

# Shortfalls and opportunities in terrestrial vertebrate species discovery

**Authors:** Mario R. Moura<sup>1,2,3</sup>, Walter Jetz<sup>1,2</sup>

## **Affiliations:**

<sup>1</sup> Department of Ecology and Evolutionary Biology, Yale University, New Haven, CT, USA

<sup>2</sup> Center for Biodiversity and Global Change, Yale University, New Haven, CT, USA

<sup>3</sup> Department of Biological Sciences, Federal University of Paraíba, Areia, PB, Brazil

\*Correspondence to: [mariormoura@gmail.com](mailto:mariormoura@gmail.com), [walter.jetz@yale.edu](mailto:walter.jetz@yale.edu)

## **Abstract:**

Meter-resolution imagery of our world and myriad biodiversity records collected through citizen scientists and automated sensors belie the fact that much of the planet's biodiversity remains undiscovered. Conservative estimates suggest only 13 to 18% of all living species may be known at this point<sup>1-4</sup>, although this number could be as low as 1.5%<sup>5</sup>. This biodiversity shortfall<sup>6,7</sup> strongly impedes the sustainable management of our planet's resources, as the potential ecological and economic relevance of undiscovered species remains unrecognized<sup>8</sup>. Here we use model-based predictions of terrestrial vertebrate species discovery to estimate future taxonomic and geographic discovery opportunities. Our model identifies distinct taxonomic and geographic unevenness in future discovery potential, with greatest opportunities for amphibians and reptiles and for Neotropical and IndoMalayan forests. Brazil, Indonesia, Madagascar, and Colombia emerge as holding greatest discovery opportunities, with a quarter of future species descriptions expected there. These findings highlight the significance of international support for taxonomic

22 initiatives and the potential of quantitative models to aid the discovery of species before their  
23 functions are lost in ignorance<sup>8</sup>. As nations draw up new policy goals under the post-2020 global  
24 biodiversity framework, a better understanding of the magnitude and geography of this known  
25 unknown is critical to inform goals and priorities<sup>9</sup> and to minimize future discoveries lost to  
26 extinction<sup>10</sup>.

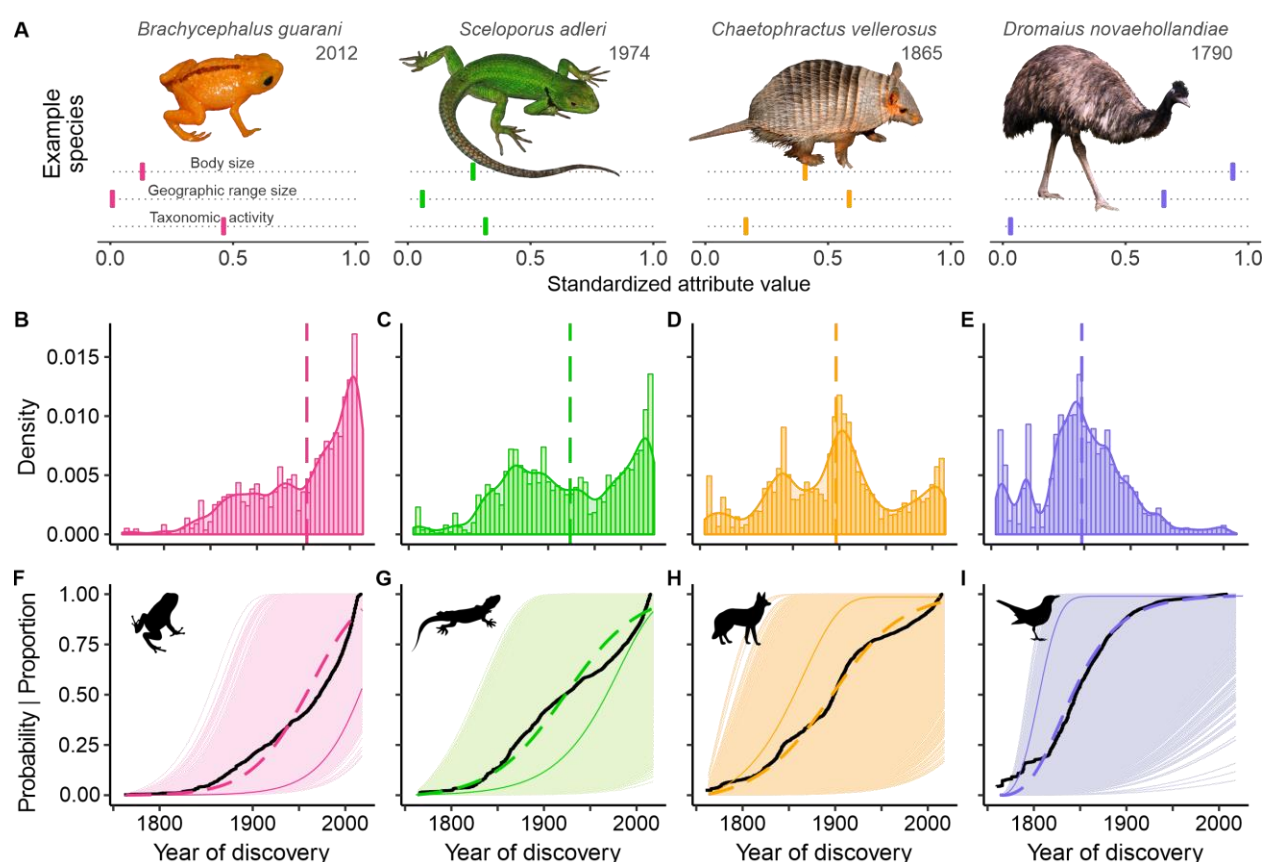
27

## 28 **Main Text**

29 Previous studies have tackled this challenge through a range of extrapolation techniques  
30 using species discovery curves and expert opinion<sup>1-3,11</sup>, but with limited detail beyond  
31 global/continental taxon percentages or counts<sup>12,13</sup>. Here we use the effects organismal  
32 characteristics have on discovery probability to provide taxonomic and geographic specificity of  
33 future species discovery<sup>14-17</sup>. For example, take one of the largest extant birds, the emu  
34 *Dromaius novaehollandiae*, which was described in 1790 in a time and region with limited  
35 taxonomic activity; centuries before a small, elusive frog species *Brachycephalus guarani*,  
36 discovered in 2012 in Brazil (Fig 1a). The difference between the two species matches previous  
37 insights about the effects of body size, range size and taxonomic activity on discovery<sup>16,18-20</sup>.  
38 We extend this comparison to eleven biological, environmental, and sociological attributes in a  
39 ‘time-to-event’ model framework to estimate the probability a given species’ discovery (event)  
40 over time<sup>21,22</sup> (Supplementary Information).

41 For all extant species, the modelling framework provides predicted discovery year and  
42 discovery probability at the present (herein 2015) based on the weighted importance of each  
43 assessed attribute. In our earlier example, *B. guarani* (described in 2012) had a 49% (95%CI: 37-  
44 64%) chance of discovery by 2015 given its attributes. Conversely, the discovery probability for

45 the emu exceeded 50% already in 1759, increasing to 100% by 2015 (95%CI: 100-100%). When  
 46 applied to 32,172 species of amphibians, reptiles, birds, and mammals, the predicted discovery  
 47 curves match observed differences in temporal description patterns in the four taxa. Most bird  
 48 species saw high discovery probabilities early on, matching the median avian description year of  
 49 1845. In contrast, half of all amphibian descriptions occurred after 1972, and modelled discovery  
 50 curves accordingly show a slow increase (Fig. 1).



51

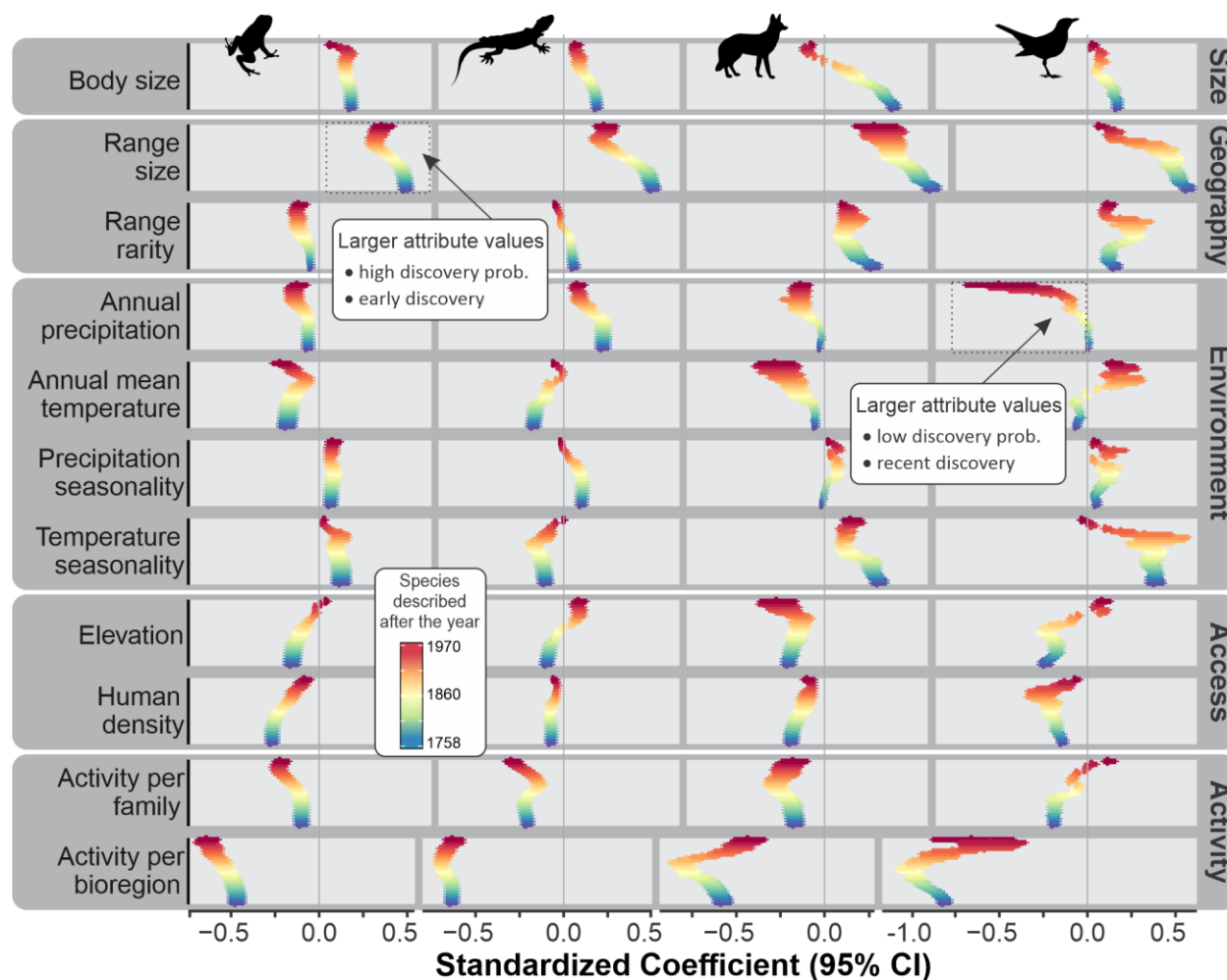
52 **Fig. 1.** Variation in observed and predicted discovery trends for the years 1759-2014 across the four terrestrial  
 53 vertebrate groups. (A) Example species and their attributes (standardized to vary from 0 and 1 in each group,  
 54 separately). (B-E) Variation in species descriptions over time. Vertical dashed lines indicate the year in which 50%  
 55 of the known species were described. (F-I) Time-to-event model-based predictions of discovery probability for each  
 56 species (light colours; solid coloured line representing example in A), and average trends across all (coloured  
 57 dashed-line). Black lines show the empirical cumulative growth of described species described across time  
 58 (expressed as proportion of known species).

59 Species' body size, geographic range size, and taxonomic activity strongly affect  
60 variation in discovery probability, but terrain and environmental conditions also matter<sup>16,19,23</sup>.  
61 Species tend to have higher discovery probability if they are large-bodied, wide-ranged, located  
62 in cold climates or characterized by, at the time, low taxonomic activity or low human density  
63 (Fig. 2). The magnitude, and sometimes also direction, of effects differs somewhat among the  
64 four groups. It also varies across time, reflecting developments in taxonomists' modes and  
65 toolbox, as well as changes in the kinds of species left to be discovered (Fig S3-S6). For  
66 example, among more recently discovered species, body size or human density have lost  
67 predictive strength. In amphibians, higher elevations are less of a constraint on discovery  
68 probability than in the past, whereas the recency of mammal discovery continues to be associated  
69 with higher elevations. Among bird species described since the mid-20<sup>th</sup> century, wetter  
70 locations have yielded later discoveries, but not so prior to that time. Notably, in amphibians and  
71 reptiles, clades and regions with more active taxonomists remained those with greatest discovery  
72 potential. This highlights how gaps in taxonomic expertise continue to limit our recognition of  
73 species.

74 Averaged across species in an assemblage or clade, the divergence of modelled discovery  
75 probability from 100% informs the portion of species yet to be discovered given past modes of  
76 description. Among vertebrate clades with >5 species, South American shrew opossums  
77 (Paucituberculata), dibamids, geckos and relatives, wall lizards and other lacertids emerge as  
78 having the greatest relative undescribed diversity (Fig. 3). Scaled by groups' species count, the  
79 models identify several frog clades, geckos and iguanas and their relatives, and snakes as the  
80 vertebrate groups with the highest expected number of future species discoveries. Among  
81 mammals, rodents and bats feature in the top ten higher-level taxa, partly reflective of the



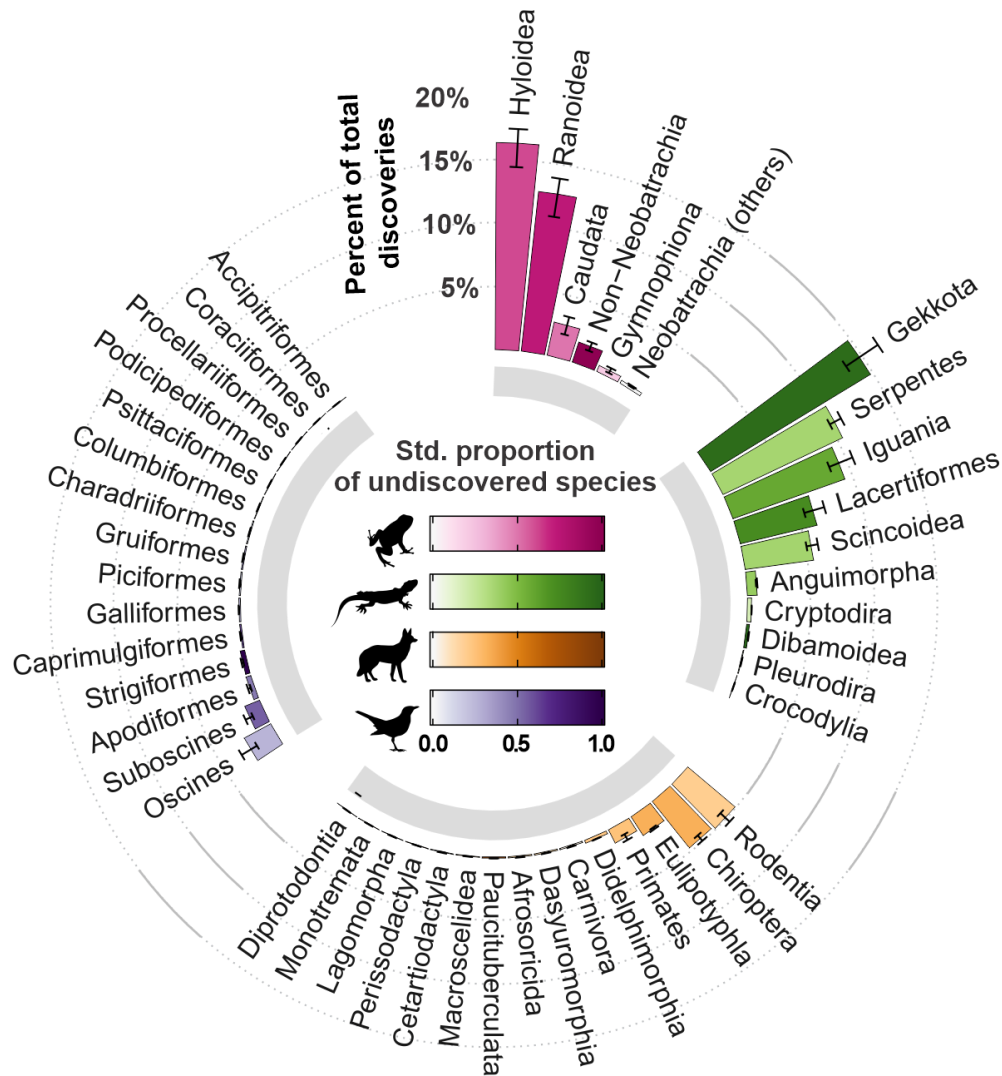
82 already large described diversity of these groups <sup>24</sup>. Cross-validation with holdout data on past  
83 discoveries indicate a strong predictive power of our framework and suggest robustness across  
84 groups (Extended Data, Figs S8, S14). While some of the identified taxonomic discovery  
85 hotspots are not unexpected, our evaluation across all terrestrial vertebrates offers a transparent  
86 quantitative comparison in support of taxonomic research priorities.



87

88 **Fig. 2.** Joint effects of species-level attributes on discovery probability over different time periods. Standardized  
89 coefficients above 0 indicate that species with high values for a given attribute had higher discovery probability  
90 (prob.) and thus were likely discovered early on. Negative standardized coefficients mean high attribute values  
91 depressed discovery probability and delayed discovery. The vertical colour gradients illustrate the variation in  
92 coefficient from all (bottom) to more recently described species as species are successively removed from the  
93 analysis. Coefficients include 95% confidence intervals as horizontal bars.

94



95 **Fig. 3.** Predicted future discovery potential across major terrestrial vertebrate taxa. Bar height indicates the  
 96 percentage of all future terrestrial vertebrate discoveries predicted to occur in the taxon, with error bars indicating  
 97 95% confidence intervals. Bar colours show the proportion of undiscovered species within each vertebrate class,  
 98 standardized to vary from 0 to 1 (see Supplementary Information). See Figs S16-S19 (Extended Data) for the full set  
 99 of discovery metrics at family level.

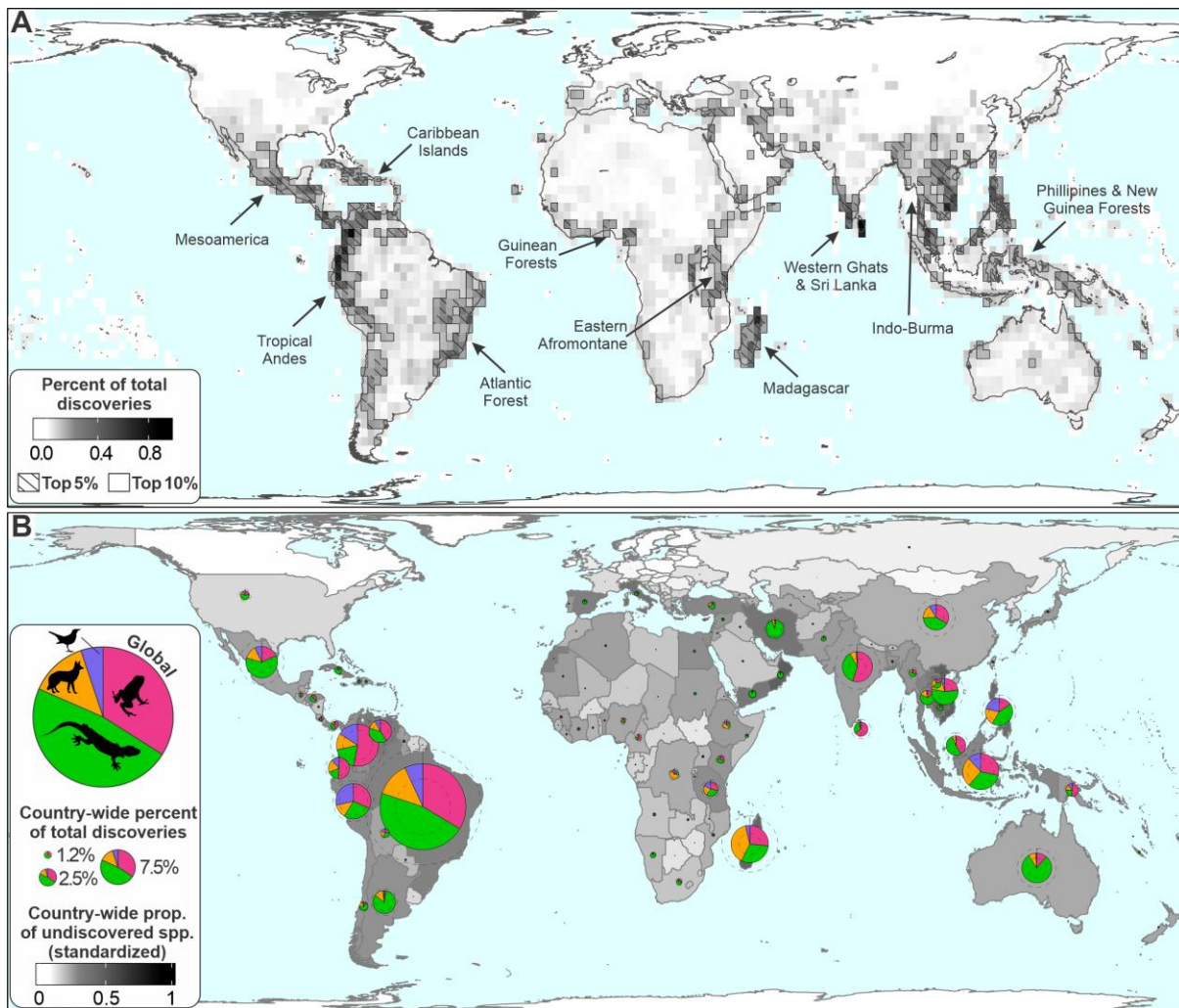
100 Both the facilitators of species discovery – such as fieldwork and systematics initiatives –  
 101 and the drivers of undiscovered species’ demise – such climate- and land-use change – are  
 102 strongly place-based<sup>25,26</sup>. We therefore extended our discovery predictions to geographic space.  
 103 We mapped attribute-driven discovery probabilities across species distributions while applying a  
 104 subsampling procedure accounting for range-size driven variation in representation<sup>27</sup>. As the

105 locations with highest number of expected future discoveries (Fig. 4), we identify the Tropical  
106 Andes and the Atlantic Forest, the Eastern Afromontane and West African Guinean Forests,  
107 Madagascar, and the Western Ghats, Sri Lanka, Indo-Burma, and Philippines and New Guinea  
108 Forests. Projected unknown species richness covaries with extant species richness (Spearman  $r =$   
109  $0.87-0.90$ ), but discovery hotspots also included locations with relative limited extant diversity  
110 such as the Southern Andes and Caatinga region. Validations with observed discoveries strongly  
111 support these projected spatial discovery patterns across different spatial resolutions (Spearman  $r$   
112  $= 0.71-0.92$  for all groups, see Extended Data Figs S9-S12, S15).

113 Expertise, support, and incentives for future discovery are ultimately tied to nations – the  
114 stewards of these unknown biological resources. Aggregating our spatial estimates to countries  
115 highlights several South American and South Asian nations and Madagascar as countries with  
116 highest projected future discoveries, i.e. greatest “discovery debt” or conversely, “biodiversity  
117 reward” (Fig. 4, and Fig. S25, Extended Data). These countries are home to a large diversity of  
118 taxa with attributes indicative of low discovery rate to date, and thus likely contain many  
119 expected future discoveries. Brazil stands out with multiple diversity centres across its large area,  
120 holding ca. 10.5–10.8% of all projected future discoveries, which largely coincides with the  
121 10.8–13.4% of ant genera discovery projected for this country<sup>13</sup>. Other top discovery debt  
122 countries for terrestrial vertebrates include Indonesia (5.0–5.7%), Madagascar (3.9–4.9%), and  
123 Colombia (4.1–4.4%).

124 Reflecting their class-wide high discovery potential, reptiles constitute the greatest  
125 portion of this future prospect. Undiscovered reptiles are expected in more arid regions, such as  
126 Australia, Iran, and Argentina, correlating with existing centres of reptile diversity and  
127 endemism<sup>28</sup>. In the tropics, many countries owe most future discoveries to amphibians,

128 particularly in southern Asia and northern South America (Extended Data, Fig. S25). Discovery  
129 potential for mammals is more limited and concentrated in recent description hotspots such as  
130 Madagascar. Compared to other taxa and reflecting their flattened description curve (Fig. 1), the  
131 discovery potential for birds is low, but given past trends the model estimates further discoveries  
132 especially in Peru, Colombia, Brazil, and Philippines.



133

134 **Fig. 4.** Global variation in predicted discovery potential, quantified as the percent of all global terrestrial vertebrate  
135 discoveries predicted to occur in a region. (A) Variation across 220km grid cells, standardized to percent of total  
136 discoveries. (B) Variation among countries, with colours showing mean discovery potential, expressed as country-  
137 wide proportion (prop.) of undiscovered species (spp.) and standardized to vary from zero to one. Pie charts  
138 illustrate the predicted distribution of discoveries among the four vertebrate classes in each country (“Global” in  
139 legend shows the global pattern); pie chart size indicates the country-wide total, with dashed grey lines indicating  
140 the 95% confidence interval. See Figs S20-S24 (Extended Data) for maps of the discovery metrics at different  
141 spatial resolutions.

