of 5027 random sites. Sites marked as 'a' through 'j' represent 10 randomly selected sites (noted as ten square plots in the main text) from this set of 5027. These points have varying road coverage rates, ranging from low to high, as illustrated in

Figure S3 (below). This figure is associated with the methods for the background to subsampling of plots and shows the locations of all plots used for the ten road coverages applied to the birdlife maps for analysis additional details are provided in Methods S5.

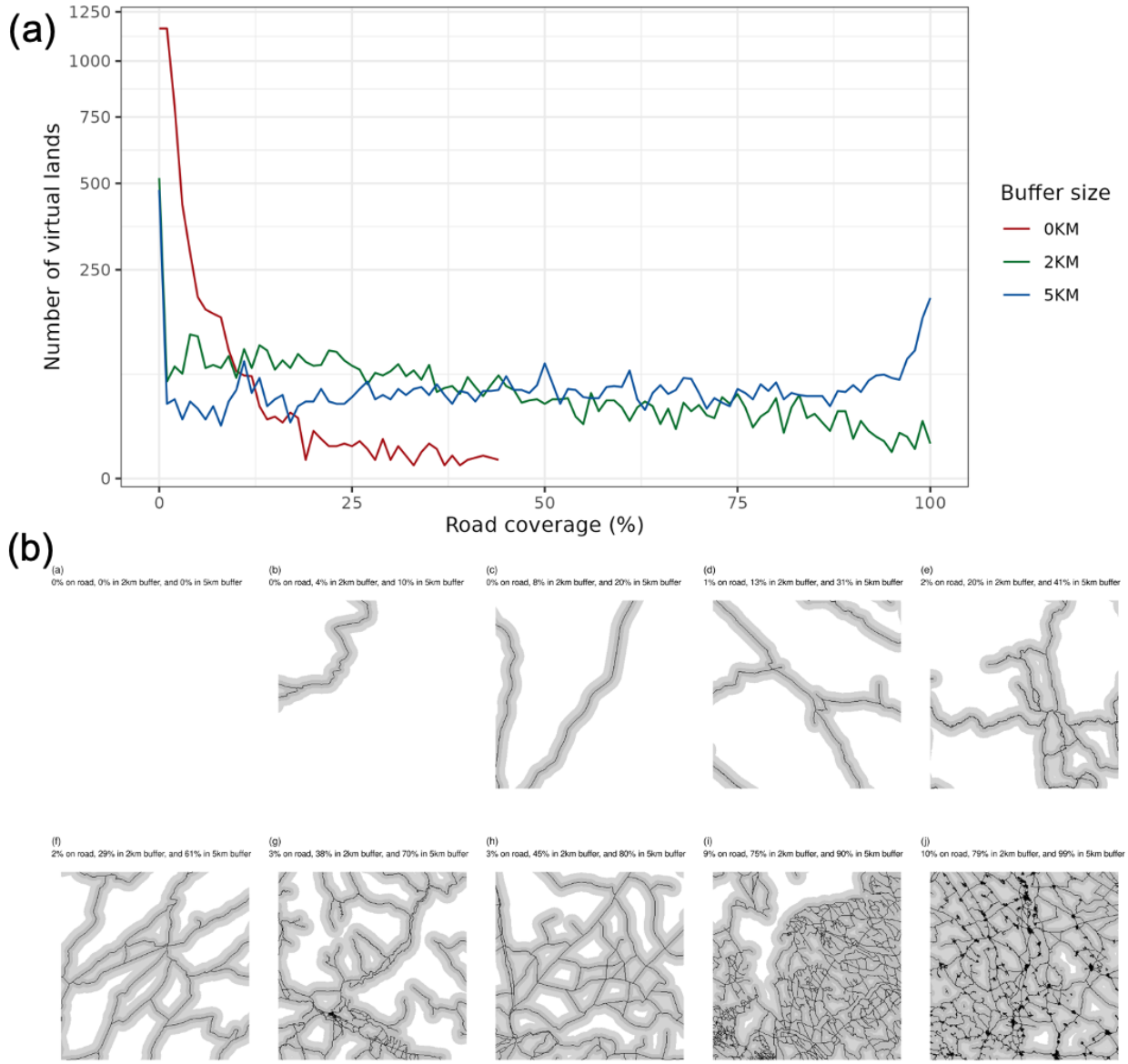


Figure S3. Random virtual lands with different road coverage gradients. (a) The road coverage of 5027 random virtual lands in different buffer sizes; (b) 10 of 5027 virtual lands with different road coverages. These road networks were super-imposed over plots with different buffer coverage, and sample-point location for subsampling of Birdlife data for bias scenarios was extracted from within these buffers. The description of how these scenarios were used is detailed in Methods S5

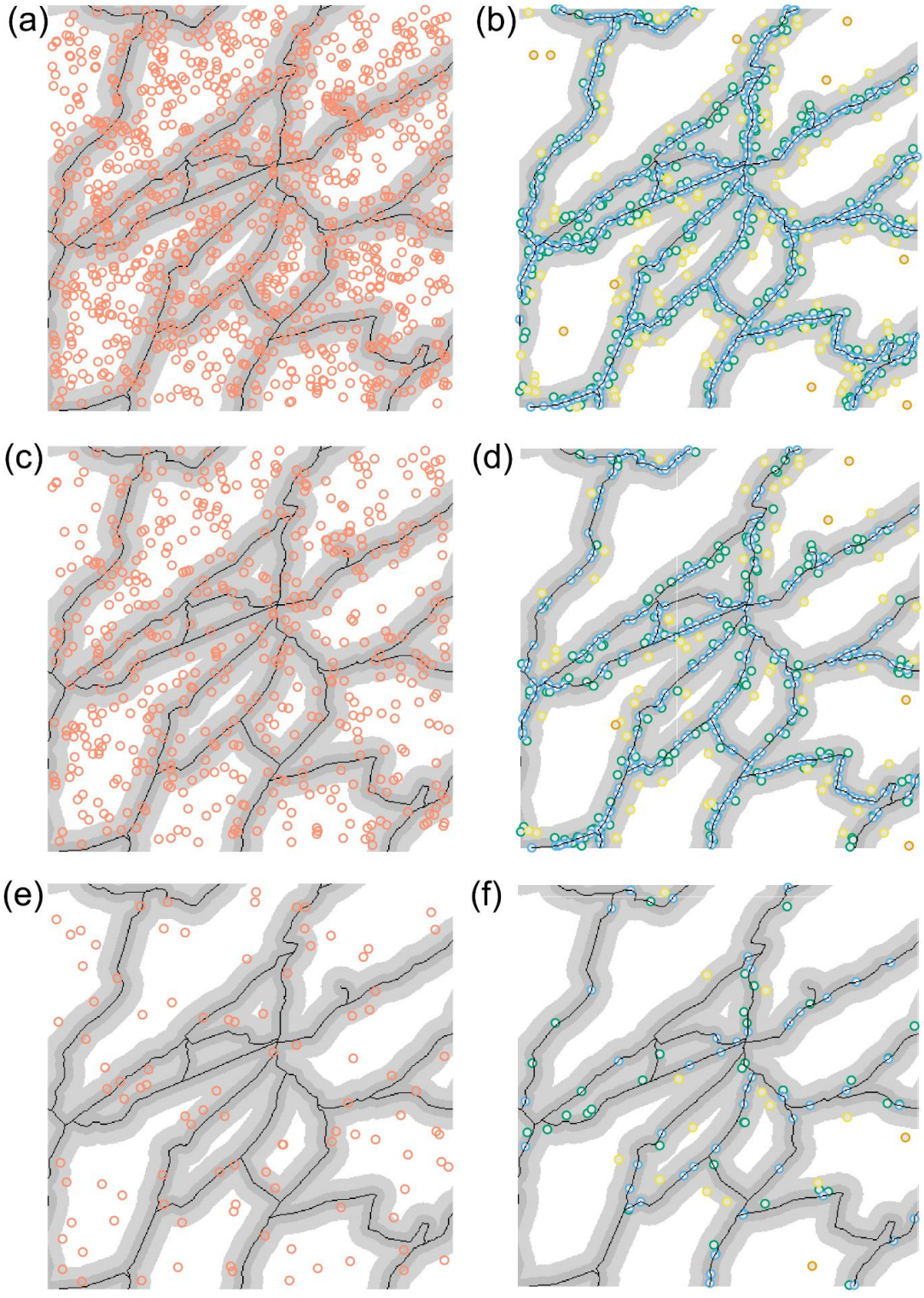


Figure S4. 1000, 500 and 100 random observing events (a, c, and e) and 1000, 500 and 100 road–based observing events (b, d, and e). 55% of them (in blue) are on the road, 33% of them (in green) are on the 2km buffer of the road, 11% (in yellow) are on the 5km buffer of the road, and 1% (in brown) are outside of the buffer of the road. This demonstrates different forms of bias in terms of sample volume, and cluster type (clustered vs random).

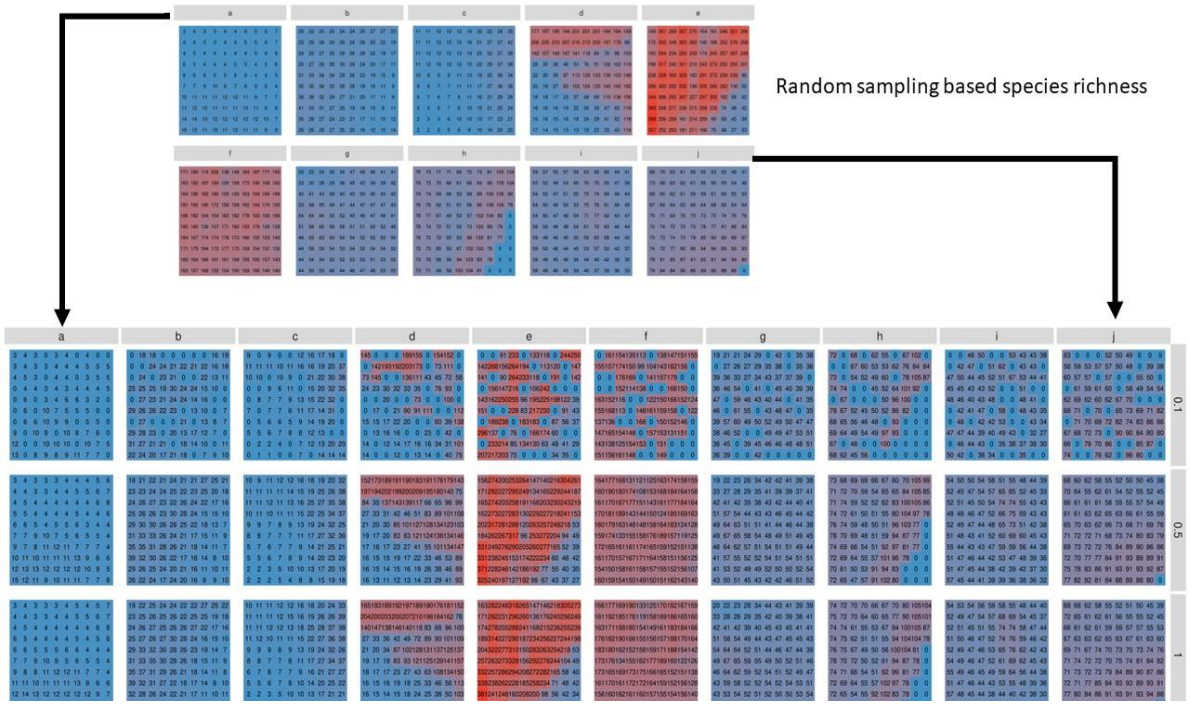


Figure S5a. Random sampling of richness grids

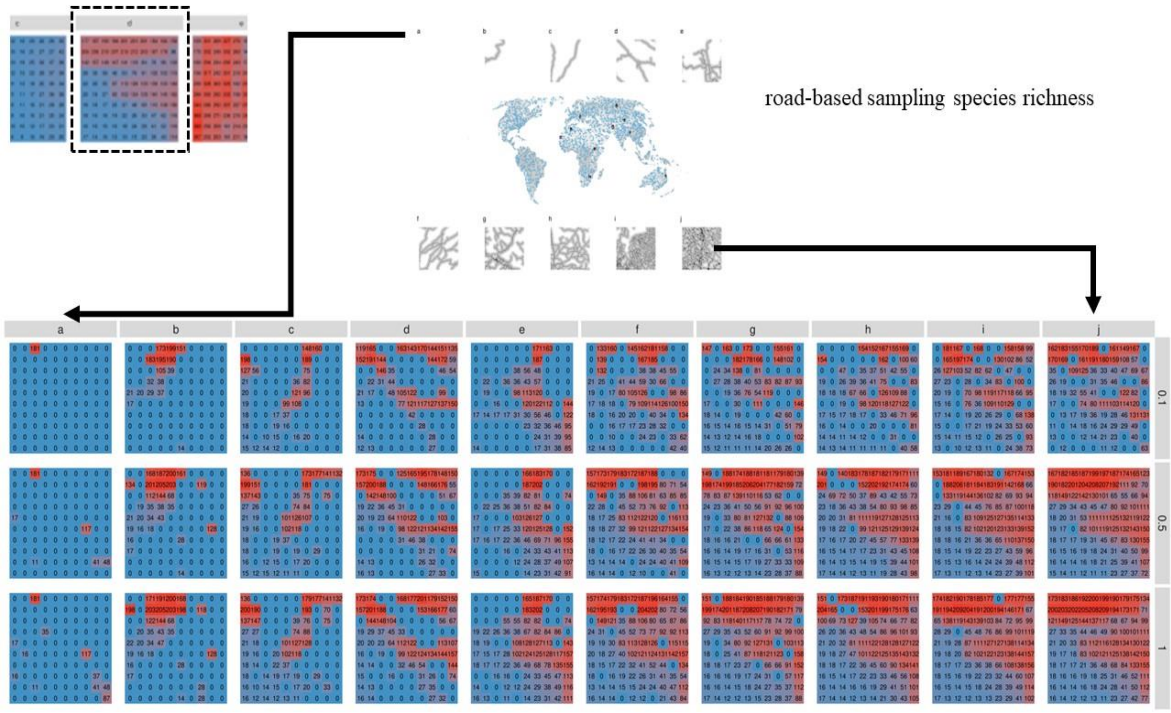


Figure S5b. Biased sampling of richness grids, blue indicates low sampling, red indicates high sampling.

S6. Applying the diversity metrics to the different scenarios- Sixth step

This step is from the sixth paragraph of the main text methods on. It details how biased data was provided to each diversity index, then assessed in relation to the "complete" richness pattern from Birdlife data.

Following all the steps above, we generated a total of 150,810 virtual scenarios by combining 10 diversity patterns × 5028 (1 random sampling + 5027 road–based sampling behaviours) × 3 sampling intensities (100, 500 and 1000). These processes are visually represented in Figure S2 and S3.

For each virtual observation dataset, we calculated several diversity metrics, including, Hurlbert's index (also often referred to as ES10, 20, 50 100 and 200; Waller 2019) using the 'entropart' package (Marcon & Herault, 2015), Shannon's index, Simpson's index, Hill numbers (Rényi diversity, scales are 0, 0.25, 0.5, 1, 2, 4, 8, 16, 32, 64, and infinity; Hill 1973) and Pielou's index with the 'vegan' package (Oksanen et al., 2013).

Furthermore, we applied the metrics above to the eBird data to map the global diversity metrics at multiple resolutions to assess how they differed from the BirdLife patterns.

For the metrics used, we summarize their definitions and characteristics here (based on Hurlbert 1971; Hill 1973; Fleishman et al., 2006; Bandeira et al., 2013; Chao et al., 2014; Colwell et al., 2014; Waller 2019), and the formulas can be found in supplemental source.

S7. Mapping diversity indices- Seventh step

Following the previous six steps we mapped out diversity by applying distribution models to the global eBird data analysed using each of the indices assessed. This involved two prime steps, firstly mapping diversity using each metric globally from the eBird data, and reclassifying each index into a equal scale to normalise data, then a second step of modelling richness from this.

7a Mapping eBird diversity indices

We mapped out alpha diversity from eBird data based on each index at resolutions of 1, 2, 5, 10, 20, and 50km. Following this, given that each index had a different numerical scale, each was reclassified into 20 bins with equal divisions for both 5 and 10km resolutions (using the reclassify function in ArcMap 10.8), to be the equivalent of 5% value intervals and enable a more direct comparison. We then quantified the coverage area of each of the 20 bins in each metric and resolution. Obviously, many different approaches can be applied here (i.e see Hughes J et al., 2023), but on a global scale finding something that reflected real–world patterns and made best use of different metrics needed to balance data availability and biases, as tested here).

7b. Projecting diversity

To project diversity beyond sampled sites, we used MaxEnt to model diversity based on various climatic drivers for the world, and climate plus landcover drivers for regions, to map

each “diversity band” for equally weighted diversity. Whilst maxent performance may also vary, given its popularity for modelling, it provides a useful approach to assess between the performance of different indices more widely. This type of approach has been used in multiple other studies as it can cross–reference the environmental drivers of diversity to project diversity in areas where data has not been collected, and where species level data to allow modelling may not exist (Orr et al., 2021; Liu et al., 2022; Potapov et al., 2023). Firstly, we downloaded a number of bioclimatic and other climate variables (see supplemental source link file) at 5km and 10km resolutions, and the 5km resolution data were partitioned by realm for separate models using the Ecoregion 2017 (WWF, 2017) data giving Palearctic, Afrotropical, Nearctic, Neotropical, North Asia, South Asia and Australasia realms (Oceania was excluded as the realm was too small). These variables were based on previous models we have run for various regions, and are known to be useful indicators of diversity (Orr et al., 2021; Liu et al., 2022). Diversity indices were then averaged for both 1) all metrics and 2) those found as good and reasonable (all except Hurlbert’s and Pielou’s indices), and one using all metrics averaged. These were then converted to point form at both resolutions giving whole integers, and values of over 16 merged for the 5km analysis because of small sample size. These were then exported to CSV form, and run in MaxEnt using default parameters and five replicates, with each diversity value treated as the equivalent of a species. The averages of the five thresholded replicates were used, and a 10–percentile training presence threshold used to classify “suitable” and “unsuitable” to give binary maps for each diversity level. These were then reclassified to give values of “0” and the appropriate diversity level (e.g for a model diversity value of 10 areas above the threshold would have a value of 10). For each scenario (global at 10km, with good and all diversity metrics, and then per realm at 5km with good and all diversity metrics) we then used the raster mosaic function in ArcMap 10.8, using the maximum value to develop two maps of diversity for each scenario (good and all diversity metrics). The mosaic function was then used to merge the good vs all metric maps, both using the sum and the mean function to map out diversity for each region and resolution (mean will average out outlying values in indices, whereas sum may capture more nuance where there are differences between indices). Visual inspection was then used to assess if the final maps represented what we know from higher resolution analyses of various regions, as well as comparison to BirdLife maps (Hughes et a., 2021b), however metrics (AUC, AIC) were not compared as the main aim was to compare between indices, rather than to develop a new map of diversity.

II-Definitions and functions of each index assessed

Various metrics have become popular for the analysis of biodiversity patterns in recent years. Yet indices have assumptions and limitations, as well as varying sensitivities. Below we outline the functions and mechanism used within each index to facilitate interpretation of methods, and enable a greater understanding of what each index does to those who are developing methods to map different facets of diversity.

Species Richness:

Definition: Species richness is a simple count of the number of different species present in a community. It is a fundamental measure of diversity, providing a basic understanding of the variety of life in a given area.
Effect of Sample Bias: Sampling biases directly affect species richness by determining which species are included in the sample. Overrepresentation of common species can inflate richness estimates, while underrepresentation of rare species may lead to an underestimation of the true species richness.

Shannon–Weaver Index

Definition: The Shannon–Weaver index measures the entropy or uncertainty in a community, incorporating both species richness and evenness. It essentially quantifies the diversity of a system by considering both the number of species present and their relative abundances.

$$D = -\sum_{i=1}^{S} p_i log(p_i)$$

p_i is the proportional abundance of species i.

Effect of Sample Bias: Sampling biases in the context of the Shannon–Weaver index can distort diversity estimates. Overrepresentation of dominant species in the sample may inflate the index, leading to an overestimation of diversity. Conversely, underrepresentation of rare species can result in underestimation, as the contribution of less common species may be overlooked.

Simpson's Index

Definition: Simpson's index assesses the probability that two individuals randomly selected from a community belong to the same species. It is a measure of dominance in a community, with higher values indicating greater dominance.

$$D = 1 - \sum_{i=1}^{s} p_i^2$$

p_i is the proportional abundance of species i.

Effect of Sample Bias: Sampling biases impact Simpson's index by skewing dominance patterns. Overrepresentation of dominant species can lead to an inflated index, suggesting higher dominance than is actually present. Conversely, underrepresentation of rare species may result in a lower index, overlooking the potential impact of less common species on community structure.

Hill Numbers:

Definition: Hill numbers are a mathematically unified family of diversity indices (differing among themselves only by a parameter q) that incorporate species richness and species relative abundances.

$$when\ q > 0\ and\ x \neq 1\ \ D_q = \left(\sum_{i=1}^{s} p_i^q\right)^{\frac{1}{1-q}}$$

$$when\ q = 1. D = exp\left(-\sum_{i=1}^{s} p_i \log\left(p_i\right)\right)$$

p_i is the proportional abundance of species i.

Effect of Sample Bias: Hill numbers are sensitive to sampling biases, especially when rare species are not adequately sampled. This can lead to an overestimation of diversity, particularly for higher–order Hill numbers, as rare species contribute disproportionately to the effective diversity.

Hulbert's index (ES):

Definition: Hulbert's index (often written as ESX) is the statistically expected number of unique species in a random sample of X occurrence records, and is an indicator of diversity. The score can be computed without random sampling, but the mean of infinite random sampling will produce the same result.

Effect of Sample Bias: Sampling biases impact Hulbert's index by potentially inflating or deflating effective diversity. Overrepresentation of common species can inflate Hulbert's index, while underrepresentation of rare species may result in an underestimation of effective diversity.

Pielou's Index (Evenness):

Definition: Pielou's index, also known as the Evenness Index or J', quantifies the evenness of species abundance distribution in a community. It assesses how equally individuals are distributed among the different species that are present.

$$D = 1 - \sum_{i=1}^{S} \left(\frac{n_i(n_i - 1)}{N(N-1)}\right)$$

S: The number of species
n_i: the abundance of the nth species
N: the total abundance of each species

Effect of Sample Bias: Sampling biases can impact Pielou's index by distorting the representation of species abundance. Overrepresentation of dominant species may artificially

increase evenness values, suggesting a more balanced community structure than actually present. Underrepresentation of certain species, especially rare ones, may result in a lower evenness index, indicating less uniformity in the distribution of individuals across species.

The match between each of these metrics under each scenario was then compared to the original Birdlife richness maps (from which the biased samples were subsampled) to provide a comparative "truth". We used the 'postResample' function in 'caret' package in R to calculate the R^2 values, to compute the correlation between each diversity metric and the patterns from Birdlife richness.

$$R^2 = \frac{\sum_i (\hat{y}_i - \bar{y})^2}{\sum_i (y_i - \bar{y})^2}$$

Which y_i is the diversity metric from BirdLife richness of a cell i, $\bar{y}$ is the mean diversity metric of all the cells in a given area, and $\hat{y}_i$ is the diversity metric calculated from the subsampled datasets representing a given bias scenarios.

III-Supplemental results

Mapping diversity with different metrics

This section provides greater descriptive detail on model performance of each metric for the "***Mapping diversity with different metrics***" section of the results. This simulation is to broadly understand the general patterns of performance of each index applied.

Performance started to increase at 2km, but major differences exist in the relative performance gain of different metrics, and Shannon's index also started to show marginal gains relative to some other metrics. This repeated at 5km, though the disparities between index performance became more demonstrable. Pielou's index also outperforms ES10 at 5km, but this trend then reverses for coarser resolutions. Shannon's index also started to show marked improvements in performance at 10km, but required a much higher road density than Simpson's index, species richness, or Hill_Inf to do so. These patterns generally held at 20km, though improvements in Shannon's index, and at very high road densities it may actually out–perform other indices at the coarsest resolution. Finally, at 50km the performance of Hill_Inf decreases, and ES10 is actually the most consistent, followed by Simpson's and then Shannon's indices (at road densities over 50%–Figure S16).

At a 5km resolution, these patterns are progressively pronounced, especially at higher road densities, and most landscapes show Pielou's index and ES10 as the most poorly ranked). These same trends exist at 10km, and by 20km only Pielou's index performs poorly. Additionally, at increasingly coarse resolutions show performance increasing at lower road densities. At the coarsest resolution (50km), the top performing indicators remained the same, but Pielou's index was often poor, and species richness and Hill_Inf varied between landscapes.

Dealing with small volumes of data

This section provides greater descriptive detail on model performance of each metric with limited data quantities for the "***Dealing with small volumes of data***" section of the results. This simulation is to explore the impact of low sample sizes on reliability of outcomes.

