nature portfolio

https://doi.org/10.1038/s41559-021-01411-5

Supplementary information

Shortfalls and opportunities in terrestrial vertebrate species discovery

In the format provided by the authors and unedited

SUPPLEMENTARY INFORMATION

Shortfalls and opportunities in terrestrial vertebrate species discovery

Mario R. Moura*, Walter Jetz*

Correspondence to: mariormoura@gmail.com, walter.jetz@yale.edu

This PDF file includes:

Supplementary Methods

Supplementary Results

Supplementary References

Supplementary Tables

Supplementary Figures

Other Supplementary Information for this manuscript include the following:

Supplementary Data

SUPPLEMENTARY METHODS

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

Model validation

We used a four-fold cross-validation approach to examine the predictive accuracy of our models. For each vertebrate group, the species were randomly partitioned into four equal parts, with three of those used as the training-fold and the fourth as validation-fold. The cross-validation process was repeated four times, with each of the four-fold subsamples used once as the validation data. For each cross-validation round, we performed the model averaging approach using the 11 predictor variables and obtained for each species, the weighted average value of the (i) discovery probability in year 2015, and (ii) estimated description date.

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

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

1) Species-level validation

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

2) Taxon-level validation

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

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

mean square error (NRMSE). These three statistics were computed for each taxonomic rank (family, order) of each of the four vertebrate groups. Computations were performed in R using the 'hydroGOF' [2] and 'stats' [3] packages.

3) Assemblage-level validation

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

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

4) Bioregion-level validation

Per bioregion results are aggregates of the assemblage-level predictions and are represented by the assemblage-level validation.

5) Country-level validation

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

Model limitations

It is important to note that the estimated time-to-event functions derived from the AFT models may overestimate the discovery probability if the empirical time-to-event functions do not yet approach asymptote [4]. Consequently, the estimated proportion of known species, *PropKnown*, for species in a taxon or assemblage will also be overestimated, ultimately leading to more conservative estimates of the *PropUnknown* and *UnknownSR*. We acknowledge that the discovery curve of amphibians and reptiles have not yet approach asymptote, as evidenced by the cumulative number of species description over time (Fig. 1, main text). That said, our model framework is subject to three major limitations.

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

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

And third, our model validation procedure is subject to two mathematical constraints. The number of estimated discoveries is affected by the size of the training-fold. In adding more species to the training-fold, we tend to increase the known richness per taxon or assemblage and consequently, the estimated unknown richness (*UnknownSR*). Moreover, the number of observed discoveries per taxon or assemblage is dependent on the size of the validation-fold; that is, the number of species left out of the model training. In increasing the size of validation-fold we tend to obtain higher values of observed discoveries, with the opposite occurring if we decrease the size of validation-fold. Therefore, the size of validation- and training-folds may affect how we rank taxa and regions according to their potential for future species discovery.

We investigate these three limitations further below.

1) Robustness to sampling condition

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

Although the theoretical background to deal with incompletely observed time-to-event data has improved in the recent decades [5], the statistical tools available to account for right truncation are either restricted to particular time-to-event models - e.g. Cox Proportional Hazard Model [7] - or have somewhat limited application due to the fail in estimator convergence [8]. Herein, we

assessed the robustness of our results to the violation of this sampling condition through a sensitivity analysis. We created 10 data subsets containing increasing levels of right truncation. To do so, we successively discarded *x*% of the most recent-described species, with *x* varying from 0 to 50, at intervals of 5%. For each data subset, we repeated the model averaging framework and registered the average weighted coefficients of each predictor variable, as well as the estimated discovery probability of each species.

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

2) Influence of the time period of species discovery

To assess the differential influence of early species discoveries in our model performance, we performed a sensitivity analysis by successively discarding previously described species. Our goal was i) to evaluate the variation of predictors across time and ii) to identify the period that offered strongest model performance for prediction. In addition to the full time period of this study (1759-2014), we defined other 22 time periods covering the interval from *d* to 2014, where *d* is a decade from 1760 to 1970 (e.g. 1760-2014, 1770-2014, …, 1970-2014). We then filtered our species dataset to include only those taxa described within each time period and repeated the

model validation framework at the level of species, taxa, and assemblages. We note that with increasing *d* the sample sizes available in the datasets decreased (Supplementary Table 3).

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

3) Influence of the size of cross validation partitions

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

SUPPLEMENTARY RESULTS

Model Limitations and Sensitivity Analyses

1) Robustness to sampling condition

We found low variation in the effect size of model coefficients if up to 30% of more recently described species were discarded before model computations (Supplementary Figure 1). The increasing of right-truncation decreased the effect size of model coefficients. Only a few covariates changed the effect direction with increasing levels of right-truncation. The model coefficients were nearly invariant for birds.

The variation in model coefficients is expected to influence the estimated discovery probability. In increasing the level of the right-truncation of the species dataset, we obtained higher values of discovery probability relative to those estimated from more complete datasets (Supplementary Figure 2). Such changes were evident mostly for amphibians and reptiles, and they were virtually nonexistent for mammals and especially birds.

Among the most consistent predictors affecting the discovery probabilities there were the (i) geographic range size, (ii) taxonomic activity per biome, and (iii) species body size. The frequency distribution of these predictor variables (Supplementary Figures 3-6) confirms the well-known trend of recent-described species to show narrower geographic ranges, smaller body sizes, and be described by more taxonomists relative to species described long ago [9–11].

It is worth noting that this sensitivity analysis indirectly considered the modelling of discovery probability for species datasets with different ending dates, in an opposite way to the analysis on the influence of the time period of species discovery (next subtopic). Here, species datasets always started in 1759 but ended at different dates, according to the percentage of recently described species discarded before computations. Thus, the ending dates were not

196 necessarily equal for a same percentage of discarded species. For instance, the first half of the
197 currently known amphibian species were described by 1972, whereas 50% of the bird diversity
198 were already known by 1845 (see histograms of description year, Figure 1).
199 Had we been able to incorporate species-level attribute of unknown species, we would have
200 found higher effect size for the most important model coefficients and therefore estimated lower
201 species discovery probabilities. Given the right-truncation nature of data used in our analysis, it
202 is likely that our discovery metrics, the *PropUnknown* and *UnknownSR* per taxon or per
203 assemblage, are underestimated. The proportion and number of unknown species we report
204 should be considered as conservative estimates, particularly for amphibians and reptiles.
205 *2) Influence of the time period of species discovery*
206 We evaluated the sensitivity of our 'discovery metrics' to the time period of species
207 discovery (e.g. 1760-2014, 1770-2014, …, 1970-2014). In discarding species described long ago,
208 most of the model coefficients decreased their effect size (Fig. 2, main text). The ability of the
209 AFT models to correctly predict the species description dates did not improve after excluding
210 earlier-described species, except in mammals (but at the cost of discarding more than 70% of all
211 known mammal species; Supplementary Information Supplementary Table 3 and Supplementary
212 Figure 7).
213 At the taxon-level, the Spearman correlation between observed and estimated discoveries
214 was roughly constant after discarding species described during the first century of the modern
215 taxonomy (Supplementary Figure 8). At the assemblage level, the model performance also
216 decreased as old described species were discarded before computations. The decline in model
217 performance was evident across all subsampling levels used to control the overrepresentation of
218 wide-ranging species (Supplementary Figures 9-12). The subsampling level of 5 occurrences per

219 species showed the best performance in explaining the observed discoveries per grid cell, a result
220 consistent across all spatial resolutions and vertebrate classes.
221 Overall, we did not obtain better measures of estimated discoveries by removing species
222 described long ago. We therefore used the complete dataset (species described from 1759 to
223 2014) to estimate the discovery probability at the species-level and to estimate the number of
224 unknown species (*UnknownSR*) at the taxon- and assemblage-level. Overall, we found a strong
225 association with description year in extracting the predicted year of discovery at the threshold of
226 0.5 of discovery probability (Spearman r = 0.65–0.81 for all groups, see Supplementary Figure
227 13).
228 *3) Influence of the size of cross-validation partitions*
229 Across taxonomic ranks, we found similar values of Spearman correlation between
230 observed and estimated discoveries, regardless of the size of the species dataset used in the
231 model training and mapping procedure (Supplementary Figure 8). The size of cross-validation
232 partitions did not influence model performance when using different time periods of species
233 discovery either (Supplementary Figure 8). The extraction of *PropUnknown* based on small
234 sample sizes (i.e., using 25% of species in the model training) were less able to properly
235 characterize discovery patterns at the taxon-level relative to large sample sizes (including 50, 75,
236 90% of species in the model training).
237 The regression slope between observed and estimated discoveries per taxon tended to
238 decrease when assessed at higher taxonomic ranks and for training-folds containing more species
239 (Supplementary Figure 8). The estimated discoveries underestimated the observed discoveries
240 across all taxonomic ranks, although such underestimation was less pronounced when using 90%
241 of species in the model training (Supplementary Table 4). After standardizing *UnknownSR* to

percent of total discoveries, we observed similar model performance for all sizes of the cross-validation partitions (Supplementary Figure 14).