142           Research on quantifying species discovery shortfalls is by definition imprecise, with  
143           absolute estimates of undiscovered species differing by orders of magnitude<sup>1,3,5</sup>. We focused on  
144           the geographic and taxonomic differentiation in discovery potential and used a well-known  
145           subset of global biodiversity to develop a generalizable framework to address this challenge. We  
146           do not expect that our discovery projections will hold up in exact form. The present estimates are  
147           a direct reflection of past description processes and their correlates, and any forward  
148           interpretation therefore needs to recognize intrinsic limitations. Despite ongoing calls for  
149           taxonomic standardization and stability<sup>29,30</sup>, species also represent scientific hypotheses that are  
150           sometimes revisited, refuted, or revalidated<sup>31,32</sup>. Our models therefore are not able to distinguish  
151           operational definitions of valid species and the potential heterogeneous associations arising from  
152           variable practices around, e.g., recognizing cryptic species or splits<sup>33</sup>. There may also be parts of  
153           the multivariate predictor space that lack data to inform the model and thus miss actual discovery  
154           opportunities. Nevertheless, extensive model validations confirmed a strong predictive ability for  
155           species discoveries and highlight the potential to increase discovery rates through the use of  
156           quantitative frameworks such as the one presented.

157           Our findings indicate that discovery gaps hinder the safeguarding and realization of  
158           biodiversity for certain kinds of species and for select places and countries much more than  
159           others. We show that specific countries require increased capacity and support to address this  
160           challenge. After centuries of efforts by biodiversity explorers and taxonomists, the catalogue of  
161           life still has too many blank pages. Extending the presented approach to other taxa has the  
162           potential to underpin taxonomic research initiatives that help speed up discovery before species  
163           are lost in ignorance<sup>34</sup>. With discussions of the Post-2020 Global Biodiversity Framework

164 ongoing, we urge intergovernmental recognition of the unevenness in this knowledge shortfall  
165 and of the growing scientific opportunities to address it.

166

## 167 **METHODS**

168 We used species-level biological, environmental, and sociological attributes to parameterize  
169 time-to-event models of discovery probability across time in birds, mammals, amphibians, and  
170 reptiles<sup>21</sup>. The models provided for each species an estimated probability of discovery in the  
171 present time given its attributes. We used these species predictions to characterize the major  
172 taxonomic and geographic groups they are part of for their future discovery potential.

173 Specifically, for major taxa and assemblages worldwide the central tendency (geometric mean)  
174 of model-driven discovery probability of its member species informed an estimate of the  
175 proportion of their known vs. yet to discovered species richness.

176

### 177 **Species data**

178 We compiled trait data for nearly all extant species of terrestrial vertebrates<sup>28,35</sup>, excluding  
179 those with uncertain geographic distribution, that is, species with occurrence reported only at the  
180 level of administrative units – country, states, etc – without precise location. We also excluded  
181 species described after 2014 to minimize potential biases from their potentially incomplete  
182 geographic characterization. Overall, our dataset comprised 32,878 species of terrestrial  
183 vertebrates: 7202 species of amphibians, 10,004 of reptiles, 5679 of mammals, and 9993 of birds  
184 (Data S1). Taxonomic nomenclature followed the same adopted in original data sources<sup>28,36–39</sup>.

185

### 186 **Species-level attributes**



187 Previous studies have shown that recently described species often have smaller body sizes  
188 and narrower geographic range compared to those described earlier<sup>16,17</sup>. But species detectability  
189 and discovery may be affected by a range of other attributes, such as the environmental and  
190 socioeconomic conditions where the species occurs, and the taxonomic knowledge of a given  
191 taxon or region<sup>16,17,40</sup>. We considered a total of eleven putative correlates of discovery  
192 probability (see Data S5):

193 (i) Body size. Larger animals are easier to detect thus being described first<sup>16,17,23,41</sup>.

194 Information on maximum body size was compiled from available data sources: amphibians  
195<sup>42</sup>, reptiles<sup>43–47</sup>, mammals<sup>38,48</sup>, and birds<sup>49</sup>. We complemented these data sources by  
196 inspecting the literature for body size information of species without data. In the end, we  
197 obtained body size data for 6869 (95.2%) species of amphibians, 9852 (99.7%) of reptiles,  
198 5208 (91.4%) of mammals, and 9123 (91.3%) of birds. For amphibians, body size was  
199 represented by snout-vent length (in mm, anurans) or total length (in mm, caecilians and  
200 salamanders). Reptiles had their body length measures converted into masses (g) using  
201 taxon-specific allometric equations available in the literature for squamates<sup>43</sup> or here  
202 developed for chelonians (Supplementary Information, Table S1). Birds and mammals had  
203 their body size represented by masses (g), as provided in the available data sources<sup>38,49</sup>.  
204 For 870 bird species with missing body size data, we used the genus-level mean body size  
205 as originally provided in<sup>49</sup>. For the remaining species without body size data, we  
206 performed a phylogenetic imputation using the fully-sampled global phylogenies made  
207 recently available<sup>37,38,50</sup> in concert with the R package *Rphylopars*<sup>51</sup>. We discarded seven  
208 reptile species due to the lack of both body size and phylogenetic data (*Agama congica*,  
209 *Bothrochilus montanus*, *Gerrhosaurus intermedius*, *Hemidactylus benguellensis*,



210 *Leptotyphlops lepezi*, *Rhoptropus benguellensis*, *R. montanus*). To account for uncertainty  
211 in the fully sampled phylogenies, we used 100 trees of each vertebrate group to perform  
212 imputations. We loaded the named species-level data (including missing values) and ran  
213 *phylopars* function using Brownian motion (BM) as model of trait evolution. We then  
214 matched the imputed values back to the taxonomy and summarized the BM imputations  
215 across the 100 trees as medians for each species. All body size measures were  $\log_{10}$   
216 transformed before phylogenetic imputations. Overall, we imputed body size for 333  
217 species of amphibians, 23 of reptiles, 492 of mammals. We assumed intraspecific variation  
218 in body size to be negligible relative to interspecific variation.

219 (ii) Geographic range size. Widely distributed species tend to be locally abundant<sup>52,53</sup> and are  
220 therefore easier to find, being described earlier than narrowly distributed species<sup>16,17,23,41</sup>.  
221 We overlaid expert-based extent-of-occurrence range maps of each species with an equal-  
222 area grid of 110 × 110 km cell size. Range maps were extracted from<sup>39,54</sup> for amphibians,<sup>28</sup>  
223 for reptiles,<sup>35,38,54</sup> for mammals, and<sup>35,36</sup> for birds. Range size was then measured as the  
224 number of grid cells intersected by each species. Only the native and breeding range of  
225 species were considered for these computations. Presence of a species in a grid cell was  
226 recorded if any part of the species distribution polygon overlapped with the grid cell.

227 (iii) Range rarity. Biodiversity researchers may prefer to work in areas with many or  
228 geographically rare species, and describe first the species from those areas<sup>25,55</sup>. A  
229 commonly used metric of rarity is the total range size rarity, also called endemism  
230 richness, defined as the sum of the inverse range sizes of all species present in a place<sup>56</sup>.  
231 To represent rarity at the species-level, we used the average endemism-richness within  
232 each species' range. However, grid cells (regions) that currently harbour many and/or rare

233 species had not always been known as such, since known richness and endemism patterns  
234 may change through time as species descriptions progressed. To better capture the  
235 variation in range rarity across time, we computed the endemism richness using only  
236 species described from 1758 to  $x$ , where  $x$  varied from 1758 to 2014. Then, for each  
237 species, we computed the average known within-range endemism richness at the year it  
238 was described.

239 (iv) Annual precipitation. Early descriptions dates are on average low for species occurring in  
240 Europe, North America, and western Asia<sup>57</sup>. These later regions have received substantial  
241 taxonomic effort, which explains their higher levels of inventory completeness relative to  
242 tropical and desert-like environments<sup>25</sup>. Thus, it is reasonable to consider that early  
243 naturalists were trained in temperate regions and therefore they explored first species from  
244 relatively dry regions<sup>16,58-60</sup>. We calculated the average annual precipitation in each equal-  
245 area grid cell at  $110 \times 110$  km of spatial resolution and then computed the average within-  
246 range annual precipitation for each species using the 1 km climatic layer from<sup>61</sup>.  
247 Computations were performed in the R software<sup>62</sup> using the *extract* function of ‘raster’  
248 package<sup>63</sup>.

249 (v) Annual mean temperature: Following the reasoning aforementioned, early naturalists were  
250 trained in temperate regions and therefore they explored first species from cold regions  
251<sup>16,58,59</sup>. We calculated the average temperature (annual mean) in each equal-area grid cell at  
252  $110 \times 110$  km of spatial resolution and then computed the average within-range temperature  
253 for each species using the 1 km climatic layer from<sup>61</sup>.

254 (vi) Precipitation seasonality. Early naturalists were trained in temperate regions and therefore  
255 they explored first species from high seasonal regions<sup>23,58</sup>. We calculated average within-

256 range precipitation seasonality (coefficient of variation) for each species using the 1 km  
257 climatic layer from <sup>61</sup>.

258 (vii) Temperature seasonality. Early naturalists were trained in temperate regions and therefore  
259 they explored first species from high seasonal regions <sup>16,58,59</sup>. We measured average within-  
260 range annual temperature seasonality (standard deviation) for each species using the 1 km  
261 climatic layer from <sup>61</sup>. Although climatic conditions might not have been necessarily stable  
262 over the last centuries, we assumed that average conditions from 1979-2015 represented  
263 similar climatic conditions from the 18th to 20th centuries. We argue that averaging  
264 climatic variables within species geographic range may both dilute the temporal variations  
265 in local climate and avoid the uncertainty associated with extrapolating fine resolution  
266 climatic data to past times of low density (or even absence) of weather stations.

267 (viii) Elevation. Mountainous regions might have limited accessibility, likely impeding early  
268 species descriptions from higher elevations <sup>25,40</sup>. Although early taxonomists and  
269 naturalists likely explored low elevation regions first, we avoided computations of  
270 minimum within-range elevation since species with coastal distribution could show biased  
271 values that do not necessarily reflect the most common elevation where they occur. We  
272 computed the mean elevation within each equal-area grid cell at 110 × 110 km of spatial  
273 resolution, using the 1 km global topography layer from <sup>64</sup>. We then extracted the average  
274 within-range elevation for each species.

275 (ix) Human density: A species may be described if human population density within its  
276 geographic range surpass a detectability threshold that enhance its discovery probability  
277 <sup>23,40</sup>. Such detectability is expected to be low before the species description (too few or  
278 even no humans overlapping the species geographic range) and irrelevant after its

279 discovery (human density might change but the formal description already happened).  
280 Thus, an informative measure of geographic range overlap with human settlements should  
281 consider the year of species description. Earlier descriptions occurred at times of low  
282 human density and much likely involved easily detectable species, whereas more recent  
283 descriptions have coincided with times of high human density. We therefore expect a  
284 positive association between human density and year of description. To quantify the  
285 influence of humans on species description, we computed the average human density  
286 within the species' range at the exact year of its description (for all species described after  
287 2000), or at the closest decade (for all species described before 2000). Historical data on  
288 human density data was obtained from <sup>65</sup> at the spatial resolution of 5 arc-min (~10 km).

289 (x) Activity per family. In general, taxonomists tend to discover the 'obvious' first and  
290 'obscure' later. Species like the emu (Fig. 1) were quickly noticed by taxonomists, even in  
291 times of low taxonomic activity (relative to current trends). However, other species exhibit  
292 high conspicuousness and likely required more attention to be taxonomically noticed. This  
293 level of attention is what we refer here as taxonomic activity. In increasing the number of  
294 active taxonomists per family, a potentially inconspicuous or cryptic species may receive  
295 enough attention to have its discovery unveiled by taxonomists. Otherwise, a species may  
296 remain unknown while the taxonomic activity within its family is kept low. Similarly to  
297 human population, the number of taxonomists has increased over time, with fewer  
298 taxonomists authoring early species descriptions <sup>66</sup>. Therefore, we expect species described  
299 long ago to show low within-family taxonomic activity, and consequently high discovery  
300 probability. We used the number of authors of each species description as a proxy for  
301 taxonomic activity. We standardized the surnames included in the authority name of each

302 species to lowercase letters without special characters. For each species, we identified the  
303 taxa described in the same family and year as the focal taxon and then calculated the  
304 aggregate number of unique active taxonomists. Because the number of active taxonomists  
305 in a given year is expected to increase as more species are described in that year, we  
306 standardized this measure by dividing it by the number of within-family descriptions in the  
307 respective year. Computations were performed using the *stringr*<sup>67</sup> and *splitstackshape*<sup>68</sup>  
308 R-packages.

309 (xi) Activity per bioregion. In a similar way to the influence of taxonomic activity per family, a  
310 new species may go unnoticed while the number of taxonomists working within its  
311 geographic range remain too low or even non-existent. Under low levels of taxonomic  
312 activity, those easy to find species are likely described first. In other words, species with  
313 high ‘taxonomic conspicuousness’ may require eyes from more taxonomists. Following the  
314 trend of increasing the number of taxonomists per species over time<sup>66</sup>, we expect early  
315 described species to show low within range taxonomic activity, in contrast to recent  
316 described species. These computations followed those for the taxon level, but with species  
317 instead subset by geography. Specifically, we used the biogeographical realm and biome  
318 classification proposed in<sup>69</sup> to compute the percentage overlap between species geographic  
319 ranges and realm-biome combination, or bioregion<sup>70</sup> (e.g. Tropical and subtropical moist  
320 broadleaf forests in the Neotropics). Each species was classified as typical of a given  
321 bioregion if it either occurred in at least 25% of the bioregion or the bioregion intersected  
322 with at least 25% of its geographic range (species could be typical of multiple bioregions).  
323 For each species, we then selected its ‘typical’ bioregion and extracted all other species  
324 described in the same year and co-occurring in the same bioregion as focal taxon. We then

325 summed the number of unique taxonomist names that described species within the selected  
326 bioregion and divided it by the number of within-bioregion descriptions to obtain a  
327 measure independent of the number species descriptions. We recognize that our metrics of  
328 taxonomic activity might be affected by duplicated names (same taxonomist entered with  
329 the different surname) or homonyms (taxonomists with the same surname), but given the  
330 spatial, taxonomic, and temporal constraints applied, we expect this issue to be negligible.

331

332 Since all our predictor variables vary over many orders of magnitude, we  $\log_{10}$  transformed  
333 them to reduce their skewness. We examined multicollinearity of the predictor variables using  
334 the Variance Inflation Factor (VIF). Predictors holding VIF values  $> 10$  are regarded as having  
335 high multicollinearity and should be excluded from the model <sup>71</sup>. As none of our predictors  
336 achieved VIF  $> 5$  (Supplementary Information, Table S2), we kept all of them for the subsequent  
337 analysis. VIF computations were performed with the ‘usdm’ R package <sup>72</sup>.

338

### 339 **Time-to-event models**

340 We used time-to-event analysis, also known as survival analysis <sup>21</sup>, to assess the effect of  
341 species-level attributes on the description rates observed in a given vertebrate class. Time-to-  
342 event analysis is commonly used in the medical, engineering, and social sciences to assess  
343 factors influencing the probability of an event (e.g., death, mechanical failure, getting a job), but  
344 has also been used in ecological studies <sup>22,73,74</sup>. In our analysis, the event of interest is the species  
345 description date and the measure of time the number of years passed until this date since 1758,  
346 the beginning of modern taxonomy through Linnaeus <sup>75</sup>. Although our species data covers the  
347 period from 1758 to 2014, we did not use species described in the year 1758 itself to avoid the

348 large number of descriptions due to Linnaeus work <sup>75</sup>. Overall, our class-level time-to-event  
349 models were informed by 7,185 species of amphibians, 9,889 of reptiles, 9,557 of birds, and  
350 5,541 of terrestrial mammals described between 1759 and 2014.

351 Specifically, we modelled time-to-description using Accelerated Failure Time (AFT)  
352 model, which is a parametric time-to-event model to evaluate covariate effects on the  
353 acceleration/deceleration of the probability of an event <sup>76</sup>. The output of the AFT model includes  
354 the probability of a given species to have remained unknown across time, i.e. its survival  
355 probability. We define the 1 minus species survival probability as species discovery probability  
356 (Fig. 1). This discovery probability is always increasing; as time moves forward, we accumulate  
357 chances to discover an unknown species.

358 We initially ran a model selection procedure to identify the family error distribution that is  
359 best suited to our time-to-description variable <sup>77</sup>. For each of the four vertebrate classes, we built  
360 null AFT models (Time to event  $\sim 1$ ) using six different family error distributions (Exponential,  
361 Weibull, Log-normal, Log-logistic, Gamma, and Gompertz) available in the *flexsurvreg* function  
362 from the ‘flexsurv’ R package <sup>78</sup>. We then identified the model offering the best error family  
363 distribution using the Bayesian Information Criterion (BIC) <sup>79</sup>. Once the best family error  
364 distribution was selected, we proceeded with the subsequent analysis using the predictor  
365 variables.

366 Given the high number of possible models using all predictor combinations ( $2^{11} - 1 = 2047$   
367 models), it may be difficult to find an overwhelmingly supported model because the best  
368 predictors (if any) will have their importance diluted among multiple models <sup>80</sup>. Hence, to  
369 incorporate the uncertainty around the variable selection procedure into our model coefficients,  
370 we passed the predictors through a model averaging procedure <sup>81</sup> and for each possible AFT



371 model obtained the standardized coefficients. Computations performed using the ‘MuMIn’<sup>82</sup> and  
372 ‘stats’<sup>62</sup> R packages.

373

### 374 **Species-level predictions**

375 For every possible AFT model we computed the discovery probability of each species in  
376 year 2015 (herein considered as ‘present time’, Data S1). To relate this value to observed  
377 description dates, we used a threshold of 0.5 to convert the estimated discovery probability for a  
378 given time step into a predicted description and extracted the corresponding year as predicted  
379 description date. We note that the discovery curves returned by an AFT model have the same  
380 shape but a different position along the time axis, the latter being determined by the covariates of  
381 each species. Because of that, using a different threshold for the binary conversion does not  
382 affect the slope of the relationship between the observed and estimated description dates, only  
383 the intercept changes if a different threshold value is applied. This procedure yielded for every  
384 species and each of the 2,047 AFT models i) a discovery probability in year 2015 and ii) a  
385 predicted description date. We weighted these metrics by the relative BIC weights (wBIC) of  
386 their models to arrive at species-level discovery metrics used in subsequent analyses.  
387 Computations performed using the ‘MuMIn’ R package<sup>82</sup>. All species-level estimates are  
388 available through Data S1 file.

389

### 390 **Taxon-level predictions**

391 We used the species-level predictions of discovery probabilities to characterize individual  
392 families and higher-level groupings for their potential for future discoveries. Specifically, we  
393 estimated the taxon-level proportion of known species to date (*PropKnown*) as the central

394 tendency (geometric mean) of discovery probability of its member species<sup>22</sup>. We also obtained  
395 the known species richness (*KnownSR*) per taxon, and used both *PropKnown* and *KnownSR* to  
396 calculate total richness (*TotalSR*) through a rule of three:  $TotalSR = KnownSR \times 100 /$   
397 *PropKnown*. If the *PropKnown* of a given taxon was 100%, then all species were expected to be  
398 described in the respective taxon, and  $TotalSR = KnownSR$ . If *PropKnown* was < 100%, then  
399  $TotalSR > KnownSR$ , and the difference between these variables represented the unknown  
400 species richness:  $UnknownSR = TotalSR - KnownSR$ . The proportion of unknown species  
401 (*PropUnknown*) was given by 1 minus *PropKnown*. We highlight that *PropKnown* (our measure  
402 of central tendency), represents a snapshot of the current biodiversity knowledge for a given  
403 species sample. In approaching saturation of species discovery in a sample, *PropKnown* is  
404 expected to shift towards 100%. We only computed these metrics for families and higher-level  
405 groupings holding five or more species.

406 For mammals and birds, we used taxonomic orders as the highest-level taxonomic rank,  
407 although we split Passeriformes birds into Oscines and Suboscines. In amphibians and reptiles,  
408 orders are highly uneven in size, which led us to use a more informative higher-level grouping  
409 for them. For amphibians, we kept the orders Gymnophiona (caecilians) and Caudata  
410 (salamanders and relatives), and followed<sup>83</sup> to divide the order Anura into four groups: (i) non-  
411 Neobatrachia (some primitive anuran families), (ii) Hyloidea taxon within Neobatrachia,  
412 including most frog species from the Nearctic and Neotropic realms, (iii) Ranoidea taxon within  
413 Neobatrachia, including most frog species from the Afrotropical, Palearctic, Indo-Malay, and  
414 Australasia realms, and (iv) other Neobatrachia (a non-monophyletic set of Neobatrachian  
415 families not included in Ranoidea or Hyloidea). For reptiles, we kept the order Crocodylia  
416 (alligators and relatives), divided the order Testudines in the suborders Pleurodira (side-necked

417 turtles) and Cryptodira (hidden-necked turtles), and followed <sup>84</sup> to split the order Squamata into  
418 seven groups: Gekkota (geckos and relatives), Iguania (iguanas, chameleons, and relatives),  
419 Scincoidea (skinks and relatives), Lacertoidea (teiids, lacertids, amphisbaenians, and relatives),  
420 Anguimorpha (glass lizards, monitors, and relatives), Dibamidae (dibamids or blind skinks, also  
421 referred as Dibamoidea), and Serpentes (snakes).

422 The model validation (see below) indicated that our framework satisfactorily identified the  
423 relative potential of taxa to hold unknown species, but it underestimated absolute values of  
424 *UnknownSR* and *PropUnknown* per taxon (Supplementary Information, Table S4; Extended  
425 Data, Fig. S8). Thus, we standardized both measures of discovery potential. For each vertebrate  
426 class, we divided the *UnknownSR* by the total number of estimated discoveries (i.e., sum of  
427 *UnknownSR* across taxa) to provide the estimated percent of total discoveries. The proportion of  
428 unknown species (*PropUnknown*) per taxa was standardized to vary between 0 and 1, by first  
429 subtracting the minimum observed for each vertebrate class and then dividing by the respective  
430 range of *PropUnknown*. The value of 1 indicated the taxon with the highest proportion of  
431 unknown species (whatever such number might be), and not necessarily a taxon with 100% of  
432 unknown species. All taxon-level estimates are available through Data S2 files.

433

### 434 **Assemblage-level predictions**

435 We followed the same rationale we used to for taxon above to estimate the variation in  
436 future discovery potential in geographic space. Specifically, we considered the proportion of  
437 species that remain to be discovered in an assemblage, *PropUnknown*, an emergent property of  
438 its species members and their attributes. This approach follows the growing recognition in trait  
439 biogeography and macroecology of the species-level drivers of larger-scale patterns <sup>14,22,85–87</sup>. We

440 used the equal-area grid cell species distribution data (see above) to derive species lists for each  
441 the four vertebrate classes for assemblages of 220, 440, and 880 km grid cell size, discarding all  
442 assemblages with less than five species. For each assemblage we then calculated *PropUnknown*  
443 and *UnknownSR* based on the discovery probabilities of its member species, the same way we  
444 did at the taxon-level.

445 Different to the by-taxon characterization, however, in the assemblage patterns wide-  
446 ranging species are overrepresented, since those are counted multiple times throughout grid cells  
447 <sup>36,88</sup>. Considering that widely distributed species tend to be described first <sup>16,17,41</sup>, this can bias  
448 assemblage measures towards lower *PropUnknown*. This unevenness in the representation of  
449 wide-ranging species can be controlled through a random subsampling approach that provides  
450 range-size controlled estimates of aggregate measures at the assemblage level <sup>27</sup>.

451 Briefly, the subsampling algorithm we applied considers the random extraction of  $x$  grid  
452 cells belonging to a given species' range. If the species' range was smaller than  $x$ , then, all grid  
453 cells are extracted for that species. The  $x$  here is analogous to the pseudoreplication level of a  
454 dataset. If  $x$  equals 1, then the geographic range of all species will be subsampled to show only  
455 one grid cell per species. The subsampling algorithm was applied to all species in a single  
456 iteration, and the subsampled geographic ranges were then overlapped for the extraction of the  
457 *PropUnknown* and *UnknownSR*. These computations were performed 100 times, and the mean  
458 value across iterations was extracted for each grid cell to represent the geographic pattern of the  
459 respective aggregate measure under the  $x$  level of pseudoreplication. Additional details on this  
460 subsampling algorithm are available in <sup>27</sup>. In this study, we used seven different levels of  $x$ : 1, 5,  
461 10, 50, 100, 200, and 500 grid cell occurrences per species, also including the observed pattern

462 using all grid cell occurrences per species. Computations performed using ‘data.table’ R package  
463 <sup>89</sup>.

464 As in many ecogeographical investigations <sup>90</sup>, this study focuses in the variation of species-  
465 level attributes across space to describe and explain biodiversity patterns. Given the dominant  
466 unavailability of spatially varying data on species-level attributes, most ecogeographical studies  
467 – including this one – assume species-level attributes to be spatially constant. While the  
468 incorporation of spatially varying covariates through hierarchical modelling remains an open  
469 avenue in trait biogeography<sup>91</sup>, we argue that it has limited influence in our results for two major  
470 reasons. First, most species in our dataset are not widely distributed, which implies less potential  
471 for biological attributes to vary in space. For instance, 50%, 70%, and 88% of species in our  
472 dataset occupy  $\leq 4$  grid cells at respectively 220, 440, and 880 km of spatial resolution (See Data  
473 Availability for raw data). Second, the subsampling algorithm we applied reduces the  
474 overrepresentation of widely distributed species, whom are the ones with highest potential to  
475 show spatially varying biological attributes.