At the 100–point level, the metrics plateau after 25 at a 1km resolution, after which most metrics perform similarly (other than the worse Pielou's and Hurlbert's indices). Performance patterns then followed a similar ranking at resolutions between 1–5, with Shannon's, Simpson's, indices and Hill_Inf performing similarly well. As the resolutions became coarser, the performance of Hurlbert's index improved, so by the 50km resolutions it actually performed better than the others, whilst species richness moved from performing well to being the second–worst metric. With 100 points, the ability of each metric to represent actual diversity patterns showed very interesting trends, with many showing reduced performance to a 25% coverage, before increasing again.

Mapping diversity

This section provides greater descriptive for the "***Mapping diversity***" section of the results. This section explores the impact of different selected variables on impacting richness patterns overall for the modelled outcomes.

As using primary data only provides an index of diversity in sampled areas, projecting that diversity requires a method to interpolate diversity. Here, we used the mean values of diversity indices to project diversity using Maxent. The analysis worked better using the global 10km data, where sample sizes are adequate for modelling even at the higher diversity levels (Figure 5), with the summed diversity between the good and all metrics making this most clear (Figure S22). Globally, drivers of low diversity included a range of drivers, though PET is one driver which decreased in importance even by medium levels of diversity. Conversely, both growing degree days and minimum precipitation initially increased and then decreased in importance, whilst Emberger's pluviothermic quotient increased in importance with diversity, as does temperature seasonality (Figure S22). Regional drivers do differ (though regional 5km models match known patterns less well), but could be useful if used with caution (see supplemental results and Figures S21a). However, for regional analysis, some landcover metrics were also used (which may be biased or shown regional differences and thus were not used in global analysis) as well as nighttime lights, though nighttime lights were removed from final models as well sampled areas biased results in initial models.

Regional drivers of diversity based on distribution modelling

This section provides further detail for the "***Mapping diversity***" section of the results, with a focus on regional differences. These results are primarily showcased in supplemental results, but highlight the nuanced outcomes we can see when we seperately analyse different regions with both different biogeographic histories and conditions, and likely different degrees of bias in the input data.

For Africa growing degree days is important at the lower levels, then minimum precipitation becomes more important, and annual precipitation also increases in importance (Figure S23). Maximum precipitation becomes progressively less important, whereas maximum precipitation becomes more important to intermediate levels, canopy is also increasingly important at higher levels and isothermality at high levels. In Australia potential evapotranspiration is important at low levels, whilst solar radiation and climate moisture index become more important. In the Indo–Malayan region temperature seasonality is initially important, but progressively less important, embergers quotient becomes less important, and solar radiation becomes more important than decreases in importance and canopy height becomes important at higher levels (though performance at high levels is less good). South America has similar drivers to Indo–Malaya, though embergers quotient becomes important at high levels, and vapour also becomes increasingly important. Likewise Nearctic (North America) and Palearctic have an increasing contribution of vapour and thermicity at higher levels. However, in North America temperature seasonality becomes more important to intermediate levels and growing degree days are also important to intermediate levels, whilst continentality decreases in importance. In the Palearctic

continentality remains fairly consistently important, whereas solar radiation becomes less important and vapour and thermicity more important.

Dealing with small volumes of data (retaining sampled and unsampled cells)

This section provides greater descriptive detail on model performance of each metric with limited data quantities for the "***Dealing with small volumes of data***" section of the results. This simulation is to explore the impact of low sample sizes on reliability of outcomes.

Many areas lack robust sampling data, thus useful metrics must also deal with different volumes of data, and not conflate diversity and sampling intensity. Based on average performance, at the lowest sampling level (100 points) metrics struggle at a 1km resolutions (Figure S24, S25), and performance is better at higher sampling intensities. Treating these values is challenging, as how to treat "null values" can dramatically alter the perception of metric performance. However metric performance in terms of rank is almost entirely consistent between scenarios (only the worst metrics vary, with Pielou's index normally performing worst, but in some cases ES10 showing the worst performance). Hill_Inf however is the most variable, sometimes ranking as the second best, and at better levels of sampling, and larger areas performing as one of the worst. Conversely Simpson's index always provides the best index, and in most cases is followed by species richness.

For individual plots however differences in metric performance does emerge. At a low sampling level (100 samples–Figure S25), at the highest resolution all metrics perform poorly, and whilst Simpson's index tends to perform better, the difference is fairly marginal. As resolution becomes coarser differences become more marked, and whilst Simpson's index tends to perform best, Hill and species richness outperform it in some instances. These patterns are more market by 10km, with Simpson's index performing best in 6 plots, Hill in two, species richness in one, and no clear best index in the last. These patterns remain fairly consistent at 20km, though Hill overtakes Simpson's index on two plots; however at a 50km level the performance of Hill can vary from the best to the worst index in different plots, as can species richness, and whilst Simpson's index normally performs well, it is frequently not the best metric. When samples are increased to 500 variability in metric performance increases, however the Simpson's index is normally in the top 3 and often the top 2, whereas the next best performing indices (Hill_Inf and species richness) can be among the best, or worst, depending on the area. At 1000 samples, performance is enhanced at smaller resolutions, performance is also generally better and more consistent, and the top performing variables are normally the same as those with lower levels of sampling. The ability of many metrics is greatly improved at this higher sampling level, though Pielou's index generally remains the poorest.

When bias vs random is examined further differences emerge, though at the 100 sample level all metrics perform universally poorly at resolutions below 10km, with only marginal increase in effectiveness between random and biased data (Figure S26). At 5km ES10, 20, Pielou's and Shannon's indices perform least well, and Simpson's index performs best, but only by a small margin. At 10km these differences become more pronounced, and the advantage of random sampling (vs biased sampling) also increases. However, the change at

20km is more pronounced, and most metrics perform far better at this level, with the difference between random and biased data also becoming clearer (Figure S26, S27). However, at 50km, the differences in performance between variables is much smaller, though at this level Hurlbert's index generally performs better than Hill numbers, and ES100 is as good as Simpson's index. At 500 samples patterns are similar, but all start at smaller areas, so even 5km is good, especially for random points, and most metrics work well at 10km and 20km, though Pielou's index is as ever universally poor and ES100, 200 also perform less well. At 50km most metrics work well, especially for random points, though the same metrics retain a slight advantage. Using 1000 samples these patterns continue grow stronger, with patterns emerging at higher resolutions, and the gap between the performance of different metrics continuing to decrease. Hulbert's index works poorly below 20km, Pielou's index remains universally poor, Shannon's index is poor below 10km, and Hill numbers and Simpson's index perform comparably in most metrics with 1000 samples.

IV-Supplemental Result Figures

Supplemental figures (below) largely disaggregate and display the common patterns of bias in the data to provide a basis for simulations of bias in the virtual plots, and assess the ability of popular metrics to overcome these biases. These all align with different sections of the main manuscript text, and the section and any associated figures are also noted below to facilitate interpretation and use.

1). Parameterising bias

a). Spatial completeness figures

Figures S6-S9 assess bias and completeness

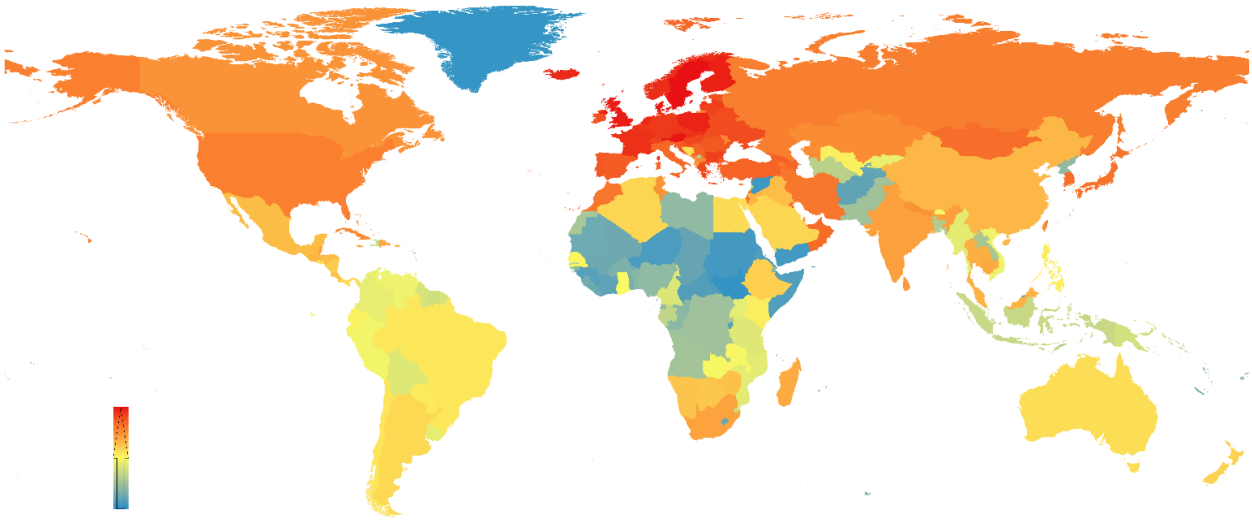


Figure S6a. Map of completeness of eBird data relative to BirdLife at a country level, (species not native to realms removed). Red-Highly complete, Yellow-intermediate completeness, Blue- low completeness. This highlights the need to understand bias to enable the meaningful analysis of global distribution patterns given the lack of data from many regions. This is discussed in depth in the "***Spatial completeness***" component of results.

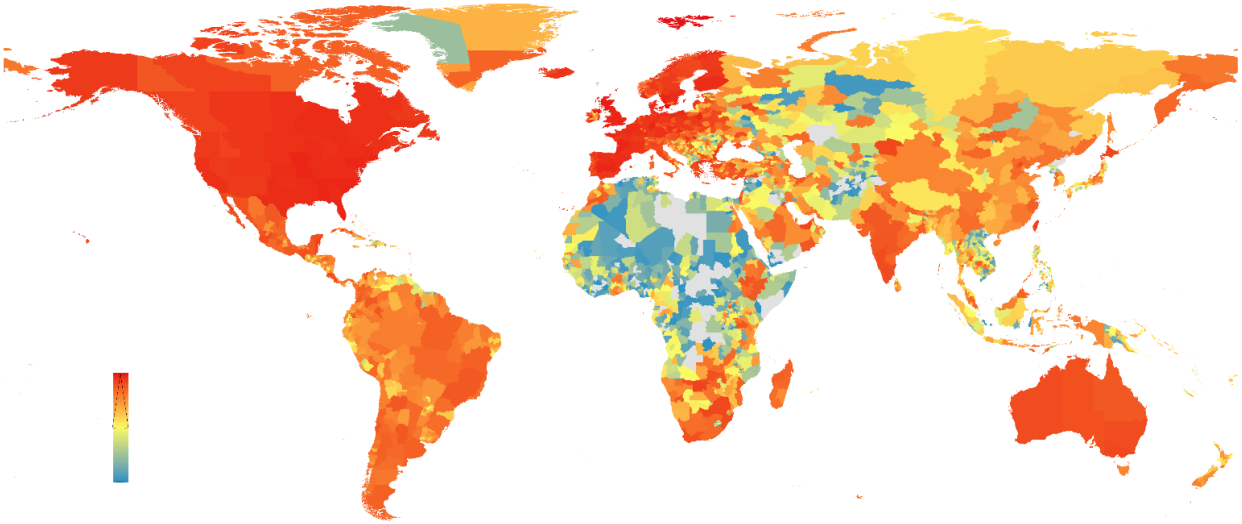


Figure S6b. Map of completeness of eBird data relative to BirdLife at a state level, (species not native to realms removed). Borders have been removed to ensure small areas are not obscured. Blue areas have lowest representativeness (especially across Africa and parts of Asia), yellow areas have medium levels of representative sampling, and red areas have the highest representative sampling. This highlights the need to understand bias to enable the meaningful analysis of global distribution patterns given the lack of data from many regions, and how when investigated at higher resolutions some areas have no data. This is discussed in depth in the "***Spatial completeness***" component of results.

In terms of completeness, Sub–Saharan Africa (48 countries) had the lowest completeness at only 68%, followed by East Asia and Pacific (36 countries) at 71%, South Asia (7/75%), Middle East and North Africa (21/75%), Latin America & Caribbean (38/84%), North America (3/85%) and finally Europe and Central Asia (54/86%). Based directly on income low income had the lowest coverage (29/59%), followed by Lower middle (49/77%), Upper middle (53/81%) and finally high income (76/83%).

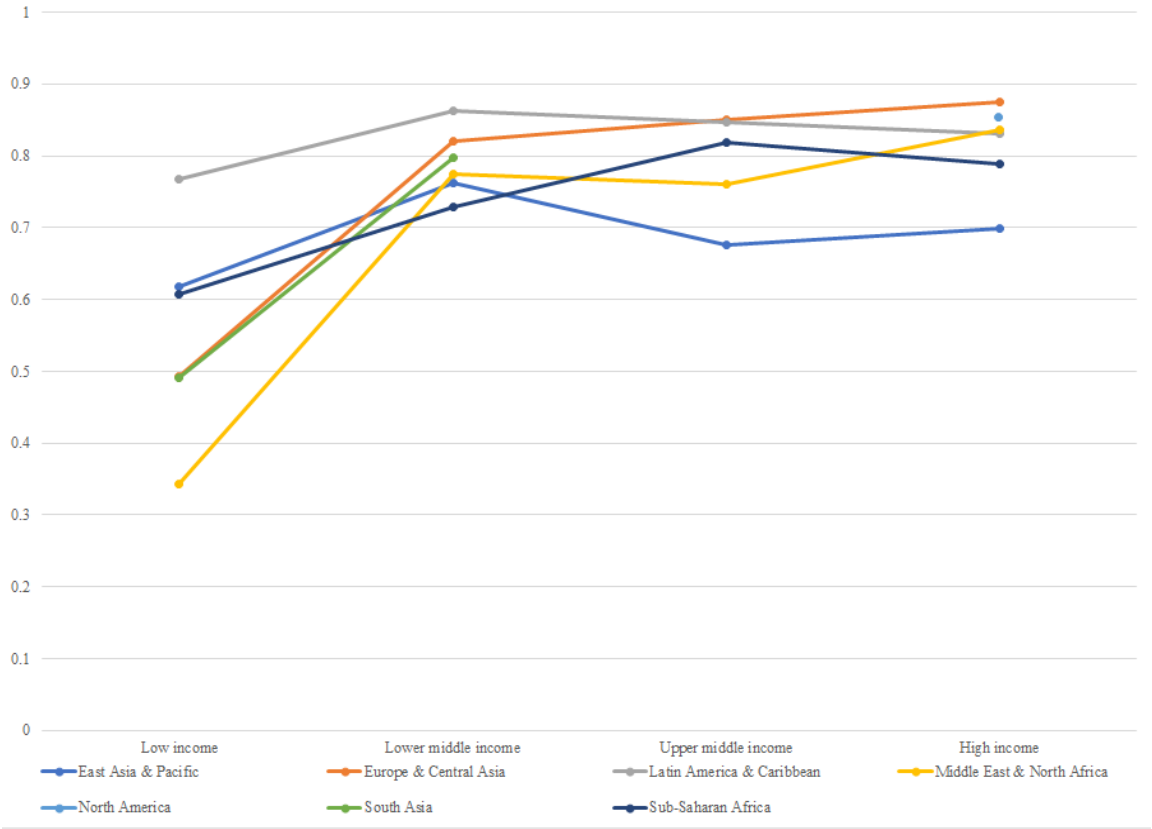


Figure S6c. Proportion of data within different degrees of completeness in different regions and income levels. Income is a major driver of data coverage globally, however the patterns are more nuanced within some regions. This is discussed in depth in the "***Spatial completeness***" component of results.

Thus in almost all cases and regions, higher income countries have greater completeness, but higher income countries in more diverse regions still have lower completeness, highlighting major data–gaps in many regions.

Completeness of eBird data relative to BirdLife data

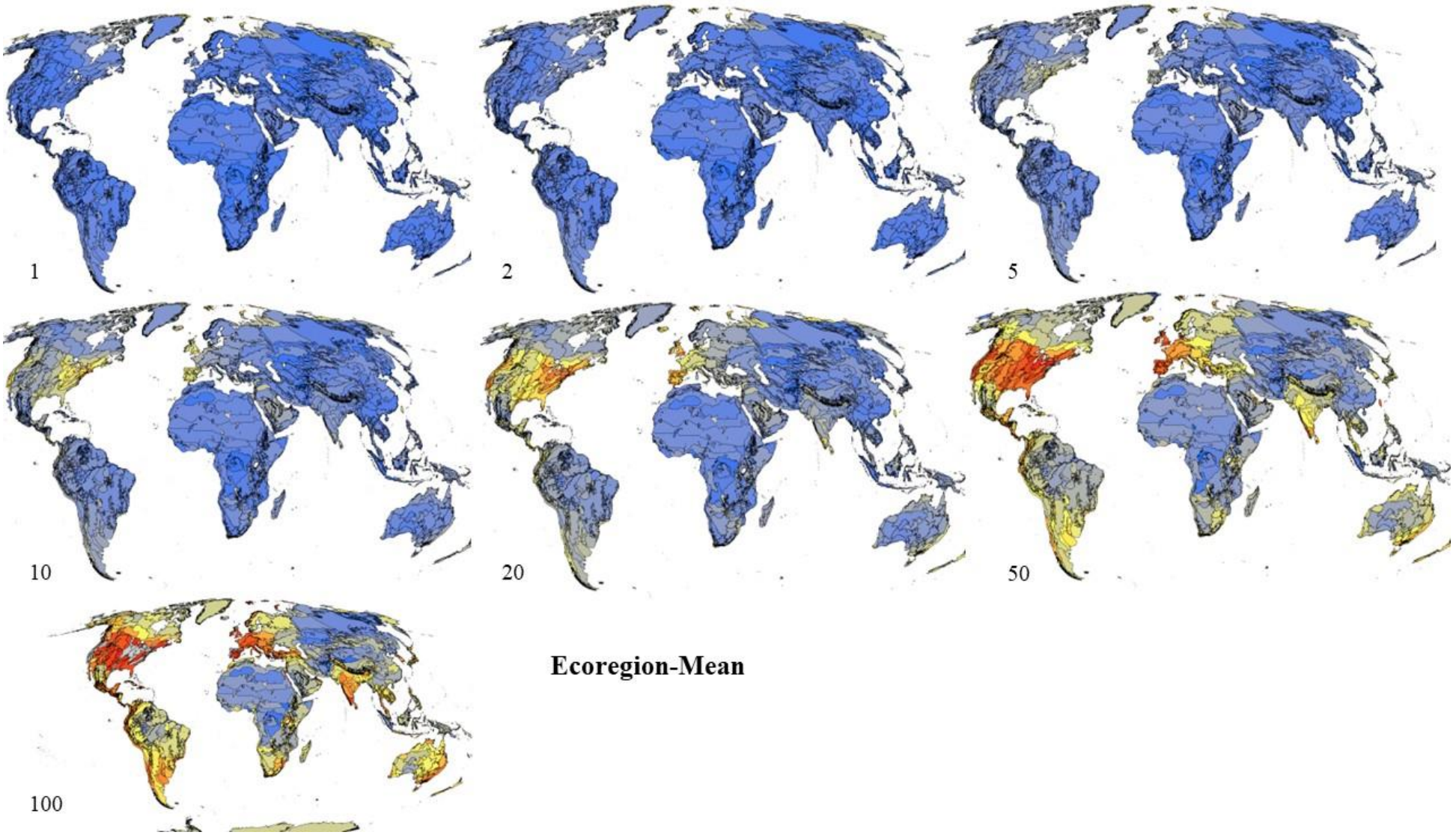


Figure S7a. Mean species completeness for each ecoregion at different resolutions from 1–100km (resolution is noted on the left of each figure). Blue has poor completeness, yellow has medium completeness and red has high completeness. This is referenced in the second paragraph of the "***Spatial completeness***" section, and details the completeness of species inclusion within eBird Point-based data per ecoregion at varying resolutions, highlighting that coarser resolutions can inflate our presumption of completeness.

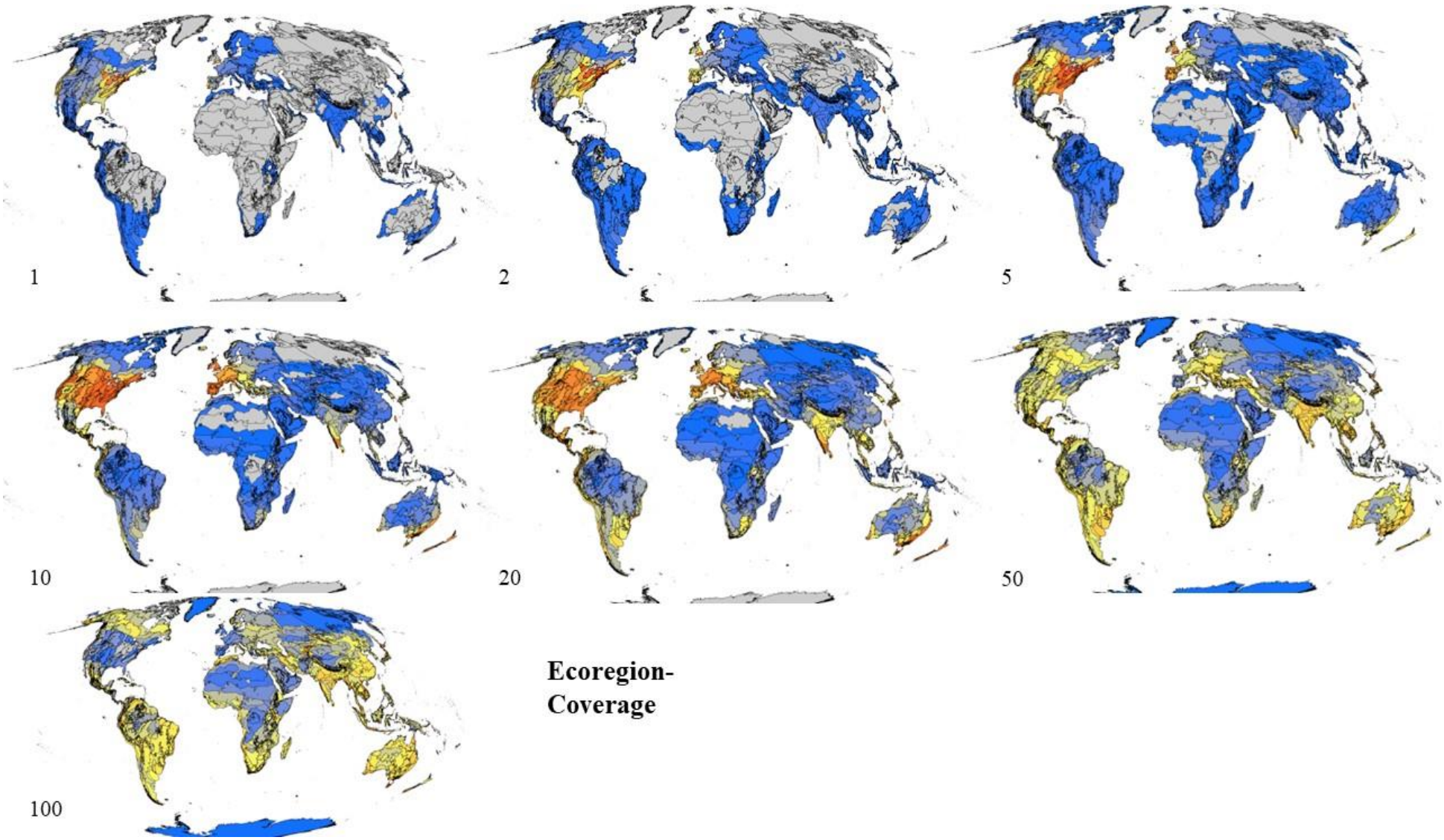


Figure S7b. Percent coverage for each ecoregion at different resolutions from 1–100km (resolution is noted on the left of each figure). Blue has poor coverage, yellow has medium coverage, and red has high coverage. This is referenced in the second paragraph of the "***Spatial completeness***" section, and details the spatial coverage of eBird Point-based data per ecoregion at varying resolutions, highlighting that coarser resolutions can inflate our presumption of completeness.