At the assemblage-level, the relationship between observed and estimated discoveries showed similar values of Spearman correlation, regardless of the size of cross-validation partitions (Supplementary Figures 9-12). Once again, subsampling 5 grid cell occurrences per species resulted in estimated discoveries that better predicted the observed discoveries per grid cell, regardless of the size of cross validation partitions. The discrepancy between the absolute number of observed and estimated discoveries per grid cell decreased when species assemblages were defined at either coarser spatial resolutions or based in training-folds including higher proportion of randomly selected species (Supplementary Table 4). After standardizing *UnknownSR* to percent of total discoveries, the relationship between estimated and observed discoveries per grid cell was also constant for all sizes of the cross-validation partitions (Supplementary Figure 15).

Overall, the estimated discoveries we obtained, either at the taxon- or assemblage-level analyses, underestimated the observed discoveries. The underestimation was higher when the discovery metrics were computed for models including fewer species (Supplementary Table 4). We therefore recommend caution in interpreting the absolute values of the estimated number of unknown species (*UnknownSR*) per taxon or per grid cell. We reinforce the strong monotonicity between the observed and estimated discoveries, which ultimately support our ability to rank taxa and regions according to their potential for the discovery of new species. We do not advise using this approach to update global numbers of unknown species, unless extensive cross validation reveal absence of under- or overestimation of *UnknownSR*.

4) Species authority name assumptions

Our time-to-description analysis is based on original species authority names and uses the year in which a given binomial was originally proposed. Authority name description years reveal patterns of species discovery over time [4,9–15], but do not account for the complicated taxonomic history of synonymizations and revalidations associated with a binomial name. A species that was recently revalidated still holds the same authority name – and therefore description year – as when it was first recognized as unique taxon. Therefore, our findings concern factors affecting the year when a species was first discovered, and not if and when a revalidation occurred.

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

Many firsthand species discoveries, i.e. species descriptions that propose new authority binomials, are reported for vertebrates every year. For example, more than 85% of amphibian species described between 1992-2003 resulted from newly proposed names, with less than 15% of descriptions concerning elevation of subspecies to species or revalidation of synonymies [21]. In reptiles, 79% of species descriptions between 1992-2017 were published outside revisionary

taxonomic studies [22]. Newly proposed binomial names are also a large part of recently described mammals. Since 2005, 1,251 new mammal species have been recognized as valid, with 42% of them comprising firsthand discoveries and 58% consisting of resurrection of synonymies or elevation of subspecies to species rank [23]. Altogether, these estimates illustrate the size of the taxonomic enterprise ahead. Concerns might go beyond differentiating firsthand discoveries from species revalidations, questioning the validity of species descriptions under the application of different species concepts [18]. To date, reflecting the ongoing debate about species concepts in biology [24], comprehensive taxonomic databases that standardize global species lists according to a single species concept remain out of reach.

SUPPLEMENTARY REFERENCES

1. Barnston, A. G. Correspondence among the Correlation, RMSE, and Heidke Forecast Verification Measures; Refinement of the Heidke Score. *Weather Forecast.* **7**, 699–709 (1992).
2. Zambrano-Bigiarini, M. hydroGOF: Goodness-of-fit functions for comparison of simulated and observed hydrological time series. (2017). doi:10.5281/zenodo.840087
3. R Core Team. R: A Language and Environment for Statistical Computing. v. 3.5.3 (2019).
4. Bebber, D. P., Marriott, F. H. C., Gaston, K. J., Harris, S. A. & Scotland, R. W. Predicting unknown species numbers using discovery curves. *Proc. R. Soc. B* **274**, 1651–1658 (2007).
5. Klein, J. P. & Moeschberger, M. L. *Survival analysis: Techniques for censored and truncated data*. *Pharmaceutical Statistics* (Springer, 2003). doi:10.1002/pst.135
6. Pante, E., Schoelinck, C. & Puillandre, N. From Integrative Taxonomy to Species

Description: One Step Beyond. *Syst. Biol.* **64**, 152–160 (2015).

7. Vakulenko-Lagun, B., Mandel, M. & Betensky, R. A. coxrt: Cox Proportional Hazards Regression for Right-Truncated Data. v. 1.0.2 (2019).

8. Mantel, N. & Myers, M. Problems of convergence of maximum likelihood iterative procedures in multiparameter situations. *J. Am. Stat. Assoc.* (1971). doi:10.1080/01621459.1971.10482289

9. Moura, M. R. *et al.* Geographical and socioeconomic determinants of species discovery trends in a biodiversity hotspot. *Biol. Conserv.* **220**, 237–244 (2018).

10. Diniz-Filho, J. A. F. *et al.* Macroecological correlates and spatial patterns of anuran description dates in the Brazilian Cerrado. *Glob. Ecol. Biogeogr.* **14**, 469–477 (2005).

11. Collen, B., Purvis, A. & Gittleman, J. L. Biological correlates of description date in carnivores and primates. *Glob. Ecol. Biogeogr.* **13**, 459–467 (2004).

12. Gaston, K. J., Blackburn, T. M. & Loder, N. Which species are described first? The case of North-American butterflies. *Biodivers. Conserv.* **4**, 119–127 (1995).

13. Colli, G. R. *et al.* In the depths of obscurity: Knowledge gaps and extinction risk of Brazilian worm lizards (Squamata, Amphisbaenidae). *Biol. Conserv.* **204**, 51–62 (2016).

14. Essl, F., Rabitsch, W., Dullinger, S., Moser, D. & Milasowszky, N. How well do we know species richness in a well-known continent? Temporal patterns of endemic and widespread species descriptions in the European fauna. *Glob. Ecol. Biogeogr.* **22**, 29–39 (2013).

15. Bebber, D. P. *et al.* Herbaria are a major frontier for species discovery. *Proc. Natl. Acad. Sci.* **107**, 22169–22171 (2010).

16. Raposo, M. A. *et al.* What really hampers taxonomy and conservation? A riposte to

Garnett and Christidis (2017). *Zootaxa* **4317**, 179 (2017).

17. Dayrat, B. Towards integrative taxonomy. *Biol. J. Linn. Soc.* **85**, 407–415 (2005).

18. Isaac, N. J. B., Mallet, J. & Mace, G. M. Taxonomic inflation: its influence on macroecology and conservation. *Trends Ecol. Evol.* **19**, 464–469 (2004).

19. Garnett, S. T. & Christidis, L. Taxonomy anarchy hampers conservation. *Nature* **546**, 25–27 (2017).

20. Bremer, K., Bremer, B., Karis, P. & Källersjö, M. Time for change in taxonomy. *Nature* **343**, 202–202 (1990).

21. Köhler, J. *et al.* New Amphibians and Global Conservation: A Boost in Species Discoveries in a Highly Endangered Vertebrate Group. *Bioscience* **55**, 693 (2005).

22. Guedes, J. J. M., Feio, R. N., Meiri, S. & Moura, M. R. Identifying factors that boost species discoveries of global reptiles. *Zool. J. Linn. Soc.* **190**, 1274–1284 (2020).

23. Burgin, C. J., Colella, J. P., Kahn, P. L. & Upham, N. S. How many species of mammals are there? *J. Mammal.* **99**, 1–14 (2018).

24. Mallet, J. Species, Concepts of. in *Encyclopedia of Biodiversity* **6**, 679–691 (Elsevier, 2013).

25. Kutner, M. H., Nachtsheim, C. J., Neter, J. & Li, W. *Applied Linear Statistical Models*. (McGraw-Hill Irwin, 2004).

SUPPLEMENTARY TABLES

Supplementary Table 1.

Taxon-specific allometric equations used to convert straight carapace length (SCL) of chelonians into body mass. Sample size used to create the equations, number of species covered by the samples, and summary statistics of each allometric equation are shown. Allometric equation: Log(BodyMass) = Intercept + Coef × Log(SCL).

Taxa	Sample Size	Number of species	Intercept	Coefficient	R^2
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
Testutinidae	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

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

Supplementary Table 2.