476 We standardized *UnknownSR* and *PropUnknown* per assemblage (Supplementary  
477 Information, Table S4; Extended Data, Fig. S8) in the same way as done for taxa. We divided the  
478 *UnknownSR* per assemblage by the sum of *UnknownSR* across assemblages to get the estimated  
479 percent of total discoveries, and standardized *PropUnknown* to vary between 0 and 1, with 1  
480 indicating the assemblage with the highest proportion of unknown species (whatever such  
481 number might be). We note that adding up the *UnknownSR* across grid cells could overestimate  
482 the total number of unknown species if unknown species occur in more than one assemblage.  
483 The 220km and 440km spatial resolutions may therefore slightly underestimate percent of total

484 discoveries, but we retained these resolutions for the visual detail they offered. All assemblage-  
485 level estimates are available through Data S3 files.

486

### 487 **Country-level predictions**

488 We used the assemblage-level predictions to compute country-wide estimates of  
489 *UnknownSR* and *PropUnknown* at each spatial resolution. For each grid cell, we quantified the  
490 proportion of landcover that overlapped with countries (a same grid cell could be assigned to  
491 multiple countries, but the proportion of landcover may differ). Country selected grid cells had  
492 their values of *UnknownSR* and *PropUnknown* weighted by the respective proportion of country  
493 landcover. Given the same spatial resolution, the sum of *UnknownSR* returns the same value  
494 when computed across either countries or assemblages. We summed the *UnknownSR* across  
495 countries to get the number of unknown species per country, and averaged the *PropUnknown* to  
496 obtain the country wide proportion of unknown species. The country-level predictions were also  
497 standardized in the same way we did for taxa and assemblages. We divided the per country  
498 *UnknownSR* by the global number of *UnknownSR* to get the estimated percent of total  
499 discoveries, and standardized *PropUnknown* to vary between 0 and 1, with 1 indicating the  
500 country with the highest proportion of unknown species. Country boundaries followed the  
501 Global Administrative Units database, version 1.6<sup>92</sup>. All country-level estimates are available  
502 through Data S4 files.

503

### 504 **REFERENCES**

- 505 1. Costello, M. J., May, R. M. & Stork, N. E. Can We Name Earth's Species Before They Go Extinct? *Science*  
506 **339**, 413–416 (2013).

- 507 2. Mora, C., Rollo, A. & Tittensor, D. P. Comment on ‘Can We Name Earth’s Species Before They Go  
508 Extinct?’ *Science* **341**, 237 (2013).
- 509 3. Mora, C., Tittensor, D. P., Adl, S., Simpson, A. G. B. & Worm, B. How Many Species Are There on Earth  
510 and in the Ocean? *PLoS Biol.* **9**, e1001127 (2011).
- 511 4. May, R. & Beverton, R. J. H. How many species? *Philos. Trans. R. Soc. London. Ser. B Biol. Sci.* **330**, 293–  
512 304 (1990).
- 513 5. Scheffers, B. R., Joppa, L. N., Pimm, S. L. & Laurance, W. F. What we know and don’t know about Earth’s  
514 missing biodiversity. *Trends Ecol. Evol.* **27**, 501–510 (2012).
- 515 6. Raven, P. H. & Wilson, E. O. A fifty-year plan for biodiversity surveys. *Science* **258**, 1099–1100 (1992).
- 516 7. Whittaker, R. J. *et al.* Conservation biogeography: Assessment and prospect. *Divers. Distrib.* **11**, 3–23  
517 (2005).
- 518 8. Hortal, J. *et al.* Seven Shortfalls that Beset Large-Scale Knowledge of Biodiversity. *Annu. Rev. Ecol. Evol.*  
519 *Syst.* **46**, 523–549 (2015).
- 520 9. Secretariat of the Convention on Biological Diversity. Guide to the global taxonomy initiative. *CBD Tech.*  
521 *Ser.* **30**, 1–195, i–viii (2010).
- 522 10. Costello, M. J., May, R. M. & Stork, N. E. Response to Comments on ‘Can We Name Earth’s Species  
523 Before They Go Extinct?’ *Science* **341**, 237 (2013).
- 524 11. Bebber, D. P., Marriott, F. H. C., Gaston, K. J., Harris, S. A. & Scotland, R. W. Predicting unknown species  
525 numbers using discovery curves. *Proc. R. Soc. B* **274**, 1651–1658 (2007).
- 526 12. Edie, S. M., Smits, P. D. & Jablonski, D. Probabilistic models of species discovery and biodiversity  
527 comparisons. *Proc. Natl. Acad. Sci.* **114**, 3666–3671 (2017).
- 528 13. Guenard, B., Weiser, M. D. & Dunn, R. R. Global models of ant diversity suggest regions where new  
529 discoveries are most likely are under disproportionate deforestation threat. *Proc. Natl. Acad. Sci.* **109**, 7368–  
530 7373 (2012).
- 531 14. Blackburn, T. M. & Gaston, K. J. What Determines the Probability of Discovering a Species - a Study of



- 532 South-American Oscine Passerine Birds. *J. Biogeogr.* **22**, 7–14 (1995).
- 533 15. Costello, M. J., Lane, M., Wilson, S. & Houlding, B. Factors influencing when species are first named and  
534 estimating global species richness. *Glob. Ecol. Conserv.* **4**, 243–254 (2015).
- 535 16. Collen, B., Purvis, A. & Gittleman, J. L. Biological correlates of description date in carnivores and primates.  
536 *Glob. Ecol. Biogeogr.* **13**, 459–467 (2004).
- 537 17. Diniz-Filho, J. A. F. *et al.* Macroecological correlates and spatial patterns of anuran description dates in the  
538 Brazilian Cerrado. *Glob. Ecol. Biogeogr.* **14**, 469–477 (2005).
- 539 18. Costello, M. J., Houlding, B. & Joppa, L. N. Further evidence of more taxonomists discovering new species,  
540 and that most species have been named: response to Bebbler *et al.* (2014). *New Phytol.* **202**, 739–740 (2014).
- 541 19. Meiri, S. Small, rare and trendy: traits and biogeography of lizards described in the 21st century. *J. Zool.*  
542 **299**, 251–261 (2016).
- 543 20. Diniz-Filho, J. A. F. *et al.* Macroecological correlates and spatial patterns of anuran description dates in the  
544 Brazilian Cerrado. *Glob. Ecol. Biogeogr.* **14**, 469–477 (2005).
- 545 21. Klein, J. P. & Moeschberger, M. L. *Survival analysis: Techniques for censored and truncated data.*  
546 *Pharmaceutical Statistics* (Springer, 2003). doi:10.1002/pst.135
- 547 22. Essl, F., Rabitsch, W., Dullinger, S., Moser, D. & Milasowszky, N. How well do we know species richness  
548 in a well-known continent? Temporal patterns of endemic and widespread species descriptions in the  
549 European fauna. *Glob. Ecol. Biogeogr.* **22**, 29–39 (2013).
- 550 23. Colli, G. R. *et al.* In the depths of obscurity: Knowledge gaps and extinction risk of Brazilian worm lizards  
551 (Squamata, Amphisbaenidae). *Biol. Conserv.* **204**, 51–62 (2016).
- 552 24. Burgin, C. J., Colella, J. P., Kahn, P. L. & Upham, N. S. How many species of mammals are there? *J.*  
553 *Mammal.* **99**, 1–14 (2018).
- 554 25. Meyer, C., Kreft, H., Guralnick, R. & Jetz, W. Global priorities for an effective information basis of  
555 biodiversity distributions. *Nat. Commun.* **6**, 8221 (2015).
- 556 26. Bellard, C. *et al.* Vulnerability of biodiversity hotspots to global change. *Glob. Ecol. Biogeogr.* **23**, 1376–

- 557 1386 (2014).
- 558 27. Quintero, I. & Jetz, W. Global elevational diversity and diversification of birds. *Nature* **555**, 246–250  
559 (2018).
- 560 28. Roll, U. *et al.* The global distribution of tetrapods reveals a need for targeted reptile conservation. *Nat. Ecol.*  
561 *Evol.* **1**, 1677–1682 (2017).
- 562 29. Garnett, S. T. & Christidis, L. Taxonomy anarchy hampers conservation. *Nature* **546**, 25–27 (2017).
- 563 30. Isaac, N. J. B., Mallet, J. & Mace, G. M. Taxonomic inflation: its influence on macroecology and  
564 conservation. *Trends Ecol. Evol.* **19**, 464–469 (2004).
- 565 31. Bremer, K., Bremer, B., Karis, P. & Källersjö, M. Time for change in taxonomy. *Nature* **343**, 202–202  
566 (1990).
- 567 32. Raposo, M. A. *et al.* What really hampers taxonomy and conservation? A riposte to Garnett and Christidis  
568 (2017). *Zootaxa* **4317**, 179 (2017).
- 569 33. Wake, D. B. Persistent Plethodontid Themes: Species, Phylogenies, and Biogeography. *Herpetologica* **73**,  
570 242–251 (2017).
- 571 34. Tedesco, P. A. *et al.* Estimating How Many Undescribed Species Have Gone Extinct. *Conserv. Biol.* **28**,  
572 1360–1370 (2014).
- 573 35. Jetz, W., McPherson, J. M. & Guralnick, R. P. Integrating biodiversity distribution knowledge: toward a  
574 global map of life. *Trends Ecol. Evol.* **27**, 151–159 (2012).
- 575 36. Jetz, W., Thomas, G. H., Joy, J. B., Hartmann, K. & Mooers, A. O. The global diversity of birds in space  
576 and time. *Nature* **491**, 444–448 (2012).
- 577 37. Jetz, W. & Pyron, R. A. The interplay of past diversification and evolutionary isolation with present  
578 imperilment across the amphibian tree of life. *Nat. Ecol. Evol.* **2**, 850–858 (2018).
- 579 38. Upham, N. S., Esselstyn, J. A. & Jetz, W. Inferring the mammal tree: Species-level sets of phylogenies for  
580 questions in ecology, evolution, and conservation. *PLOS Biol.* **17**, e3000494 (2019).
- 581 39. González-del-Pliego, P. *et al.* Phylogenetic and Trait-Based Prediction of Extinction Risk for Data-Deficient  
582 Amphibians. *Curr. Biol.* **29**, 1557-1563.e3 (2019).

- 583 40. Moura, M. R. *et al.* Geographical and socioeconomic determinants of species discovery trends in a  
584 biodiversity hotspot. *Biol. Conserv.* **220**, 237–244 (2018).
- 585 41. Gaston, K. J., Blackburn, T. M. & Loder, N. Which species are described first? The case of North-American  
586 butterflies. *Biodivers. Conserv.* **4**, 119–127 (1995).
- 587 42. Oliveira, B. F., São-Pedro, V. A., Santos-Barrera, G., Penone, C. & Costa, G. C. AmphiBIO, a global  
588 database for amphibian ecological traits. *Sci. Data* **4**, 170123 (2017).
- 589 43. Feldman, A., Sabath, N., Pyron, R. A., Mayrose, I. & Meiri, S. Body sizes and diversification rates of  
590 lizards, snakes, amphisbaenians and the tuatara. *Glob. Ecol. Biogeogr.* **25**, 187–197 (2016).
- 591 44. Hallmann, K. & Griebeler, E. M. An exploration of differences in the scaling of life history traits with body  
592 mass within reptiles and between amniotes. *Ecol. Evol.* **8**, 5480–5494 (2018).
- 593 45. Slavenko, A., Itescu, Y., Ihlow, F. & Meiri, S. Home is where the shell is: predicting turtle home range  
594 sizes. *J. Anim. Ecol.* **85**, 106–114 (2016).
- 595 46. Regis, K. W. & Meik, J. M. Allometry of sexual size dimorphism in turtles: a comparison of mass and  
596 length data. *PeerJ* **5**, e2914 (2017).
- 597 47. Itescu, Y., Karraker, N. E., Raia, P., Pritchard, P. C. H. & Meiri, S. Is the island rule general? Turtles  
598 disagree. *Glob. Ecol. Biogeogr.* **23**, 689–700 (2014).
- 599 48. Upham, N. S., Esselstyn, J. A. & Jetz, W. Ecological causes of uneven diversification and richness in the  
600 mammal tree of life. *bioRxiv* (2019). doi:10.1101/504803
- 601 49. Wilman, H. *et al.* EltonTraits 1.0: Species-level foraging attributes of the world’s birds and mammals.  
602 *Ecology* **95**, 2027–2027 (2014).
- 603 50. Tonini, J. F. R., Beard, K. H., Ferreira, R. B., Jetz, W. & Pyron, R. A. Fully-sampled phylogenies of  
604 squamates reveal evolutionary patterns in threat status. *Biol. Conserv.* (2016).  
605 doi:10.1016/j.biocon.2016.03.039
- 606 51. Goolsby, E. W., Bruggeman, J. & Ané, C. Rphylopars : fast multivariate phylogenetic comparative methods  
607 for missing data and within-species variation. *Methods Ecol. Evol.* **8**, 22–27 (2017).
- 608 52. Gaston, K. J., Blackburn, T. M. & Lawton, J. H. Interspecific Abundance-Range Size Relationships: An  
609 Appraisal of Mechanisms. *J. Anim. Ecol.* **66**, 579 (1997).
- 610 53. Borregaard, M. K. & Rahbek, C. Causality of the Relationship between Geographic Distribution and Species

- 611 Abundance. *Q. Rev. Biol.* **85**, 3–25 (2010).
- 612 54. IUCN - International Union for Conservation of Nature. IUCN Red List of Threatened Species. *Version*  
613 *2018* [www.iucnredlist.org](http://www.iucnredlist.org) (2018).
- 614 55. Freitag, S., Hobson, C., Biggs, H. C. & Jaarsveld, A. S. Testing for potential survey bias: the effect of roads,  
615 urban areas and nature reserves on a southern African mammal data set. *Anim. Conserv.* **1**, 119–127 (1998).
- 616 56. Kier, G. & Barthlott, W. Measuring and mapping endemism and species richness: a new methodological  
617 approach and its application on the flora of Africa. *Biodivers. Conserv.* **10**, 1513–1529 (2001).
- 618 57. Vilela, B. & Villalobos, F. letsR: a new R package for data handling and analysis in macroecology. *Methods*  
619 *Ecol. Evol.* **6**, 1229–1234 (2015).
- 620 58. Papavero, N. *Essays on the History of Neotropical Dipterology: with special reference to collectors: 1750-*  
621 *1905: Vol. I.* (Museu de Zoologia da Universidade de São Paulo, 1971). doi:10.5962/bhl.title.101715
- 622 59. Baselga, A., Lobo, J. M., Hortal, J., Jiménez-Valverde, A. & Gómez, J. F. Assessing alpha and beta  
623 taxonomy in eupelmid wasps: determinants of the probability of describing good species and synonyms. *J.*  
624 *Zool. Syst. Evol. Res.* **48**, 40–49 (2010).
- 625 60. Yang, W., Ma, K. & Kreft, H. Environmental and socio-economic factors shaping the geography of floristic  
626 collections in China. *Glob. Ecol. Biogeogr.* **23**, 1284–1292 (2014).
- 627 61. Karger, D. N. *et al.* Climatologies at high resolution for the earth's land surface areas. *Sci. Data* **4**, 170122  
628 (2017).
- 629 62. R Core Team. R: A Language and Environment for Statistical Computing. v. 3.5.3 (2019).
- 630 63. Hijmans, R. J. raster: Geographic Data Analysis and Modeling. <https://cran.r-project.org/package=raster>  
631 (2015).
- 632 64. Amatulli, G. *et al.* A suite of global, cross-scale topographic variables for environmental and biodiversity  
633 modeling. *Sci. Data* **5**, 180040 (2018).
- 634 65. Klein Goldewijk, K., Beusen, A., Van Drecht, G. & De Vos, M. The HYDE 3.1 spatially explicit database  
635 of human-induced global land-use change over the past 12,000 years. *Glob. Ecol. Biogeogr.* **20**, 73–86  
636 (2011).
- 637 66. Joppa, L. N., Roberts, D. L. & Pimm, S. L. The population ecology and social behaviour of taxonomists.  
638 *Trends Ecol. Evol.* **26**, 551–553 (2011).

- 639 67. Wickham, H. stringr: Simple, Consistent Wrappers for Common String Operations. R package version 1.3.1.  
640 <http://stringr.tidyverse.org> (2018).
- 641 68. Mahto, A. splitstackshape: Stack and Reshape Datasets After Splitting Concatenated Values. R package  
642 version 1.4.6. <http://github.com/mrdwab/splitstackshape> (2018).
- 643 69. Dinerstein, E. *et al.* An Ecoregion-Based Approach to Protecting Half the Terrestrial Realm. *Bioscience* **67**,  
644 534–545 (2017).
- 645 70. Jetz, W. & Fine, P. V. A. Global Gradients in Vertebrate Diversity Predicted by Historical Area-Productivity  
646 Dynamics and Contemporary Environment. *PLoS Biol.* **10**, e1001292 (2012).
- 647 71. Kutner, M. H., Nachtsheim, C. J., Neter, J. & Li, W. *Applied Linear Statistical Models*. (McGraw-Hill  
648 Irwin, 2004).
- 649 72. Naimi, B. usdm: Uncertainty Analysis for Species Distribution Models. [https://cran.r-](https://cran.r-project.org/package=usdm)  
650 [project.org/package=usdm](https://cran.r-project.org/package=usdm) (2017).
- 651 73. Bebber, D. P. *et al.* Herbaria are a major frontier for species discovery. *Proc. Natl. Acad. Sci.* **107**, 22169–  
652 22171 (2010).
- 653 74. Guedes, J. J. M., Feio, R. N., Meiri, S. & Moura, M. R. Identifying factors that boost species discoveries of  
654 global reptiles. *Zool. J. Linn. Soc.* in press (2020). doi:10.1093/zoolinnean/zlaa029
- 655 75. von Linné, C. *Caroli Linnaei...Systema naturae per regna tria naturae: secundum classes, ordines, genera,*  
656 *species, cum characteribus, differentiis, synonymis, locis.* (Impensis Direct. Laurentii Salvii, 1758).  
657 doi:10.5962/bhl.title.542
- 658 76. Harrell, F. E. *Regression Modeling Strategies*. (Springer New York, 2001). doi:10.1007/978-1-4757-3462-1
- 659 77. George, B., Seals, S. & Aban, I. Survival analysis and regression models. *J. Nucl. Cardiol.* **21**, 686–694  
660 (2014).
- 661 78. Jackson, C. flexsurv : A Platform for Parametric Survival Modeling in R. *J. Stat. Softw.* **70**, 1–33 (2016).
- 662 79. Burnham, K. P. & Anderson, D. R. *Model Selection and Multimodel Inference*. (Springer New York, 2004).  
663 doi:10.1007/b97636
- 664 80. Burnham, K. P. & Anderson, D. R. *Model Selection and Multimodel Inference: A Practical Information-*  
665 *Theoretic Approach. Ecological Modelling* **172**, (Springer, 2002).
- 666 81. Johnson, J. B. & Omland, K. S. Model selection in ecology and evolution. *Trends Ecol. Evol.* **19**, 101–108

- 667 (2004).
- 668 82. Barton, K. MuMIn: Multi-Model Inference. R package version 1.43.6. 1–74 (2019). Available at:  
669 <https://cran.r-project.org/package=MuMIn>.
- 670 83. Alexander Pyron, R. & Wiens, J. J. A large-scale phylogeny of Amphibia including over 2800 species, and a  
671 revised classification of extant frogs, salamanders, and caecilians. *Mol. Phylogenet. Evol.* **61**, 543–583  
672 (2011).
- 673 84. Pyron, R. A., Burbrink, F. T. & Wiens, J. J. A phylogeny and revised classification of Squamata, including  
674 4161 species of lizards and snakes. *BMC Evol. Biol.* **13**, 93 (2013).
- 675 85. Fisher, D. O. & Blomberg, S. P. Correlates of rediscovery and the detectability of extinction in mammals.  
676 *Proc. R. Soc. B Biol. Sci.* **278**, 1090–1097 (2011).
- 677 86. Jetz, W., Sekercioglu, C. H. & Böhning-Gaese, K. The Worldwide Variation in Avian Clutch Size across  
678 Species and Space. *PLoS Biol.* **6**, e303 (2008).
- 679 87. Jetz, W. & Rubenstein, D. R. Environmental Uncertainty and the Global Biogeography of Cooperative  
680 Breeding in Birds. *Curr. Biol.* **21**, 72–78 (2011).
- 681 88. Jetz, W. & Rahbek, C. Geographic range size and determinants of avian species richness. *Science* **297**,  
682 1548–1551 (2002).
- 683 89. Dowle, M. & Srinivasan, A. data.table: Extension of ‘data.frame’. R package version 1.12.4. (2019).  
684 Available at: <https://cran.r-project.org/package=data.table>.
- 685 90. Gaston, K. J., Chown, S. L. & Evans, K. L. Ecogeographical rules: elements of a synthesis. *J. Biogeogr.* **35**,  
686 483–500 (2008).
- 687 91. Violle, C., Reich, P. B., Pacala, S. W., Enquist, B. J. & Kattge, J. The emergence and promise of functional  
688 biogeography. *Proc. Natl. Acad. Sci.* **111**, 13690–13696 (2014).
- 689 92. GDAM. Database of Global Administrative Areas, version 3.6. (2019). Available at: <http://www.gadm.org>.

690 **ACKNOWLEDGMENTS**

691 We are grateful to S. Meiri, D.S. Rinnan, G. Reygondeau, N. Upham, M. Costello, D. Wake, and  
692 Joaquín Hortal for providing helpful comments on the research or manuscript drafts. We thank  
693 C. Haddad, L.C. Márquez, G. Singh, and A.F. Meyer for providing pictures of the example  
694 species in Figure 1. This work was produced, in part, with the support of the National  
695 Geographic Society. W.J. also acknowledges support from the E.O. Wilson Biodiversity  
696 Foundation and Half-Earth Project, NSF grant DEB-1441737 and NASA Grants  
697 80NSSC17K0282 and 80NSSC18K0435.

698

699 **AUTHOR CONTRIBUTIONS**

700 MRM and WJ conceived the study; MRM analysed the data, MRM and WJ developed the  
701 figures, MRM and WJ wrote the text.

702

703 **COMPETING INTERESTS**

704 Authors declare no competing interests.

705

706 **DATA AVAILABILITY**

707 All datasets of species distributions used in the analyses are available at Map of Life  
708 (<https://mol.org>). R-scripts will be made publicly available on publication.



709 **SUPPLEMENTARY INFORMATION**

710 **SUPPLEMENTARY METHODS**

711 We performed model validations to verify if our model framework were able to satisfactorily  
712 predict species discoveries based on random species subsets left out of the model training. We  
713 also performed extensive sensitivity analysis to assess limitations of our approach to three issues  
714 detailed in the following.

715

716 **Model validation**

717 We used a four-fold cross-validation approach to examine the predictive accuracy of our  
718 models. For each vertebrate group, the species were randomly partitioned into four equal parts,  
719 with three of those used as the training-fold and the fourth as validation-fold. The cross-  
720 validation process was repeated four times, with each of the four-fold subsamples used once as  
721 the validation data. For each cross-validation round, we performed the model averaging approach  
722 using the 11 predictor variables and obtained for each species, the weighted average value of the  
723 (i) discovery probability in year 2014, and (ii) estimated description date.

724 For each training-fold, we obtained the coefficients of all possible AFT models (2,047  
725 models covering all predictor combinations), as well as their BIC weights (wBIC). We used the  
726 model coefficients of each training-fold to predict the description dates of the 25% of species left  
727 out of the model fitting (validation-fold). These predictions were obtained for each one of the  
728 2,047 AFT models, in each training-fold. For each species, we then calculated the weighted  
729 average of the predicted description dates using wBIC of each model as relative weights. These  
730 species-level predictions of the description date were therefore independent from the model  
731 fitting.

732 At this point, we proceeded with model validation at three different levels: species, taxon,  
733 assemblage.

#### 734 1) *Species-level validation*

735 We assessed the accuracy of the predictions of description dates from the average weighted  
736 AFT model with three different statistics. (i) The Spearman correlation measured our ability to  
737 correctly rank species according to their discovery year. (ii) The slope of the linear regression  
738 between observed and predicted description dates assessed under- or overestimation of absolute  
739 values. (iii) The normalized root mean square error (NRMSE) measured the divergence of  
740 predictions from observations <sup>1</sup>. NRMSE is given in percentage, where lower values indicate less  
741 residual variance [ $\text{NRMSE} = (\sqrt{\sum_1^N (\hat{x}_i - x_i)^2 / N}) / (x_{\max} - x_{\min})$ ]. These three statistics were  
742 computed for each the four vertebrate groups. Computations were performed in R using the  
743 ‘hydroGOF’ <sup>2</sup> and ‘stats’ <sup>3</sup> packages.

#### 744 2) *Taxon-level validation*

745 For each cross-validation round, we calculated taxon-level estimates of *UnknownSR* as  
746 explained above. We averaged the outputs of the four cross-validation rounds to get the  
747 estimated *UnknownSR* (hereafter, estimated discoveries). Using the 25% of the species left out of  
748 the model, we obtained the observed number of unknown species (observed discoveries) per  
749 taxon, that is, the species richness per taxon based on the validation fold.