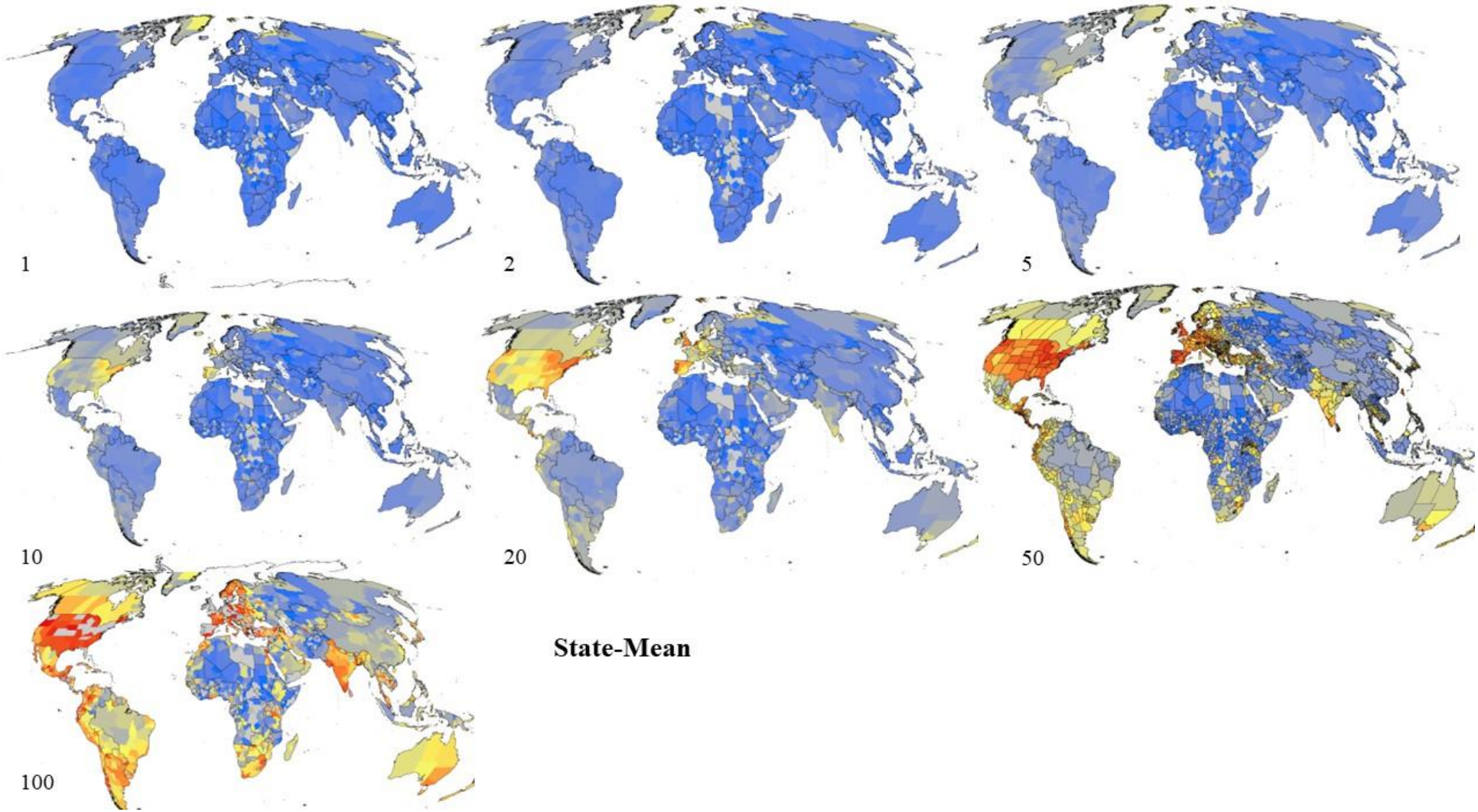


Figure S8a. Mean species completeness for each state at different resolutions from 1–100km (resolution is noted on the left of each figure). Blue has poor completeness, yellow has medium completeness, and red has high completeness. This is referenced in the second paragraph of the "***Spatial completeness***" section. It highlights that at higher resolutions very few areas have a good coverage of species recorded by eBird point-based data.

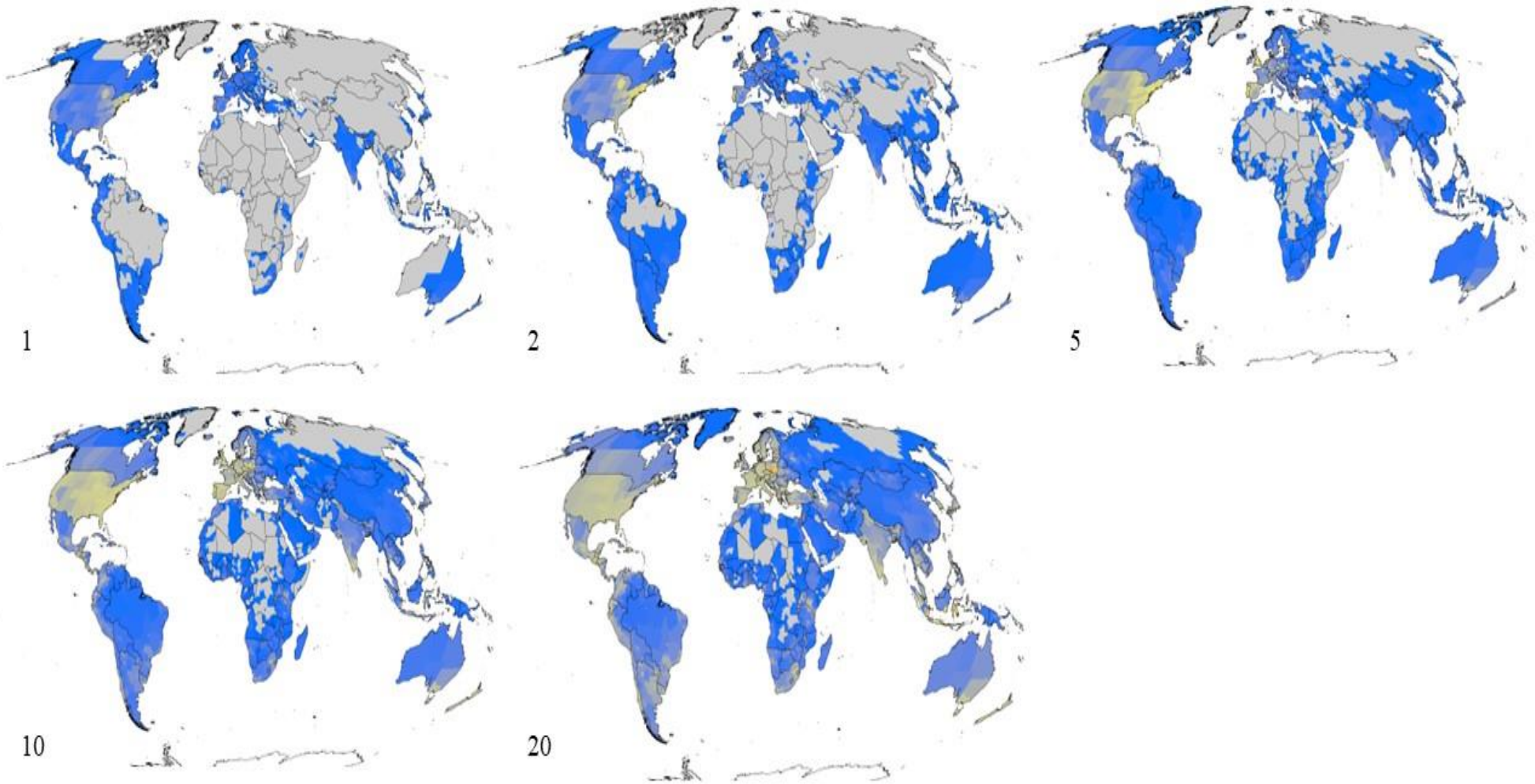


Figure S8b. Spatial coverage for each state at different resolutions from 1–20km (resolution is noted on the left of each figure). Blue has poor coverage, yellow has medium coverage, and red has high coverage. This is referenced in the second paragraph of the "***Spatial completeness***" section, it highlights the low degree of data coverage of almost everywhere outside the United States and Europe based on where observations have been made.

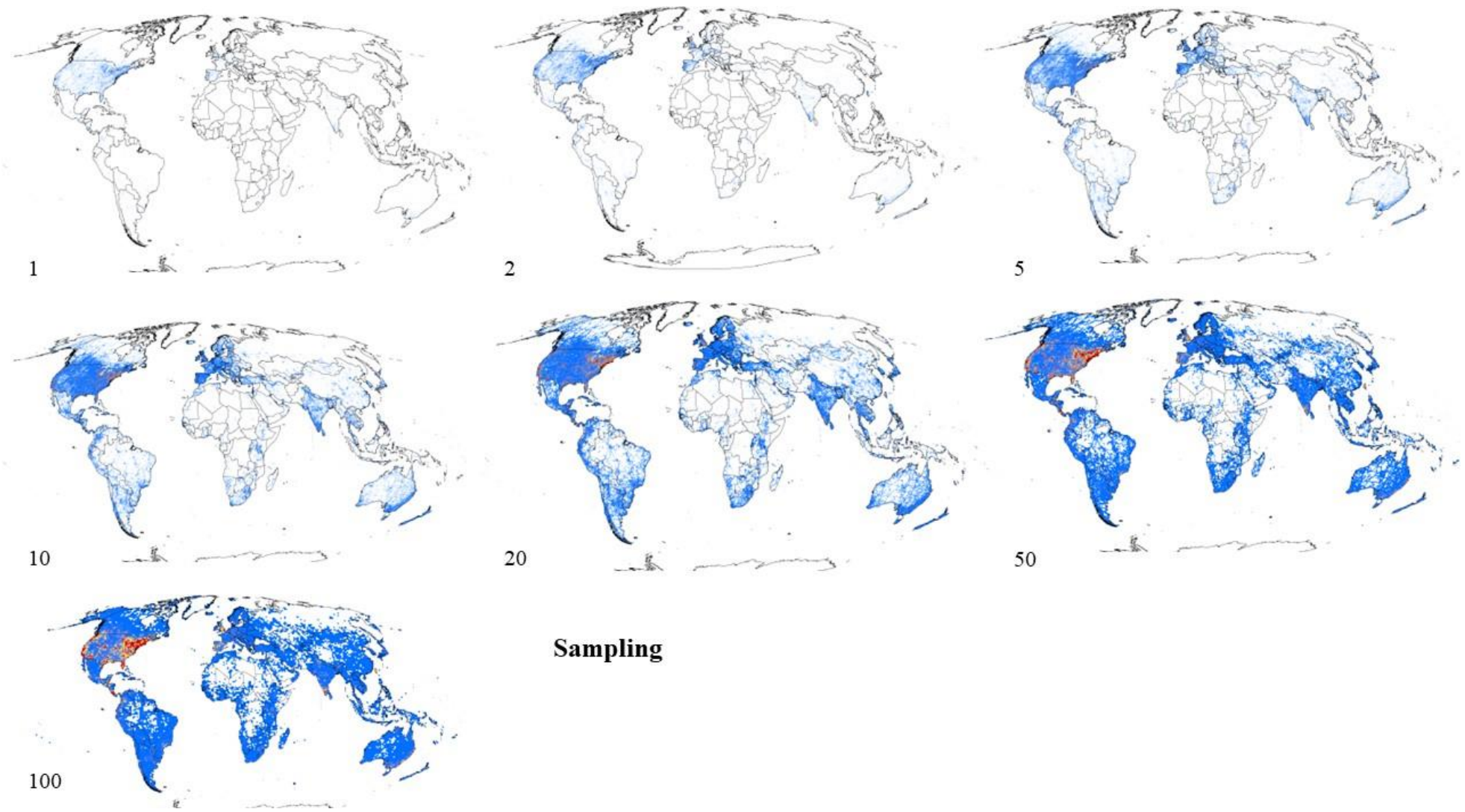


Figure S9. Sampling levels at different resolutions from 1–100km (resolution is noted on the left of each figure). Blue has poor sampling, yellow has medium sampling, and red has high sampling. This is referenced in the second paragraph of the "***Spatial completeness***" section, by plotting point density we can obtain an even higher resolution understanding in the major spatial gaps in data collection globally, and how coarser resolutions may provide a false perception of the degree of data coverage.

b). Spatial biases figures

Overall proximity trends

Figures S10- S14 are for the Spatial biases part of the manuscript

Species level and country responses to roads

Whilst most data is collected from near roads, that is not true for all species. Specialist species in particular may still require more pristine and intact habitat, meaning their inclusion in global datasets may be limited without specialist efforts. This means great caution is needed when projecting point-based samples from global databases as generalists will be over-representative as they are disproportionately abundant in disturbed habitats. These results are mentioned in brief in the "***Spatial biases***" section of the main text, but provided with more detail below.

Most species have most records within a short distance of the road, this is not to say that they do not have records outside this region but the majority of data is from areas adjacent to roads. For example, over 80% of countries have at least 19% of species with all records within 2km of a built-up area, and 41.3% of species within 5km, whereas this is 16 and 40.4% for roads. For 90% of all records, 80% of countries have 24% of species within 2km of an urban area, 47.4% in 5km and 53.2% within 5km of a road. At 75% of records, this becomes 38%, 60% and 72%, and at 60% of records, the numbers are 47%, 69.3% and 82%. Thus, whilst most records for most species are biased, this varies considerably by country based on physical geography.

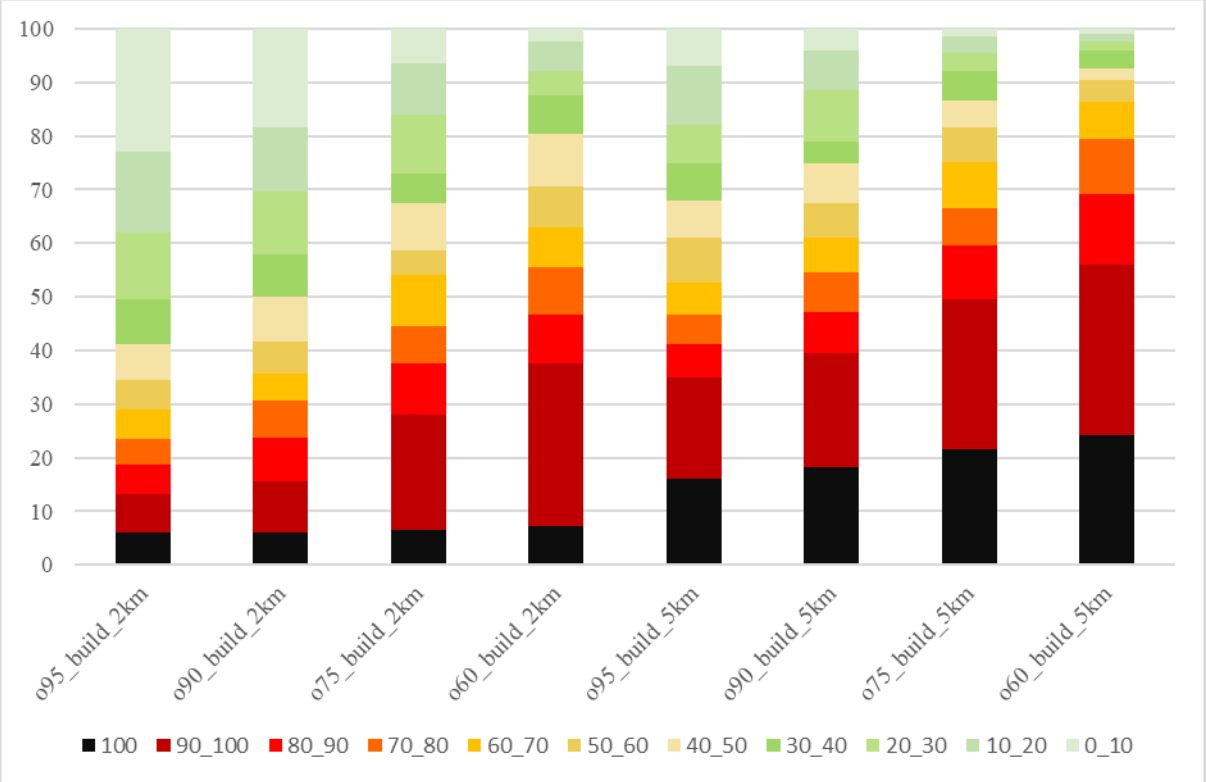


Figure S10a. Percentage of species in percentages of countries with different proportions of records within set distances of built–up areas. 100 indicates 100% of species had 100% of their records within the set distance. See the "***Spatial biases***" section of the main text.

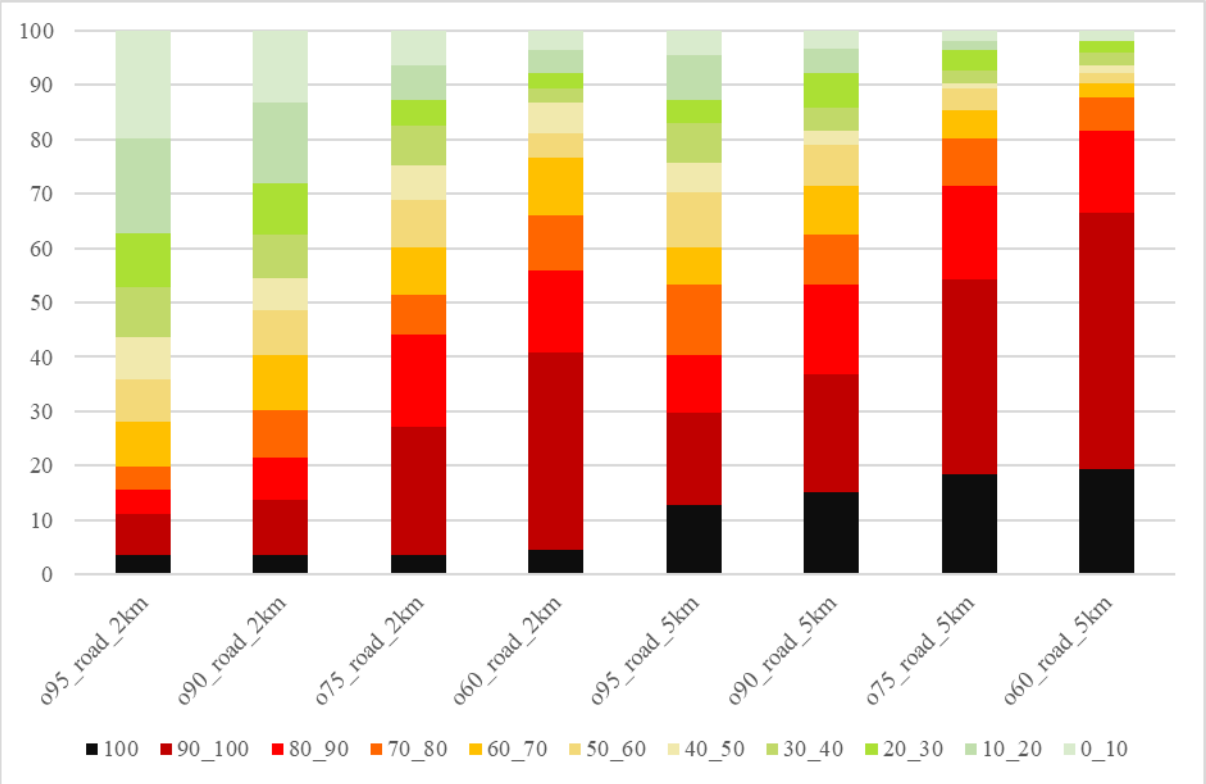


Figure S10b. Percentage of species in percentages of countries with different proportions of records within set distances of roads. 100 indicates 100% of species had 100% of their records within the set distance. See the "***Spatial biases***" section of the main text, particularly the initial part of the section.

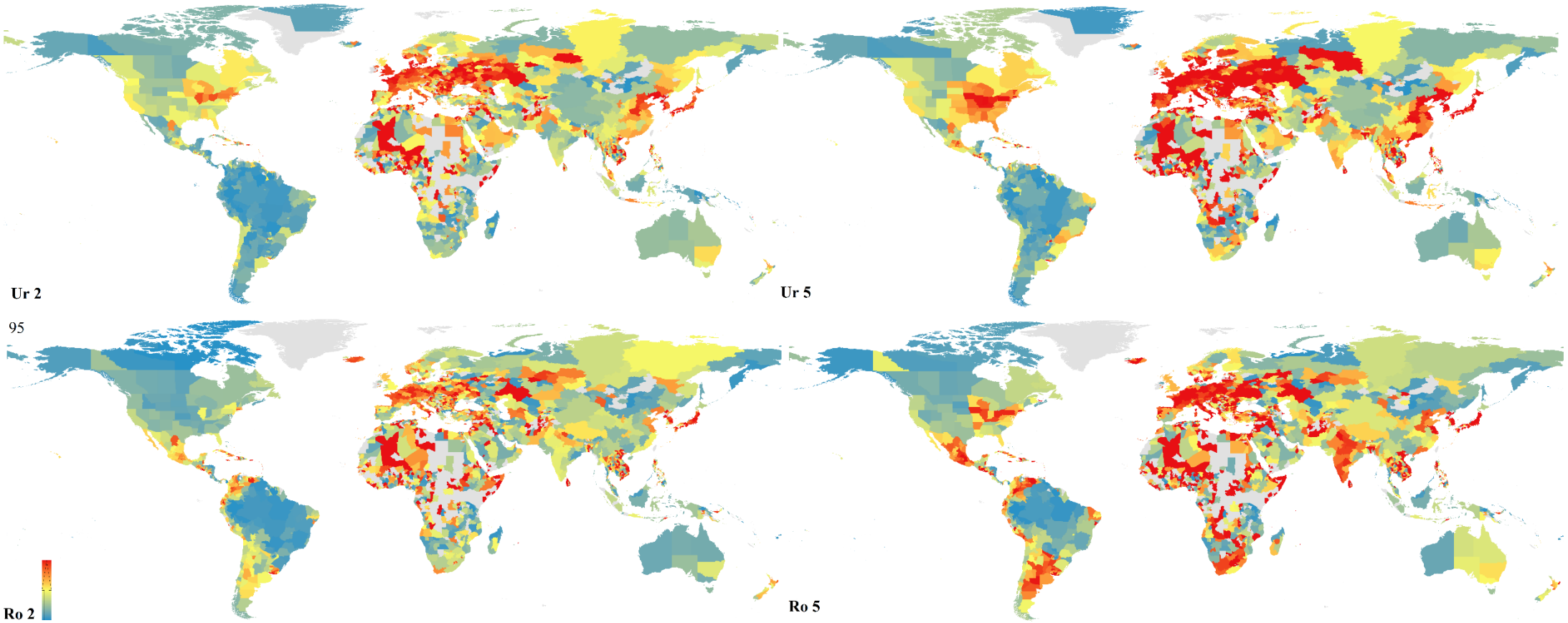


Figure S11a. Percentage of species with at least 95% of points within 2/5km of an urban area or road at a state level. Blue–low, yellow–medium, red–high. Ur-Urban, Ro-Road. See the "***Spatial biases***" section of the main text, this highlights the distribution of bias, and where data will show greatest biases towards developed and accessible regions.

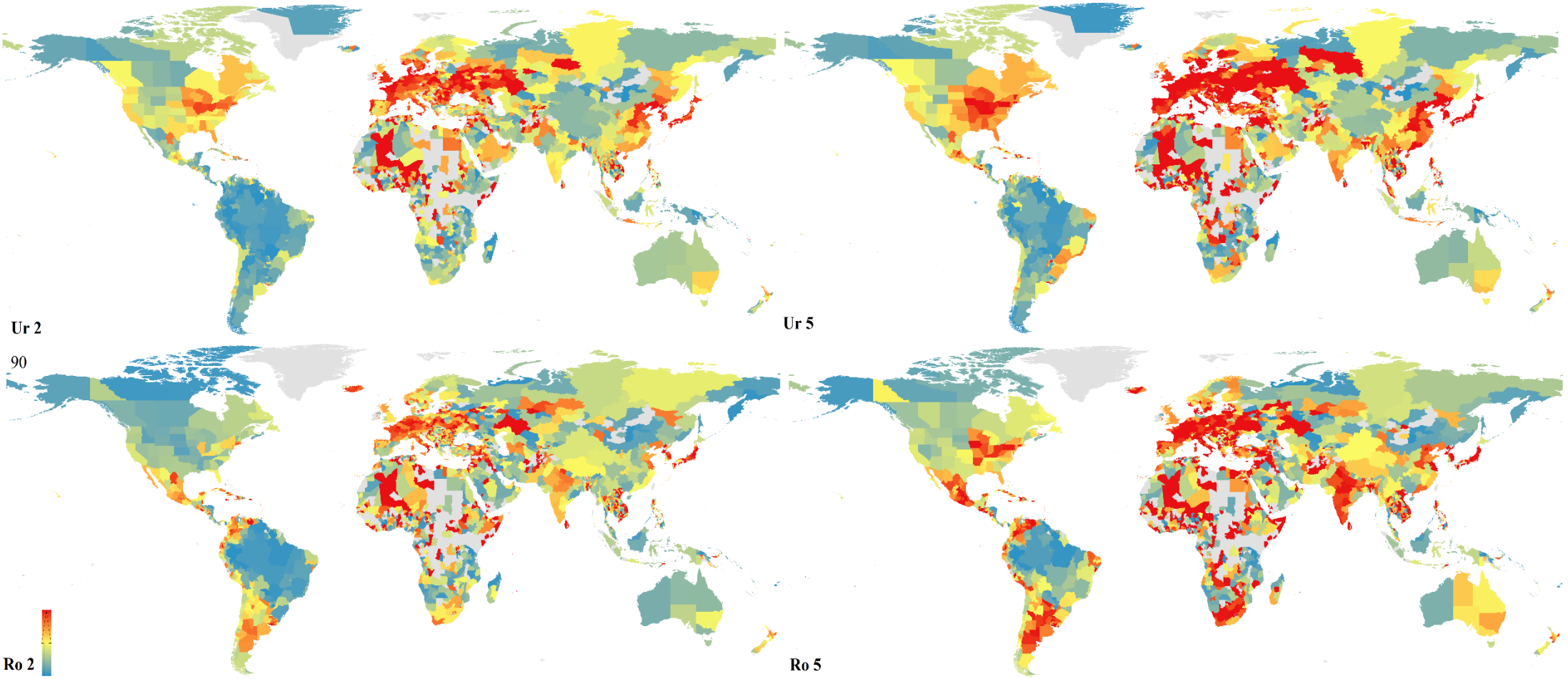


Figure S11b. Percentage of species with at least 90% of points within 2/5km of an urban area or road at a state level. Blue–low, yellow–medium, red–high. Ur-Urban, Ro-Road. See the "***Spatial biases***" section of the main text, this highlights the distribution of bias, and where data will show greatest biases towards developed and accessible regions.