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

Variable	Amphibians	Reptiles	Mammals	Birds
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

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

Supplementary Table 3.

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

Time Period	Amphibians	Reptiles	Mammals	Birds
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[a]	3150 (64.3)	2911 (87.7)	860 (76.3)	149 (99)
>1990 – 2014[a]	2598 (75.9)	2387 (91.3)	699 (83.2)	102 (99.6)
>2000 – 2014[a]	1748 (92.9)	1687 (96.8)	494 (93.6)	36 (100)
>2010 – 2014[a]	515 (100)	645 (100)	184 (100)	0 (100)

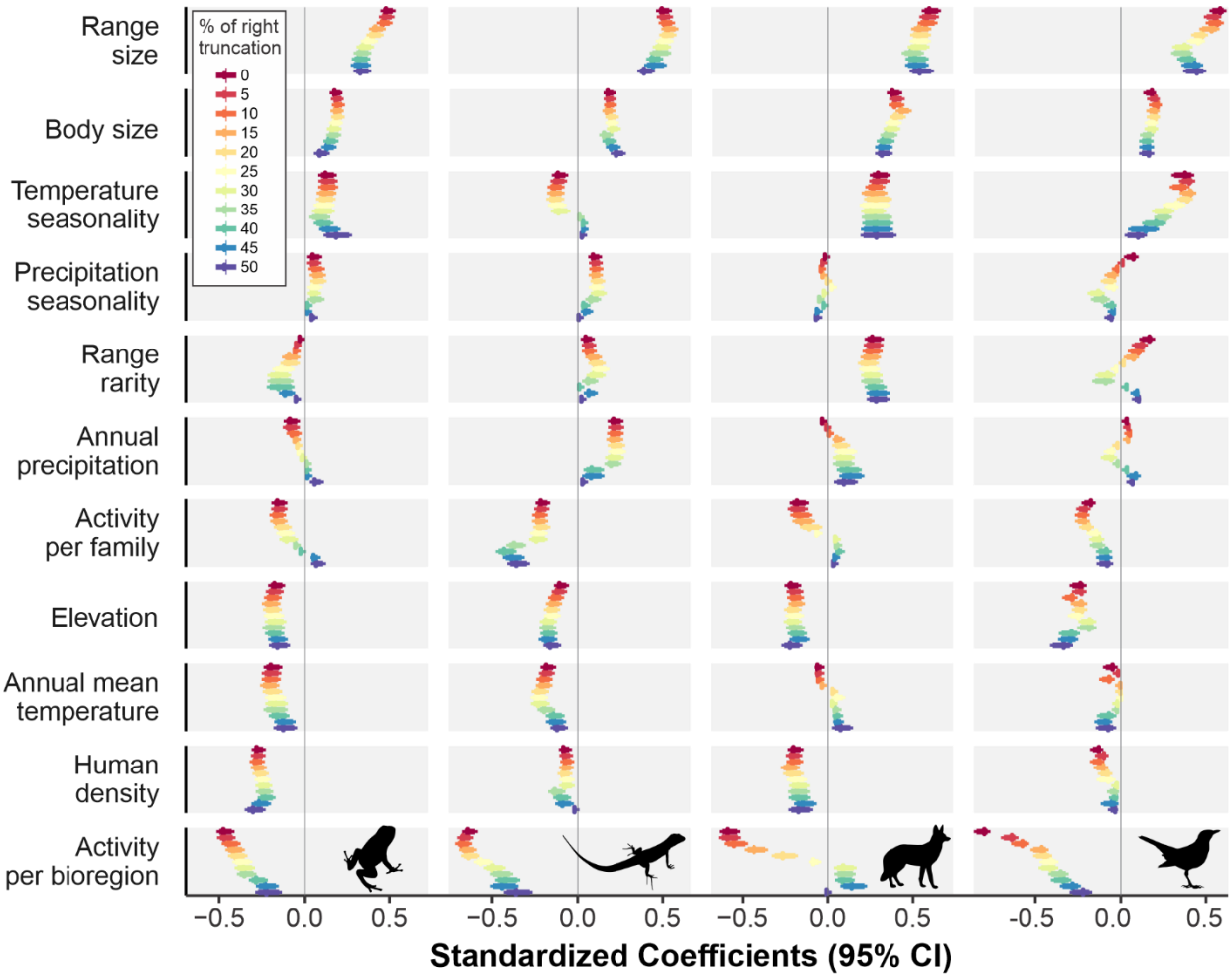
[a] Time periods not used in the sensitivity analysis.

Supplementary Table 4.

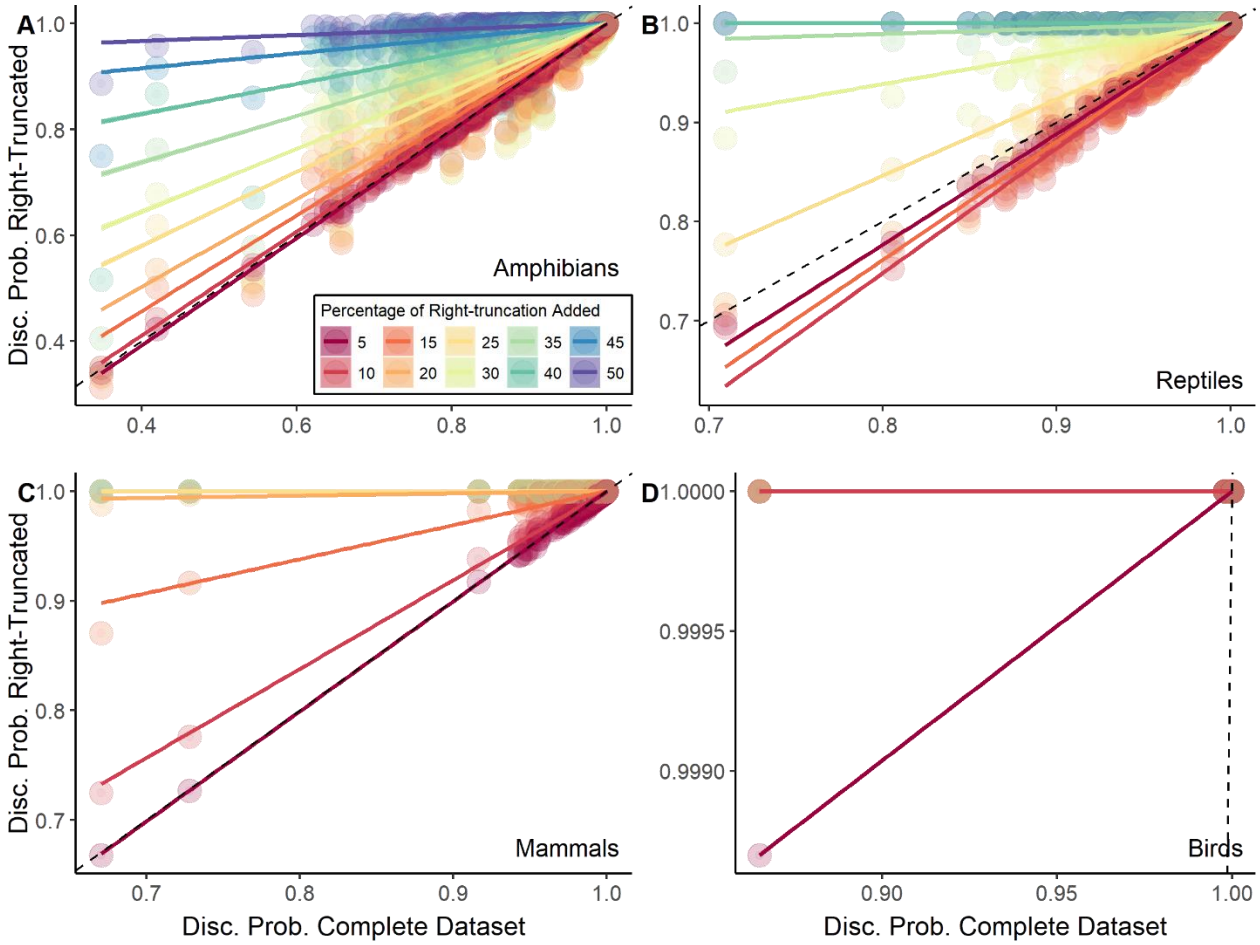
Validation-fold and grouping dependence of deviation from a 1:1 relationship. Values shown how many times, on median, the observed discoveries were higher than the estimated discoveries. Results shown for the complete time period of species discoveries (1759-2014). At the assemblage level, only the subsampling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