750 To evaluate predictive accuracy, we calculated three different statistics using the observed  
751 and estimated discoveries per taxon. (i) The Spearman correlation measured our ability to  
752 correctly rank taxonomic genera, families, and orders with respect to the number of unknown  
753 species. (ii) The slope of the linear regression between observed and estimated discoveries was  
754 used to check for under- or overestimation of the estimated discoveries. (iii) The normalized root

755 mean square error (NRMSE). These three statistics were computed for each taxonomic rank  
756 (family, order) of each of the four vertebrate groups. Computations were performed in R using  
757 the ‘hydroGOF’<sup>2</sup> and ‘stats’<sup>3</sup> packages.

### 758 *3) Assemblage-level validation*

759 This validation followed that conducted at the taxon level. Here, the outputs of the four  
760 cross-validation rounds were averaged to get estimated *UnknownSR* (estimated discoveries) per  
761 assemblage grid cell. We used the 25% of species in the validation fold to extract the observed  
762 number of unknown species (observed discoveries) per assemblage grid cell.

763 The relationship between observed and estimated discoveries per grid cell was assessed  
764 through three different statistics: (i) Spearman correlation, (ii) slope of the linear regression, and  
765 the (iii) normalized root mean square error (NRMSE). These three statistics were computed for  
766 each subsampling level (using 1, 5, 10, 50, 100, 200 and 500 grid cell occurrences per species)  
767 and the observed data (all grid cell occurrences per species), at the three spatial resolutions (grain  
768 sizes 220, 440, and 880 km), and for each vertebrate group. Computations were performed in R  
769 using the ‘hydroGOF’<sup>2</sup> and ‘stats’<sup>3</sup> packages.

### 770 *4) Country-level validation*

771 Per country results are aggregates of the assemblage-level predictions, and are represented  
772 by the assemblage-level validation.

773

### 774 **Model limitations**

775 It is important to note that the estimated time-to-event functions derived from the AFT  
776 models may overestimate the discovery probability if the empirical time-to-event functions do  
777 not yet approach asymptote<sup>4</sup>. Consequently, the estimated proportion of known species,

778 *PropKnown*, for species in a taxon or assemblage will also be overestimated, ultimately leading  
779 to more conservative estimates of the *PropUnknown* and *UnknownSR*. We acknowledge that the  
780 discovery curve of amphibians and reptiles have not yet approach asymptote, as evidenced by the  
781 cumulative number of species description over time (Fig. 1, main text). That said, our model  
782 framework is subject to three major limitations.

783 First, time-to-event data are often incompletely observed, in which case the data may be  
784 considered censored and/or truncated. Broadly speaking, truncation is related to the study design  
785 and is further divided in two types. Left truncation arises from the specification of a minimum  
786 entry time for the sampling units. If the event occurs before the minimum entry time, those  
787 sampling units will never enter the study. Right truncation occurs when sampling units are only  
788 observable if they have experienced the event <sup>5</sup>. Our species data shows this latter feature. The  
789 sampling condition imposed to our data may affect our estimates of discovery probability.

790 Second, over the temporal scale considered in this study (almost 260 years), taxonomists  
791 have dealt with different issues to describe new species. For example, earlier naturalists crossed  
792 oceans on ships to find unknown taxa in regions previously considered highly remote. In  
793 contrast, modern taxonomists have been required to use multiple tools to accumulate more  
794 evidences to provide highly detailed descriptions of new species <sup>6</sup>. Although these different  
795 technological contexts are important to the discovery process, they are difficult to measure and  
796 incorporate into our models. Both predictors and model performance may vary through time. In  
797 using only species described more recently as input data, we might get different results.

798 And third, our model validation procedure is subject to two mathematical constraints. The  
799 number of estimated discoveries is affected by the size of the training-fold. In adding more  
800 species to the training-fold, we tend to increase the known richness per taxon or assemblage and

801 consequently, the estimated unknown richness (*UnknownSR*). Moreover, the number of observed  
802 discoveries per taxon or assemblage is dependent on the size of the validation-fold; that is, the  
803 number of species left out of the model training. In increasing the size of validation-fold we tend  
804 to obtain higher values of observed discoveries, with the opposite occurring if we decrease the  
805 size of validation-fold. Therefore, the size of validation- and training-folds may affect how we  
806 rank taxa and regions according to their potential for future species discovery.

807 We investigate these three limitations further below.

### 808 *1) Robustness to sampling condition*

809 The right truncation feature creates a sampling condition,  $S_i \leq T_{max}$ , where  $S_i$  is the  
810 taxonomic age of species  $i$  (i.e., the number of years since 1758), and  $T_{max}$  equals the temporal  
811 range of this study ( $2014 - 1758 = 256$  years). Only species with the taxonomic age  $\leq 256$   
812 entered the study. Underlying this sampling condition is the assumption that the chance of an  
813 event at the time  $T > T_{max}$  is zero [ $P(T > T_{max}) = 0$ ]. This condition may be especially relevant if  
814  $T_{max}$  is too short, which could lead to an underrepresentation of recent-described species.  
815 Consequently, the importance of species-level attributes may depart from their true effect,  
816 leading to biased estimates of discovery probability.

817 Although the theoretical background to deal with incompletely observed time-to-event data  
818 has improved in the recent decades<sup>5</sup>, the statistical tools available to account for right truncation  
819 are either restricted to particular time-to-event models - e.g. Cox Proportional Hazard Model<sup>7</sup> -  
820 or have somewhat limited application due to the fail in estimator convergence<sup>8</sup>. Herein, we  
821 assessed the robustness of our results to the violation of this sampling condition through a  
822 sensitivity analysis. We created 10 data subsets containing increasing levels of right truncation.  
823 To do so, we successively discarded  $x\%$  of the most recent-described species, with  $x$  varying

824 from 0 to 50, at intervals of 5%. For each data subset, we repeated the model averaging  
825 framework and registered the average weighted coefficients of each predictor variable, as well as  
826 the estimated discovery probability of each species.

827 This sensitivity analysis is intended to answer three questions. First, how the effect size of  
828 predictors varies when the AFT models are trained with datasets holding higher levels of right-  
829 truncation? Second, how the changes in effect size (if any) affect the estimation of species'  
830 discovery probability? To answer this latter question is important to compare the discovery  
831 probabilities for the same set of species. For this purpose, we used the oldest half of the known  
832 species, since these species were included in all data subsets. And third, what kind of the species-  
833 level attributes are expected for species yet to be described, and how those attributes would  
834 affect the estimation of discovery probability?

#### 835 *2) Influence of the time period of species discovery*

836 To assess the differential influence of early species discoveries in our model performance,  
837 we performed a sensitivity analysis by successively discarding previously described species. Our  
838 goal was i) to evaluate the variation of predictors across time and ii) to identify the period that  
839 offered strongest model performance for prediction. In addition to the full time period of this  
840 study (1759-2014), we defined other 22 time periods covering the interval from  $d$  to 2014, where  
841  $d$  is a decade from 1760 to 1970 (e.g. 1760-2014, 1770-2014, ..., 1970-2014). We then filtered  
842 our species dataset to include only those taxa described within each time period and repeated the  
843 model validation framework at the level of species, taxa, and assemblages. We note that with  
844 increasing  $d$  the sample sizes available in the datasets decreased (Table S3).

845 For the sensitivity analysis at the species-level, we investigated the relationship between  
846 observed and predicted description dates. At the level of taxa and assemblages, the sensitivity

847 analyses dealt with the association between observed and estimated discoveries per taxonomic  
848 rank and assemblage grid cell, respectively. We obtained three statistics of model evaluation for  
849 the relationship of interest in each time period: (i) Spearman  $r$ , (ii) regression slope, and (iii)  
850 NRMSE. After identifying which time period returned the best model performance (see  
851 Supplementary Text section), we ran an additional analysis using the dataset for the best time  
852 period to obtain the species-level discovery metrics used to characterize discovery trends per  
853 taxon and per assemblage.

### 854 *3) Influence of the size of cross validation partitions*

855 To elucidate if the mathematical constraints of our model validation affected our results, we  
856 repeated the procedure using four different sizes of validation- and training-folds. More  
857 specifically, we included 25, 50, 75, and 90% of randomly selected species in the training-fold,  
858 while keeping each respective complement (75, 50, 25, and 10% of species) in the validation-  
859 fold. We then repeated the model validation procedure at the level of taxa and assemblages, and  
860 registered three different statistics to assess the relationship between observed and estimated  
861 discoveries: (i) Spearman correlation, (ii) slope of the linear regression, and the (iii) normalized  
862 root mean square error (NRMSE). This sensitivity analysis was also repeated across the different  
863 time periods of species discovery discussed above. Computations were performed in R using the  
864 ‘hydroGOF’<sup>2</sup> and ‘stats’<sup>3</sup> packages.

865

## 866 **SUPPLEMENTARY RESULTS**

### 867 **Model Limitations and Sensitivity Analyses**

#### 868 *1) Robustness to sampling condition*

869 We found low variation in the effect size of model coefficients if up to 30% of more  
870 recently described species were discarded before model computations (Extended Data, Fig. S1).  
871 The increasing of right-truncation decreased the effect size of model coefficients. Only a few  
872 covariates changed the effect direction with increasing levels of right-truncation. The model  
873 coefficients were nearly invariant for birds.

874 The variation in model coefficients is expected to influence the estimated discovery  
875 probability. In increasing the level of the right-truncation of the species dataset, we obtained  
876 higher values of discovery probability relative to those estimated from more complete datasets  
877 (Extended Data, Fig. S2). Such changes were evident mostly for amphibians and reptiles, and  
878 they were virtually nonexistent for mammals and especially birds.

879 Among the most consistent predictors affecting the discovery probabilities there were the (i)  
880 geographic range size, (ii) taxonomic activity per biome, and (iii) species body size. The  
881 frequency distribution of these predictor variables (Extended Data, Figs S3-S6) confirms the  
882 well-known trend of recent-described species to show narrower geographic ranges, smaller body  
883 sizes, and be described by more taxonomists relative to species described long ago<sup>9-11</sup>.

884 It is worth noting that this sensitivity analysis indirectly considered the modelling of  
885 discovery probability for species datasets with different ending dates, in an opposite way to the  
886 analysis on the influence of the time period of species discovery (next subtopic). Here, species  
887 datasets always started in 1759 but ended at different dates, according to the percentage of  
888 recently described species discarded before computations. Thus, the ending dates were not  
889 necessarily equal for a same percentage of discarded species. For instance, the first half of the  
890 currently known amphibian species were described by 1972, whereas 50% of the bird diversity  
891 were already known by 1845 (see histograms of description year, Figure 1).



892 Had we been able to incorporate species-level attribute of unknown species, we would have  
893 found higher effect size for the most important model coefficients and therefore estimated lower  
894 species discovery probabilities. Given the right-truncation nature of data used in our analysis, it  
895 is likely that our discovery metrics, the *PropUnknown* and *UnknownSR* per taxon or per  
896 assemblage, are underestimated. The proportion and number of unknown species we report  
897 should be considered as conservative estimates, particularly for amphibians and reptiles.

## 898 2) *Influence of the time period of species discovery*

899 We evaluated the sensitivity of our ‘discovery metrics’ to the time period of species  
900 discovery (e.g. 1760-2014, 1770-2014, ..., 1970-2014). In discarding species described long ago,  
901 most of the model coefficients decreased their effect size (Fig. 2, main text). The ability of the  
902 AFT models to correctly predict the species description dates did not improve after excluding  
903 earlier-described species, except in mammals (but at the cost of discarding more than 70% of all  
904 known mammal species; Table S3, Extended Data, Fig. S7).

905 At the taxon-level, the Spearman correlation between observed and estimated discoveries  
906 was roughly constant after discarding species described during the first century of the modern  
907 taxonomy (Extended Data, Fig. S8). At the assemblage level, the model performance also  
908 decreased as old described species were discarded before computations. The decline in model  
909 performance was evident across all subsampling levels used to control the overrepresentation of  
910 wide-ranging species (Extended Data, Figs S9-S12). The subsampling level of 5 occurrences per  
911 species showed the best performance in explaining the observed discoveries per grid cell, a result  
912 consistent across the all spatial resolutions and vertebrate classes.

913 Overall, we did not obtain better measures of estimated discoveries by removing species  
914 described long ago. We therefore used the complete dataset (species described from 1759 to

915 2014) to estimate the discovery probability at the species-level and to estimate the number of  
916 unknown species (*UnknownSR*) at the taxon- and assemblage-level. Overall, we found a strong  
917 association with description year in extracting the predicted year of discovery at the threshold of  
918 0.5 of discovery probability (Spearman  $r = 0.65$ – $0.81$  for all groups, see Extended Data, Fig.  
919 S13).

### 920 3) Influence of the size of cross-validation partitions

921 Across taxonomic ranks, we found similar values of Spearman correlation between  
922 observed and estimated discoveries, regardless of the size of the species dataset used in the  
923 model training and mapping procedure (Extended Data, Fig. S8). The size of cross-validation  
924 partitions did not influence model performance when using different time periods of species  
925 discovery either (Extended Data, Fig. S8). The extraction of *PropUnknown* based on small  
926 sample sizes (i.e., using 25% of species in the model training) were less able to properly  
927 characterize discovery patterns at the taxon-level relative to large sample sizes (including 50, 75,  
928 90% of species in the model training).

929 The regression slope between observed and estimated discoveries per taxon tended to  
930 decrease when assessed at higher taxonomic ranks and for training-folds containing more species  
931 (Extended Data, Fig. S8). The estimated discoveries underestimated the observed discoveries  
932 across all taxonomic ranks, although such underestimation were less pronounced when using  
933 90% of species in the model training (Table S4). After standardizing *UnknownSR* to percent of  
934 total discoveries, we observed similar model performance for all sizes of the cross-validation  
935 partitions (Extended Data, Fig. S14).

936 At the assemblage-level, the relationship between observed and estimated discoveries  
937 showed similar values of Spearman correlation, regardless of the size of cross-validation

938 partitions (Extended Data, Figs S9-S12). Once again, subsampling 5 grid cell occurrences per  
939 species resulted in estimated discoveries that better predicted the observed discoveries per grid  
940 cell, regardless of the size of cross validation partitions. The discrepancy between the absolute  
941 number of observed and estimated discoveries per grid cell decreased when species assemblages  
942 were defined at either coarser spatial resolutions or based in training-folds including higher  
943 proportion of randomly selected species (Table S4). After standardizing *UnknownSR* to percent  
944 of total discoveries, the relationship between estimated and observed discoveries per grid cell  
945 was also constant for all sizes of the cross-validation partitions (Extended Data, Fig. S15).

946 Overall, the estimated discoveries we obtained, either at the taxon- or assemblage-level  
947 analyses, underestimated the observed discoveries. The underestimation was higher when the  
948 discovery metrics were computed for models including fewer species (Table S4). We therefore  
949 recommend caution in interpreting the absolute values of the estimated number of unknown  
950 species (*UnknownSR*) per taxon or per grid cell. We reinforce the strong monotonicity between  
951 the observed and estimated discoveries, which ultimately support our ability to rank taxa and  
952 regions according to their potential for the discovery of new species. We do not advise using this  
953 approach to update global numbers of unknown species, unless extensive cross validation reveal  
954 absence of under- or overestimation of *UnknownSR*.

#### 955 4) *Species authority name assumptions*

956 Our time-to-description analysis is based on original species authority names and uses the  
957 year in which a given binomial was originally proposed. Authority name description years reveal  
958 patterns of species discovery over time<sup>4,9-15</sup>, but do not account for the complicated taxonomic  
959 history of synonymizations and revalidations associated with a binomial name. A species that  
960 was recently revalidated still holds the same authority name – and therefore description year – as

961 when it was first recognized as unique taxon. Therefore, our findings concern factors affecting  
962 the year when a species was first discovered, and not if and when a revalidation occurred.

963 It is worth noting that a species description is a scientific hypothesis <sup>16</sup>, which can be  
964 revisited if more data become available, as often illustrated in integrative taxonomy studies <sup>17</sup>. It  
965 is not uncommon for broad studies such as taxonomic reviews to split a previously described  
966 species into multiple species, and in some cases, to resurrect synonyms or elevate subspecies to  
967 species rank. Although such ‘splitting’ might sometimes be viewed as undesirable <sup>18,19</sup>, when  
968 driven by scientific insights is a vital part of the taxonomic knowledge evolution, as taxonomies  
969 are not static over time <sup>16,20</sup>. Among 149 integrative taxonomy studies recently published in  
970 vertebrates (including fish), 40% consider all species as valid without changes, 31% pointed out  
971 at least one undescribed species but did not formally describe it, and 30% described at least one  
972 new species <sup>6</sup>. Thus, at least among vertebrates, the identification of new lineages (if any) is not  
973 necessarily followed by the proposition of new names.

974 Many firsthand species discoveries, i.e. species descriptions that propose new authority  
975 binomials, are reported for vertebrates every year. For example, more than 85% of amphibian  
976 species described between 1992-2003 resulted from newly proposed names, with less than 15%  
977 of descriptions concerning elevation of subspecies to species or revalidation of synonymies <sup>21</sup>. In  
978 reptiles, 79% of species descriptions between 1992-2017 were published outside revisionary  
979 taxonomic studies <sup>22</sup>. Newly proposed binomial names are also a large part of recently described  
980 mammals. Since 2005, 1,251 new mammal species have been recognized as valid, with 42% of  
981 them comprising firsthand discoveries and 58% consisting of resurrection of synonymies or  
982 elevation of subspecies to species rank <sup>23</sup>. Altogether, these estimates illustrate the size of the  
983 taxonomic enterprise ahead. Concerns might go beyond differentiating firsthand discoveries

984 from species revalidations, questioning the validity of species descriptions under the application  
985 of different species concepts<sup>18</sup>. To date, reflecting the ongoing debate about species concepts in  
986 biology<sup>24</sup>, comprehensive taxonomic databases that standardize global species lists according to  
987 a single species concept remain out of reach.

988

## 989 SUPPLEMENTARY REFERENCES

- 990 1. Barnston, A. G. Correspondence among the Correlation, RMSE, and Heidke Forecast Verification  
991 Measures; Refinement of the Heidke Score. *Weather Forecast.* **7**, 699–709 (1992).
- 992 2. Zambrano-Bigiarini, M. hydroGOF: Goodness-of-fit functions for comparison of simulated and observed  
993 hydrological time series. (2017). doi:10.5281/zenodo.840087
- 994 3. R Core Team. R: A Language and Environment for Statistical Computing. v. 3.5.3 (2019).
- 995 4. Bebber, D. P., Marriott, F. H. C., Gaston, K. J., Harris, S. A. & Scotland, R. W. Predicting unknown species  
996 numbers using discovery curves. *Proc. R. Soc. B* **274**, 1651–1658 (2007).
- 997 5. Klein, J. P. & Moeschberger, M. L. *Survival analysis: Techniques for censored and truncated data.*  
998 *Pharmaceutical Statistics* (Springer, 2003). doi:10.1002/pst.135
- 999 6. Pante, E., Schoelinck, C. & Puillandre, N. From Integrative Taxonomy to Species Description: One Step  
1000 Beyond. *Syst. Biol.* **64**, 152–160 (2015).
- 1001 7. Vakulenko-Lagun, B., Mandel, M. & Betensky, R. A. coxrt: Cox Proportional Hazards Regression for  
1002 Right-Truncated Data. v. 1.0.2 (2019).
- 1003 8. Mantel, N. & Myers, M. Problems of convergence of maximum likelihood iterative procedures in  
1004 multiparameter situations. *J. Am. Stat. Assoc.* (1971). doi:10.1080/01621459.1971.10482289
- 1005 9. Collen, B., Purvis, A. & Gittleman, J. L. Biological correlates of description date in carnivores and primates.  
1006 *Glob. Ecol. Biogeogr.* **13**, 459–467 (2004).
- 1007 10. Diniz-Filho, J. A. F. *et al.* Macroecological correlates and spatial patterns of anuran description dates in the  
1008 Brazilian Cerrado. *Glob. Ecol. Biogeogr.* **14**, 469–477 (2005).
- 1009 11. Moura, M. R. *et al.* Geographical and socioeconomic determinants of species discovery trends in a  
1010 biodiversity hotspot. *Biol. Conserv.* **220**, 237–244 (2018).

- 1011 12. Gaston, K. J., Blackburn, T. M. & Loder, N. Which species are described first? The case of North-American  
1012 butterflies. *Biodivers. Conserv.* **4**, 119–127 (1995).
- 1013 13. Colli, G. R. *et al.* In the depths of obscurity: Knowledge gaps and extinction risk of Brazilian worm lizards  
1014 (Squamata, Amphisbaenidae). *Biol. Conserv.* **204**, 51–62 (2016).
- 1015 14. Essl, F., Rabitsch, W., Dullinger, S., Moser, D. & Milasowszky, N. How well do we know species richness  
1016 in a well-known continent? Temporal patterns of endemic and widespread species descriptions in the  
1017 European fauna. *Glob. Ecol. Biogeogr.* **22**, 29–39 (2013).
- 1018 15. Bebber, D. P. *et al.* Herbaria are a major frontier for species discovery. *Proc. Natl. Acad. Sci.* **107**, 22169–  
1019 22171 (2010).
- 1020 16. Raposo, M. A. *et al.* What really hampers taxonomy and conservation? A riposte to Garnett and Christidis  
1021 (2017). *Zootaxa* **4317**, 179 (2017).
- 1022 17. Dayrat, B. Towards integrative taxonomy. *Biol. J. Linn. Soc.* **85**, 407–415 (2005).
- 1023 18. Isaac, N. J. B., Mallet, J. & Mace, G. M. Taxonomic inflation: its influence on macroecology and  
1024 conservation. *Trends Ecol. Evol.* **19**, 464–469 (2004).
- 1025 19. Garnett, S. T. & Christidis, L. Taxonomy anarchy hampers conservation. *Nature* **546**, 25–27 (2017).
- 1026 20. Bremer, K., Bremer, B., Karis, P. & Källersjö, M. Time for change in taxonomy. *Nature* **343**, 202–202  
1027 (1990).
- 1028 21. Köhler, J. *et al.* New Amphibians and Global Conservation: A Boost in Species Discoveries in a Highly  
1029 Endangered Vertebrate Group. *Bioscience* **55**, 693 (2005).
- 1030 22. Guedes, J. J. M., Feio, R. N., Meiri, S. & Moura, M. R. Identifying factors that boost species discoveries of  
1031 global reptiles. *Zool. J. Linn. Soc.* in press (2020). doi:10.1093/zoolinnean/zlaa029
- 1032 23. Burgin, C. J., Colella, J. P., Kahn, P. L. & Upham, N. S. How many species of mammals are there? *J.*  
1033 *Mammal.* **99**, 1–14 (2018).
- 1034 24. Mallet, J. Species, Concepts of. in *Encyclopedia of Biodiversity* **6**, 679–691 (Elsevier, 2013).
- 1035 25. Kutner, M. H., Nachtsheim, C. J., Neter, J. & Li, W. *Applied Linear Statistical Models*. (McGraw-Hill  
1036 Irwin, 2004).
- 1037

1038 **SUPPLEMENTARY TABLES**

1039

1040 *Table S1.*

1041 Taxon-specific allometric equations used to convert straight carapace length (SCL) of chelonians

1042 into body mass. Sample size used to create the equations, number of species covered by the

1043 samples, and summary statistics of each allometric equation are shown. Allometric equation:

1044  $\text{Log}(\text{BodyMass}) = \text{Intercept} + \text{Coef} \times \text{Log}(\text{SCL})$ .

<b>Taxa</b>	<b>Sample Size</b>	<b>Number of species</b>	<b>Intercept</b>	<b>Coefficient</b>	<b>R<sup>2</sup></b>
Chelidae	52	42	-3.915	2.991	0.947
Emydidae	82	57	-3.420	2.814	0.950
Kinosternidae	28	27	-3.235	2.715	0.913
Podocnemididae	14	8	-3.757	2.930	0.952
Testudinidae	96	52	-3.473	2.885	0.956
Cryptodira*	326	218	-3.389	2.832	0.928
Pleurodira*	68	49	-3.922	2.995	0.967

1045 \* Only used when a family-level equation was not available.

1046 *Table S2.*

1047 Variance Inflation Factor (VIF) for species-level attributes of terrestrial vertebrates using the full  
1048 dataset (species described from 1759 to 2014). VIF measures the multicollinearity of variables  
1049 included in a model, and it varies from 1 (no multicollinearity) to +Inf. VIF values > 10 reflect  
1050 high multicollinearity <sup>25</sup>.

<b>Variable</b>	<b>Amphibians</b>	<b>Reptiles</b>	<b>Mammals</b>	<b>Birds</b>
Annual mean temperature	2.695	2.364	2.590	2.615
Annual precipitation	2.661	2.251	2.424	2.817
Body size	1.125	1.192	1.074	1.072
Elevation	2.048	1.872	1.578	1.858
Human density	1.368	1.442	1.423	1.410
Precipitation seasonality	1.631	1.663	1.702	2.116
Range size	1.516	2.158	1.913	2.390
Range rarity	1.752	1.606	1.674	1.932
Activity per bioregion	2.155	2.184	2.454	1.619
Activity per family	1.922	1.953	2.311	1.322
Temperature seasonality	1.908	2.620	3.646	4.098

1051 All variables were log10 transformed before computation of the Variance Inflation Factor.

1052



1053 *Table S3.*

1054 Total number of species described within the different time periods. The number between  
 1055 parentheses indicates the percentage of species discarded if such decades are left out of the  
 1056 species dataset. Only species included in our dataset were counted.