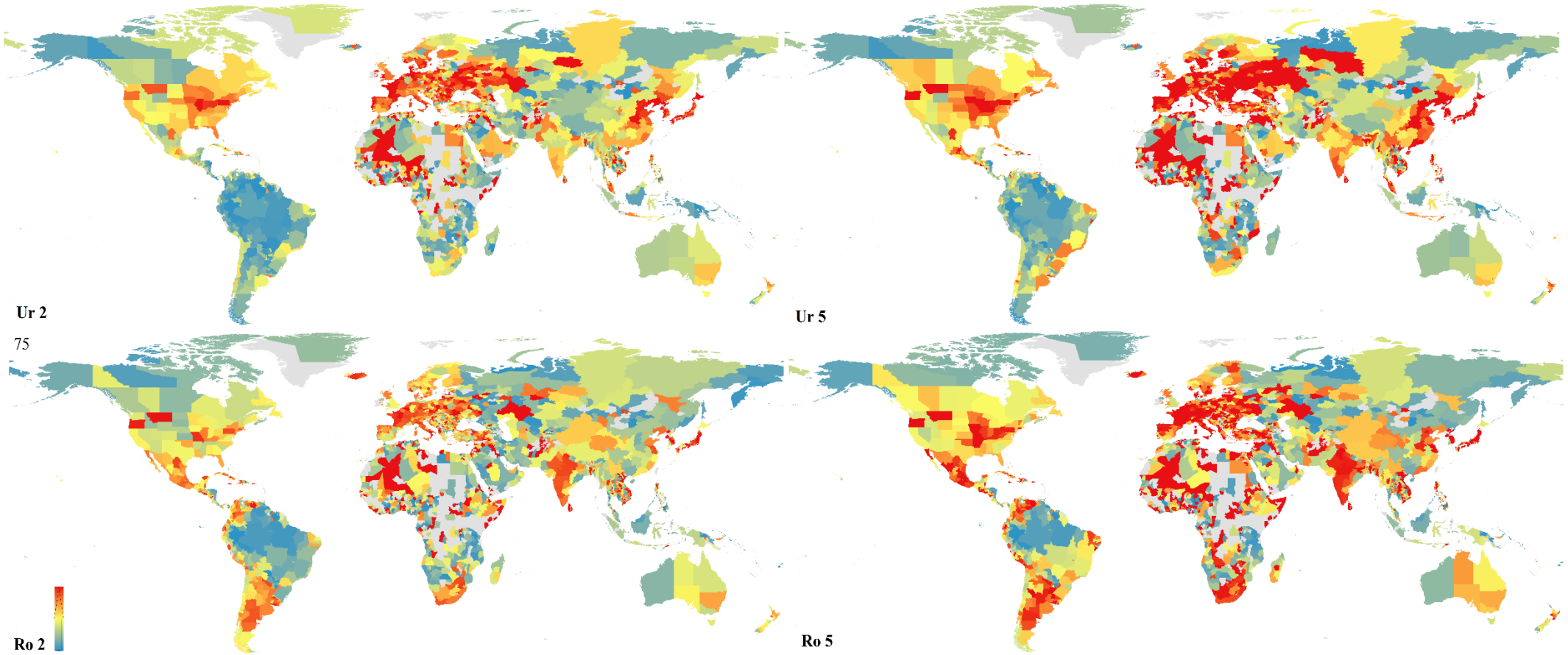


Figure S11c. Percentage of species with at least 75% of points within 2/5km of an urban area or road at a state level. Blue–low, yellow–medium, red–high. Ur-Urban, Ro-Road. See the "***Spatial biases***" section of the main text.

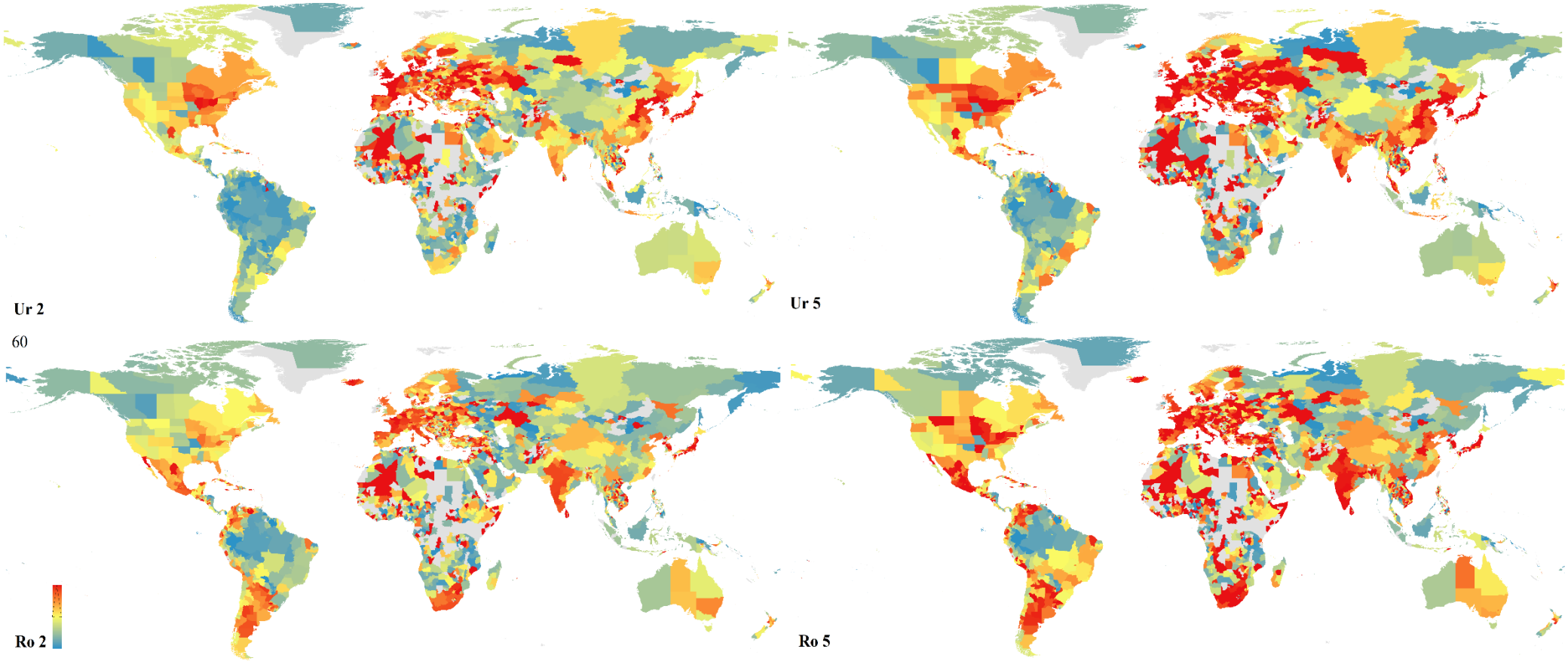


Figure S11d. Percentage of species with at least 60% of points within 2/5km of an urban area or road at a state level. Blue–low, yellow–medium, red–high. Ur-Urban, Ro-Road. See the "***Spatial biases***" section of the main text.

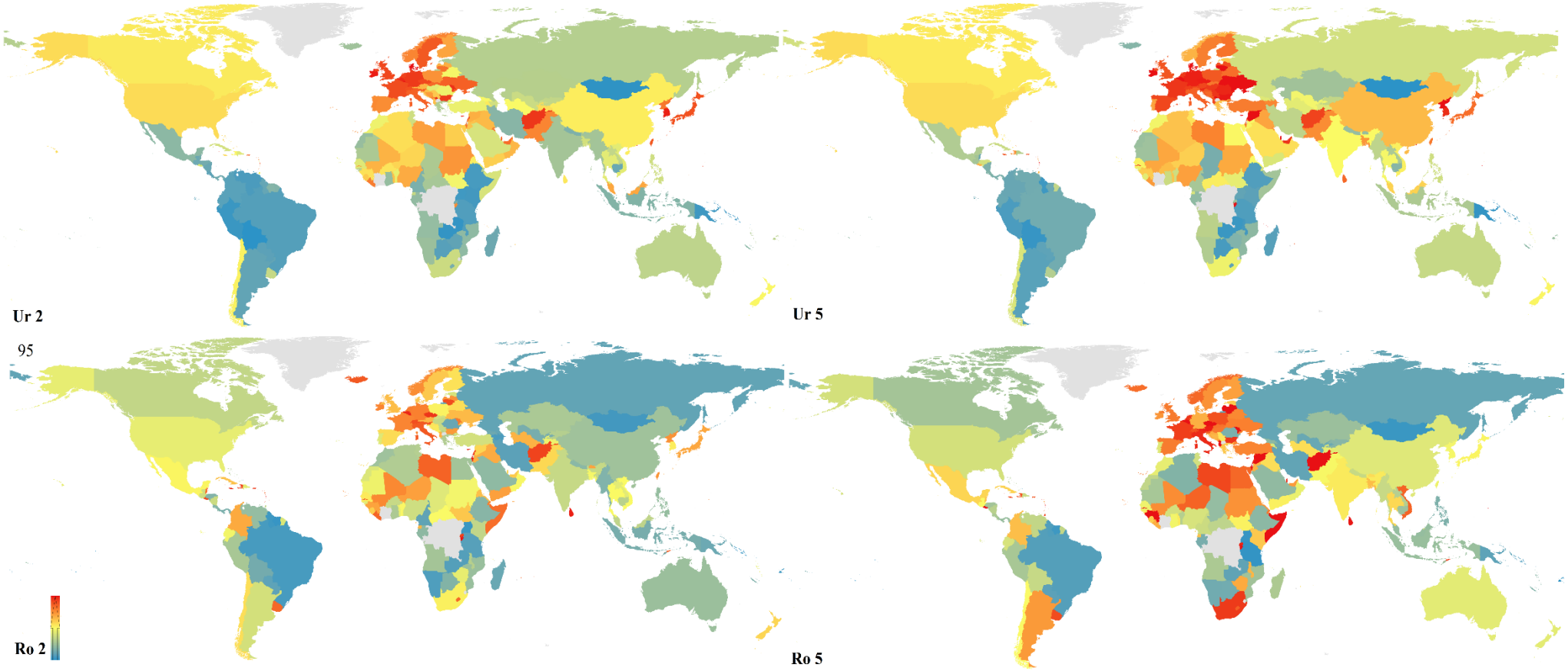


Figure S12a. Percentage of species with at least 95% of points within 2/5km of an urban area or road at a country level. Blue–low, yellow–medium, red–high. Ur-Urban, Ro-Road. Urban areas are likely to have the greatest degree of habitat modification, therefore "red" areas are the most likely to show highly biases data to only having data from the most disturbed areas. See the "***Spatial biases***" section of the main text.

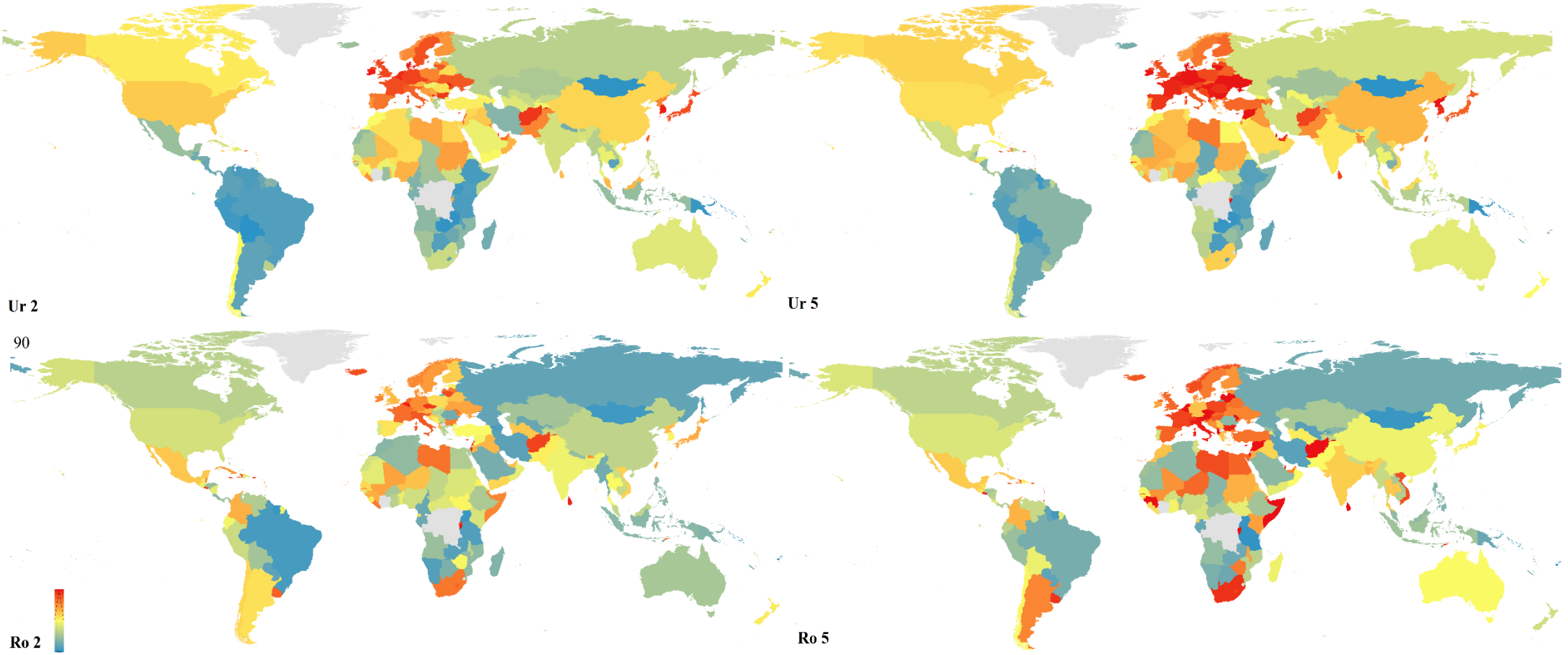


Figure S12b. Percentage of species with at least 90% of points within 2/5km of an urban area or road at a country level. Blue–low, yellow–medium, red–high. Ur-Urban, Ro-Road. See the "***Spatial biases***" section of the main text.

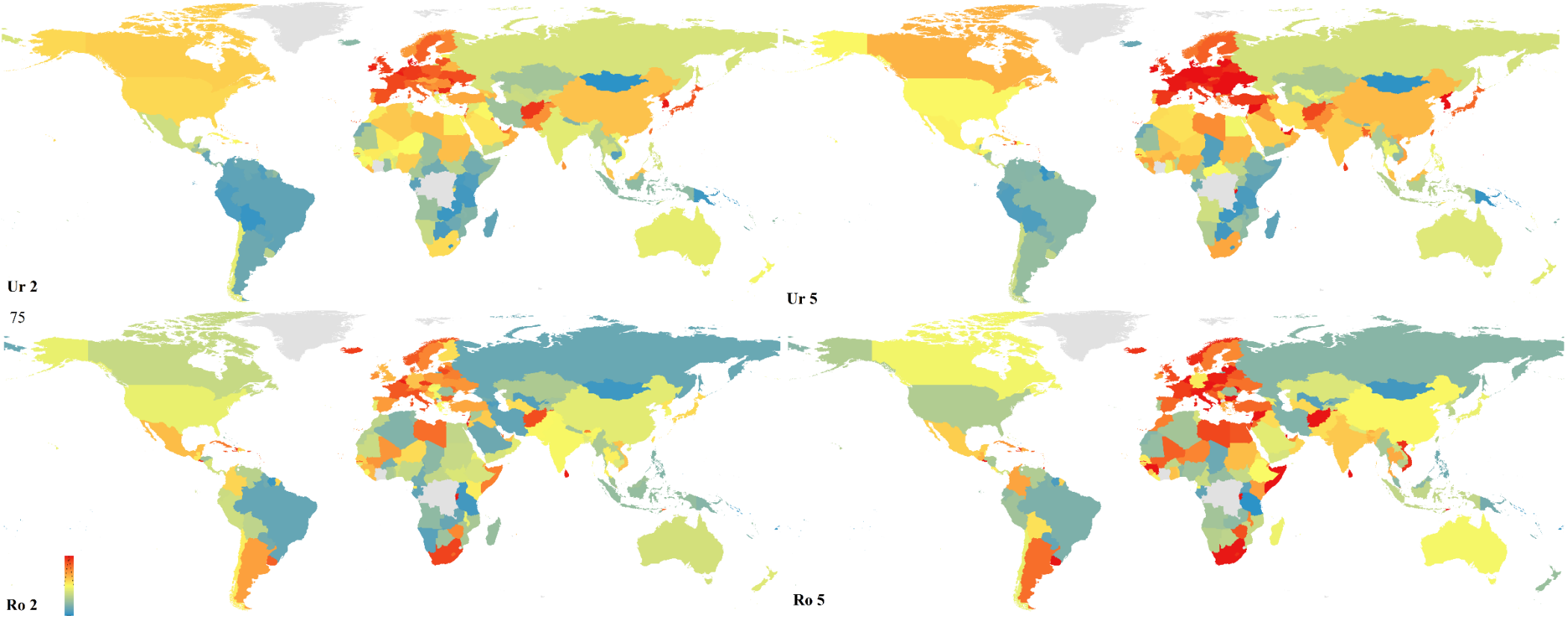


Figure S12c. Percentage of species with at least 75% of points within 2/5km of an urban area or road at a country level. Blue–low, yellow–medium, red–high. Ur-Urban, Ro-Road. See the "***Spatial biases***" section of the main text.

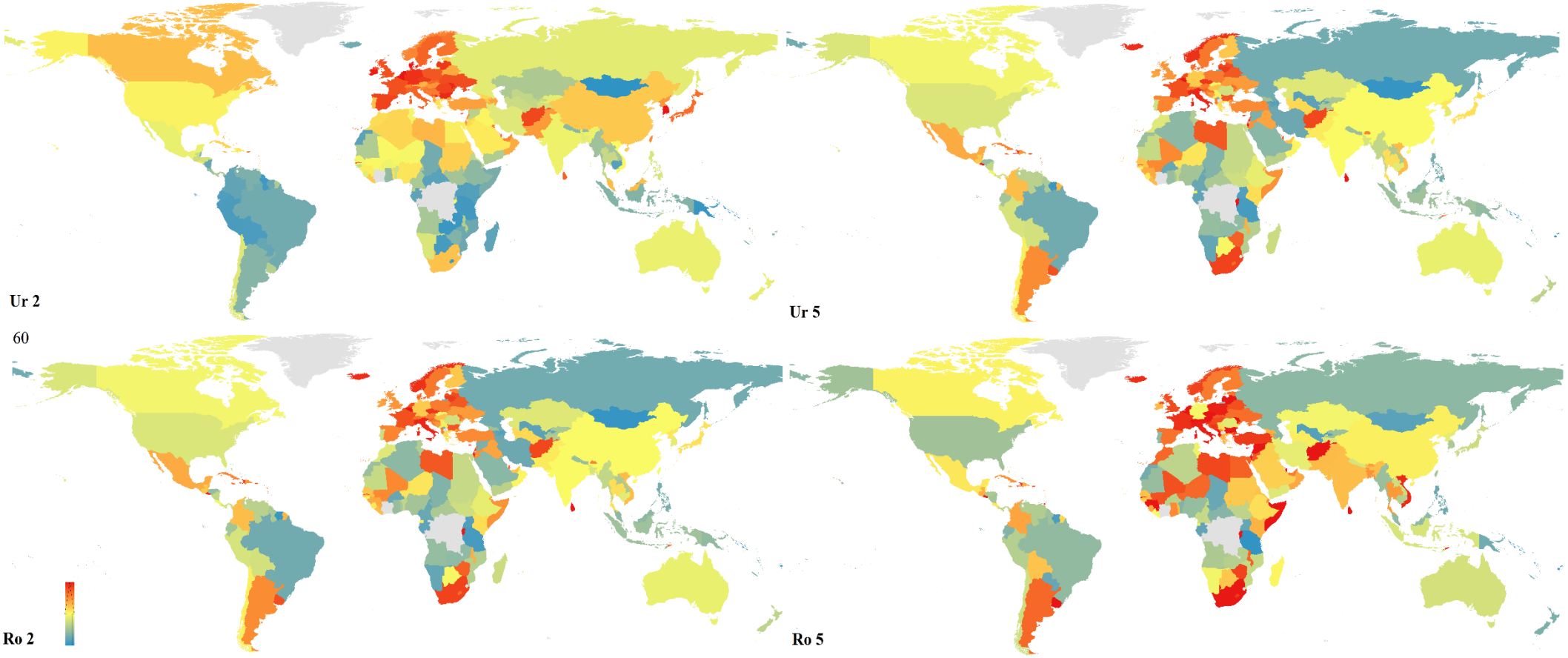


Figure S12d. Percentage of species with at least 60% of points within 2/5km of an urban area or road at a country level. Blue–low, yellow–medium, red–high. Ur-Urban, Ro-Road. See the "***Spatial biases***" section of the main text.

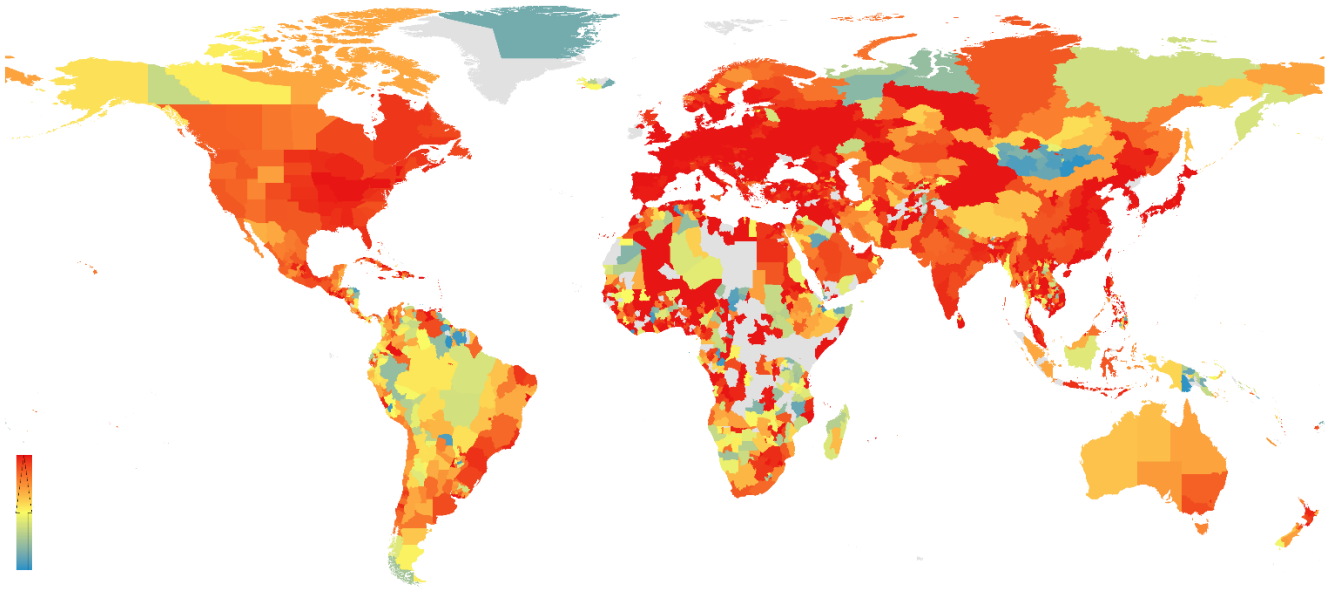


Figure S13a. Percentage of records within 5km of an urban area at a state level. Blue–low, yellow–medium, red–high. See the "***Spatial biases***" section of the main text.