Grouping level	Validation-fold size	Amphibians	Reptiles	Mammals	Birds
Taxon-level					
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
Family	**Average**	**6.899**	**7.613**	**7.047**	**14.737**
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
Higher-level grouping	**Average**	**10.862**	**11.307**	**12.120**	**35.563**
Assemblage-level					
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
220 km	**Average**	**2.814**	**3.754**	**2.754**	**3.323**
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
440 km	**Average**	**5.112**	**7.012**	**5.457**	**6.778**
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
880 km	**Average**	**8.810**	**10.951**	**12.345**	**15.960**

SUPPLEMENTARY FIGURES

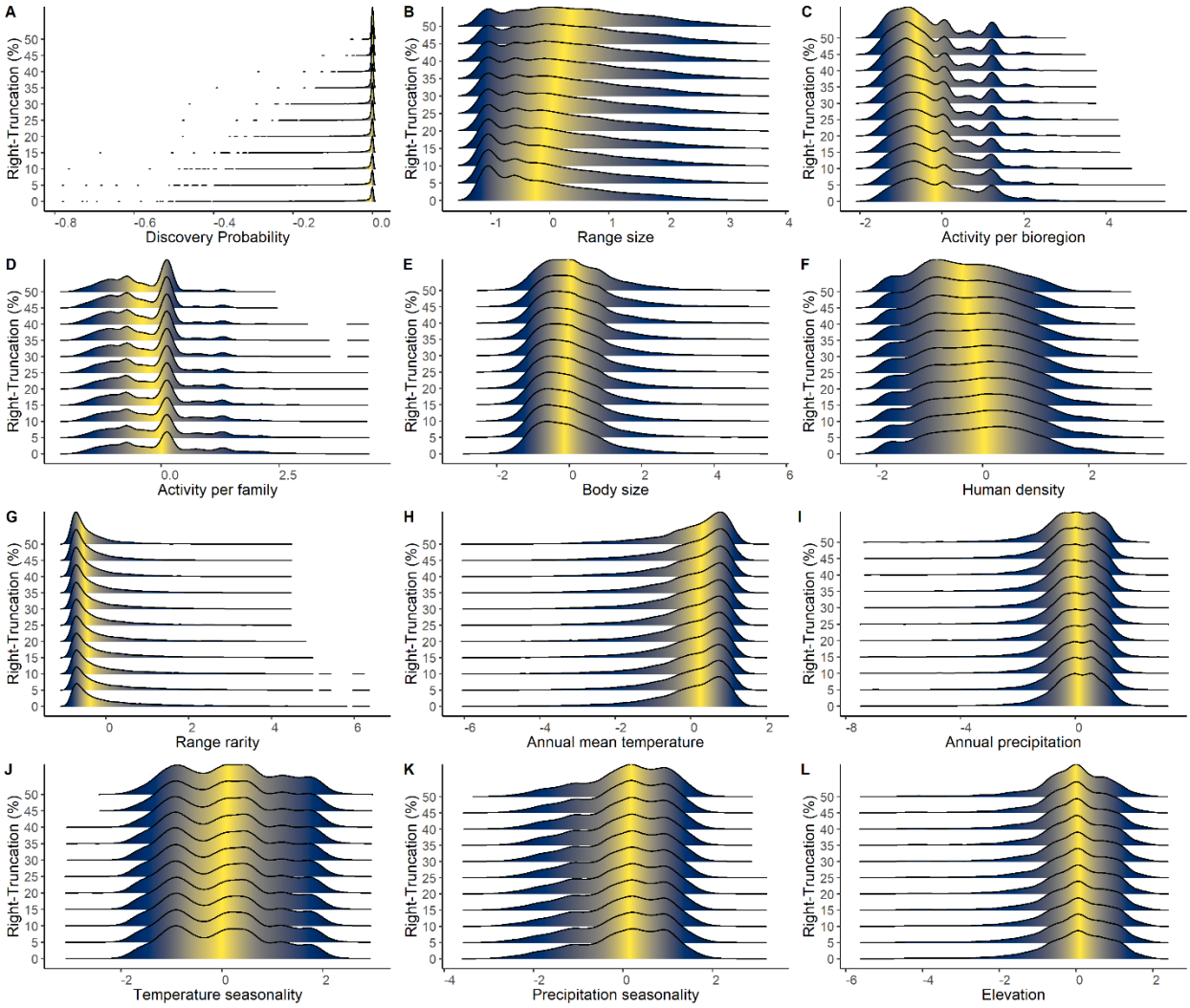


Supplementary Figure 1.
Standardized coefficients of the average weighted accelerated failure time (AFT) model computed for species datasets with increasing levels of right-truncation. Line colours indicate the percentage of recent-described species left out of the model (recent-described species were successively discarded). The horizontal bars denote the 95% confidence intervals around each coefficient. Standardized coefficients above 0 indicate that species with high values for a given attribute had higher discovery probability (prob.) and thus were likely discovered early on. Negative standardized coefficients mean high attribute values depressed discovery probability and delayed discovery.



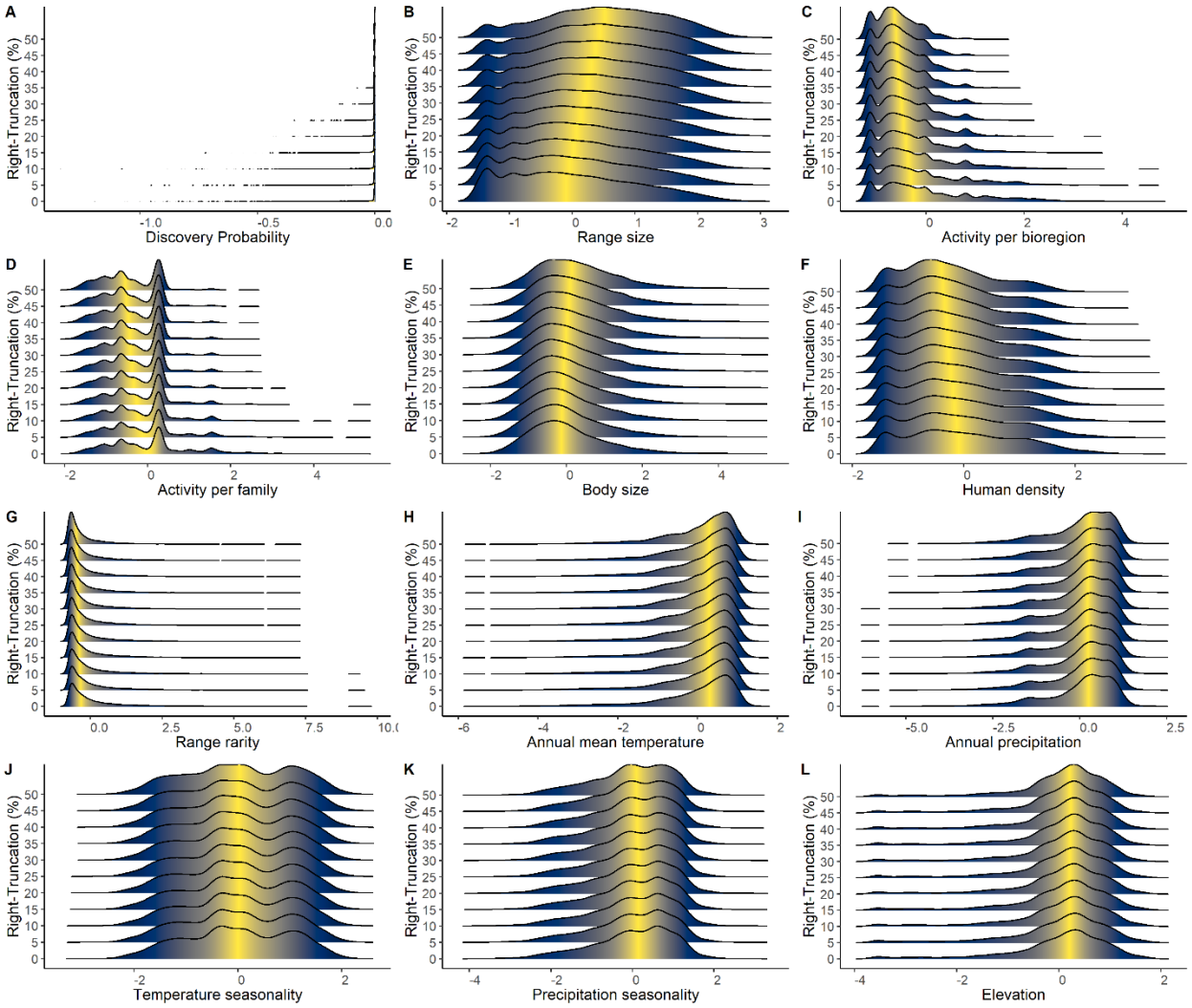
Supplementary Figure 2.