<b>Time Period</b>	<b>Amphibians</b>	<b>Reptiles</b>	<b>Mammals</b>	<b>Birds</b>
1758 – 2014	7268 (0.2)	10063 (2.4)	5700 (1.1)	9993 (4.4)
>1760 – 2014	7251 (0.5)	9948 (3.5)	5561 (1.6)	9555 (7.7)
>1770 – 2014	7235 (0.5)	9905 (6)	5501 (1.7)	9224 (9.2)
>1780 – 2014	7233 (0.6)	9888 (6.8)	5359 (2.2)	9075 (16)
>1790 – 2014	7225 (0.9)	9843 (7.9)	5312 (2.4)	8392 (16.9)
>1800 – 2014	7205 (1.1)	9818 (9)	5249 (3.6)	8300 (18.6)
>1810 – 2014	7191 (1.2)	9705 (11.6)	5186 (4.2)	8136 (24.2)
>1820 – 2014	7180 (2.2)	9640 (15.9)	5040 (6.6)	7573 (33.3)
>1830 – 2014	7108 (3.3)	9398 (21.8)	4794 (10.8)	6670 (44.3)
>1840 – 2014	7025 (4.6)	8972 (27.2)	4459 (13.6)	5564 (55.4)
>1850 – 2014	6931 (6.7)	8691 (30.4)	4152 (18.8)	4460 (64)
>1860 – 2014	6778 (10)	8176 (34.1)	3970 (25.9)	3595 (71.2)
>1870 – 2014	6539 (13)	7457 (38.1)	3758 (30.9)	2877 (78.9)
>1880 – 2014	6322 (16.5)	6957 (42.3)	3531 (35.8)	2109 (83.7)
>1890 – 2014	6069 (20.3)	6459 (51.7)	3291 (41.5)	1630 (88.3)
>1900 – 2014	5794 (23.3)	5883 (61.8)	2752 (45.7)	1173 (91.6)
>1910 – 2014	5577 (26.5)	5468 (68.8)	2179 (49.2)	842 (93.5)
>1920 – 2014	5343 (31)	5114 (73)	1780 (52.6)	647 (95.3)
>1930 – 2014	5016 (35.6)	4769 (76.7)	1541 (57.2)	466 (96.7)
>1940 – 2014	4680 (38.6)	4304 (78.3)	1326 (59.5)	325 (97.3)
>1950 – 2014	4460 (42.7)	4078 (80.4)	1235 (62.1)	273 (97.8)
>1960 – 2014	4162 (49)	3809 (82.3)	1119 (66.5)	221 (98.2)
>1970 – 2014	3705 (56.7)	3372 (84.9)	1007 (71.1)	182 (98.5)
>1980 – 2014 <sup>a</sup>	3150 (64.3)	2911 (87.7)	860 (76.3)	149 (99)
>1990 – 2014 <sup>a</sup>	2598 (75.9)	2387 (91.3)	699 (83.2)	102 (99.6)
>2000 – 2014 <sup>a</sup>	1748 (92.9)	1687 (96.8)	494 (93.6)	36 (100)
>2010 – 2014 <sup>a</sup>	515 (100)	645 (100)	184 (100)	0 (100)

1057 <sup>a</sup> Time periods not used in the sensitivity analysis.

1058 *Table S4.*

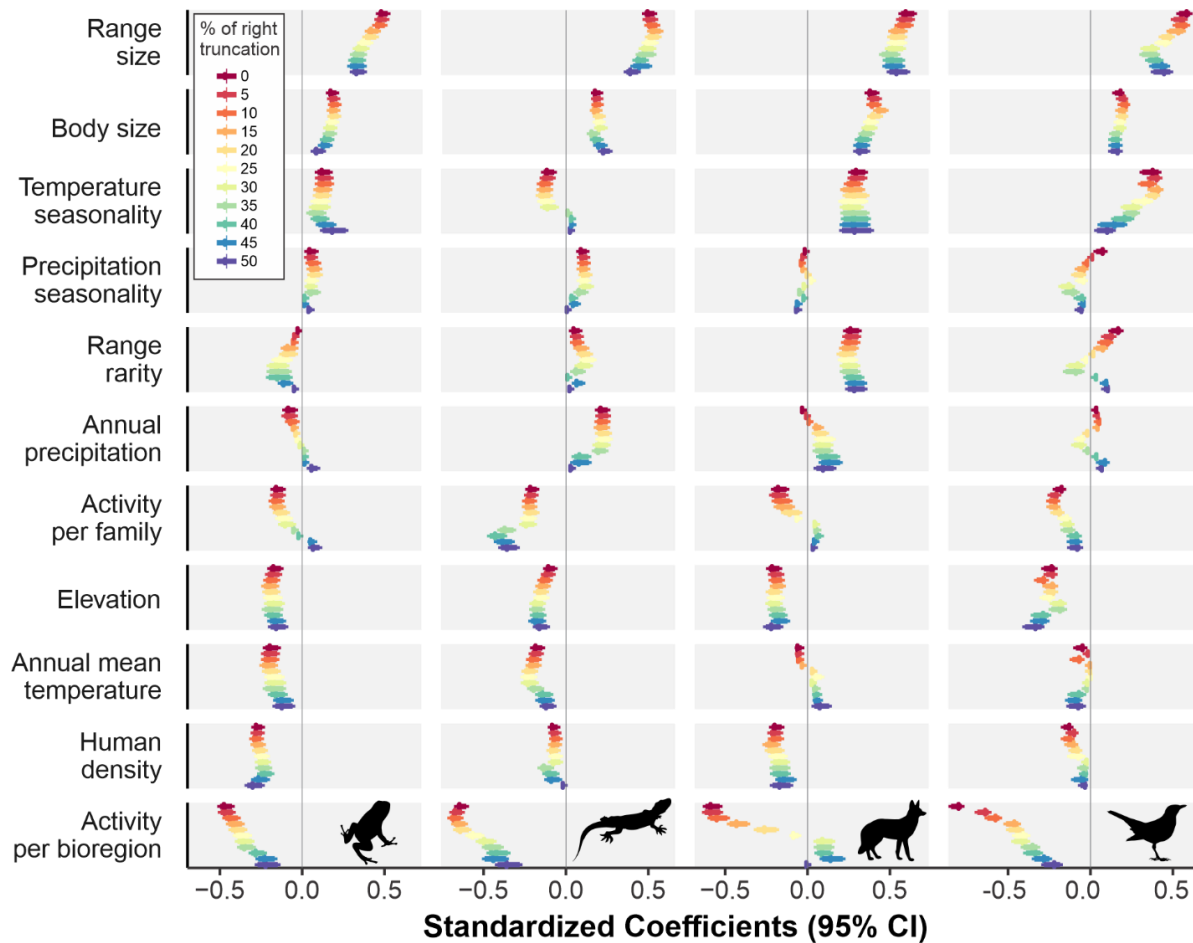
1059 Validation-fold and grouping dependence of deviation from a 1:1 relationship. Values shown  
 1060 how many times, on median, the observed discoveries were higher than the estimated  
 1061 discoveries. Results shown for the complete time period of species discoveries (1759-2014). At  
 1062 the assemblage level, only the subsampling level of 5 occurrences per species is shown.

Grouping level	Validation-fold size	Amphibians	Reptiles	Mammals	Birds
<b>Taxon-level</b>					
Family	75%	15.922	17.560	14.072	27.250
Family	50%	7.312	8.270	8.329	18.500
Family	25%	3.025	3.275	3.996	9.000
Family	10%	1.338	1.349	1.791	4.200
<b>Family</b>	<b>Average</b>	<b>6.899</b>	<b>7.613</b>	<b>7.047</b>	<b>14.737</b>
<b>Higher-level grouping</b>					
Higher-level grouping	75%	28.986	27.605	30.016	68.500
Higher-level grouping	50%	9.887	11.843	11.402	46.000
Higher-level grouping	25%	3.423	4.312	5.172	20.361
Higher-level grouping	10%	1.152	1.469	1.888	7.392
<b>Higher-level grouping</b>	<b>Average</b>	<b>10.862</b>	<b>11.307</b>	<b>12.120</b>	<b>35.563</b>
<b>Assemblage-level</b>					
220 km	75%	3.894	6.396	3.695	4.736
220 km	50%	3.068	4.147	2.957	3.662
220 km	25%	2.271	2.458	2.313	2.683
220 km	10%	2.022	2.017	2.051	2.211
<b>220 km</b>	<b>Average</b>	<b>2.814</b>	<b>3.754</b>	<b>2.754</b>	<b>3.323</b>
440 km	75%	9.430	14.658	9.687	11.627
440 km	50%	5.882	7.795	6.412	8.120
440 km	25%	3.097	3.524	3.487	4.577
440 km	10%	2.079	2.069	2.241	2.788
<b>440 km</b>	<b>Average</b>	<b>5.112</b>	<b>7.012</b>	<b>5.457</b>	<b>6.778</b>
880 km	75%	18.794	25.063	24.776	29.200
880 km	50%	9.964	11.985	14.881	19.592
880 km	25%	4.269	4.526	6.744	10.296
880 km	10%	2.211	2.229	2.981	4.751
<b>880 km</b>	<b>Average</b>	<b>8.810</b>	<b>10.951</b>	<b>12.345</b>	<b>15.960</b>

1063

1064 **SUPPLEMENTARY FIGURES**

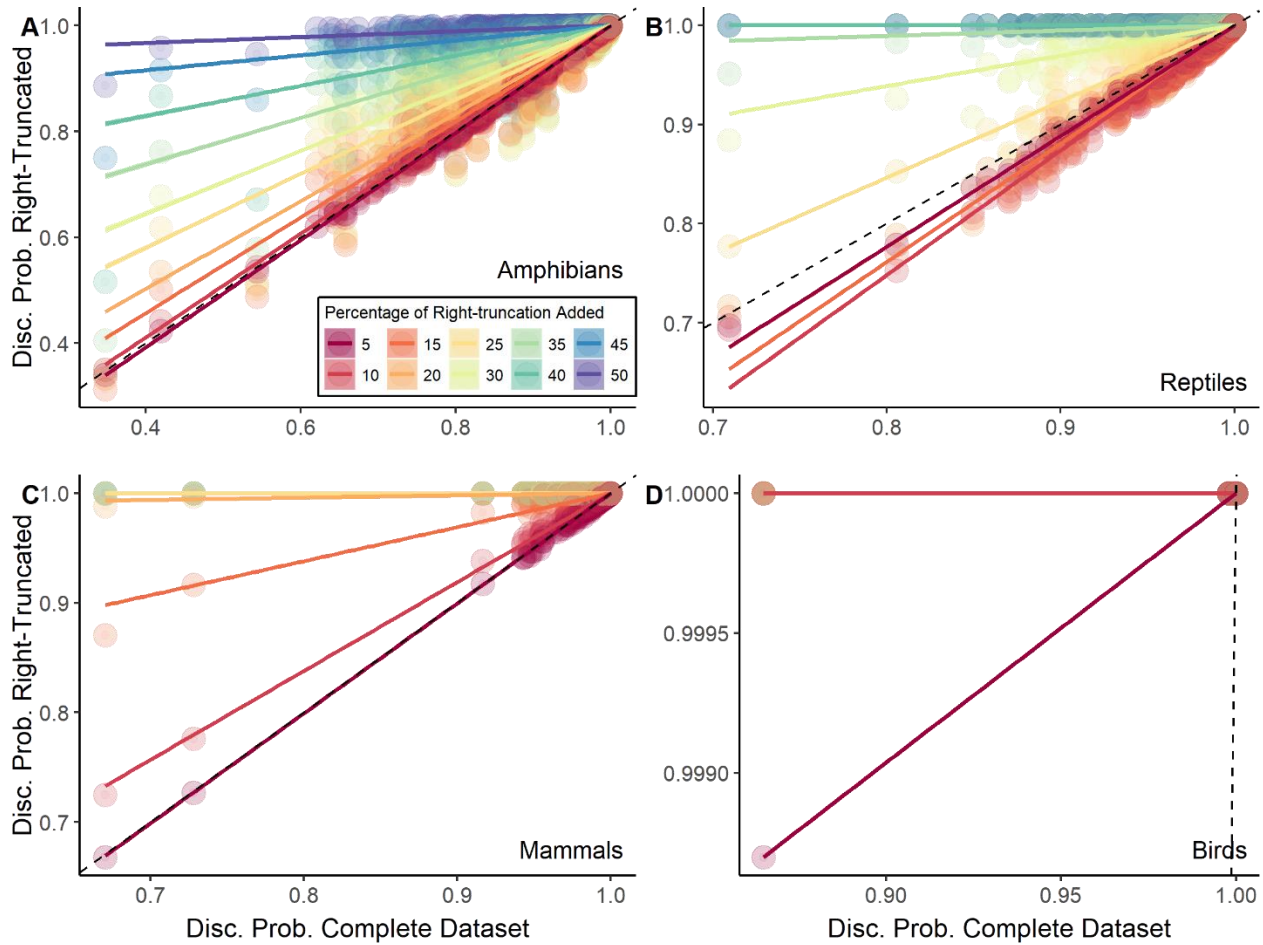
1065



1066

1067 *Fig. S1.*

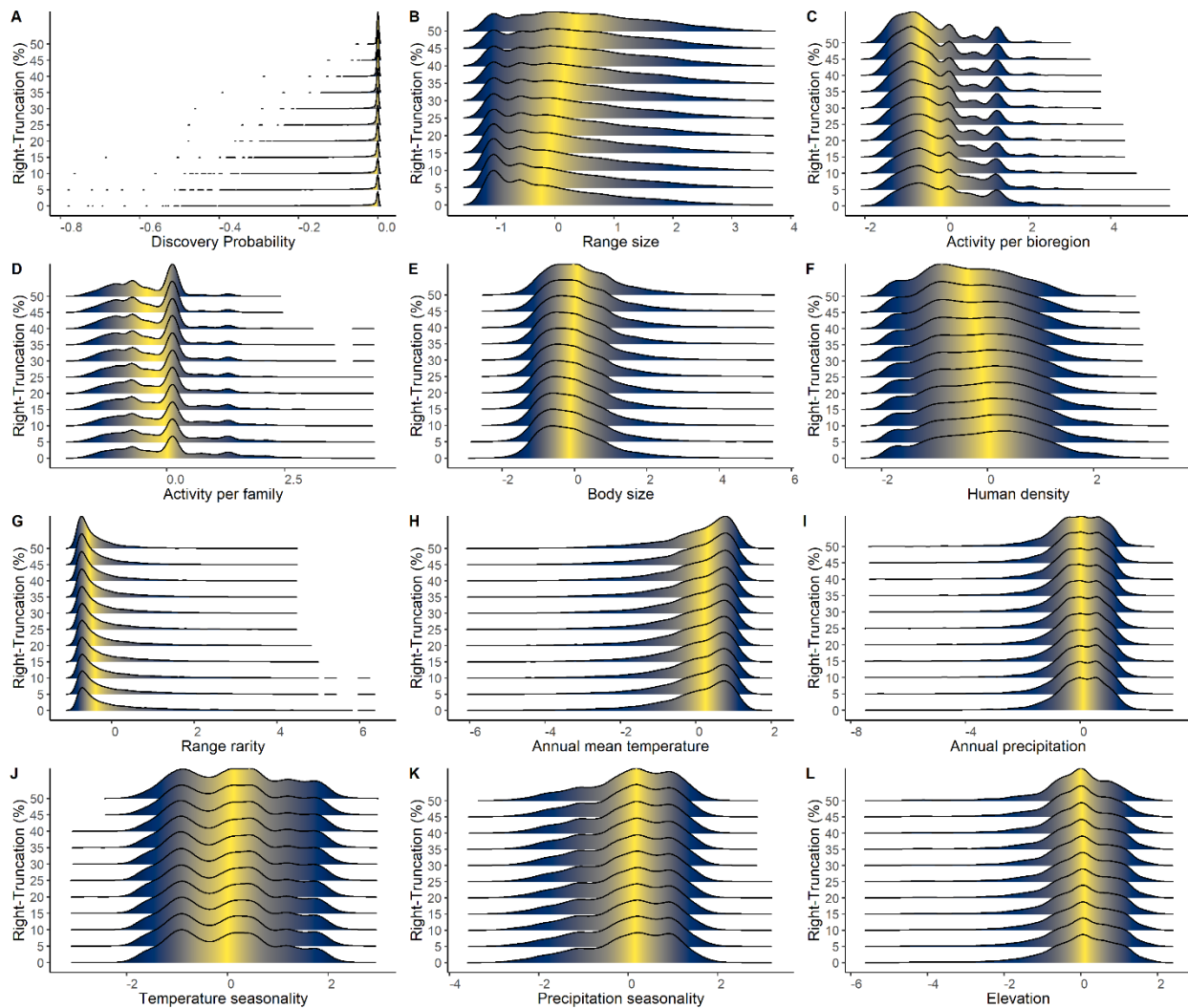
1068 Standardized coefficients of the average weighted accelerated failure time (AFT) model  
1069 computed for species datasets with increasing levels of right-truncation. Line colours indicate the  
1070 percentage of recent-described species left out of the model (recent-described species were  
1071 successively discarded). The horizontal bars denote the 95% confidence intervals around each  
1072 coefficient. Standardized coefficients above 0 indicate that species with high values for a given  
1073 attribute had higher discovery probability (prob.) and thus were likely discovered early on.  
1074 Negative standardized coefficients mean high attribute values depressed discovery probability  
1075 and delayed discovery.



1076  
1077

*Fig. S2.*

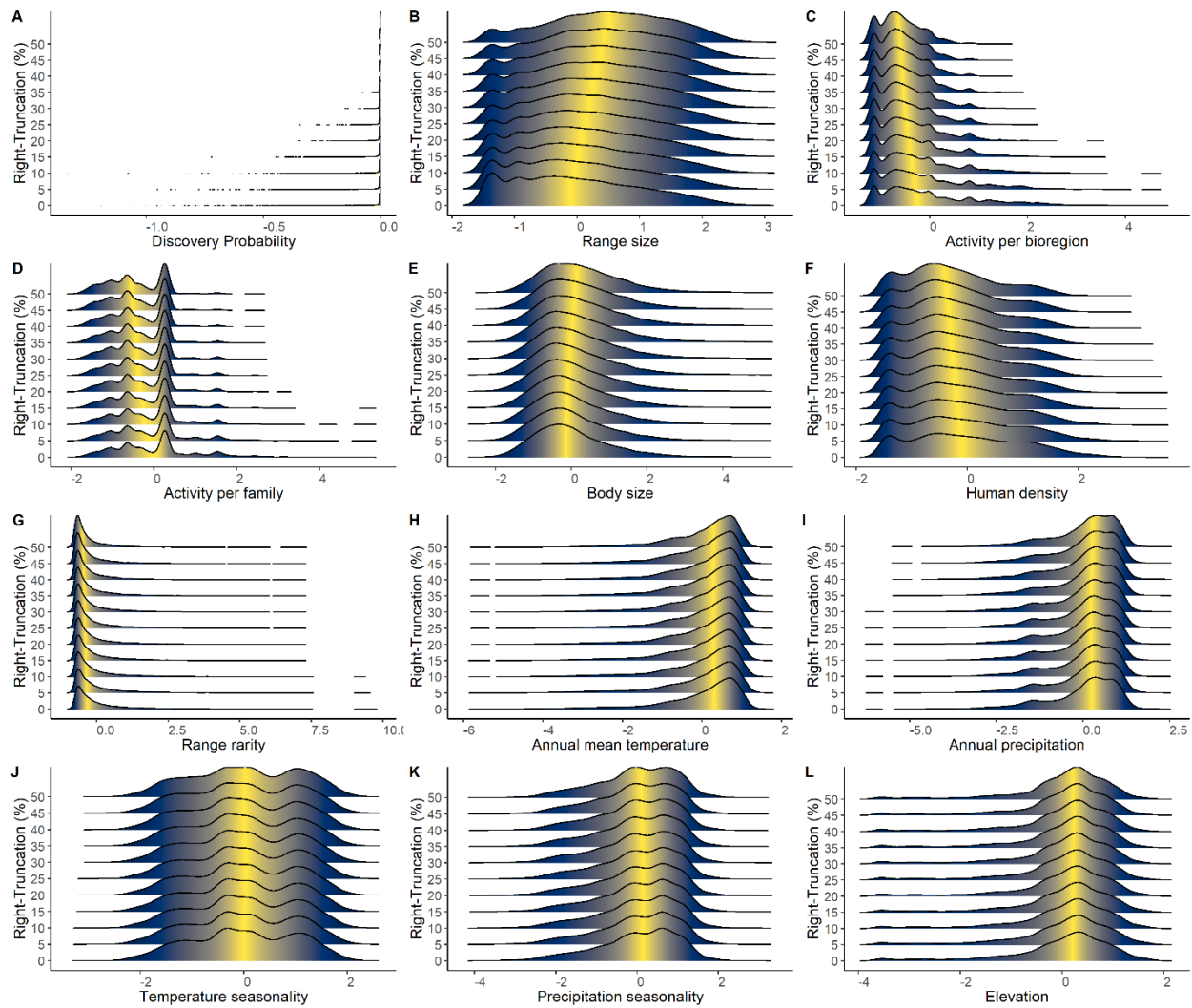
1078 Relationship between values of discovery probability computed using the complete (x-axis) and right-truncated (y-  
1079 axis) datasets. For each vertebrate group, only the oldest half of the known species is represented due to their  
1080 presence in all data subsets with different levels of right-truncation. The dashed line indicates the line of equality. In  
1081 decreasing the level of right-truncation (increasing completeness) of species dataset, the discovery probabilities tend  
1082 to be lower.



1083  
1084

*Fig. S3.*

1085 Frequency distribution of species-level attributes for data subsets with different levels of right-  
1086 truncation for amphibians. (A) Discovery probability. (B) Range size = number of  $110 \times 110$  km  
1087 grid cells occupied by the species' range. (C) Activity/bioregion = number of taxonomists per  
1088 species in the bioregions in which it typically occurs at the year of the species' description. (D)  
1089 Activity/family = number of taxonomists per species in a family at the year of species'  
1090 description. (E) Body size = maximum body size. (F) Human density = Within-range human  
1091 population density at the year of species' description. (G) Range rarity = within-range endemism  
1092 richness at the year of the species' description. (H) Annual mean temperature = within-range  
1093 annual mean temperature. (I) Annual precipitation = within-range annual precipitation. (J)  
1094 Temperature seasonality = within-range temperature seasonality. (K) Precipitation seasonality =  
1095 within-range precipitation seasonality. (L) Mean elevation = within-range mean elevation. In all  
1096 plots, the variable was  $\log_{10}$  transformed to increase readability. The colour gradient is centred in  
1097 the median value.