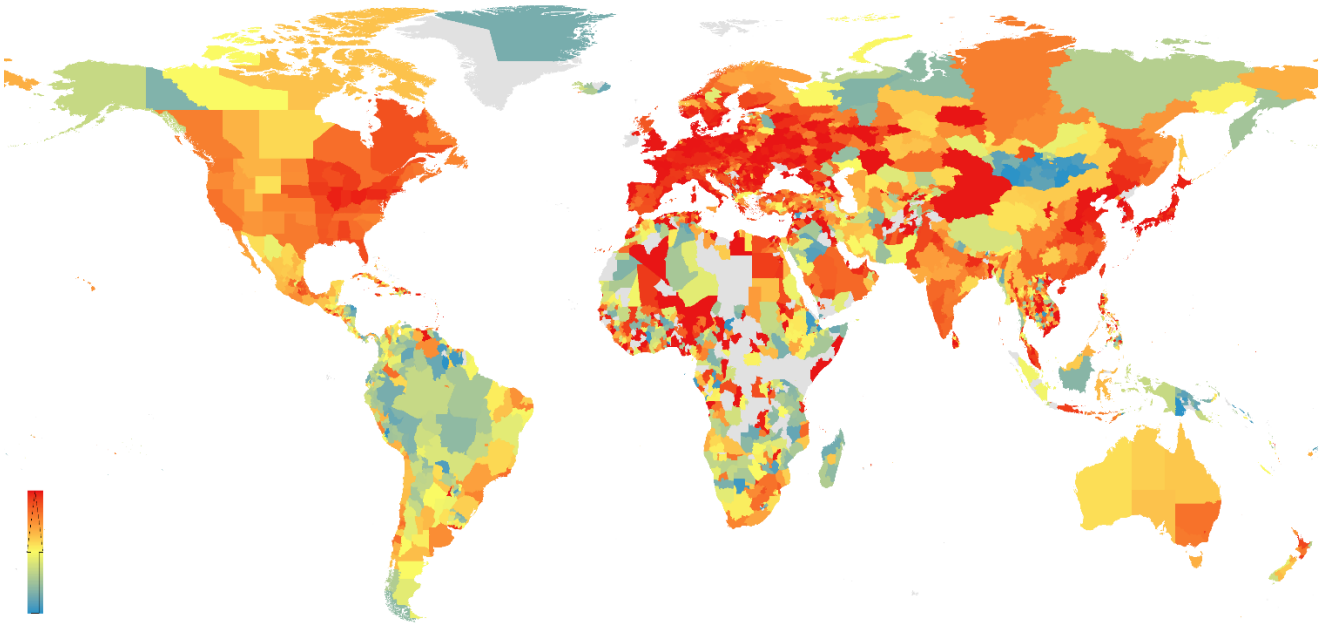


Figure S13b. Percentage of records within 2km of an urban area at a state level. Blue–low, yellow–medium, red–high. See the "***Spatial biases***" section of the main text.

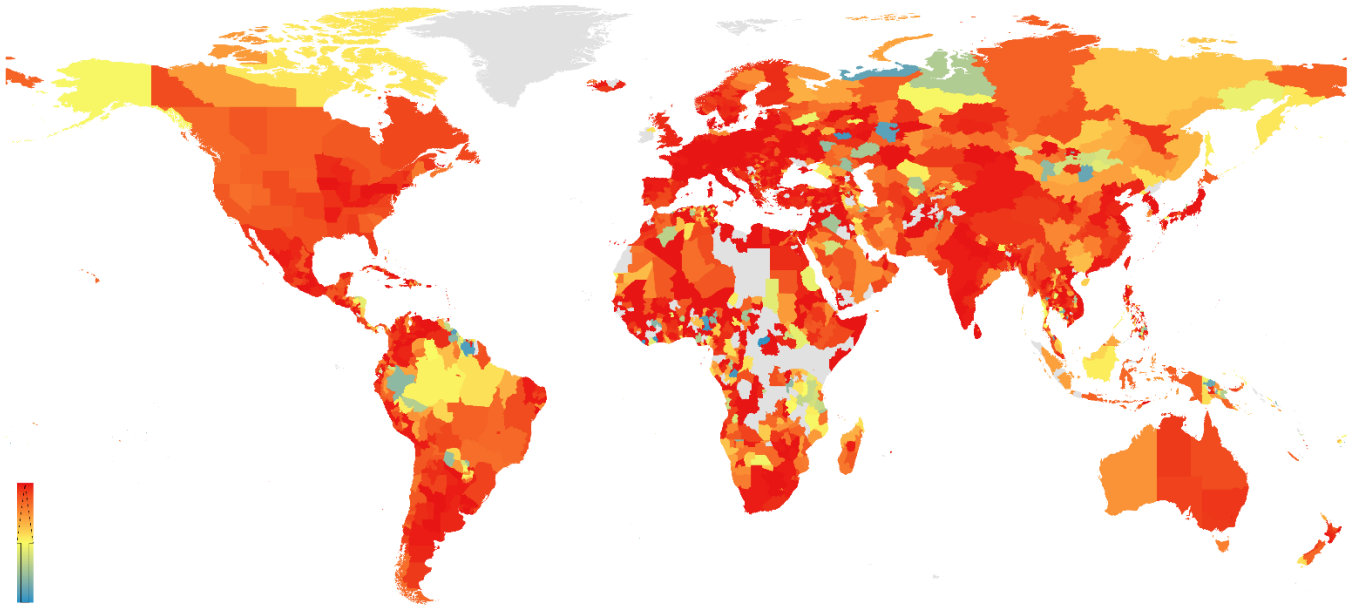


Figure S13c. Percentage of records within 5km of a road area. Blue–low, yellow–medium, red–high. See the "***Spatial biases***" section of the main text.

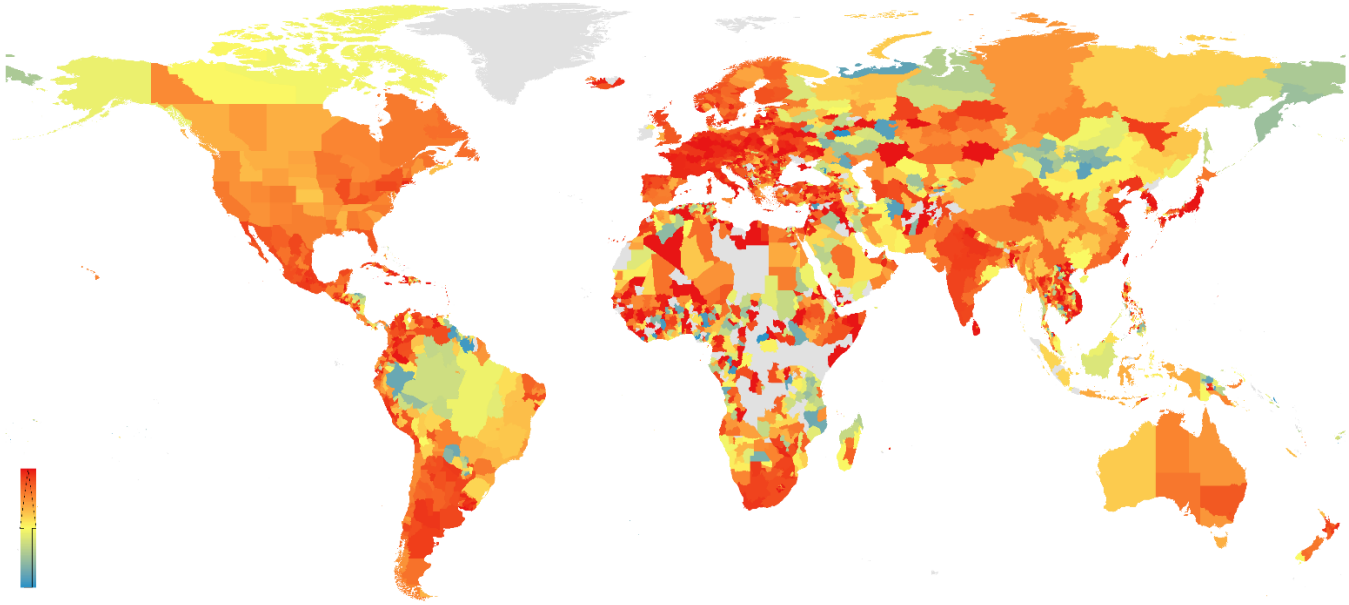


Figure S13d. Percentage of records within 2km of a road area Blue–low, yellow–medium, red–high. See the "***Spatial biases***" section of the main text.

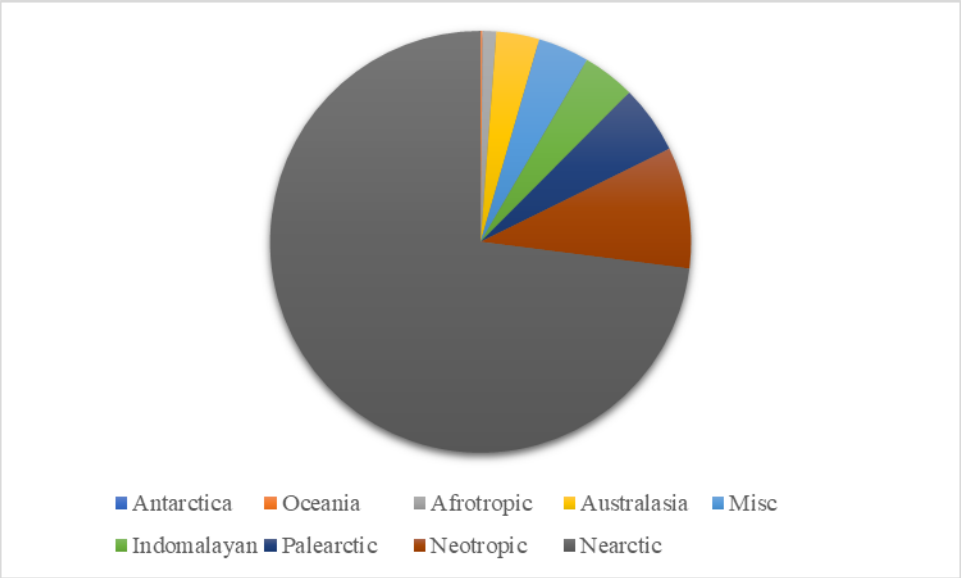


Figure S14. Percentage of bird records per realm. Only 4 species comprise 8% of all data, while 39 species account for 43% of all records. There are 10,359 species in total (0.38% of species makeup 43% of records). It also shows that like the maps, the eBird data is overwhelmingly dominated by North America. This limits the ability to use such data in other regions, as they have a much lower degree of inclusion. See the "***Spatial biases***" section of the main text.

2). Understanding how indices overcome biases

c). Mapping diversity with different metrics figures and table

Figures S15 and S16, in addition to Table S1 are associated with the "***Mapping diversity with different metrics***" section and assess the basic performance of each metric. These are also associated with Figure 3 and Figure 4 in the main text and rank and assess the performance of each index examined.

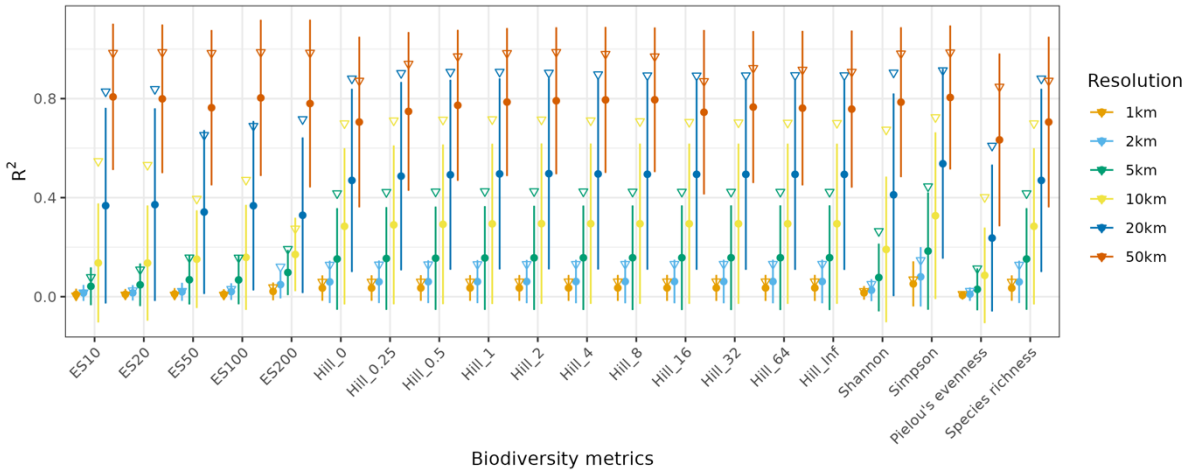


Figure S15a. Overall performance of each metric at each resolution. Triangles indicate random sampling, while circles indicate biased sampling.

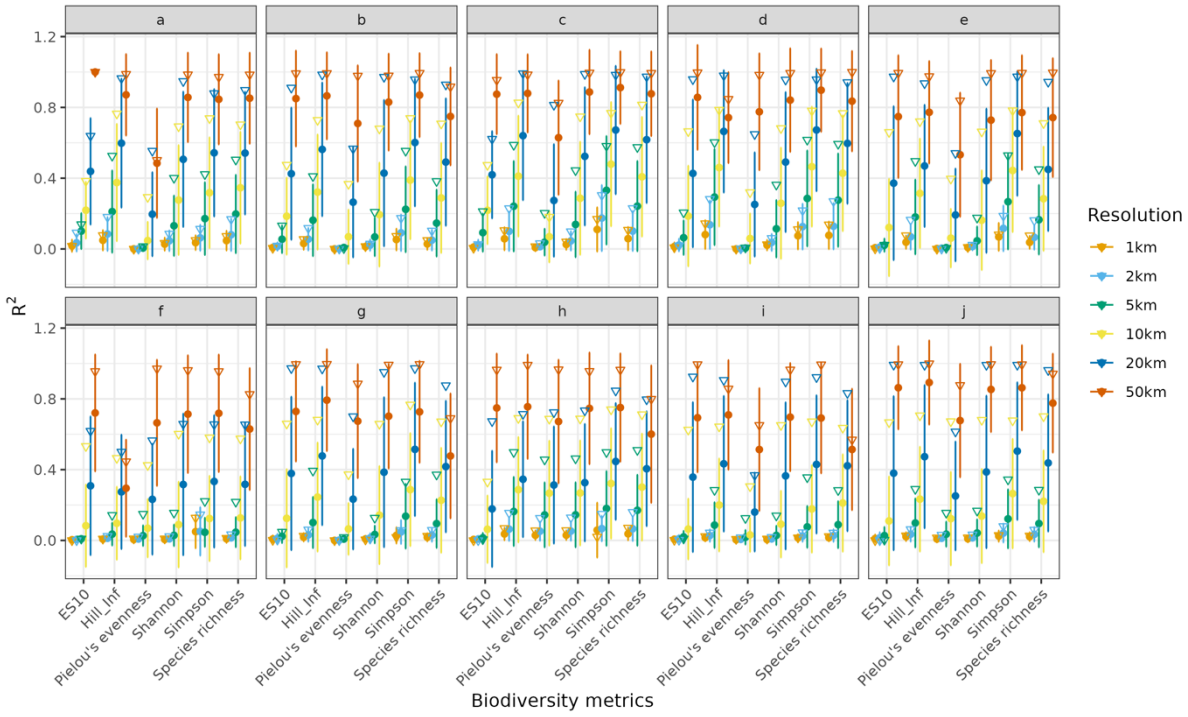


Figure S15b. Performance of each metric for each plot at each resolution. Triangles indicate random sampling, while circles indicate biased sampling.

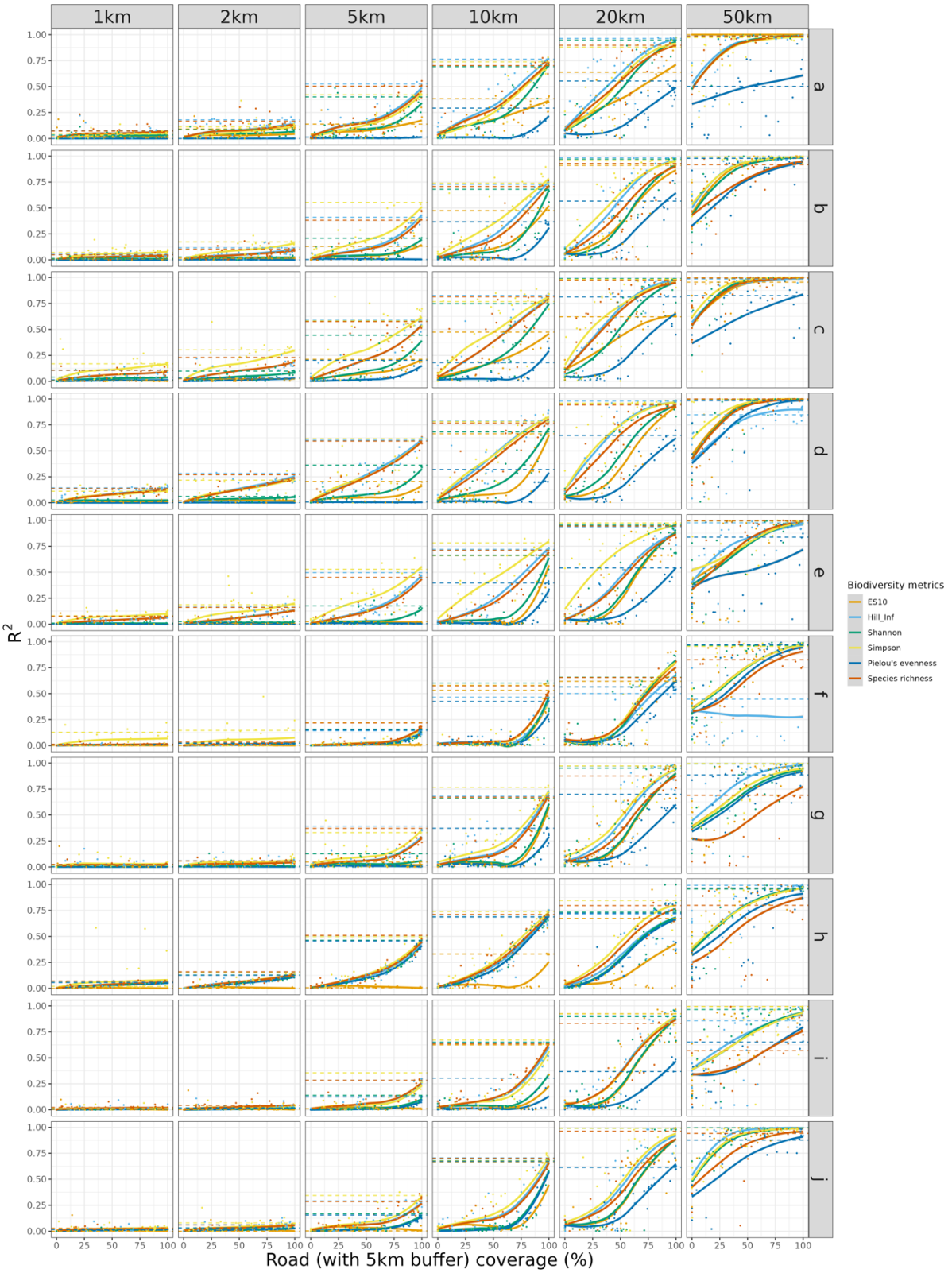


Figure S16. Performance for each plot (A-J) of each metric at each resolution and different densities of road coverage. ES is also known as Hurlbert's index. This highlights the variability between each of the ten plots examined.

Table S1. Average ranks– overall performance. All indices were ranked for performance under each scenario using the biased and random (unbiased) data to assess performance across resolutions. This is used in "***Mapping diversity with different metrics***" to assess the consistency of performance of each of the metrics

	mean_overall	stdev_overall		mean_bias	stdev_bias		mean_random	stdev_random
Simp	2.08	2.19	Simp	1.33	8.63	Hill1	1.50	7.13
Hill1	2.83	2.69	Hill4	2.67	5.89	Hill2	2.67	6.83
Hill2	3.00	1.81	Hill8	2.67	1.17	Simp	2.83	7.19
Hill4	3.50	2.43	Hill2	3.33	6.60	Hill0.50	3.33	6.85
Hill0.50	4.00	4.07	Hill1	4.17	2.34	Hill0.25	4.00	6.47
Hill8	4.58	3.40	Hill0.50	4.67	7.44	Hill4	4.33	5.02
Hill0.25	5.08	5.43	Hill32	5.50	6.46	Hill8	6.50	4.52
Hill32	6.50	5.07	Hill64	5.83	4.13	Hill0	7.00	4.27
Hill64	6.75	5.45	Hill0.25	6.17	3.37	Hill32	7.50	0.55
Hill_Inf	7.00	5.83	Hill_Inf	6.17	2.16	SR	7.50	1.51
Hill16	7.75	7.03	Hill16	7.17	1.21	Hill64	7.67	3.14
Hill0	8.92	6.37	Hill0	10.83	1.21	Hill_Inf	7.83	3.89
SR	9.42	6.58	SR	11.33	8.01	Shann	7.83	6.62
Shann	10.00	5.95	ES100	11.50	5.43	Hill16	8.33	4.97
ES100	12.17	6.45	ES10	12.00	5.95	ES100	12.83	5.28
ES10	13.50	7.70	Shann	12.17	6.46	ES200	13.50	5.60
ES200	13.58	4.64	ES20	13.50	3.25	ES20	14.67	7.49
ES20	14.08	6.11	ES200	13.67	0.52	ES10	15.00	2.99
ES50	15.50	4.93	ES50	15.83	1.63	ES50	15.17	1.21
Pielou	18.50	1.62	Pielou	19.33	7.71	Pielou	17.67	5.21

d). Dealing with small volumes of data figures.

Figures S17-S20 are linked to the "***Dealing with small volumes of data***" section of text, as well as Figure 4 from the main text. Please see text noting each number of sample points in the main text to find the associated plots below.

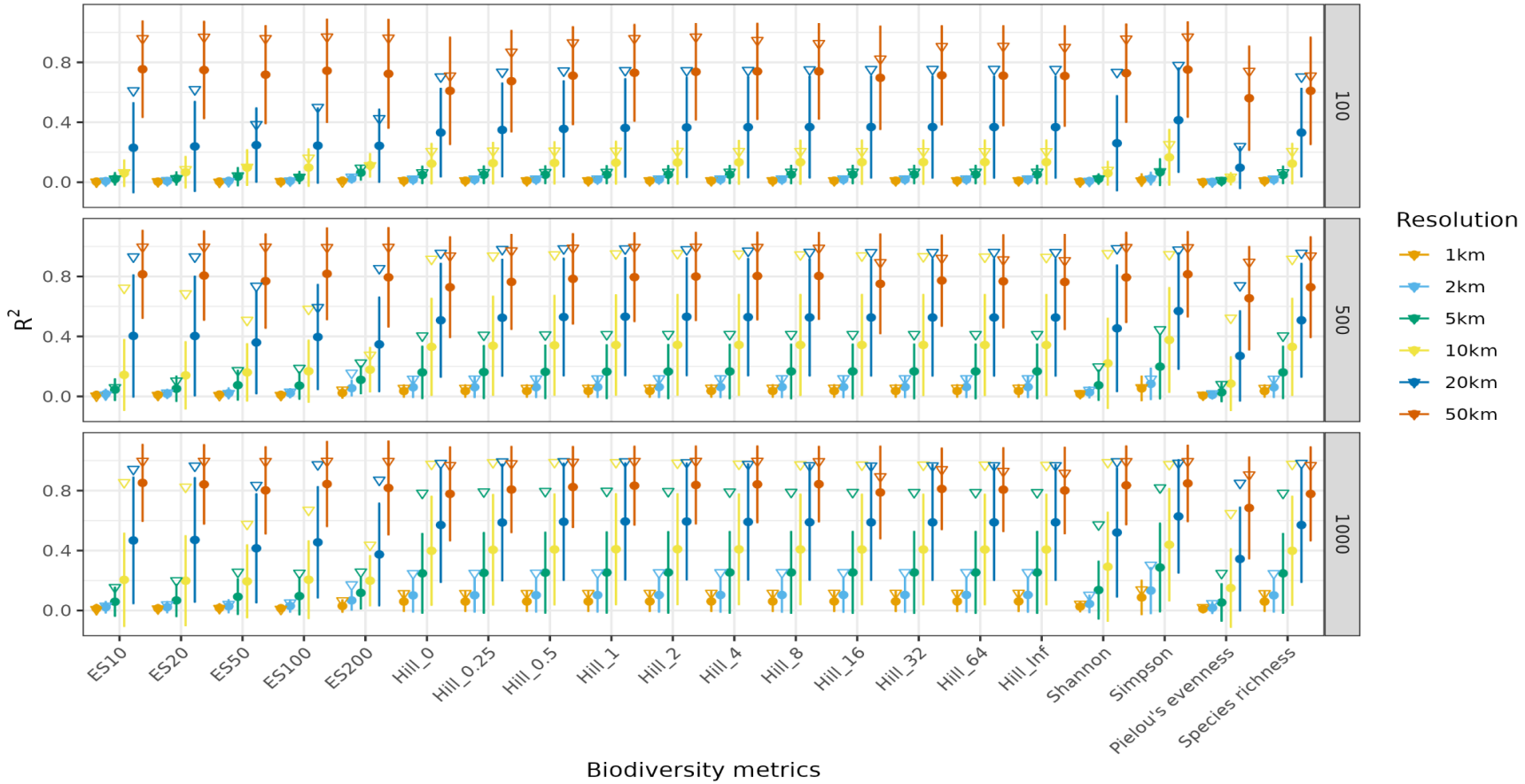


Figure S17. Overall performance of each metric at each resolution, different densities of road coverage, and different sample sizes, including 100, 500 and 1000 records. Triangles indicate random sampling, while circles indicate biased samples. ES is also known as Hurlbert's index. See "***Dealing with small volumes of data***" where random vs biased sampling is discussed.