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



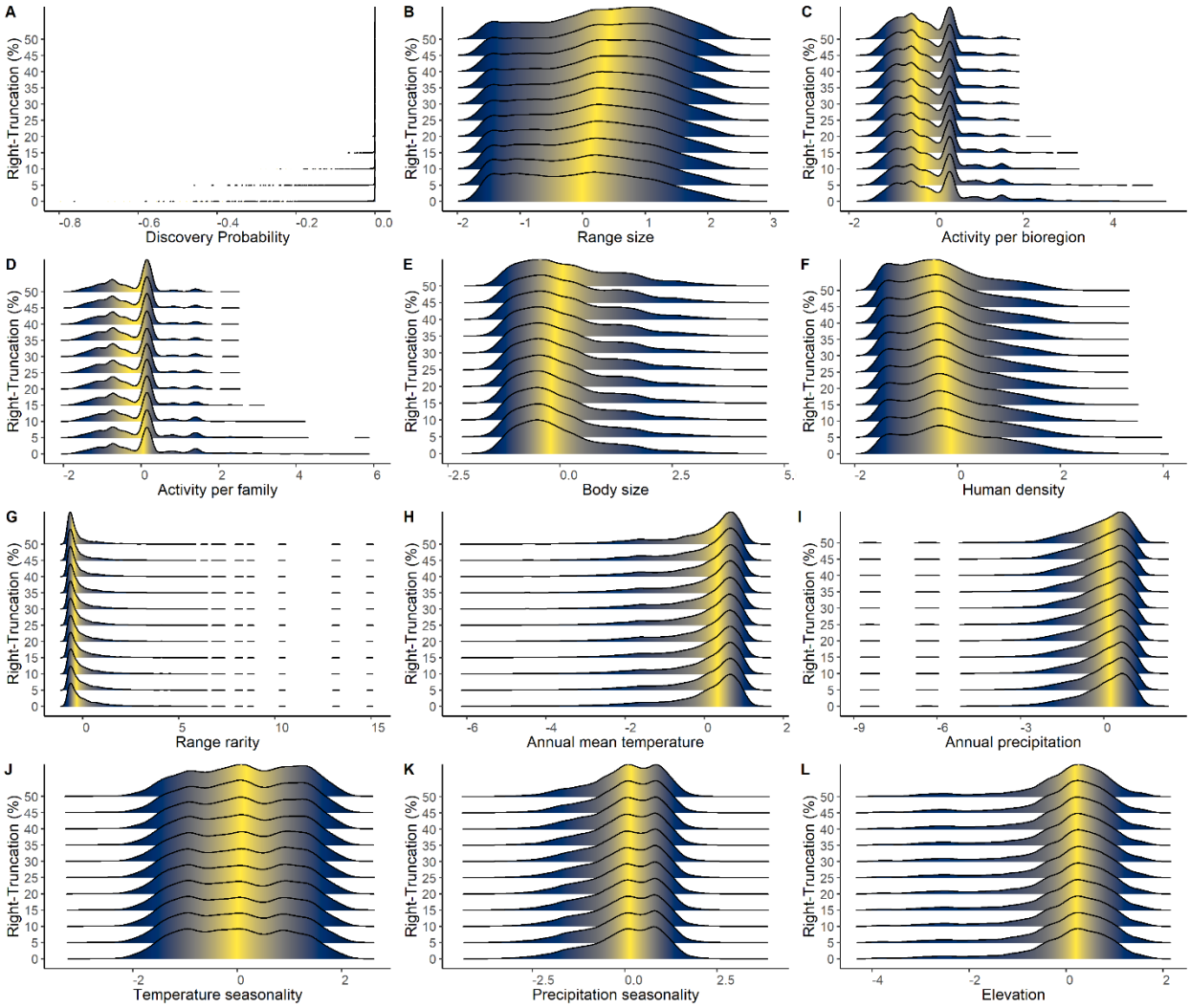
Supplementary Figure 3.

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



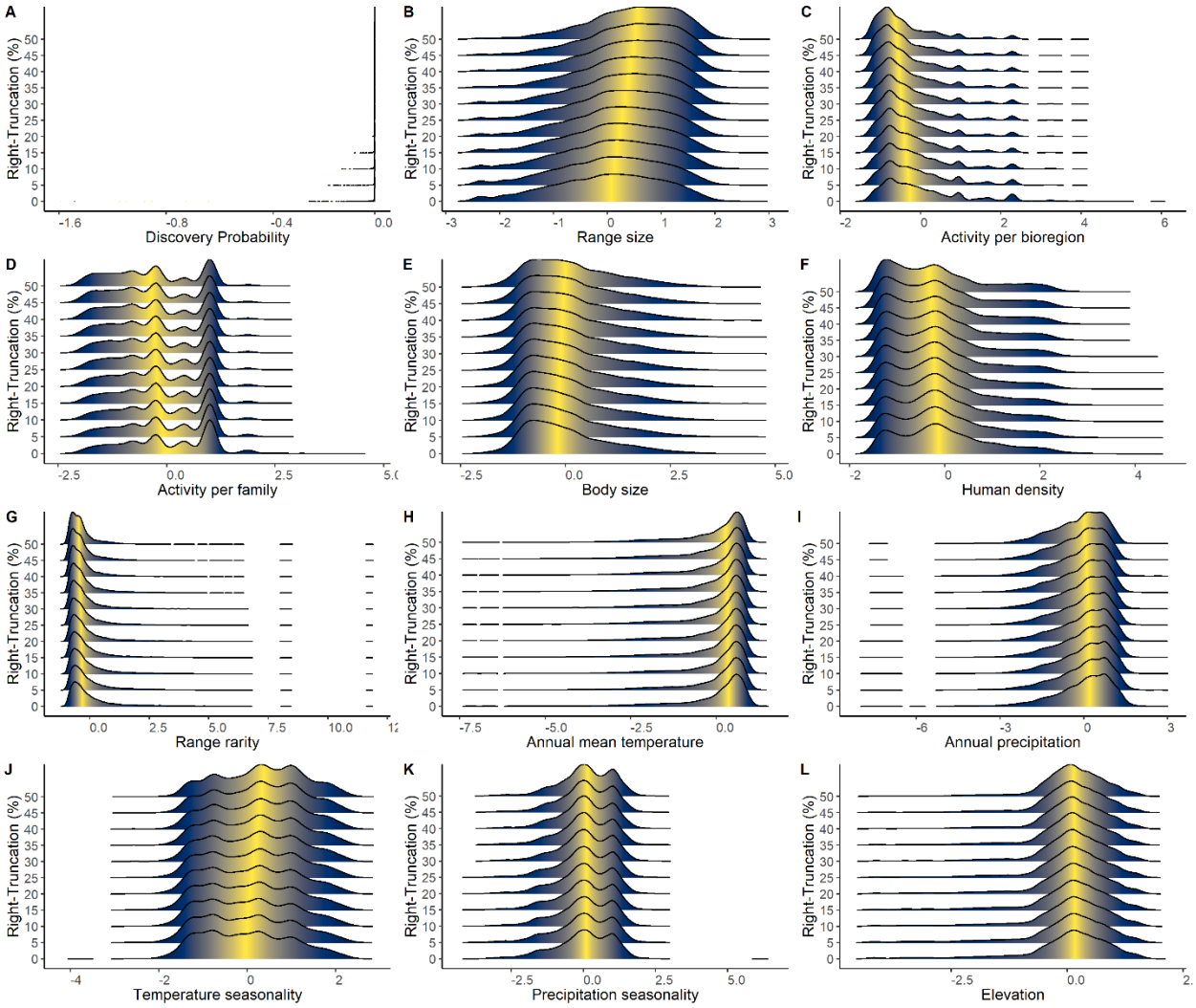
Supplementary Figure 4.