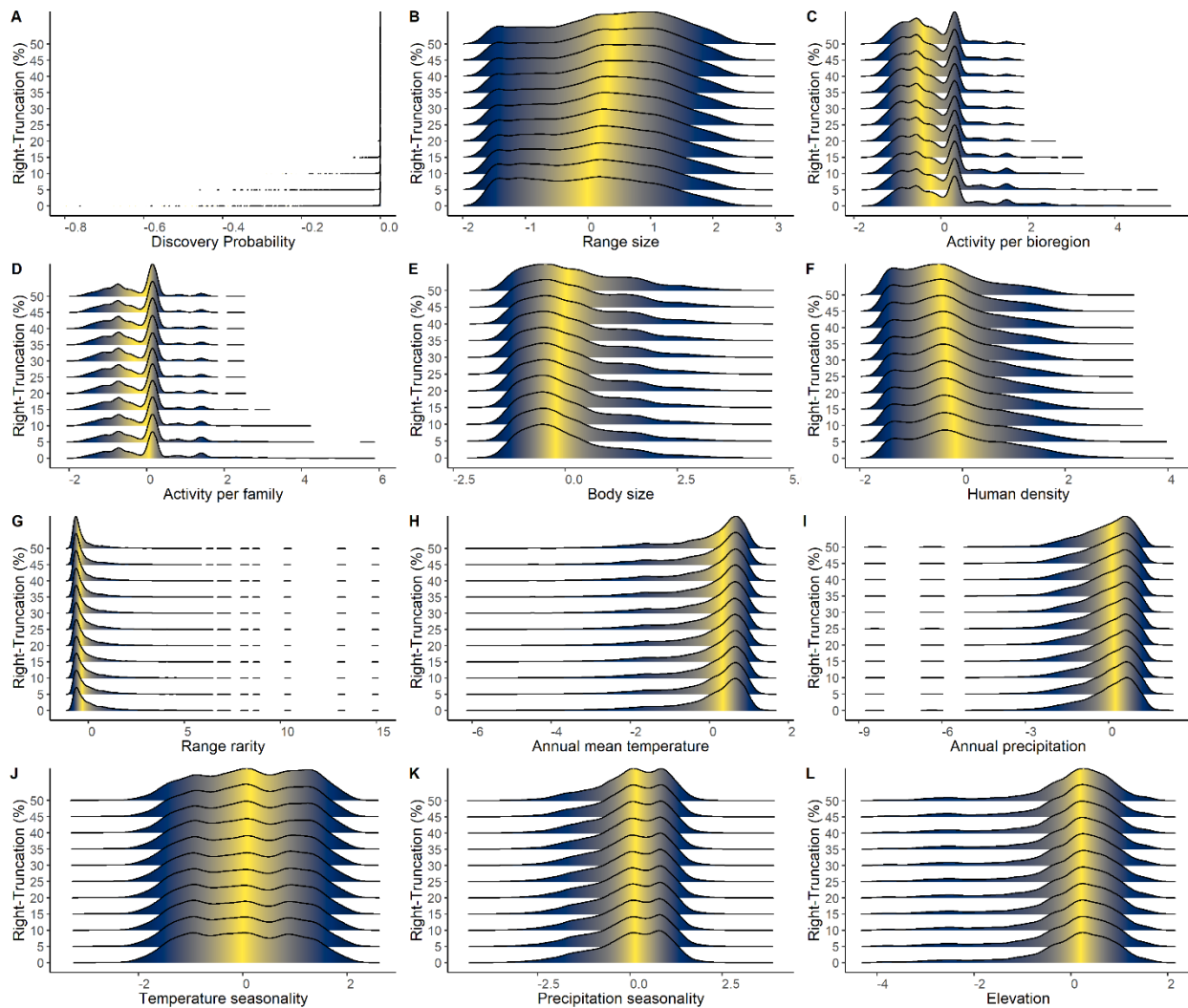


1098  
1099

*Fig. S4.*

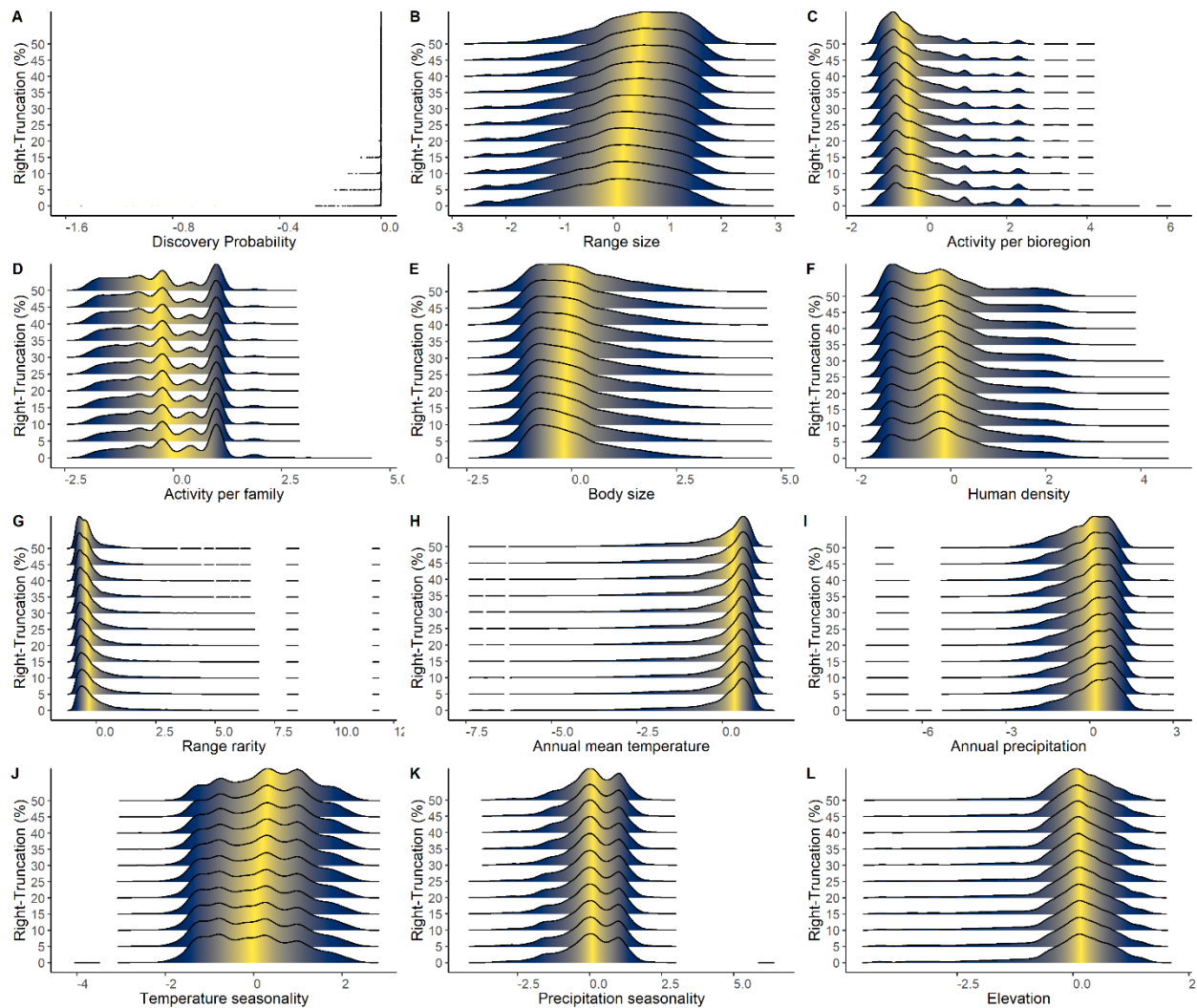
1100 Frequency distribution of species-level attributes for data subsets with different levels of right-  
 1101 truncation for reptiles. (A) Discovery probability. (B) Range size = number of  $110 \times 110$  km grid  
 1102 cells occupied by the species' range. (C) Activity/bioregion = number of taxonomists per species  
 1103 in the bioregions in which it typically occurs at the year of the species' description. (D)  
 1104 Activity/family = number of taxonomists per species in a family at the year of species'  
 1105 description. (E) Body size = maximum body size. (F) Human density = Within-range human  
 1106 population density at the year of species' description. (G) Range rarity = within-range endemism  
 1107 richness at the year of the species' description. (H) Annual mean temperature = within-range  
 1108 annual mean temperature. (I) Annual precipitation = within-range annual precipitation. (J)  
 1109 Temperature seasonality = within-range temperature seasonality. (K) Precipitation seasonality =  
 1110 within-range precipitation seasonality. (L) Mean elevation = within-range mean elevation. In all  
 1111 plots, the variable was  $\log_{10}$  transformed to increase readability. The colour gradient is centred in  
 1112 the median value.





1113  
1114 *Fig. S5.*

1115 Frequency distribution of species-level attributes for data subsets with different levels of right-  
 1116 truncation for mammals. (A) Discovery probability. (B) Range size = number of  $110 \times 110$  km  
 1117 grid cells occupied by the species' range. (C) Activity/bioregion = number of taxonomists per  
 1118 species in the bioregions in which it typically occurs at the year of the species' description. (D)  
 1119 Activity/family = number of taxonomists per species in a family at the year of species'  
 1120 description. (E) Body size = maximum body size. (F) Human density = Within-range human  
 1121 population density at the year of species' description. (G) Range rarity = within-range endemism  
 1122 richness at the year of the species' description. (H) Annual mean temperature = within-range  
 1123 annual mean temperature. (I) Annual precipitation = within-range annual precipitation. (J)  
 1124 Temperature seasonality = within-range temperature seasonality. (K) Precipitation seasonality =  
 1125 within-range precipitation seasonality. (L) Mean elevation = within-range mean elevation. In all  
 1126 plots, the variable was  $\log_{10}$  transformed to increase readability. The colour gradient is centred in  
 1127 the median value.

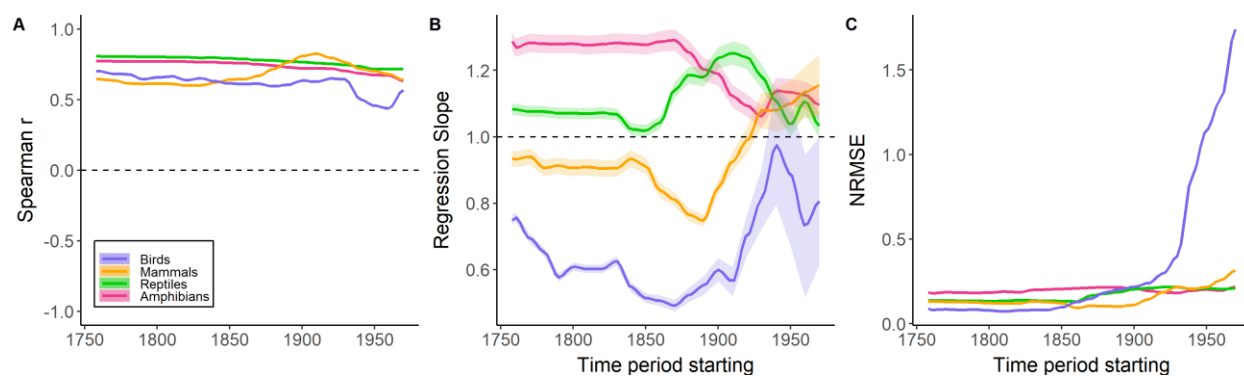


1128

1129 *Fig. S6.*

1130 Frequency distribution of species-level attributes for data subsets with different levels of right-  
1131 truncation for birds. (A) Discovery probability. (B) Range size = number of  $110 \times 110$  km grid  
1132 cells occupied by the species' range. (C) Activity/bioregion = number of taxonomists per species  
1133 in the bioregions in which it typically occurs at the year of the species' description. (D)  
1134 Activity/family = number of taxonomists per species in a family at the year of species'  
1135 description. (E) Body size = maximum body size. (F) Human density = Within-range human  
1136 population density at the year of species' description. (G) Range rarity = within-range endemism  
1137 richness at the year of the species' description. (H) Annual mean temperature = within-range  
1138 annual mean temperature. (I) Annual precipitation = within-range annual precipitation. (J)  
1139 Temperature seasonality = within-range temperature seasonality. (K) Precipitation seasonality =  
1140 within-range precipitation seasonality. (L) Mean elevation = within-range mean elevation. In all  
1141 plots, the variable was  $\log_{10}$  transformed to increase readability. The colour gradient is centred in  
1142 the median value.





1143

1144

*Fig. S7.*

1145

1146

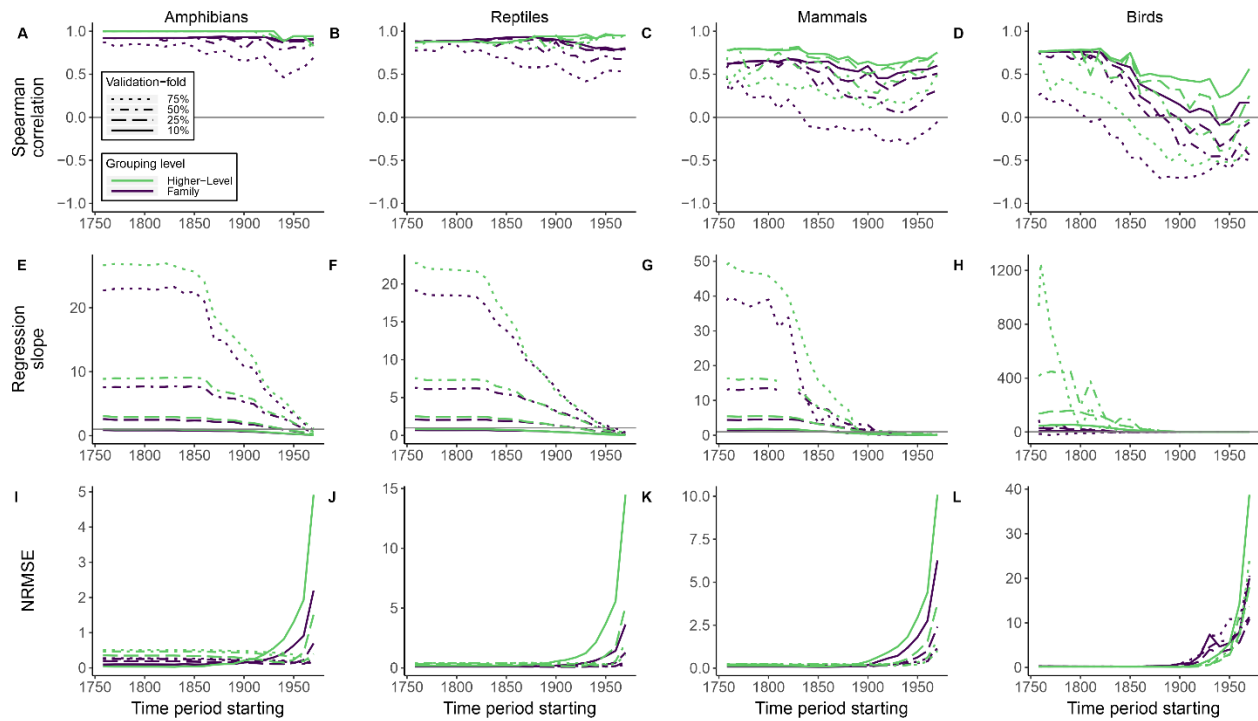
1147

1148

1149

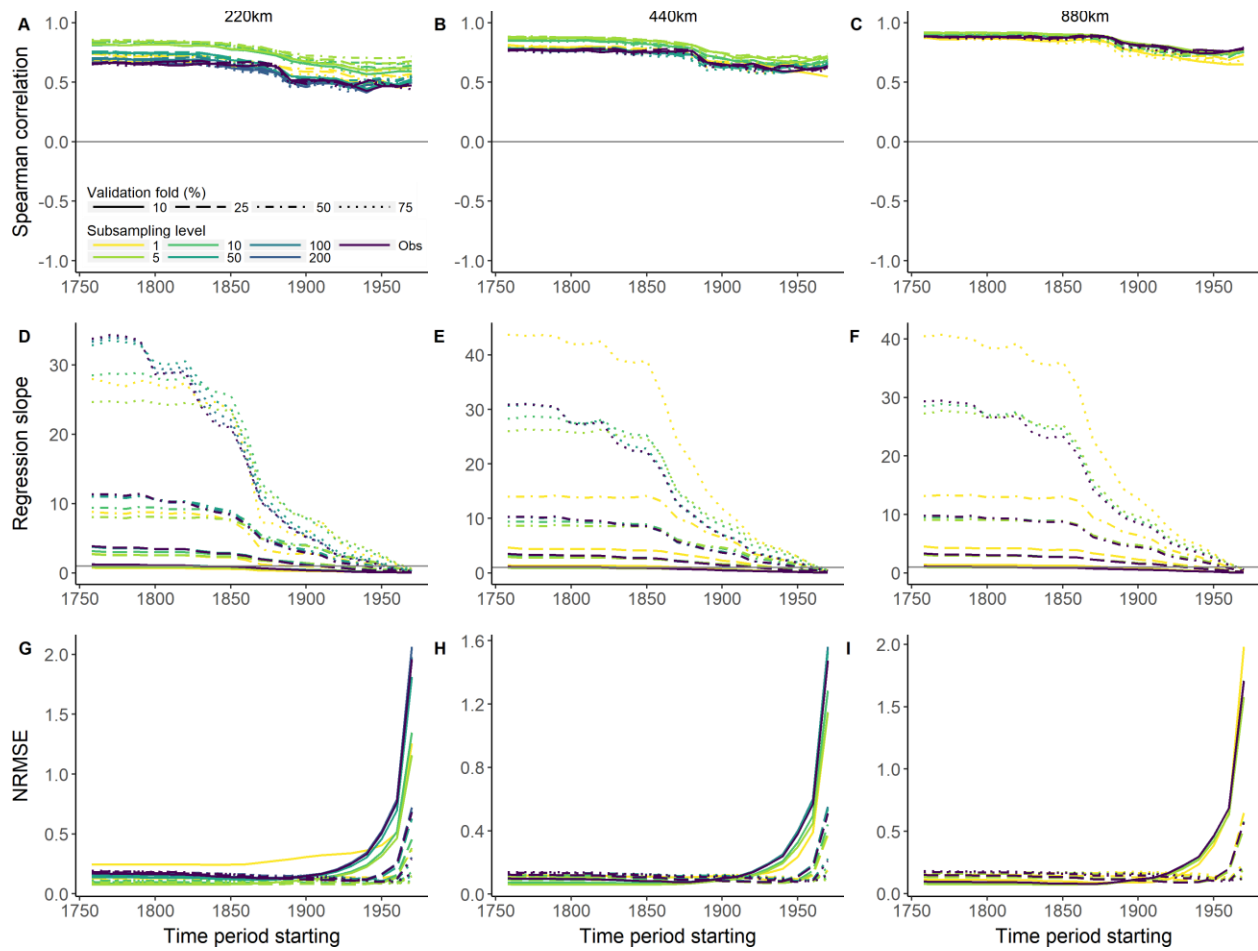
1150

**Statistic metrics derived from the sensitivity analysis at the species level.** All statistics were computed between predicted and observed year of discovery across species. Results based on model trained with 75% of species and 25% of species used as holdout data. Line colours denote different vertebrate groups. (A) Spearman correlation, (B) Regression slope, (C) Normalized Root Mean Square Error (NRMSE).



1151  
1152 *Fig. S8.*

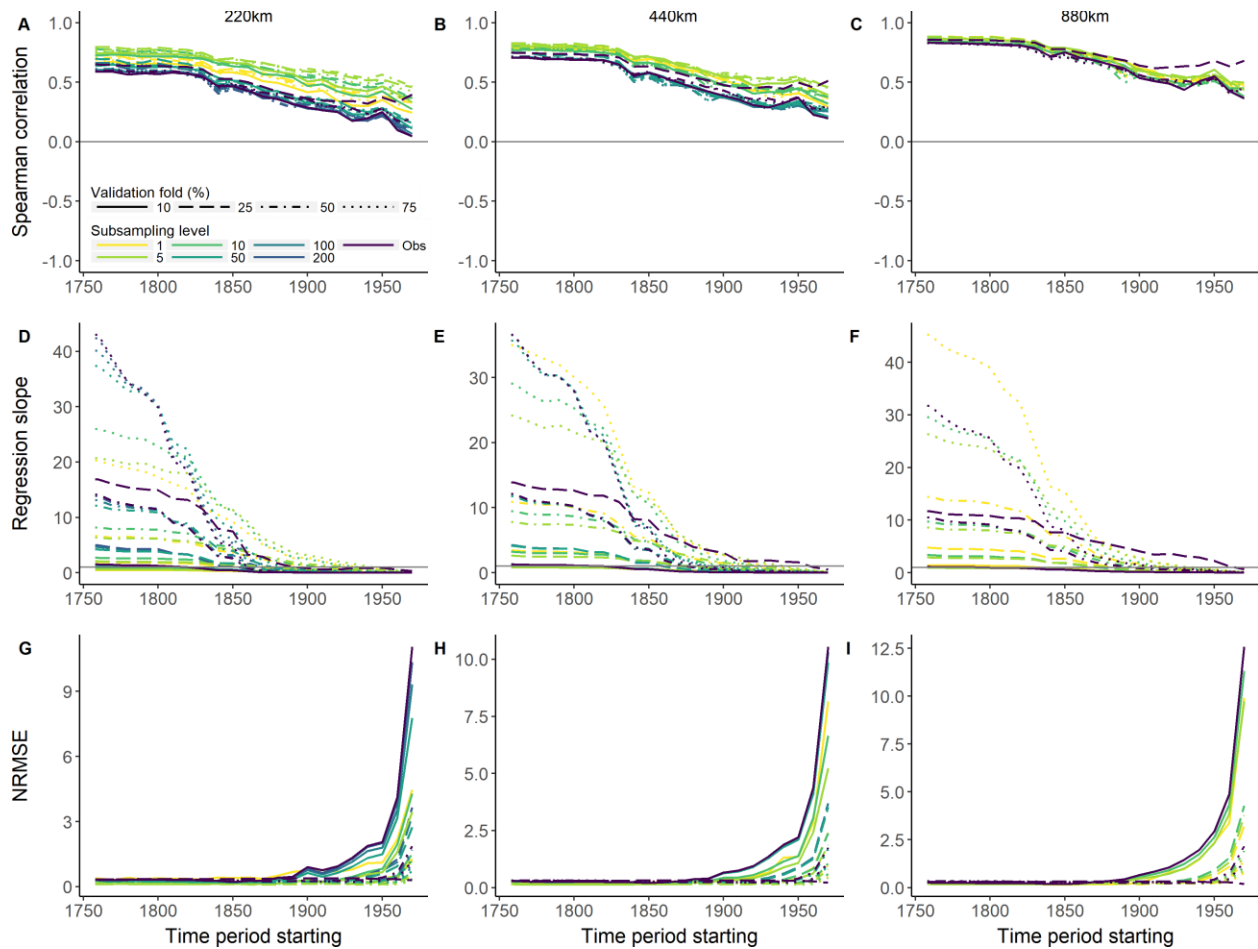
1153 **Statistic metrics derived from the sensitivity analysis at the taxon level using different sizes**  
1154 **of validation-fold.** All statistics were computed between estimated and observed discoveries  
1155 across taxa. Line types denote different sizes of the validation-fold. Line colours indicate  
1156 different taxonomic ranks. (A-D) Spearman correlation, (E-H) Regression slope, (I-L)  
1157 Normalized Root Mean Square Error (NRMSE). The size of training-fold (25, 50, 75, 90% of  
1158 species) is the complement of the respective validation-fold size (75, 50, 25, 10% of species).  
1159 Confidence intervals were omitted to increase readability.  
1160



1161  
1162

*Fig. S9.*

1163 **Statistic metrics derived from the sensitivity analysis using amphibian species.** All statistics  
1164 were computed between estimated and observed discoveries across grid cells. Line colours  
1165 denote the subsampling level adopted to control overrepresentation of wide-ranging species.  
1166 Panel columns refer to statistics calculated at different spatial resolutions (220, 440, 880 km).  
1167 (A-C) Spearman correlation, (D-F) Regression slope, (G-I) Normalized Root Mean Square Error  
1168 (NRMSE). Confidence intervals were omitted to increase readability.

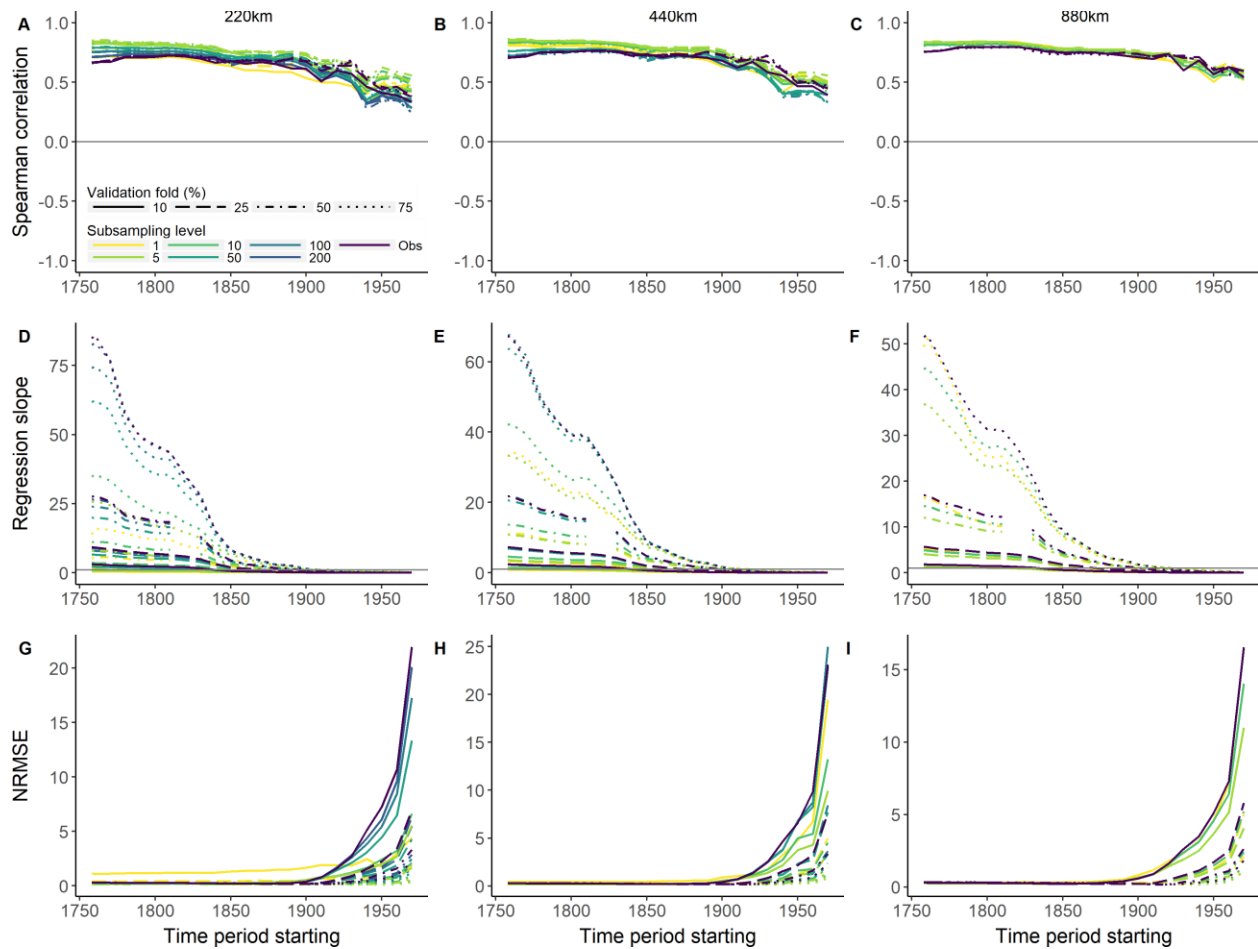


1169

1170

*Fig. S10.*

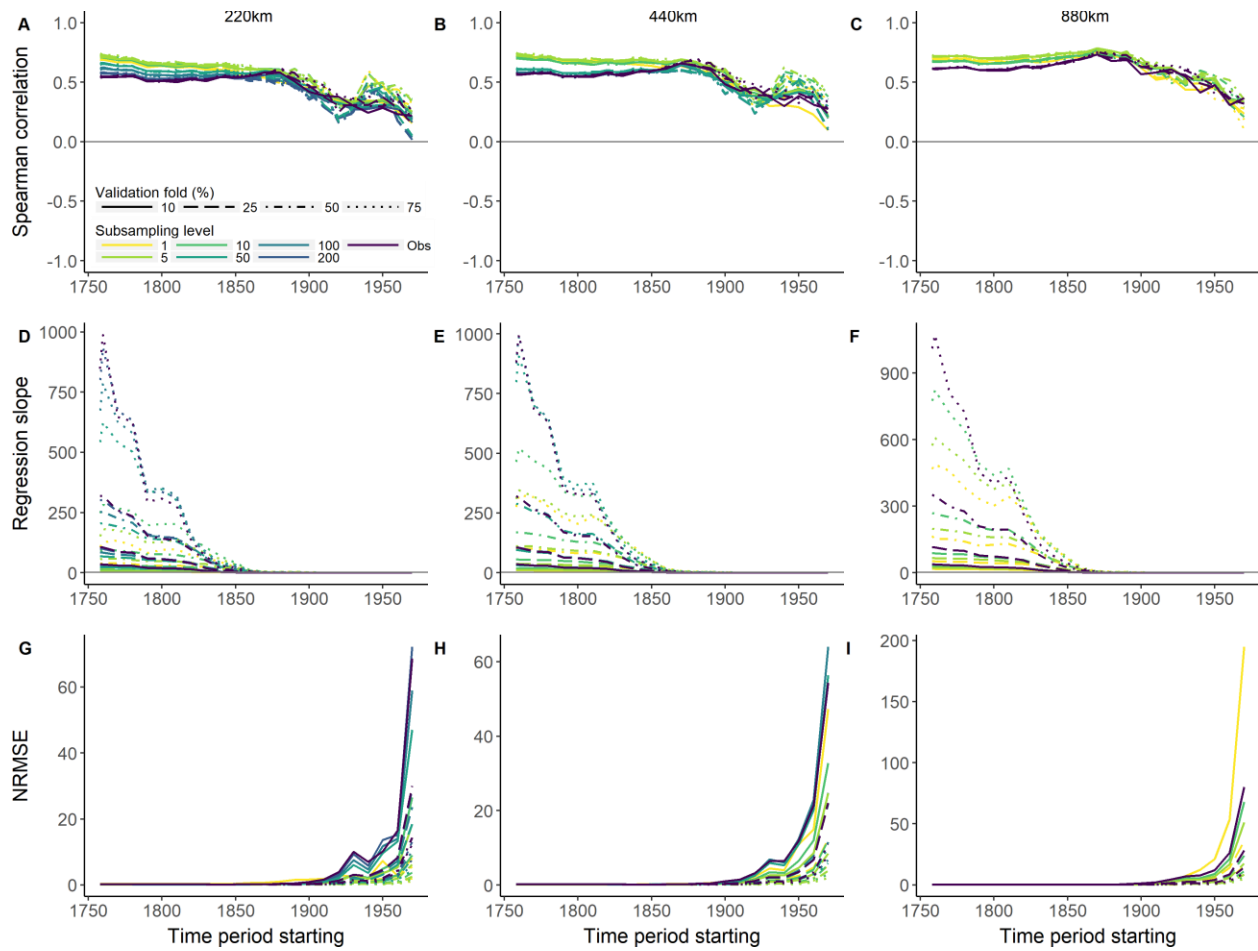
1171 **Statistic metrics derived from the sensitivity analysis using reptile species.** All statistics were  
1172 computed between estimated and observed species across grid cells. Line colours denote the  
1173 subsampling level adopted to control overrepresentation of wide-ranging species. Panel columns  
1174 refer to statistics calculated at different spatial resolutions (220, 440, 880 km). (A-C) Spearman  
1175 correlation, (D-F) Regression slope, (G-I) Normalized Root Mean Square Error (NRMSE).  
1176 Confidence intervals were omitted to increase readability.