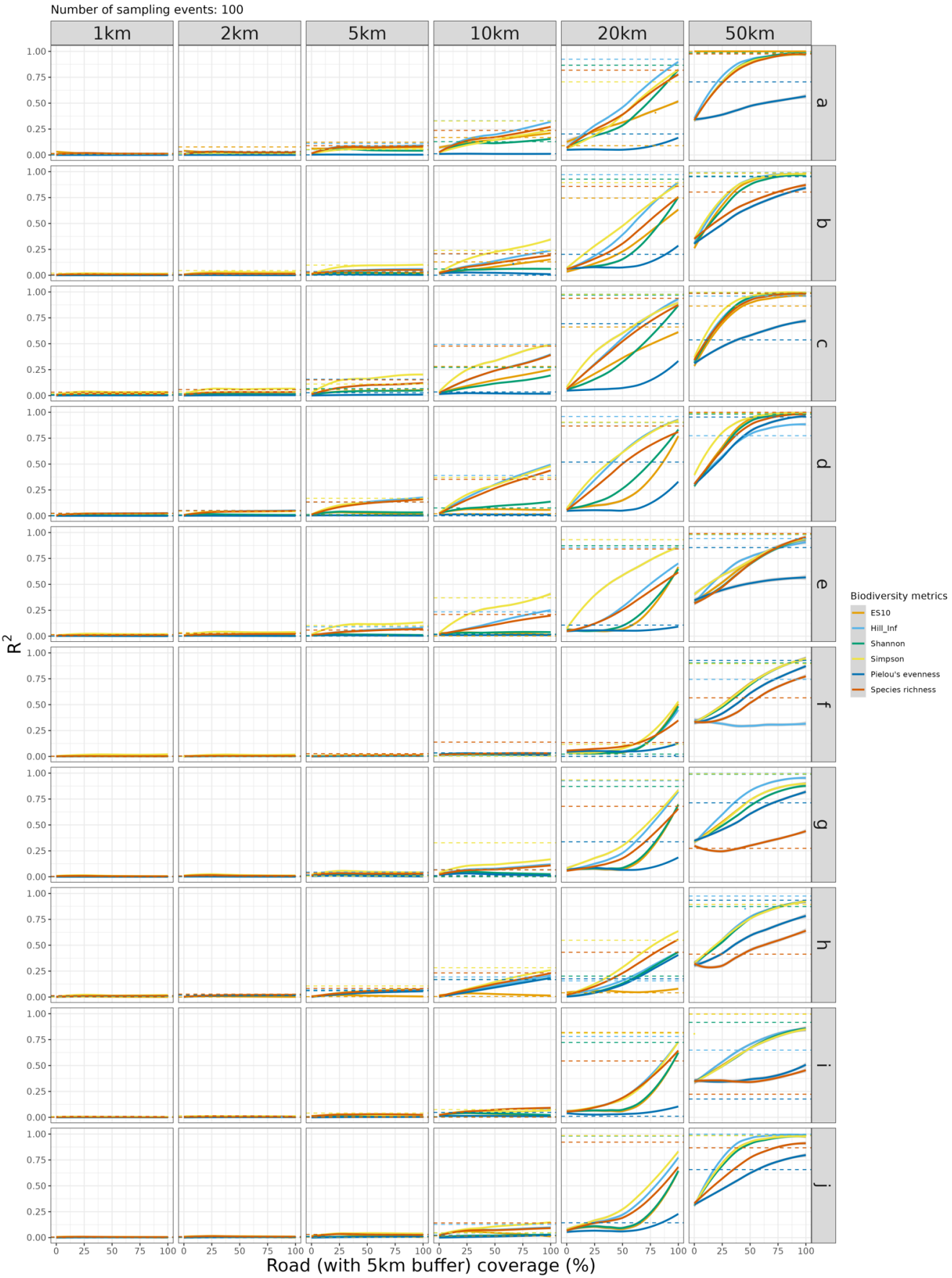


Figure S18. Performance of each metric at each resolution in each plot (A-J), different densities of road coverage, and different sample sizes, including 100 records. ES is also known as Hurlbert's index. See "***Dealing with small volumes of data***". As these plots only have 100 samples this is a fairly low data volume to attempt to reconstruct diversity patterns.

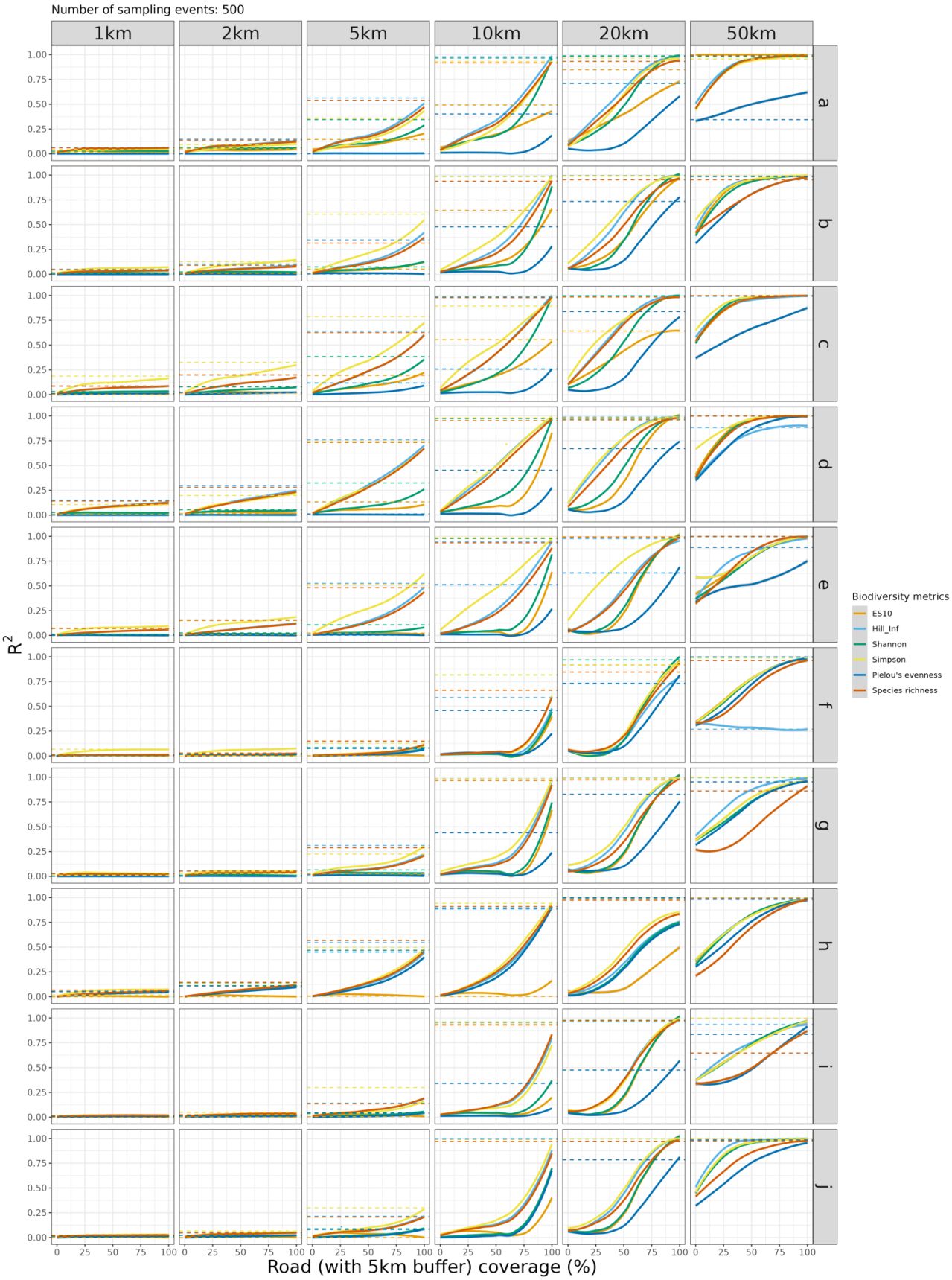


Figure S19. Performance of each metric at each resolution in each plot (A-J), different densities of road coverage, and different sample sizes, including 500 records. ES is also known as Hurlbert's index. See "***Dealing with small volumes of data***". As these plots have 500 samples this is an intermediate data volume to attempt to reconstruct diversity patterns.

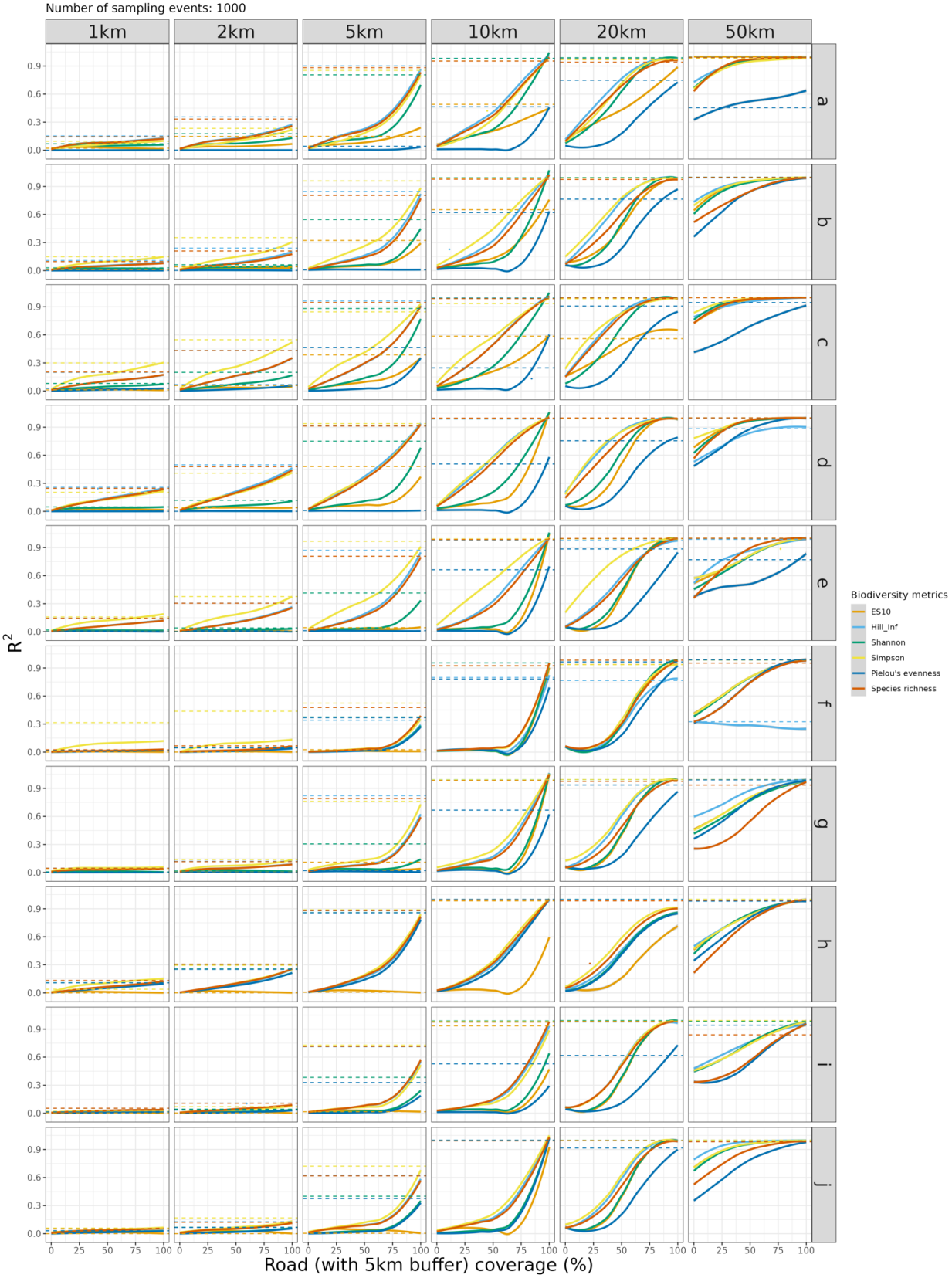


Figure S20. Performance of each metric at each resolution in each plot (A-J), different densities of road coverage, and different sample sizes, including 1000 records. ES is also known as Hurlbert's index. See "***Dealing with small volumes of data***". As these plots have 1000 samples this is a fairly high data volume to attempt to reconstruct diversity patterns.

3). Mapping and Modelling diversity

e). Mapping diversity figures.

The below section is linked to the "***Mapping diversity***" part of the main text and includes Figures S21 and S22. These figures assess the accuracy of projection of each index by looking at the percentage area at each category of richness (very low- very high) and to assess different combinations of indices, and models at both a 5 and 10km resolution globally. It should be noted that these predictions are based on the same biased data highlighted clearly by Figure S14, and thus the ability to fairly accurately predict diversity to less sampled regions and overcome biases is a mark of reasonably good performance, though biases are further discussed below, and in the main text.

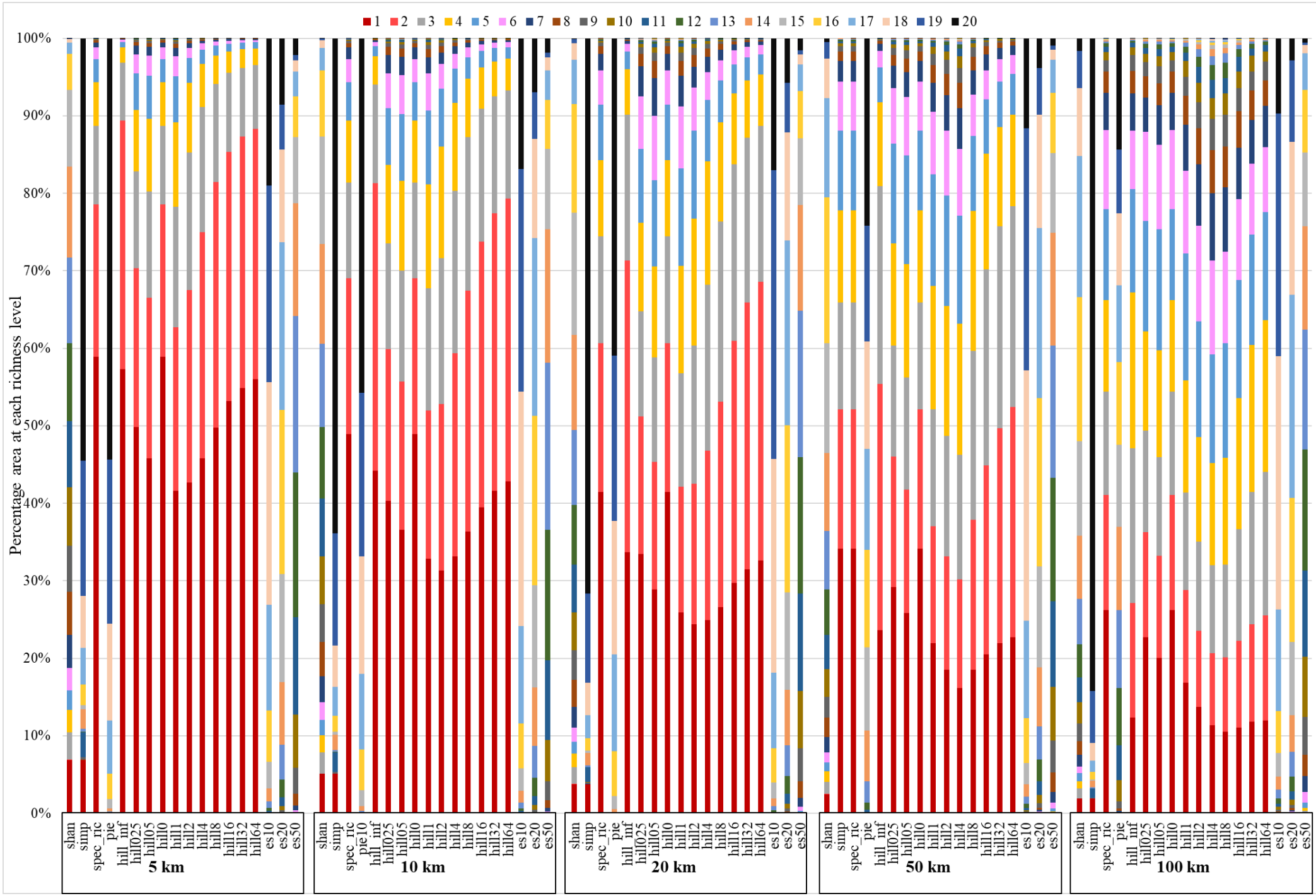

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
Percentage area at each richness level
shan
simp
spec_ric
pie
hill_inf
hill025
hill05
hill0
hill1
hill2
hill4
hill8
hill116
hill32
hill64
es10
es20
es50
5 km
10 km
20 km
50 km
100 km

Figure S21. The percentage area of each level of diversity from 1 (low diversity) to 20 (highest) at resolutions of 1, 2, 5, 10, 20, and 50km based on eBird data, which was used as basis to subsample for the Maxent modelling. Generally higher resolutions are preferred for models, as high resolutions better represent ecological patterns, (such as 5–10km), whereas coarse resolutions may include highly heterogeneous regions. Small areas projected with the lowest diversity may indicate over–prediction as low levels of diversity are expected in some regions. This figure is noted in the "***Mapping diversity***" section of the main text.

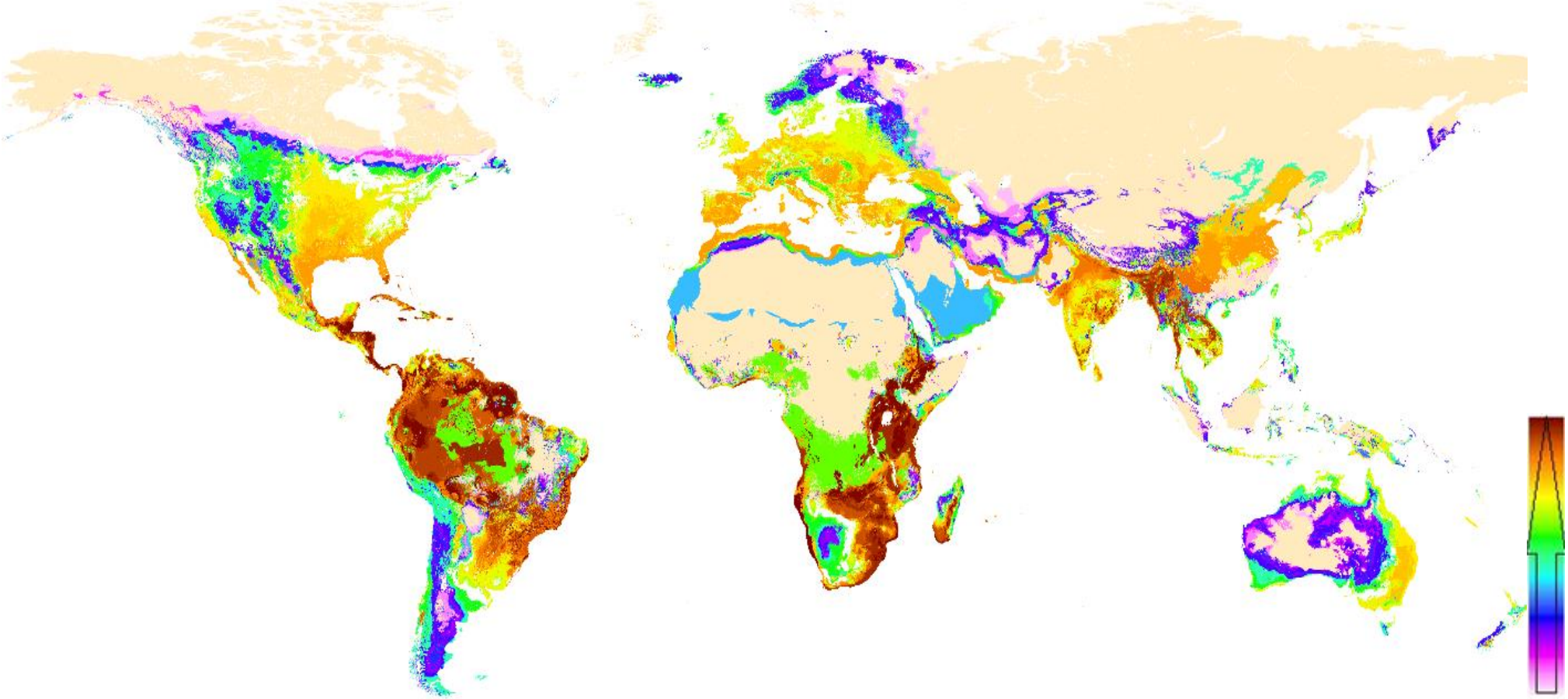


Figure S22a. Mapped diversity assembled from each of the regions modelled independently at 5km, many regions worked less well, though mainland Southeast Asia may have worked better. See Figure 5 for the 10km resolution.

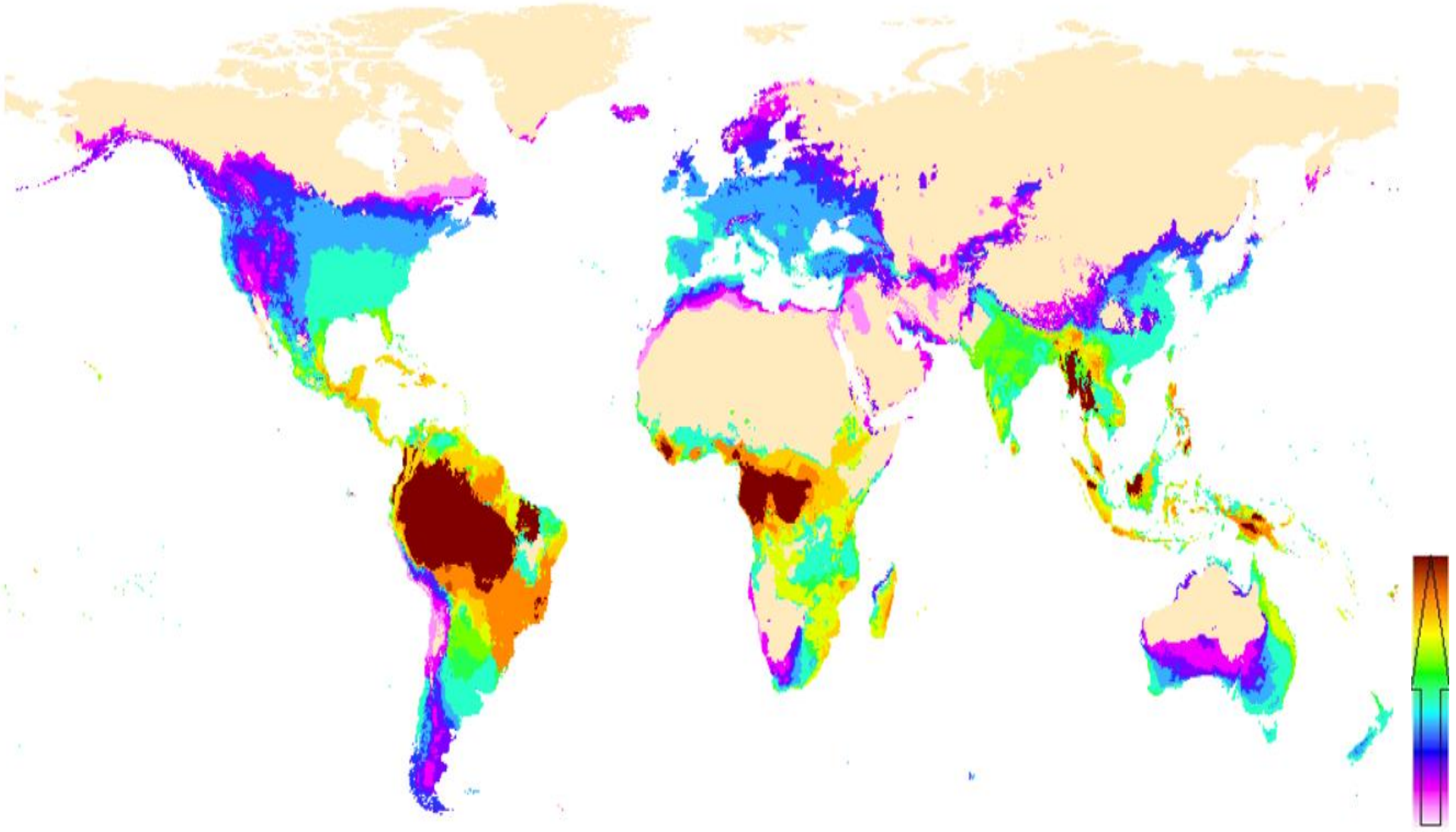


Figure S22b. Mapped diversity based on “good” metrics at 10km, a more nuanced version combining good and all metrics with different weightings performed best overall (Figure 5).