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



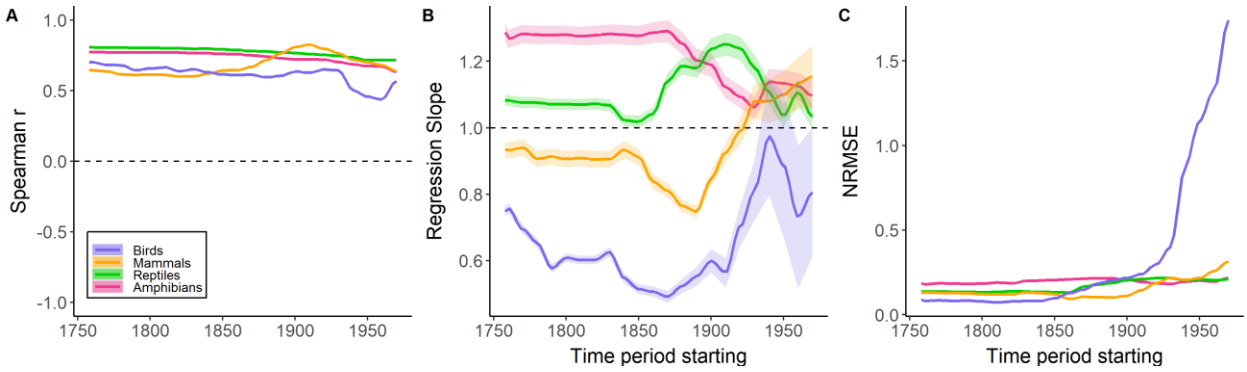
Supplementary Figure 5.

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



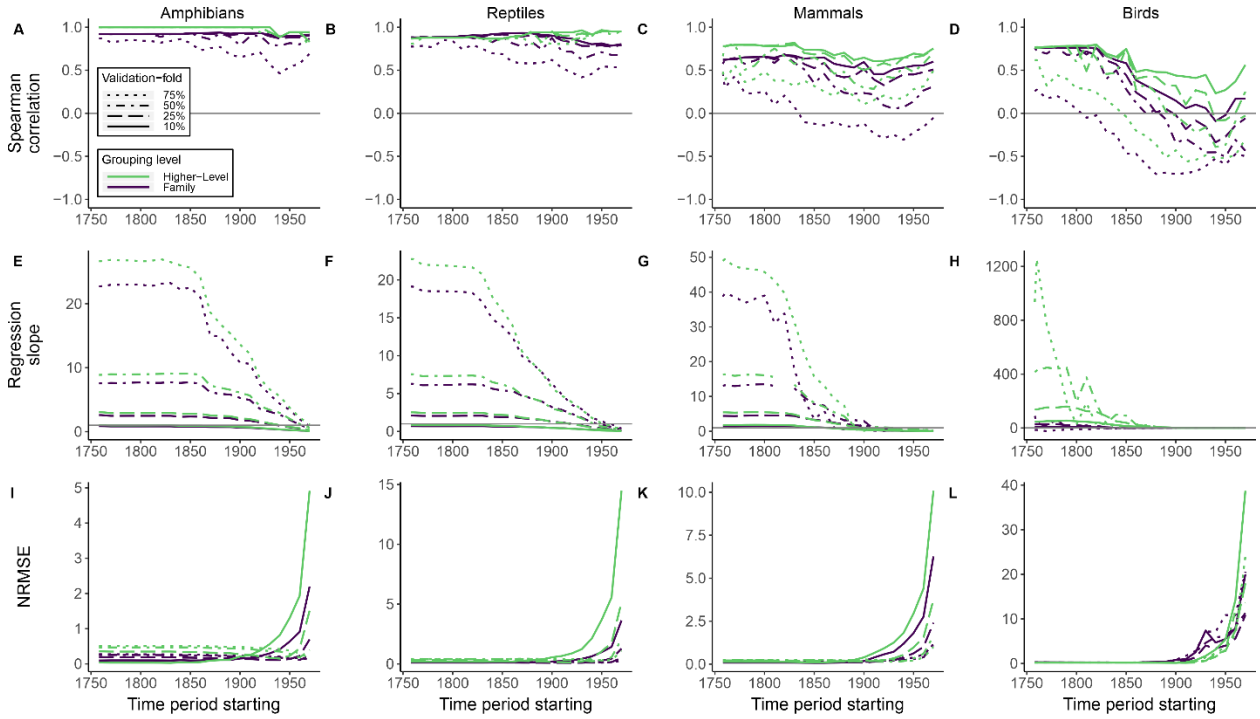
Supplementary Figure 6.

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



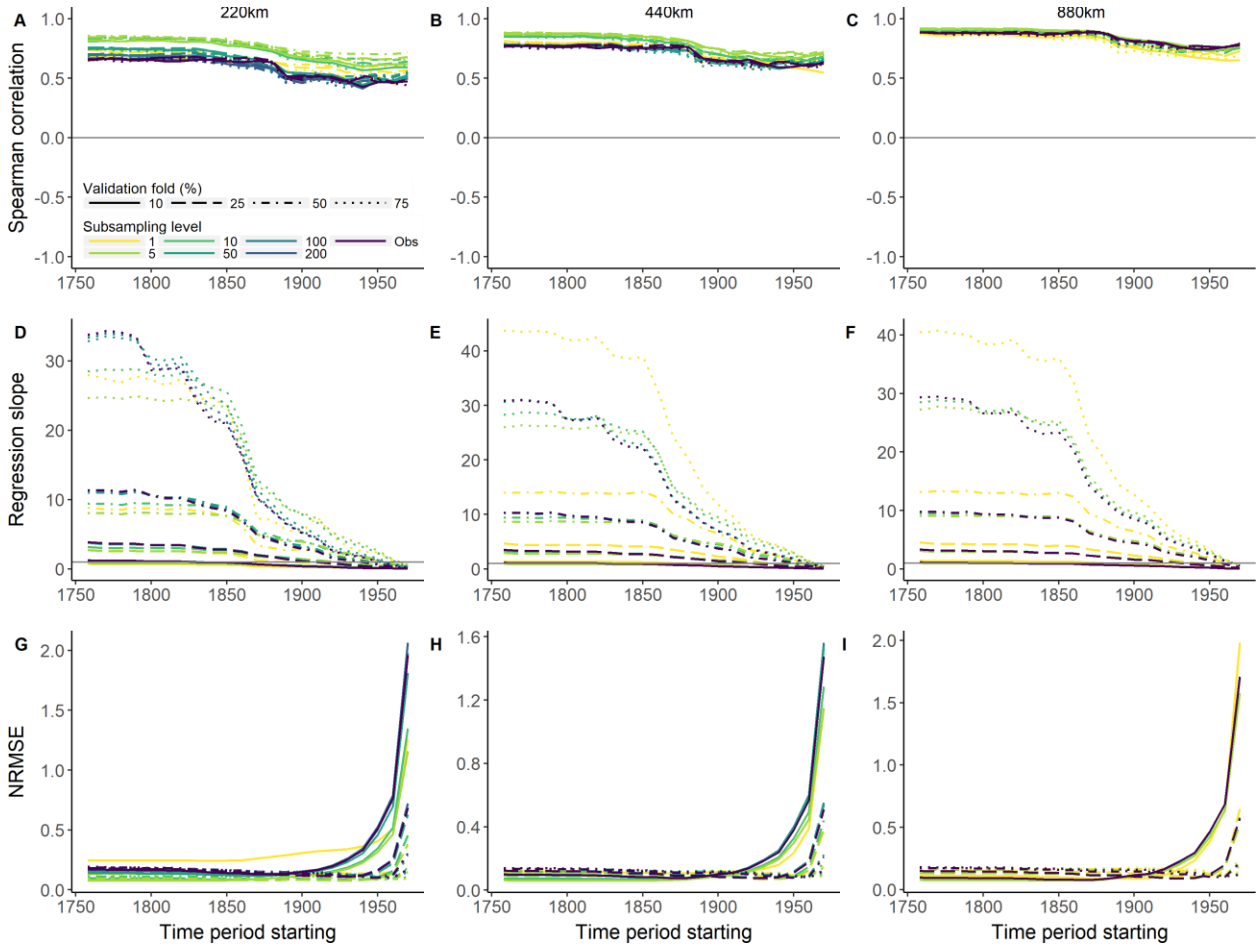
Supplementary Figure 7.

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).