1177  
1178

*Fig. S11.*

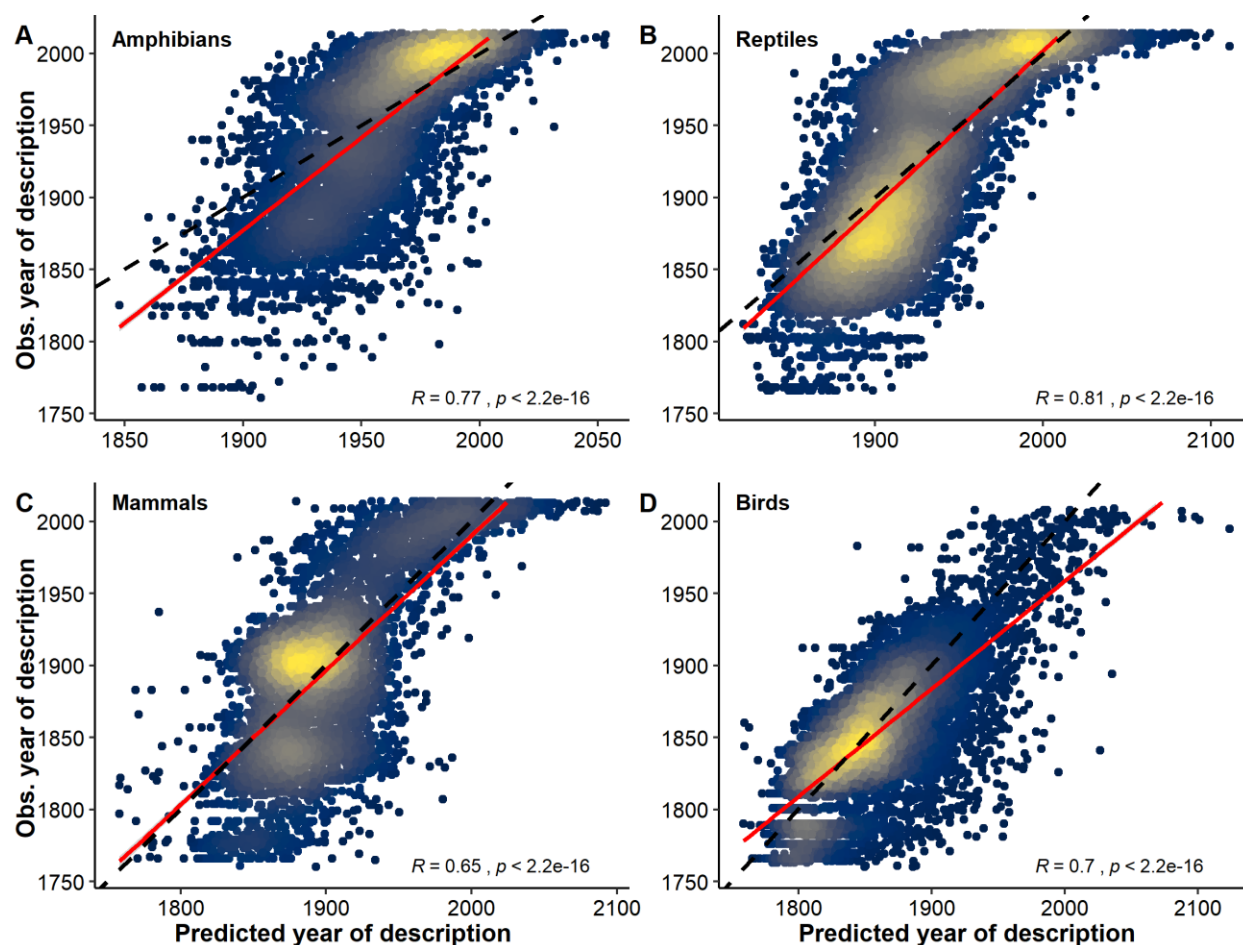
1179 **Statistic metrics derived from the sensitivity analysis using mammal species.** All statistics  
1180 were computed between estimated and observed species across grid cells. Line colours denote  
1181 the subsampling level adopted to control overrepresentation of wide-ranging species. Panel  
1182 columns refer to statistics calculated at different spatial resolutions (220, 440, 880 km). (A-C)  
1183 Spearman correlation, (D-F) Regression slope, (G-I) Normalized Root Mean Square Error  
1184 (NRMSE). Confidence intervals were omitted to increase readability.



1185  
1186

*Fig. S12.*

1187 **Statistic metrics derived from the sensitivity analysis using bird species.** All statistics were  
1188 computed between estimated and observed species across grid cells. Line colours denote the  
1189 subsampling level adopted to control overrepresentation of wide-ranging species. Panel columns  
1190 refer to statistics calculated at different spatial resolutions (220, 440, 880 km). (A-C) Spearman  
1191 correlation, (D-F) Regression slope, (G-I) Normalized Root Mean Square Error (NRMSE).  
1192 Confidence intervals were omitted to increase readability.

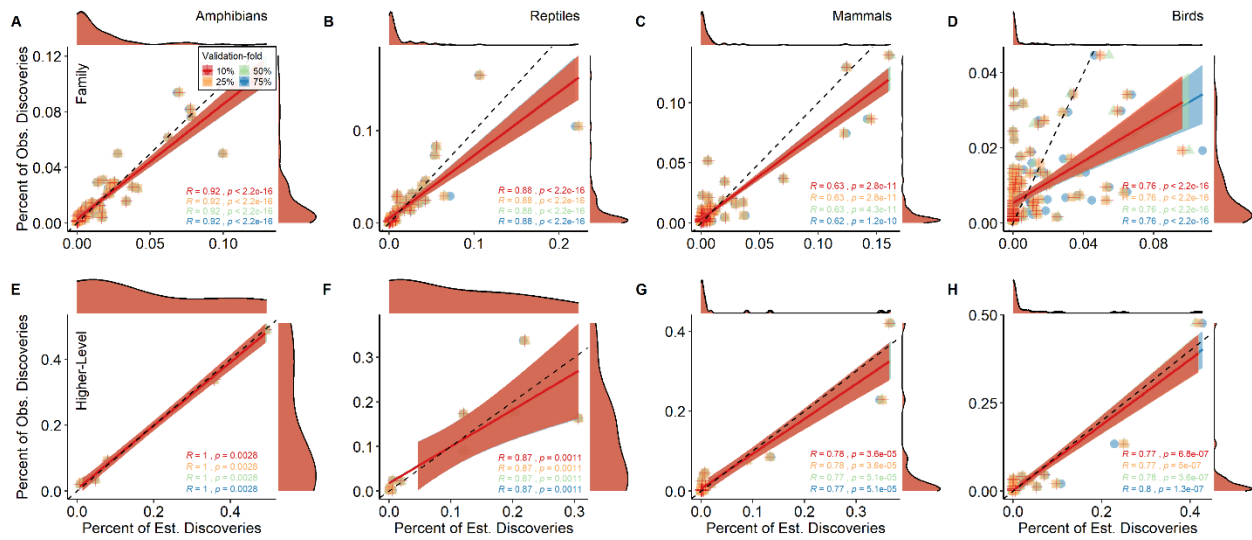


1193

1194

*Fig. S13.*

1195 Relationship between predicted and observed year of description for terrestrial vertebrates. Plots  
1196 include known species described from 1759 to 2014. (A) Amphibians, (B) Reptiles, (C)  
1197 Mammals, (D) Birds. Models were first trained using 75% of the data, and then applied to the  
1198 validation-fold (independent data) to obtain the predicted year of discovery for each species. The  
1199 dashed line indicates the line of equality. R values inside plots denote the Spearman correlation  
1200 between observed and predicted year of description.  
1201

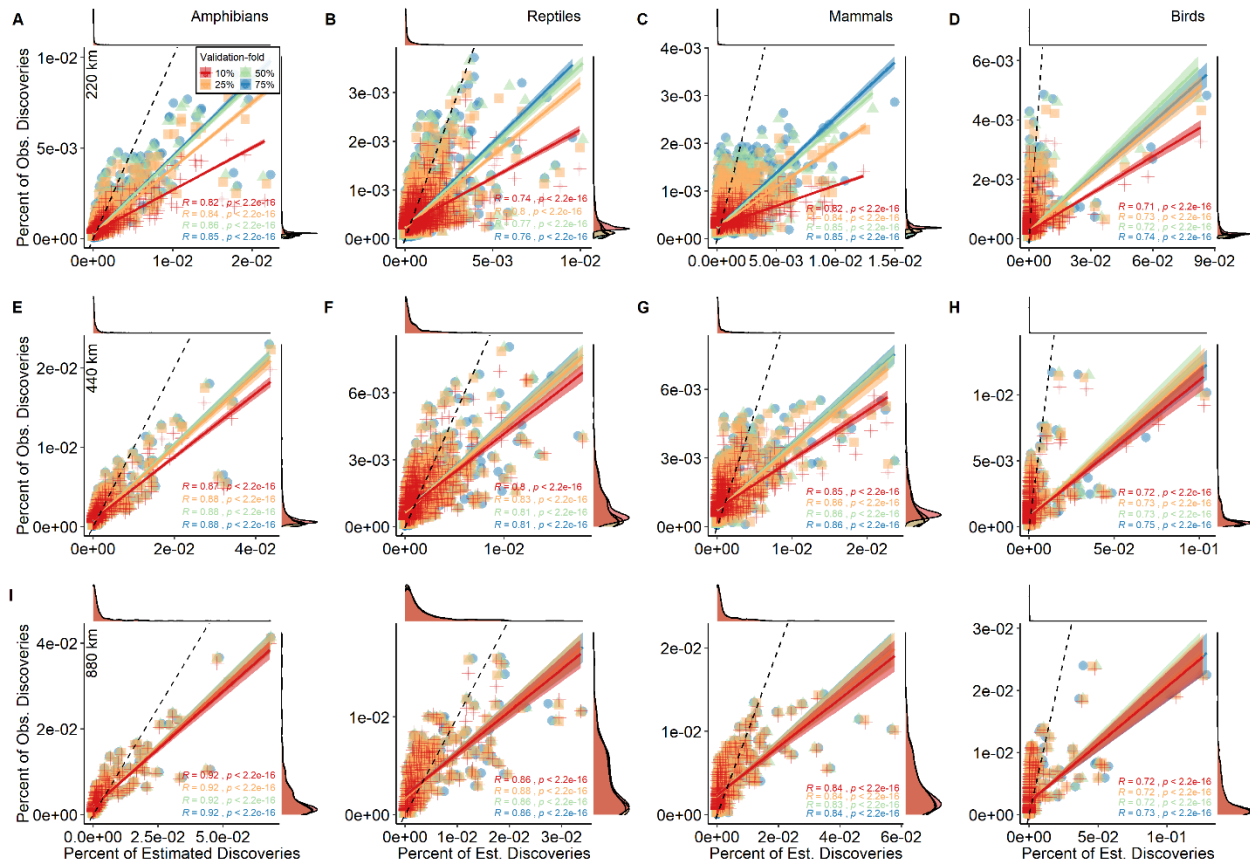


1202  
1203

*Fig. S14.*

1204 **Relationship between estimated and observed discoveries at the taxon-level.** Colours denote  
 1205 different sizes of the validation-fold. Columns represent different vertebrate groups and rows  
 1206 indicate different taxonomic ranks. The dashed line indicates the line of equality. (A, E, I)  
 1207 Amphibians. (B, F, J) Reptiles. (C, G, K) Mammals. (D, H, L) Birds. See ‘Model validation’  
 1208 section for details on the highest-level taxonomic rank used. R values inside plots denote the  
 1209 Spearman correlation between estimated and observed discoveries. Plots include known species  
 1210 described from 1759 to 2014.

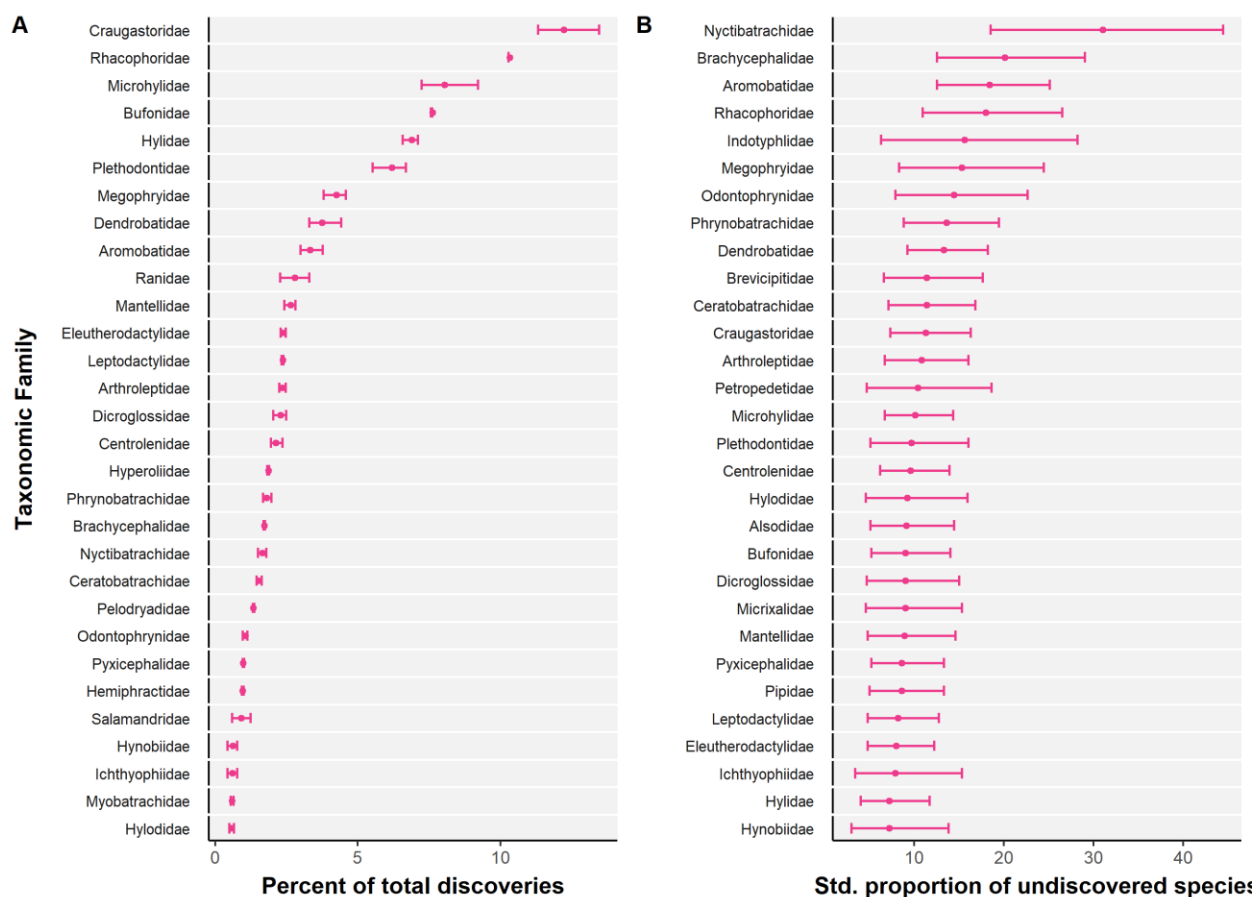




1211

1212 *Fig. S15.*

1213 **Relationship between estimated and observed discoveries at the assemblage-level.** Colours  
 1214 denote different sizes of the validation-fold. Columns represent different vertebrate groups and  
 1215 rows indicate assemblages (grid cells) defined at different spatial resolutions. The dashed line  
 1216 indicates the line of equality. (A, E, I, M) Amphibians. (B, F, J, N) Reptiles. (C, G, K, O)  
 1217 Mammals. (D, H, L, P) Birds. Only the subsampling level of 5 is shown. R values inside plots  
 1218 denote the Spearman correlation between estimated and observed discoveries.  
 1219

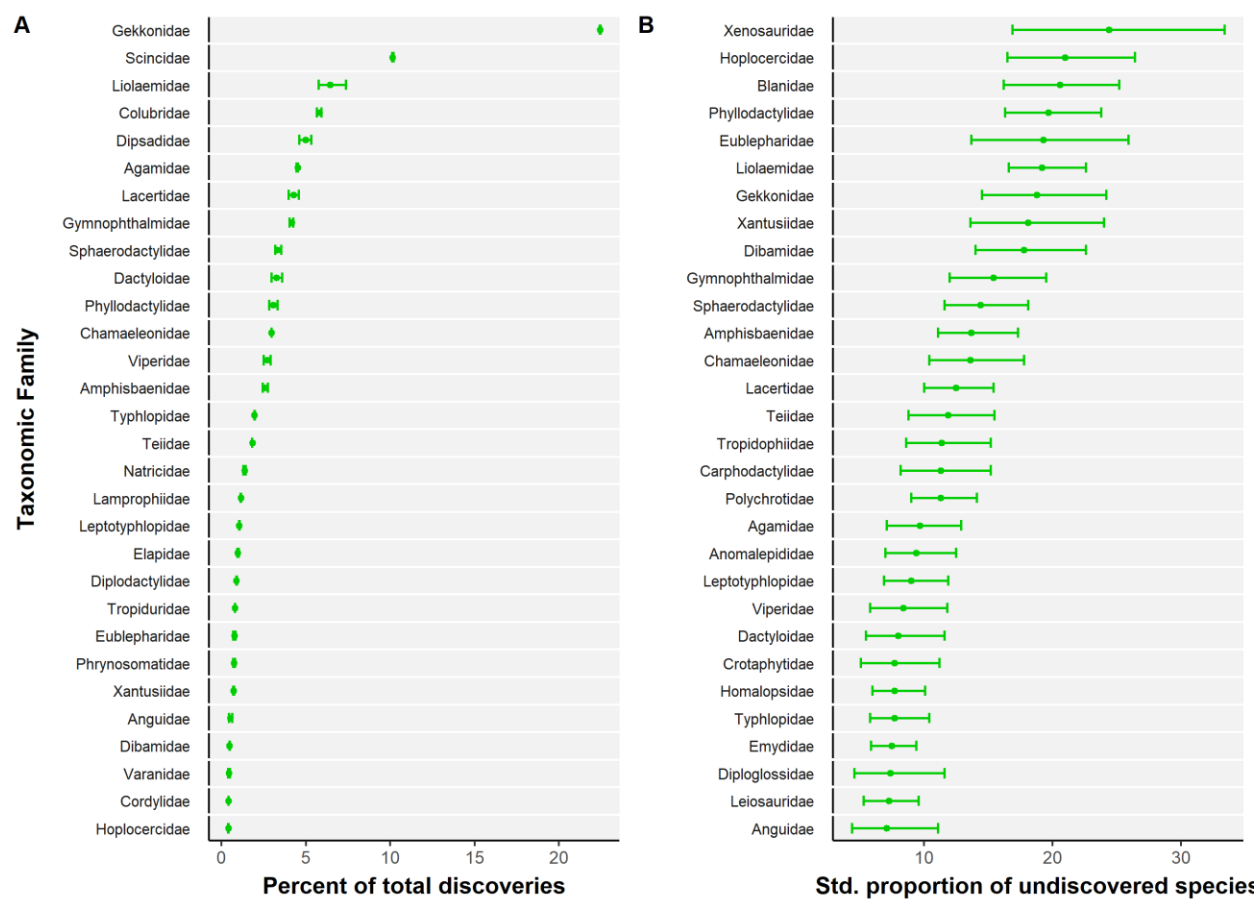


1220

1221 *Fig. S16.*

1222 **Top 30 amphibian families with highest potential for species discoveries.** (A) Taxa with highest percent of total  
 1223 discoveries. (B) Taxa with highest standardized proportion of undiscovered species. The horizontal lines denote the  
 1224 95% confidence intervals.

1225

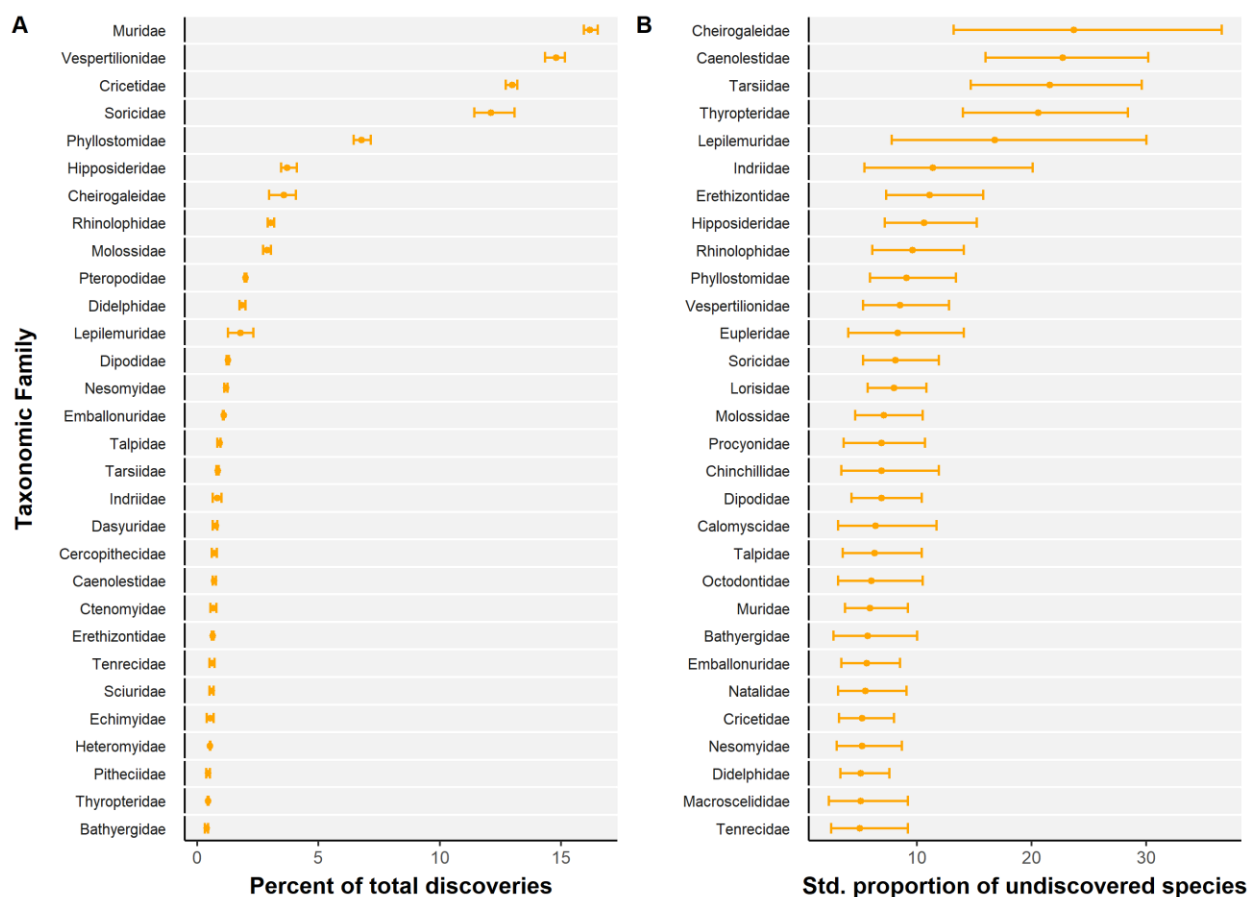


1226

1227 *Fig. S17.*

1228 **Top 30 reptile families with highest potential for species discoveries.** (A) Taxa with highest percent of total  
 1229 discoveries. (B) Taxa with highest standardized proportion of undiscovered species. The horizontal lines denote the  
 1230 95% confidence intervals.