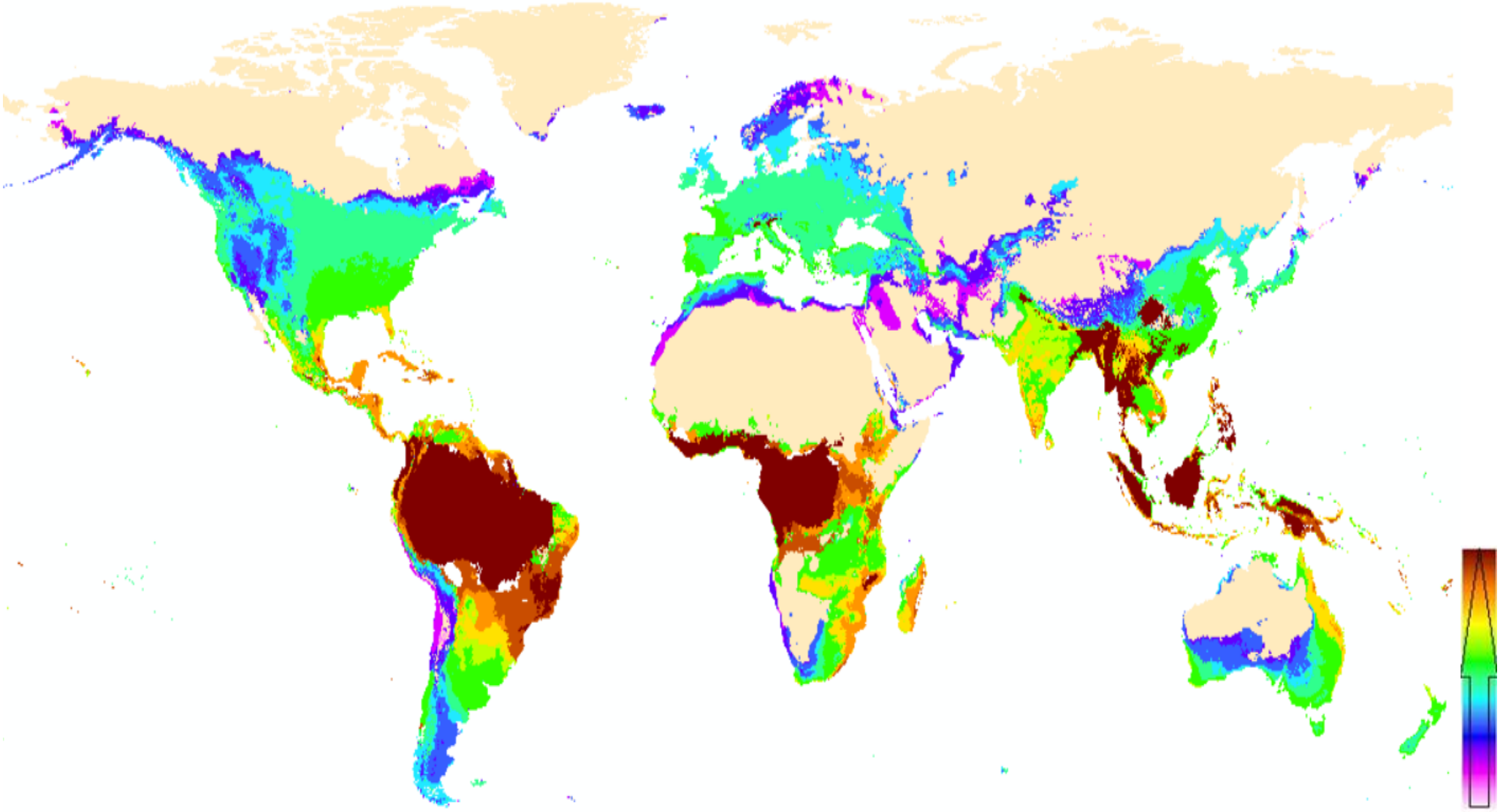
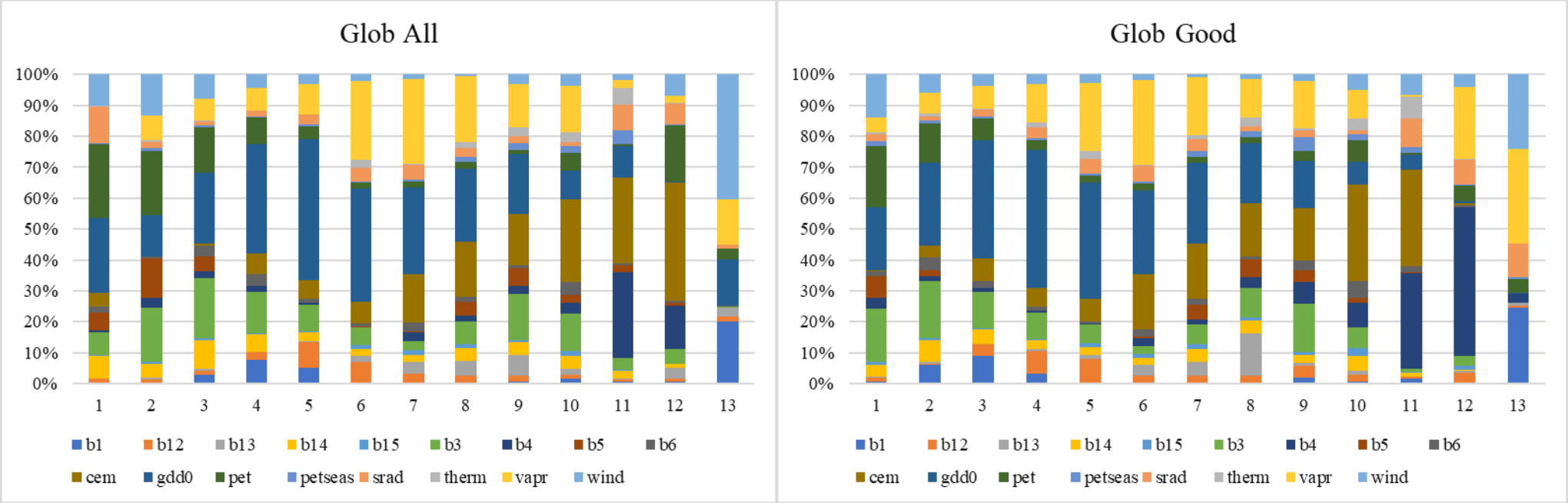


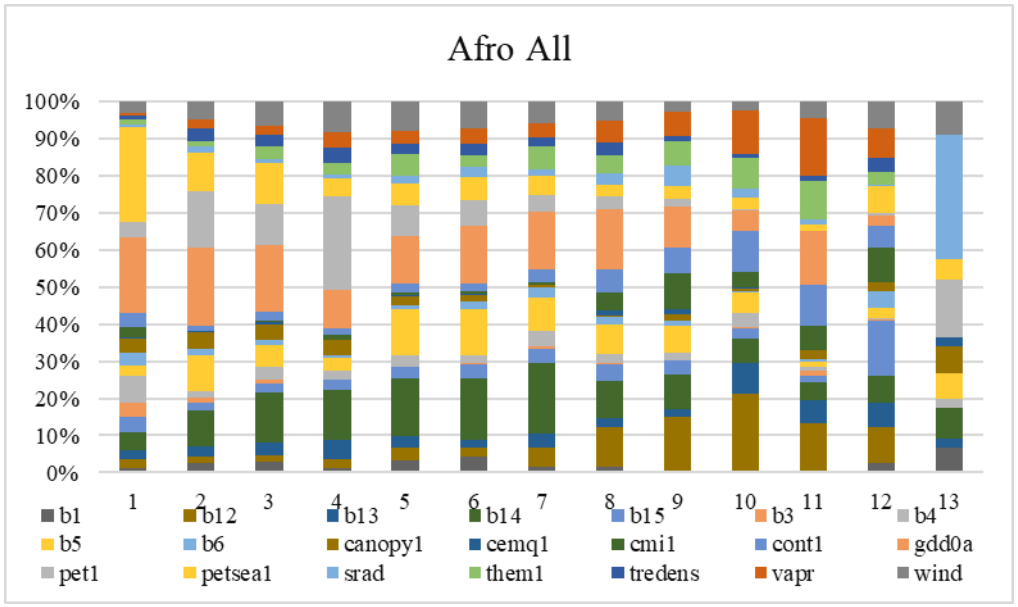
Figure S22c. Mapped diversity based on “all” metrics at 10km, a more nuanced version combining good and all metrics with different weightings performed best overall (Figure 5).

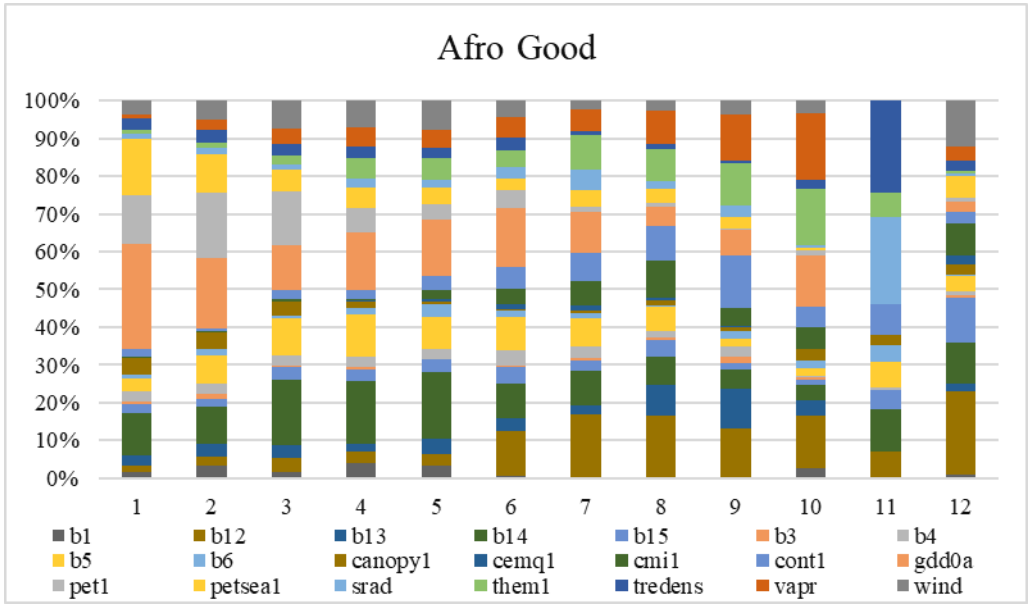
4). Supplemental Analysis

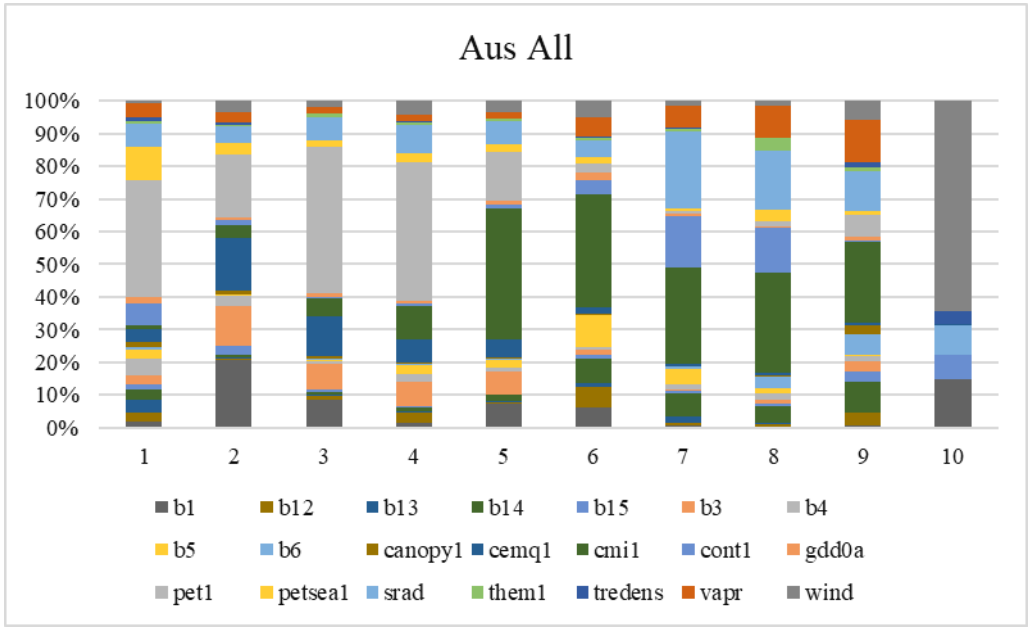
f). Regional assessments and driver analysis figures.

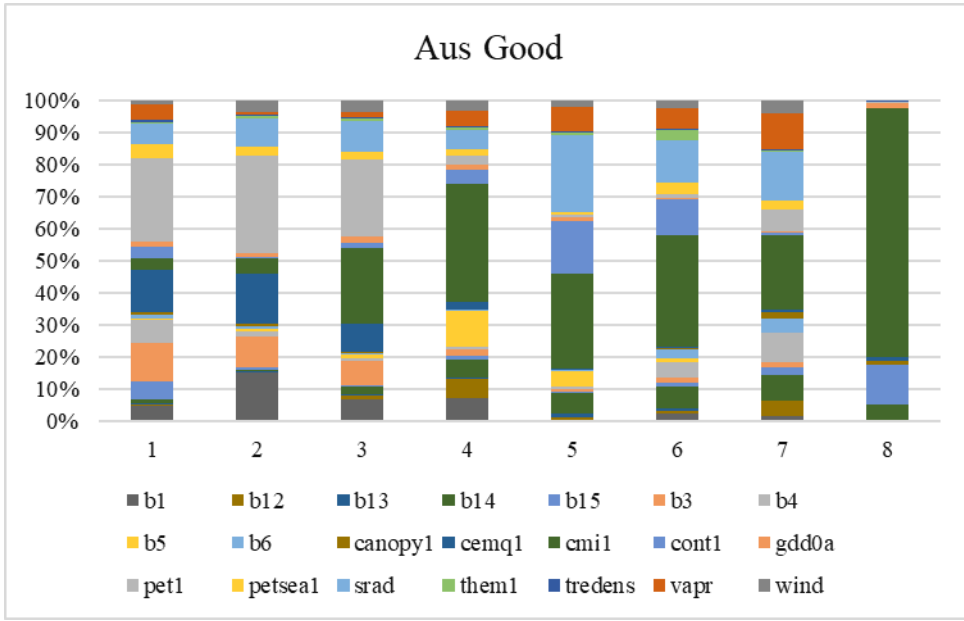
Figures S21-S23 provides analysis of the drivers of low-high diversity for all metrics, and the various combinations of metrics was conducted globally and regionally to understand the nuanced drivers of richness. This section is only discussed in the supplemental results text "***Regional drivers of diversity based on distribution modelling***".

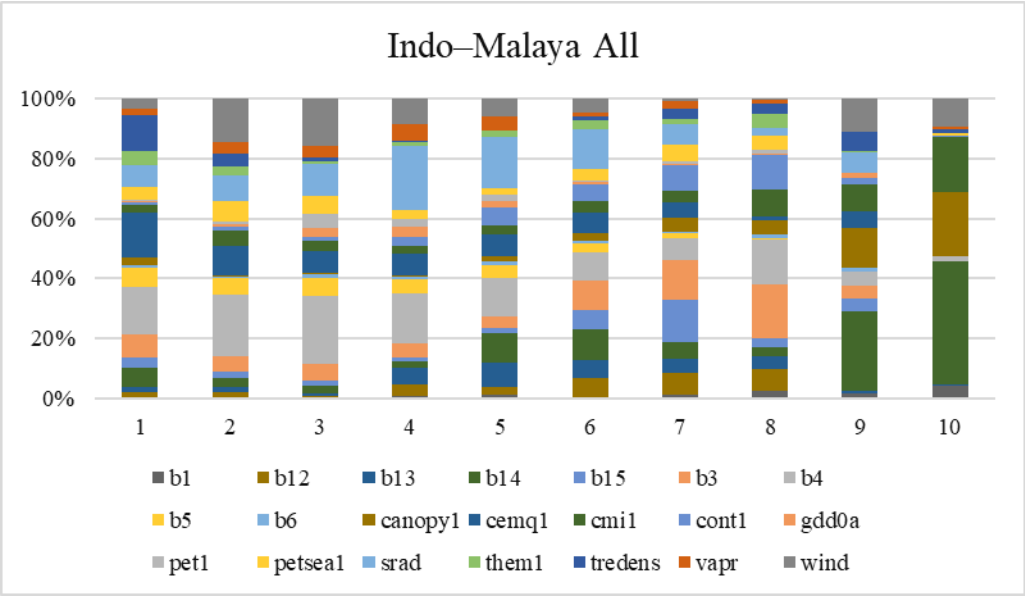


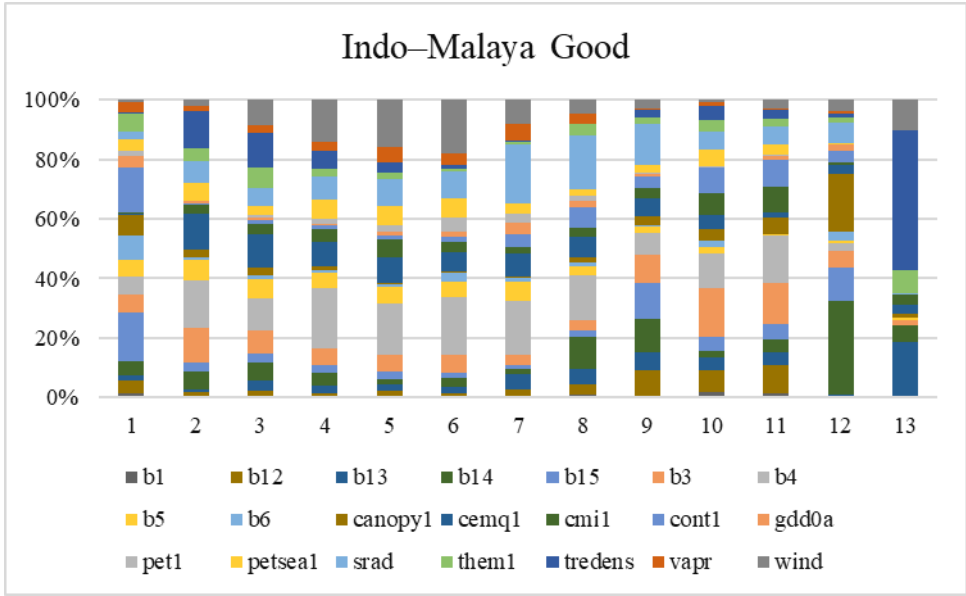

Afro All
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
1 2 3 4 5 6 7 8 9 10 11 12 13
b1 b12 b13 b14 b15 b3 b4
b5 b6 canopy1 cemq1 cmi1 cont1 gdd0a
pet1 petsea1 srad them1 tredens vapr wind

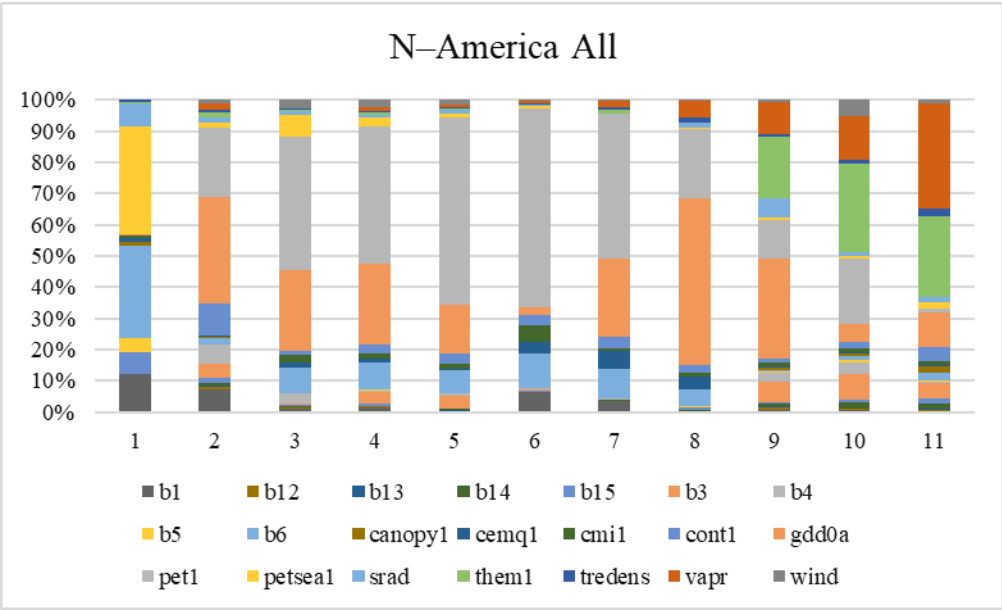

Afro Good
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
1 2 3 4 5 6 7 8 9 10 11 12
b1 b12 b13 b14 b15 b3 b4
b5 b6 canopy1 cemq1 cmi1 cont1 gdd0a
pet1 petsea1 srad them1 tredens vapr wind


Aus All
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
1 2 3 4 5 6 7 8 9 10
b1 b12 b13 b14 b15 b3 b4
b5 b6 canopy1 cemq1 cmi1 cont1 gdd0a
pet1 petsea1 srad them1 tredens vapr wind


Aus Good
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
1 2 3 4 5 6 7 8
b1 b12 b13 b14 b15 b3 b4
b5 b6 canopy1 cemq1 cmi1 cont1 gdd0a
pet1 petsea1 srad them1 tredens vapr wind


Indo–Malaya All
100%
80%
60%
40%
20%
0%
1 2 3 4 5 6 7 8 9 10
b1 b12 b13 b14 b15 b3 b4
b5 b6 canopy1 cemq1 cmi1 cont1 gdd0a
pet1 petsea1 srad them1 tredens vapr wind


Indo–Malaya Good
100%
80%
60%
40%
20%
0%
1 2 3 4 5 6 7 8 9 10 11 12 13
b1 b12 b13 b14 b15 b3 b4
b5 b6 canopy1 cemq1 cmi1 cont1 gdd0a
pet1 petsea1 srad them1 tredens vapr wind


N–America All
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
1 2 3 4 5 6 7 8 9 10 11
b1 b12 b13 b14 b15 b3 b4
b5 b6 canopy1 cemq1 cmi1 cont1 gdd0a
pet1 petsea1 srad them1 tredens vapr wind

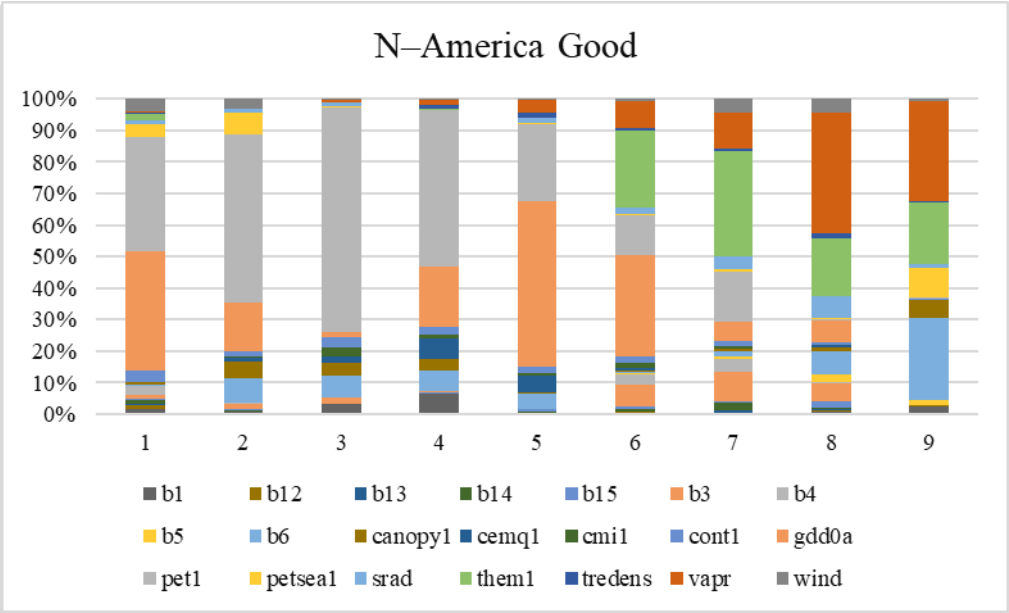

N–America Good
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
1 2 3 4 5 6 7 8 9
b1 b12 b13 b14 b15 b3 b4
b5 b6 canopy1 cemq1 cmi1 cont1 gdd0a
pet1 petsea1 srad them1 tredens vapr wind

Figure S23. Diversity drivers globally (at 10km) and regionally (at 5km), with diversity shown at successive diversity levels.

g). Dealing with small volumes of data (retaining sampled and unsampled cells) figures.