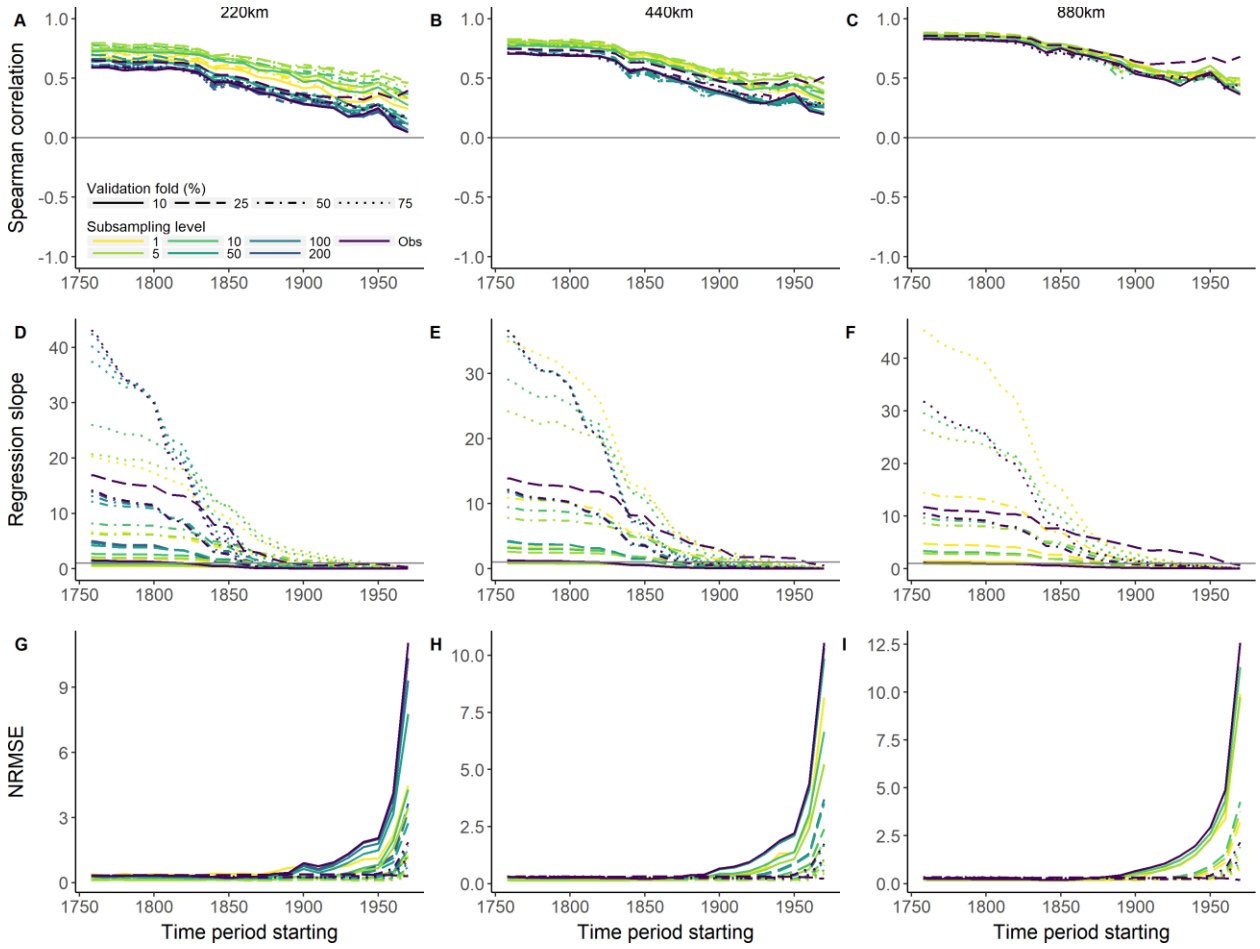
Supplementary Figure 8.

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



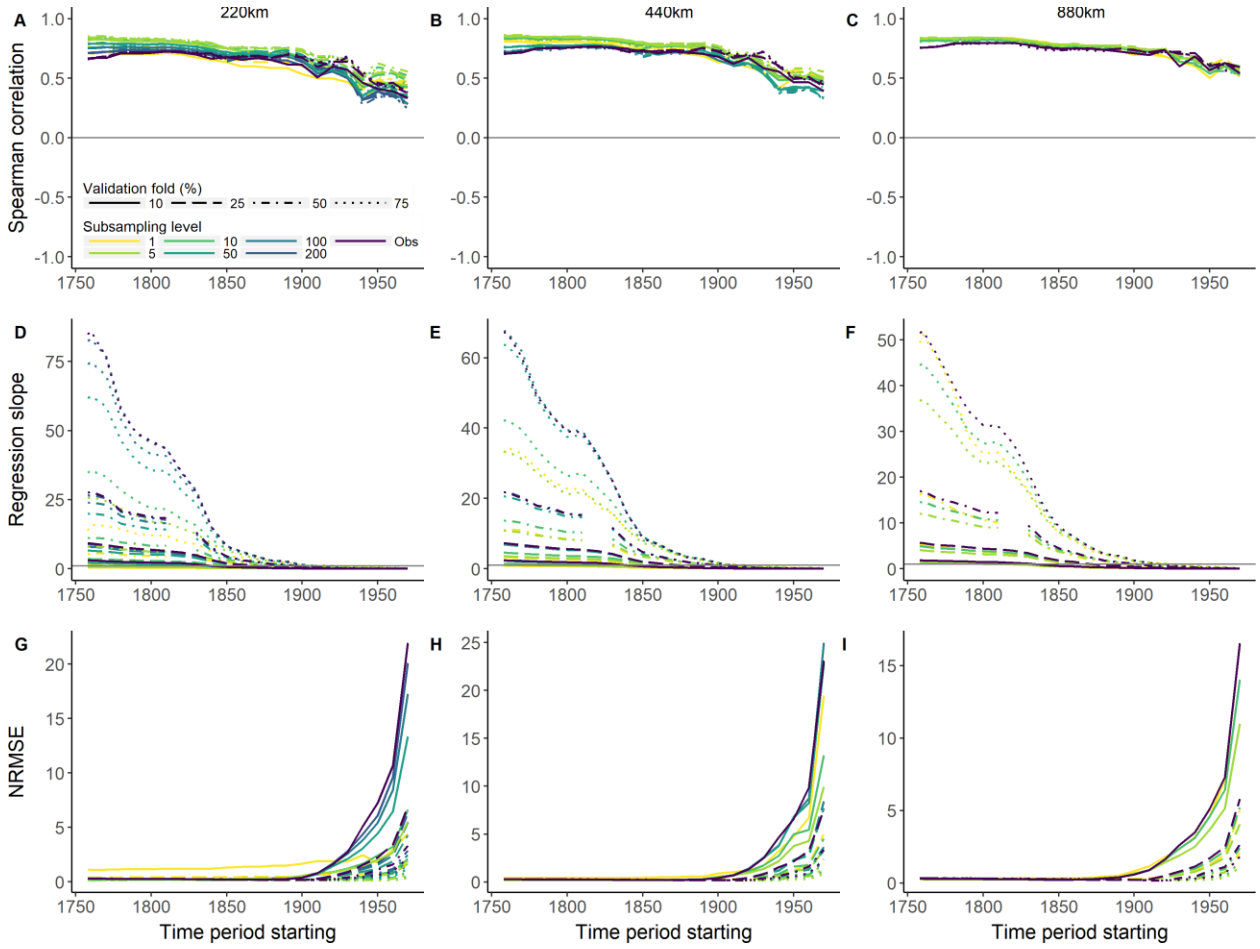
Supplementary Figure 9.