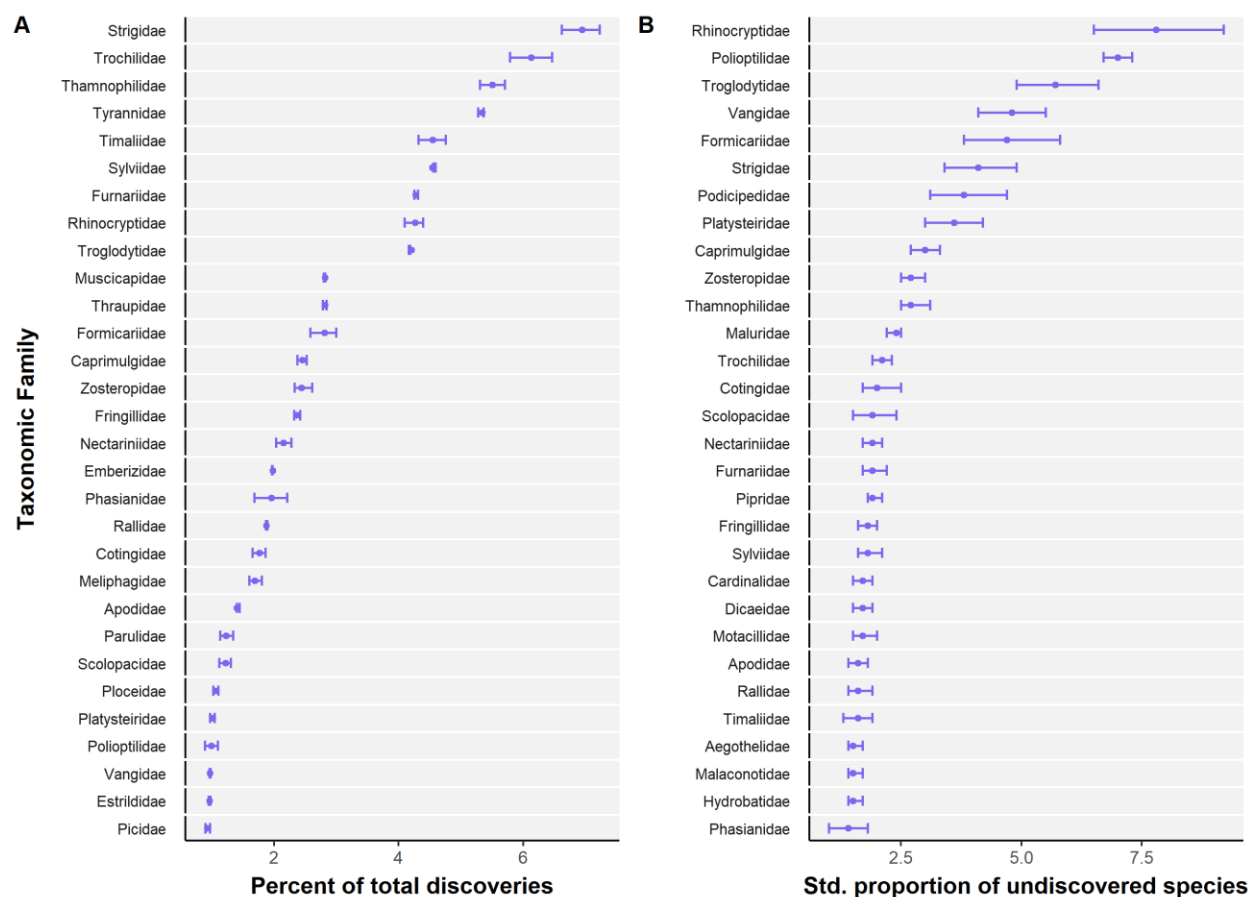
1231



1232  
1233

*Fig. S18.*

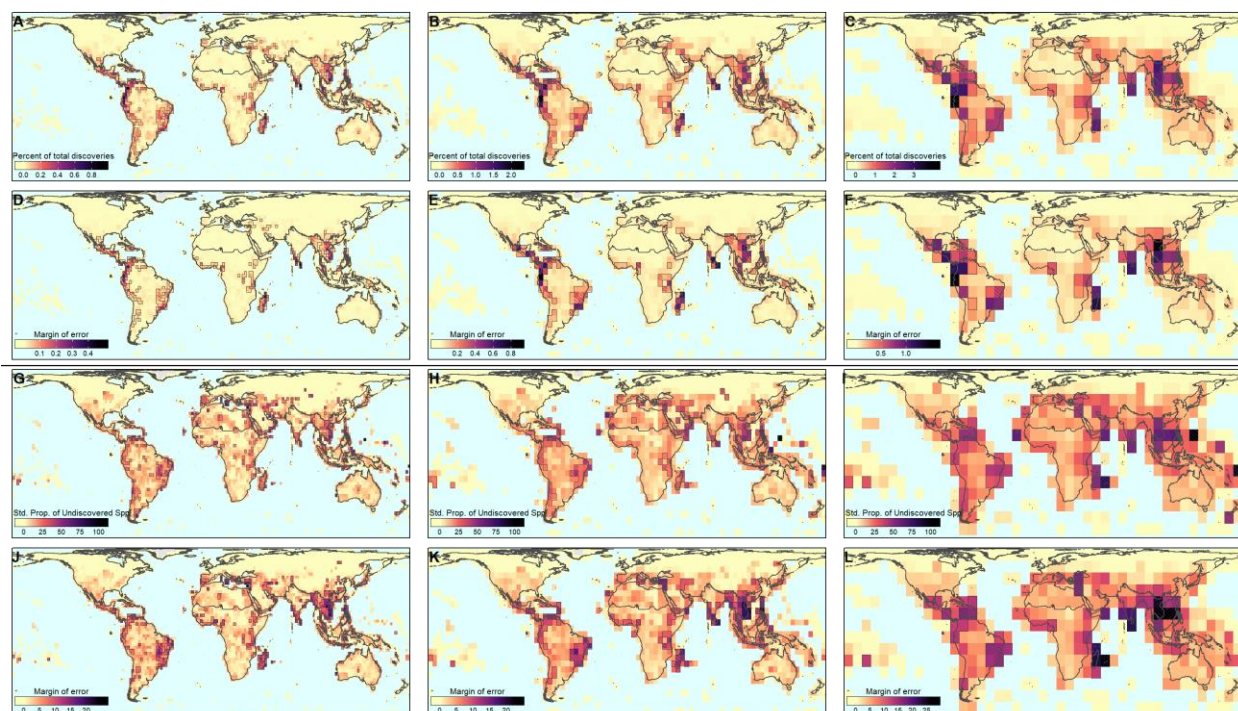
1234 **Top 30 mammal families with highest potential for species discoveries.** (A) Taxa with highest percent of total  
1235 discoveries. (B) Taxa with highest standardized proportion of undiscovered species. The horizontal lines denote the  
1236 95% confidence intervals.  
1237



1238

1239 *Fig. S19.*

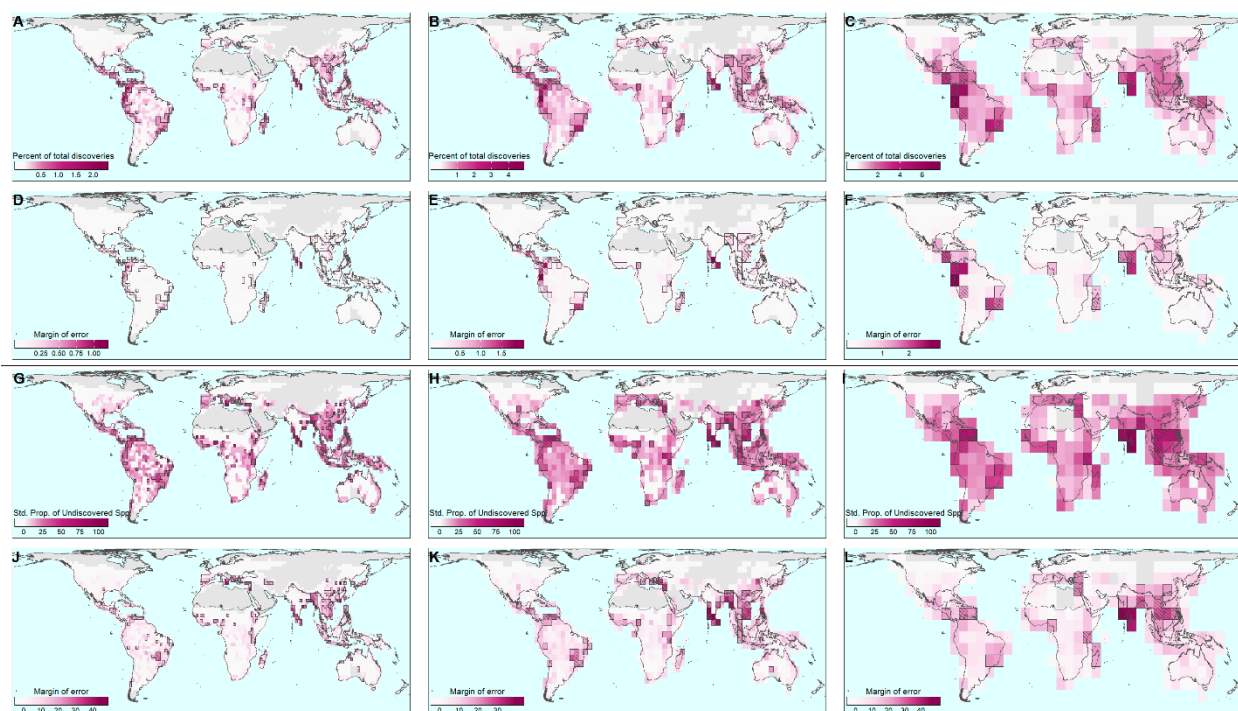
1240 **Top 30 bird families with highest potential for species discoveries.** (A) Taxa with highest percent of total  
 1241 discoveries. (B) Taxa with highest standardized proportion of undiscovered species. The horizontal lines denote the  
 1242 95% confidence intervals.  
 1243



1244  
1245

*Fig. S20.*

1246 Geographical discovery patterns for terrestrial vertebrates at different spatial resolutions. (A-C) Percent of total  
1247 predicted discoveries across grid cells and their respective (D-F) uncertainty ( $\pm$  margin of error). (G-I) Standardized  
1248 proportion of undiscovered species across grid cells and their respective (J-L) uncertainty ( $\pm$  margin of error).  
1249 Outlined and hatched regions designate grid cells holding values within respectively the top 10% and top 5% of the  
1250 mapped metric. Maps drawn at spatial resolutions of 220, 440, 880 km.

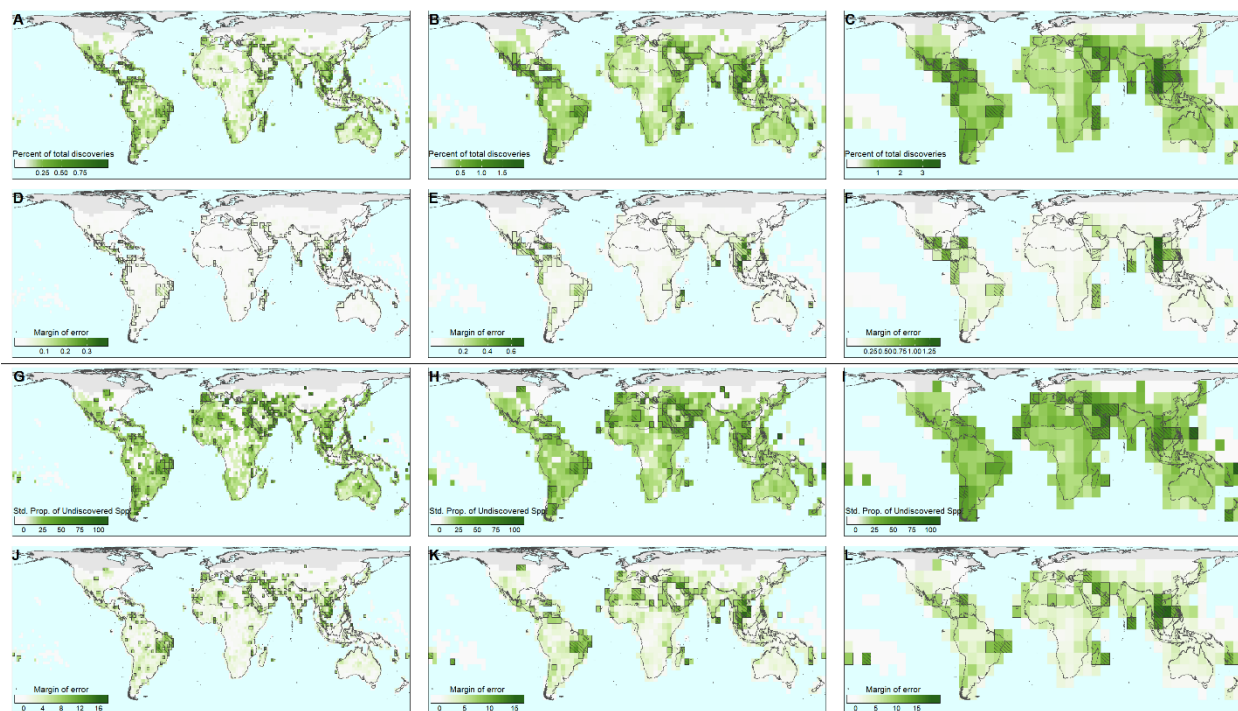


1251  
1252

*Fig. S21.*

1253 Geographical discovery patterns for amphibians at different spatial resolutions. (A-C) Percent of total discoveries  
1254 across grid cells and their respective (D-F) uncertainty ( $\pm$  margin of error). (G-I) Standardized proportion of  
1255 undiscovered species across grid cells and their respective (J-L) uncertainty ( $\pm$  margin of error). Outlined and  
1256 hatched regions designate grid cells holding values within respectively the top 10% and top 5% of the mapped  
1257 metric. Maps drawn at spatial resolutions of 220, 440, 880 km.



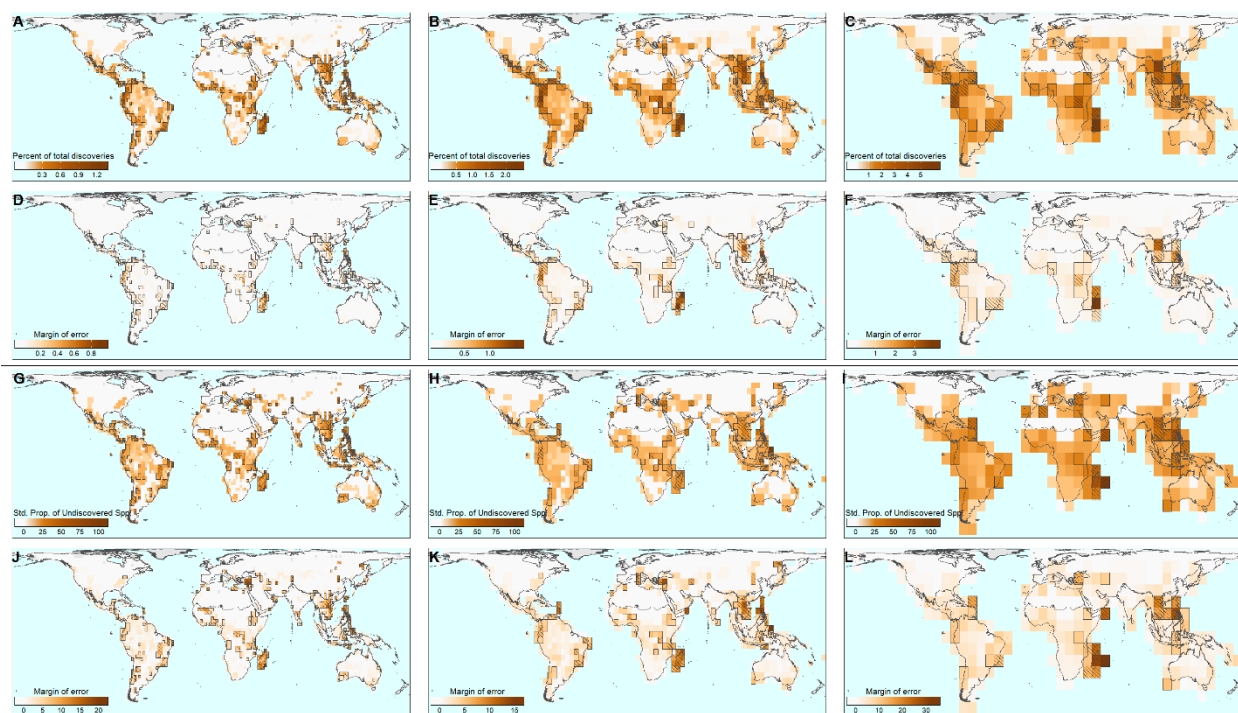


1258  
1259

*Fig. S22.*

1260 Geographical discovery patterns for reptiles at different spatial resolutions. (A-C) Percent of total discoveries across  
1261 grid cells and their respective (D-F) uncertainty ( $\pm$  margin of error). (G-I) Standardized proportion of undiscovered  
1262 species across grid cells and their respective (J-L) uncertainty ( $\pm$  margin of error). Outlined and hatched regions  
1263 designate grid cells holding values within respectively the top 10% and top 5% of the mapped metric. Maps drawn  
1264 at spatial resolutions of 220, 440, 880 km.

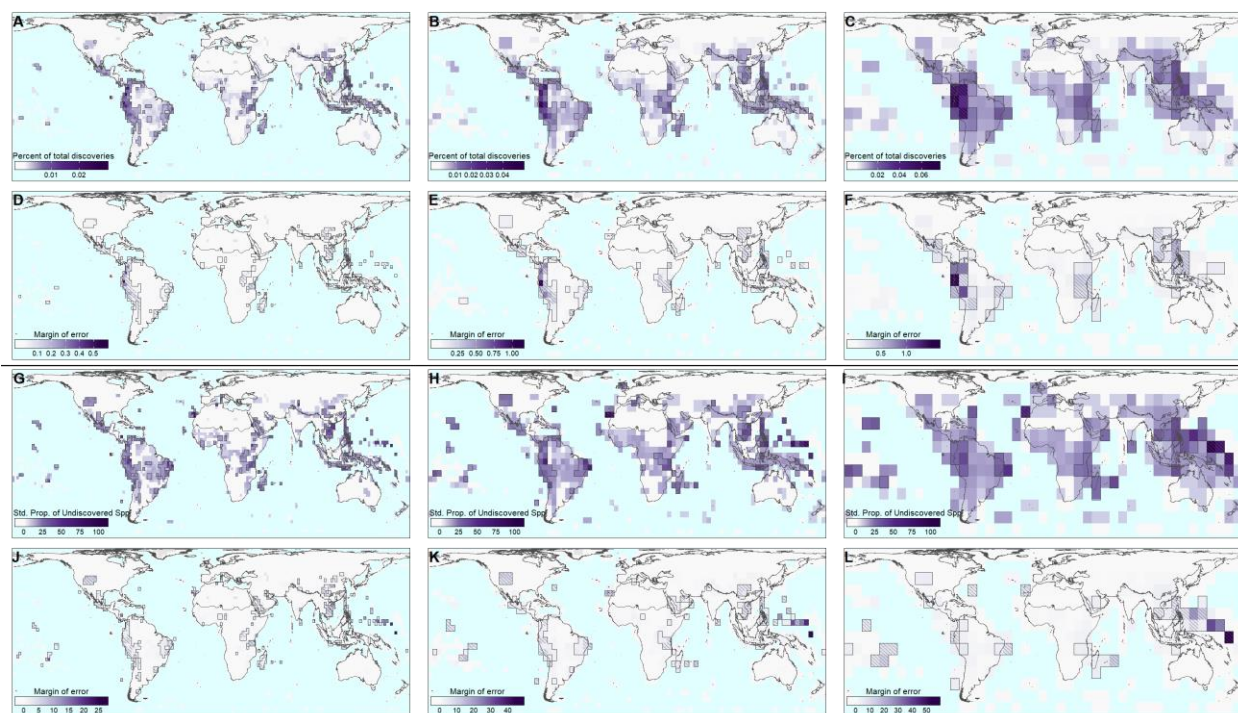




1265  
1266

*Fig. S23.*

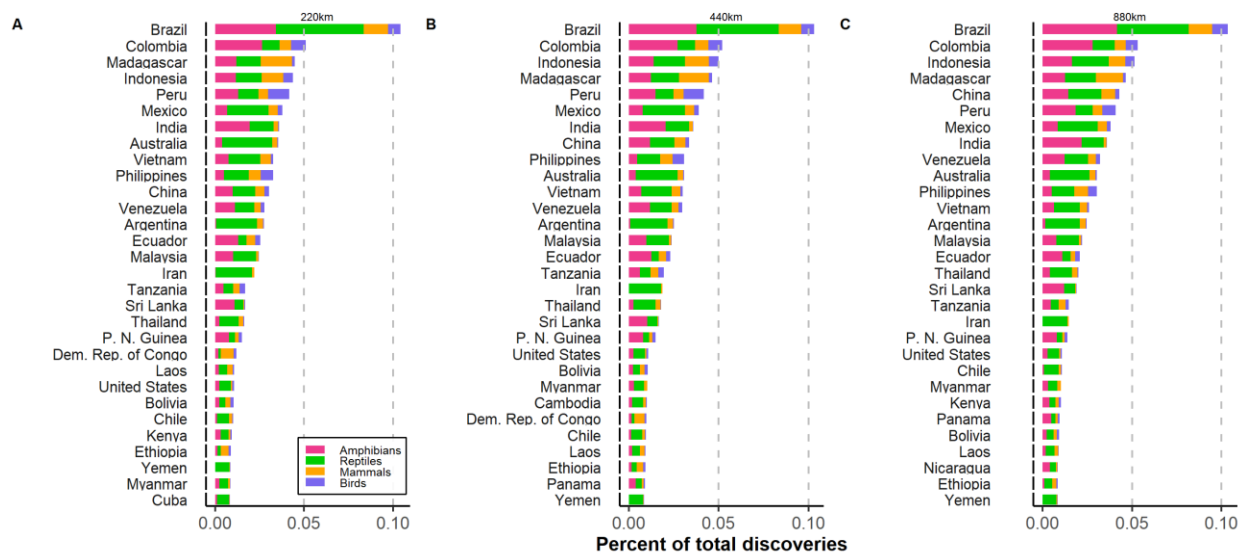
1267 Geographical discovery patterns for mammals at different spatial resolutions. (A-C) Percent of total discoveries  
1268 across grid cells and their respective (D-F) uncertainty ( $\pm$  margin of error). (G-I) Standardized proportion of  
1269 undiscovered species across grid cells and their respective (J-L) uncertainty ( $\pm$  margin of error). Outlined and  
1270 hatched regions designate grid cells holding values within respectively the top 10% and top 5% of the mapped  
1271 metric. Maps drawn at spatial resolutions of 220, 440, 880 km.



1272

1273 *Fig. S24.*

1274 Geographical discovery patterns for birds at different spatial resolutions. (A-C) Percent of total discoveries across  
1275 grid cells and their respective (D-F) uncertainty ( $\pm$  margin of error). (G-I) Standardized proportion of undiscovered  
1276 species across grid cells and their respective (J-L) uncertainty ( $\pm$  margin of error). Outlined and hatched regions  
1277 designate grid cells holding values within respectively the top 10% and top 5% of the mapped metric. Maps drawn  
1278 at spatial resolutions of 220, 440, 880 km.



1279

1280 *Fig. S25.*

1281 **Top 30 countries with higher percent of total discoveries.** Country-wide percent of total discoveries extracted  
 1282 from assemblages defined at (A) 220 km, (B) 440 km, and (C) 880 km of spatial resolution.