Figures S24- S27 show the difference in performance analysis when unsampled cells are not properly accounted for, highlighting the importance for careful analysis of such cells (and not assuming they have a default of zero). Overall patterns are generally maintained, but caution is needed when applying any such analysis to ensure these results are dealt with carefully and appropriately. These are detailed in supplemental results "***Dealing with small volumes of data (retaining sampled and unsampled cells)"*** and are not noted in the main text as these are here as a cautionary comparison.

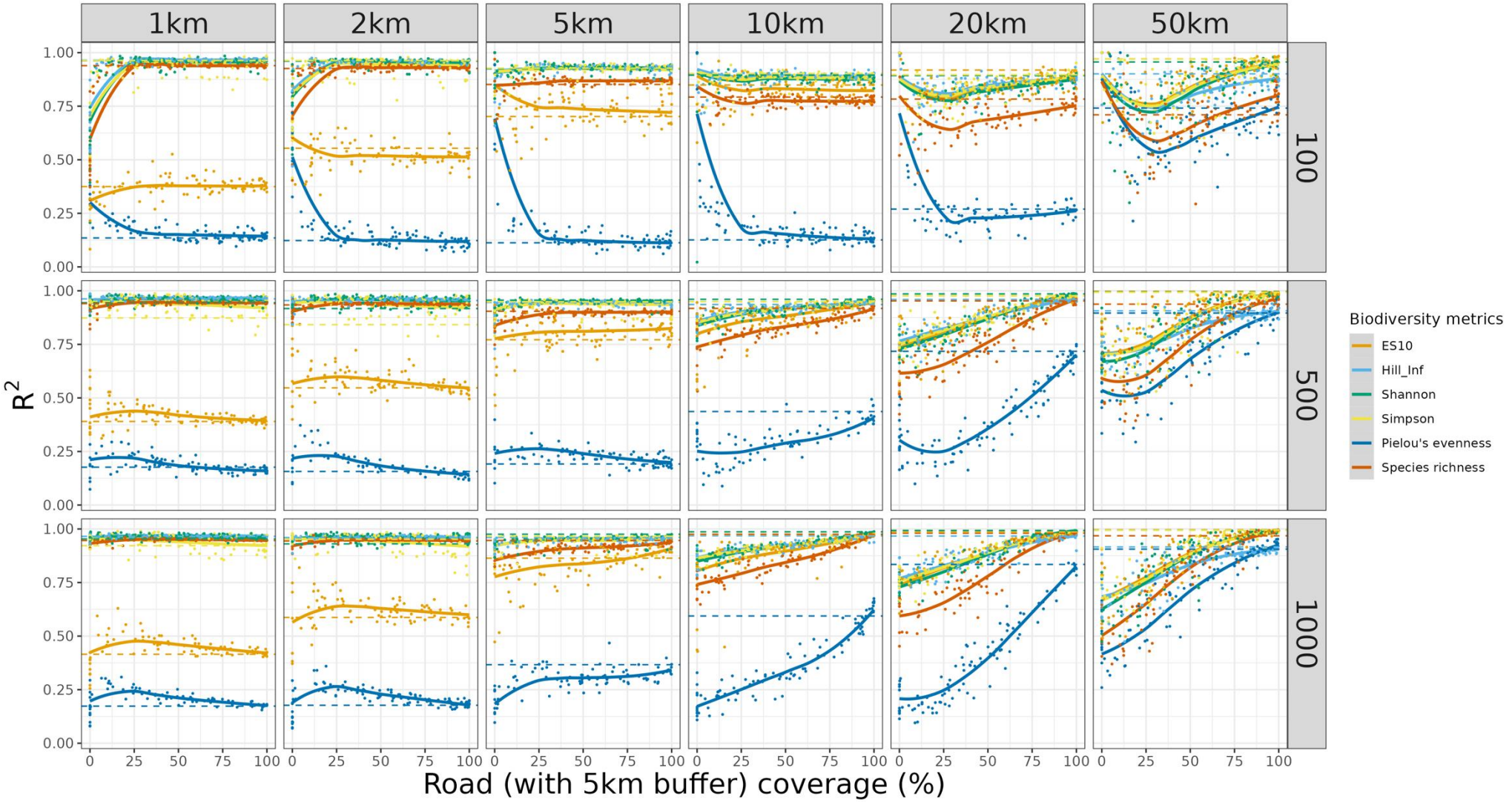
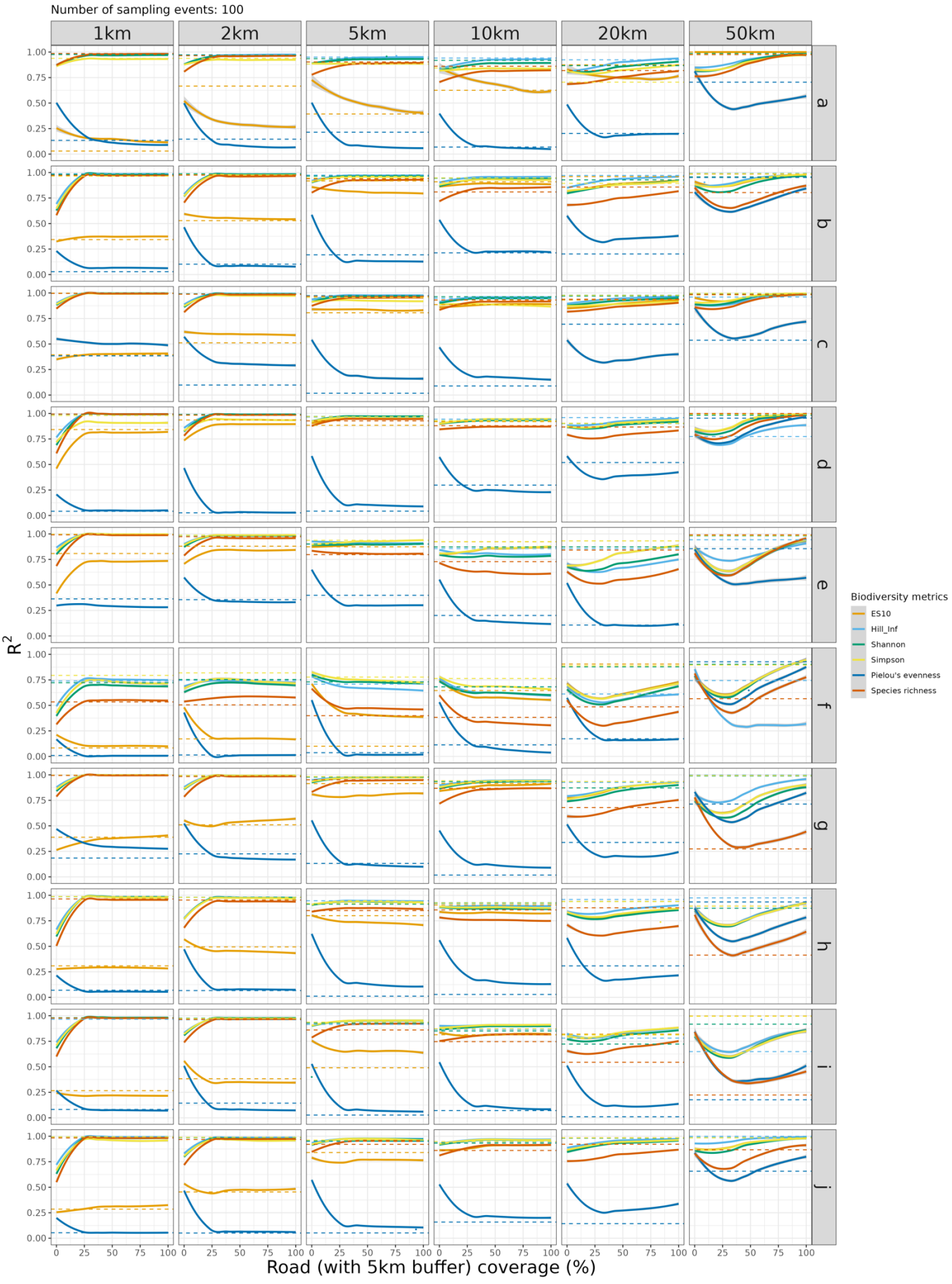
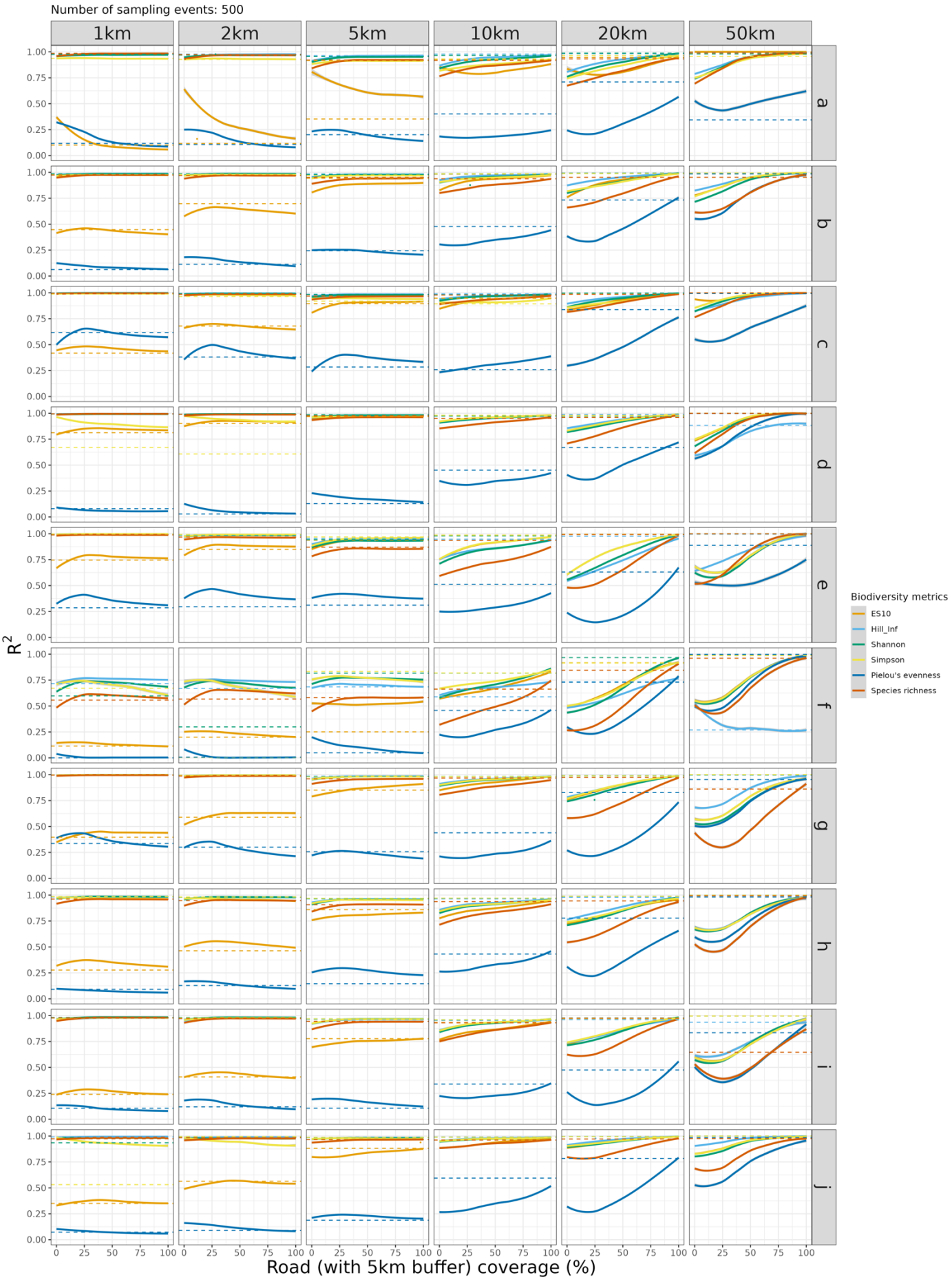


Figure S24. Overall performance of each metric at each resolution, different densities of road coverage, and different sample sizes, including 100 records, 500 records and 1000 records (whilst not accounting for unsampled cells).


Number of sampling events: 100
1km
2km
5km
10km
20km
50km
a
b
c
d
e
f
g
h
i
j
R2
1.00
0.75
0.50
0.25
0.00
0
25
50
75
100
Road (with 5km buffer) coverage (%)
Biodiversity metrics
ES10
Hill_Inf
Shannon
Simpson
Pielou's evenness
Species richness


Number of sampling events: 500
1km
2km
5km
10km
20km
50km
a
b
c
d
e
f
g
h
i
j
R2
1.00
0.75
0.50
0.25
0.00
0
25
50
75
100
Road (with 5km buffer) coverage (%)
Biodiversity metrics
ES10
Hill_Inf
Shannon
Simpson
Pielou's evenness
Species richness

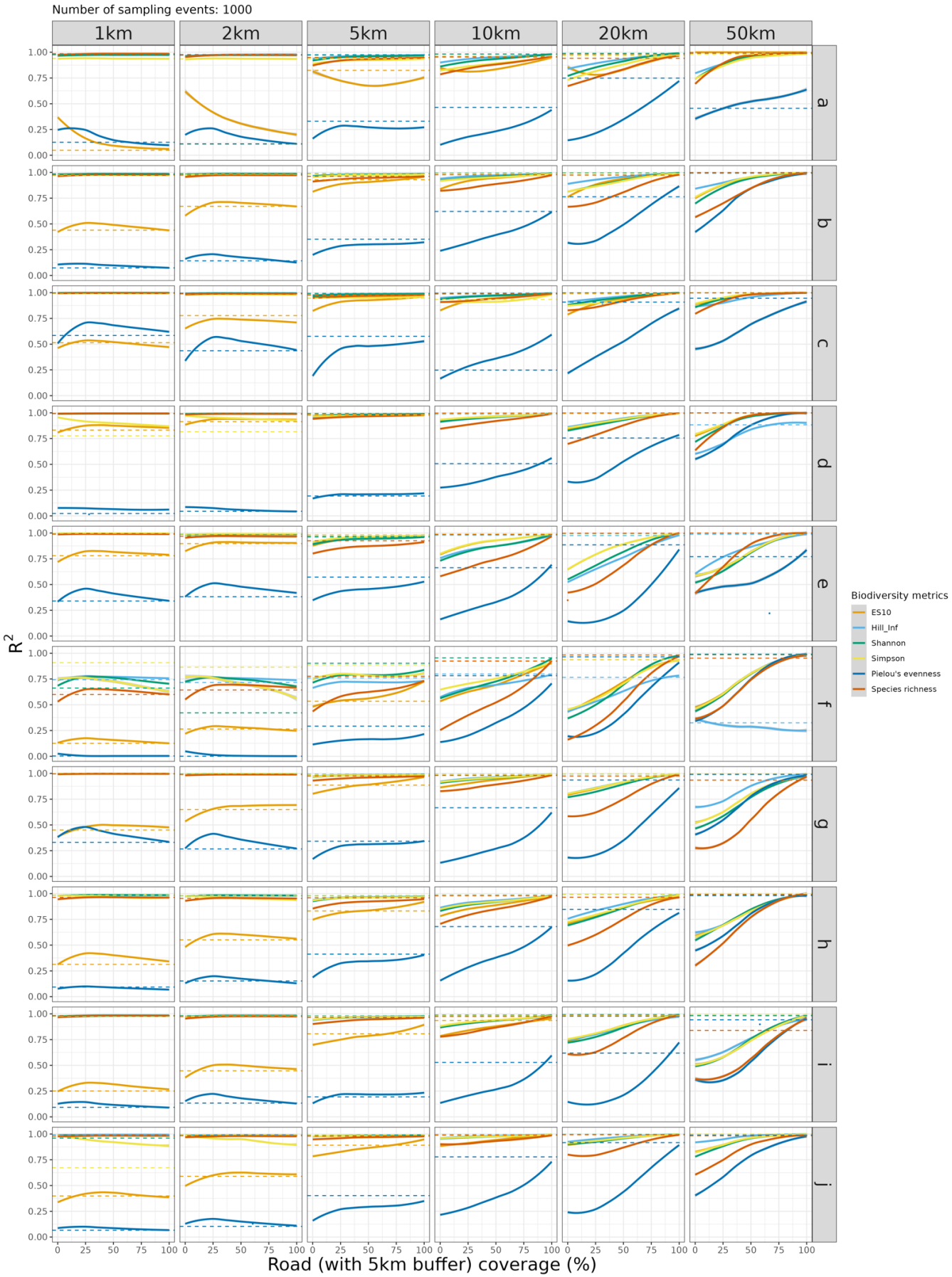


Figure S25. Performance for each plot of each metric at each resolution, different densities of road coverage, and different sample sizes, including 100 records, 500 records and 1000 records (whilst not accounting for unsampled cells).

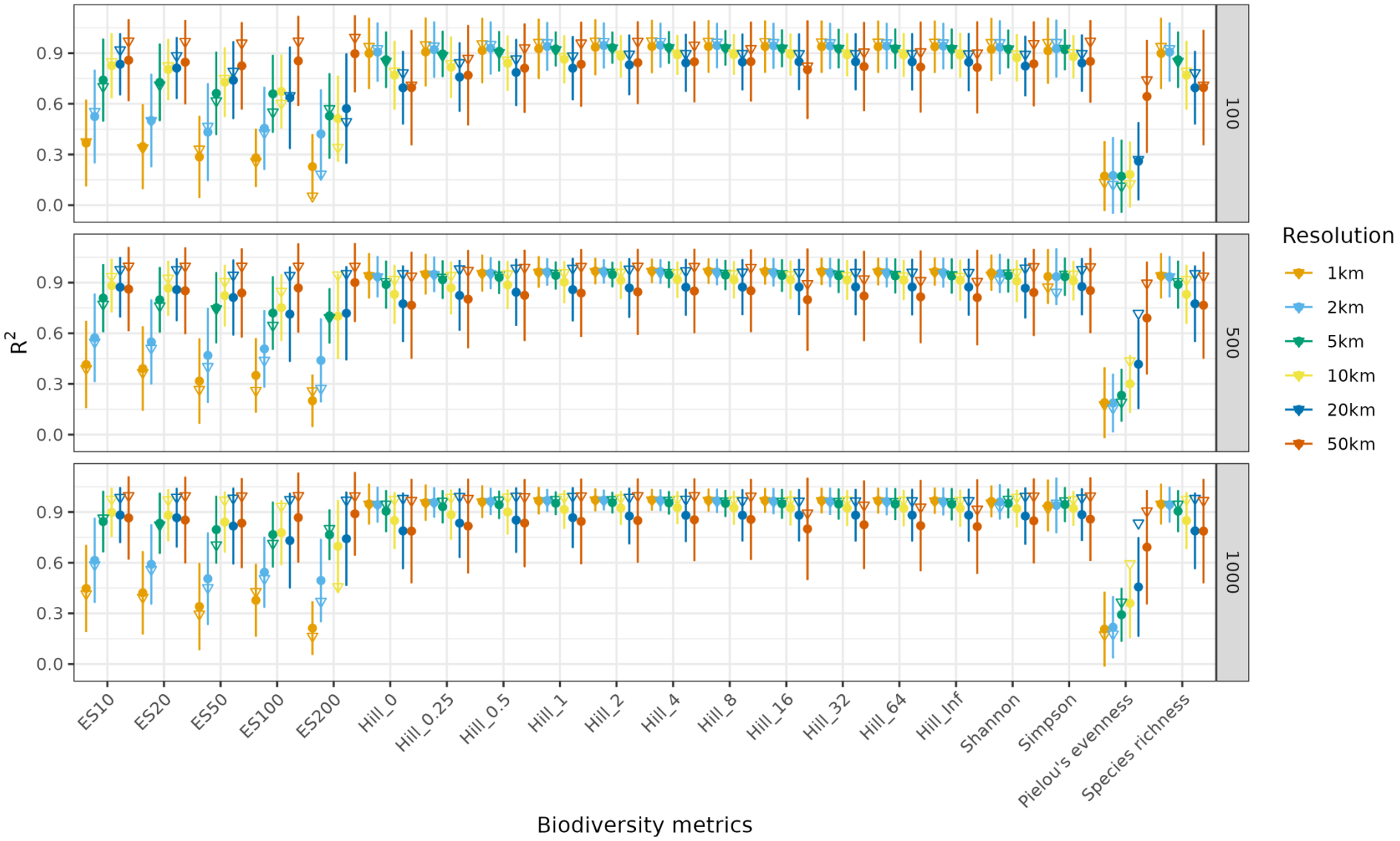


Figure S26. Overall performance of each metric at each resolution, different densities of road coverage, and different sample sizes, including 100 records, 500 records and 1000 records (whilst not accounting for unsampled cells).

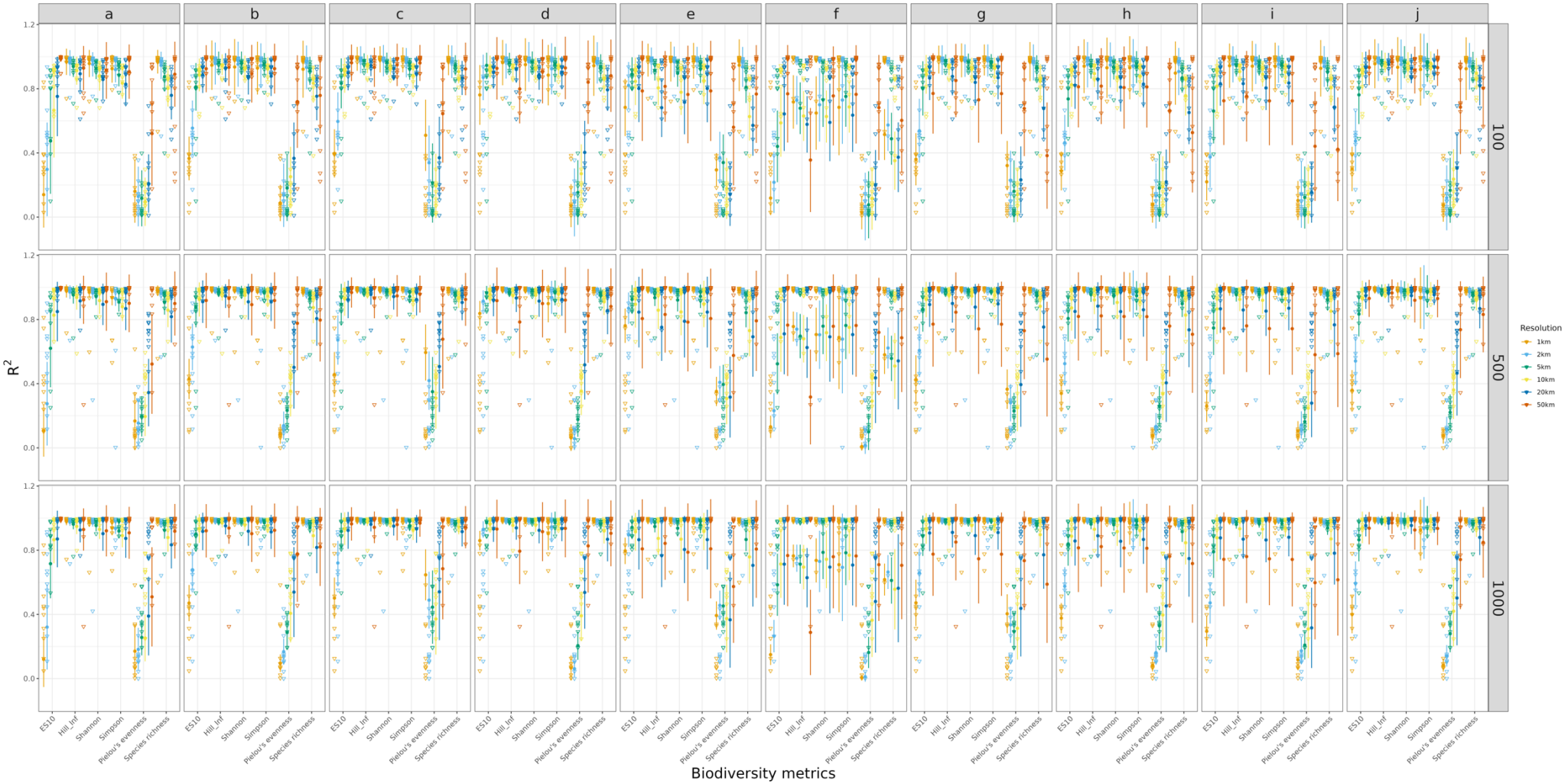


Figure S27. Performance for each plot of each metric at each resolution, different densities of road coverage, and different sample sizes, including 100 records, 500 records. Triangles indicate random sampling; circles indicate biased samples (whilst not accounting for unsampled cells).