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



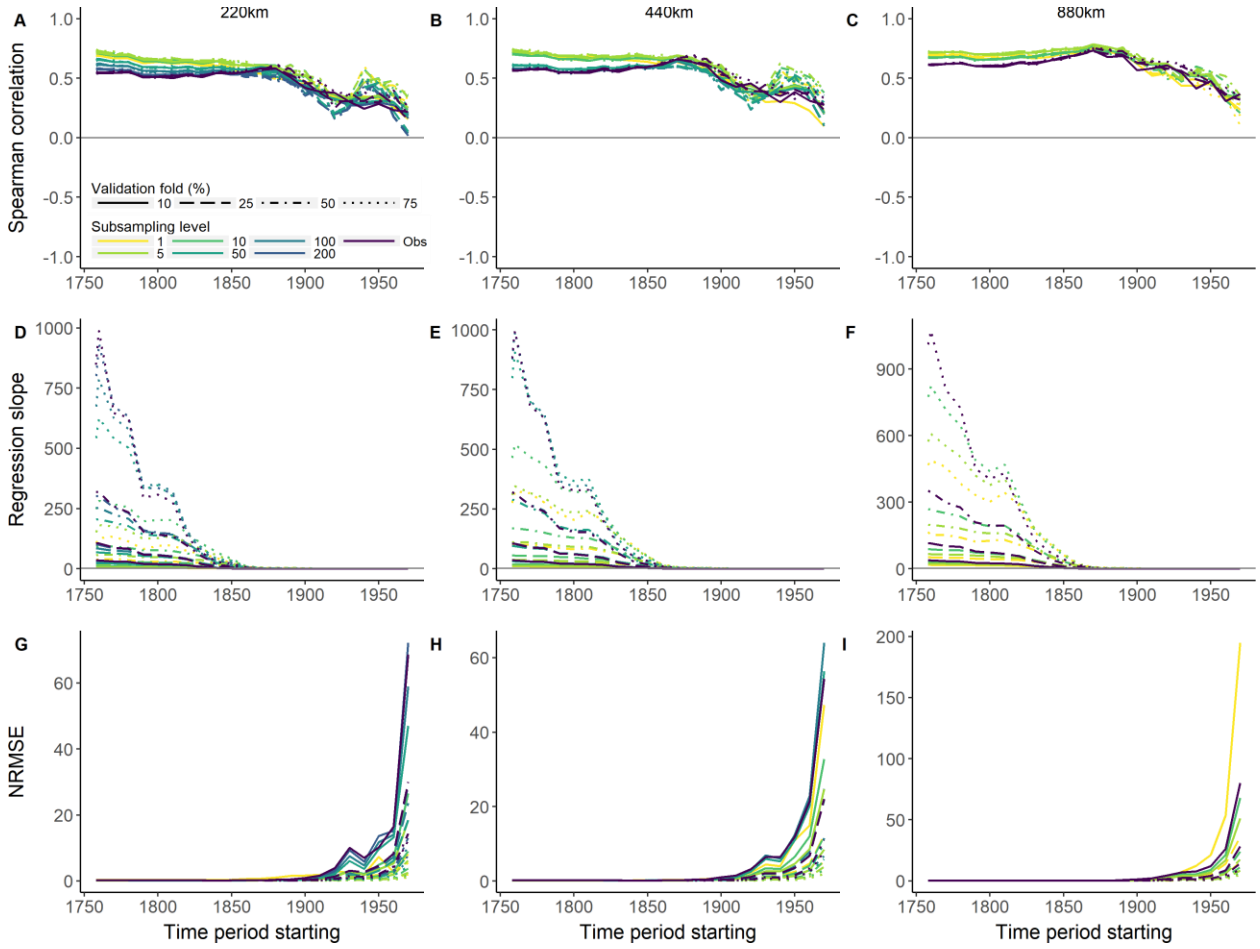
Supplementary Figure 10.

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



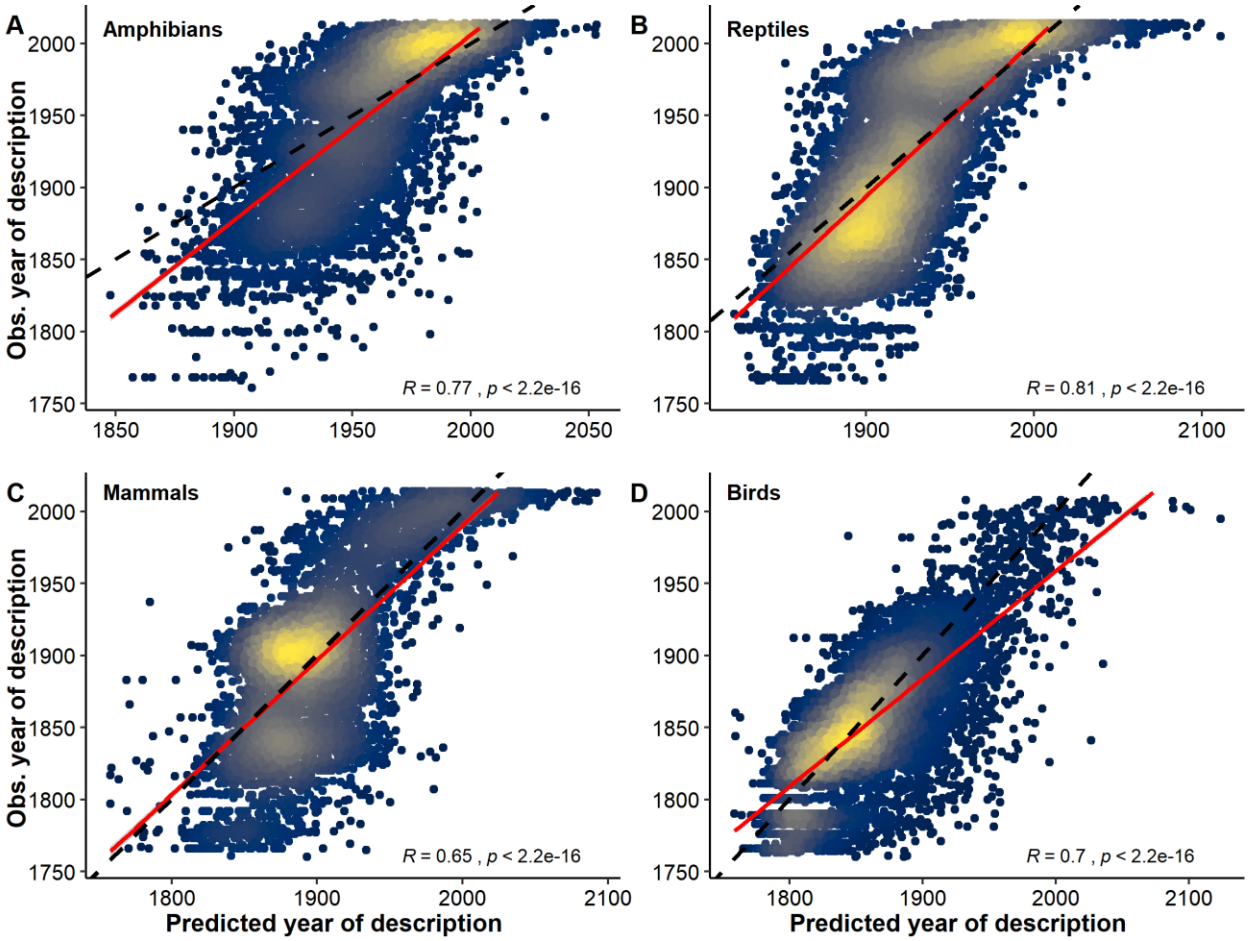
Supplementary Figure 11.

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



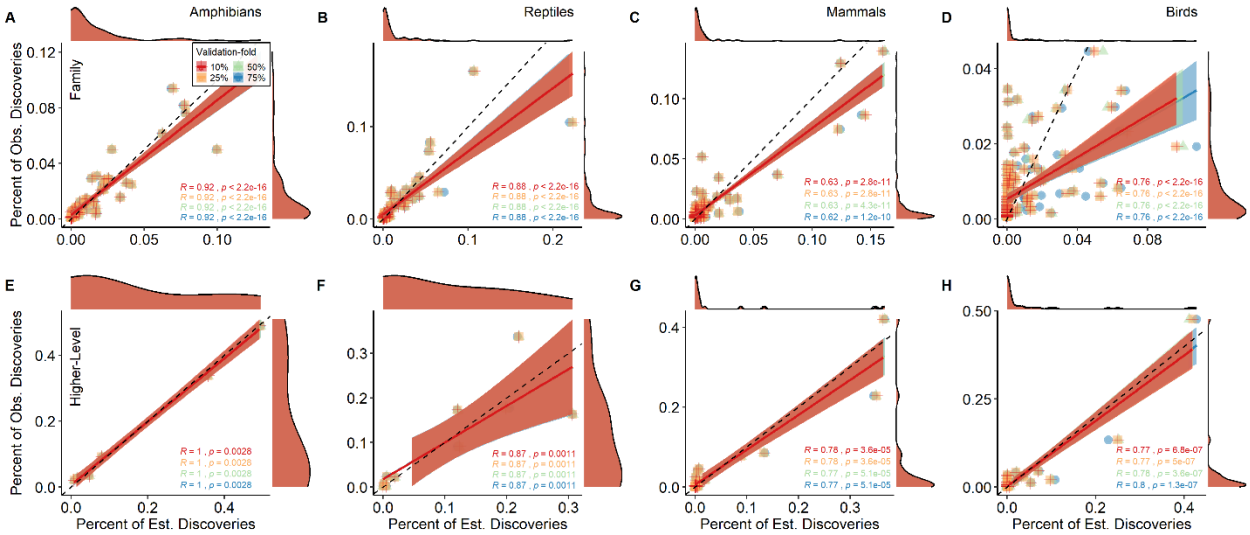
Supplementary Figure 12.

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



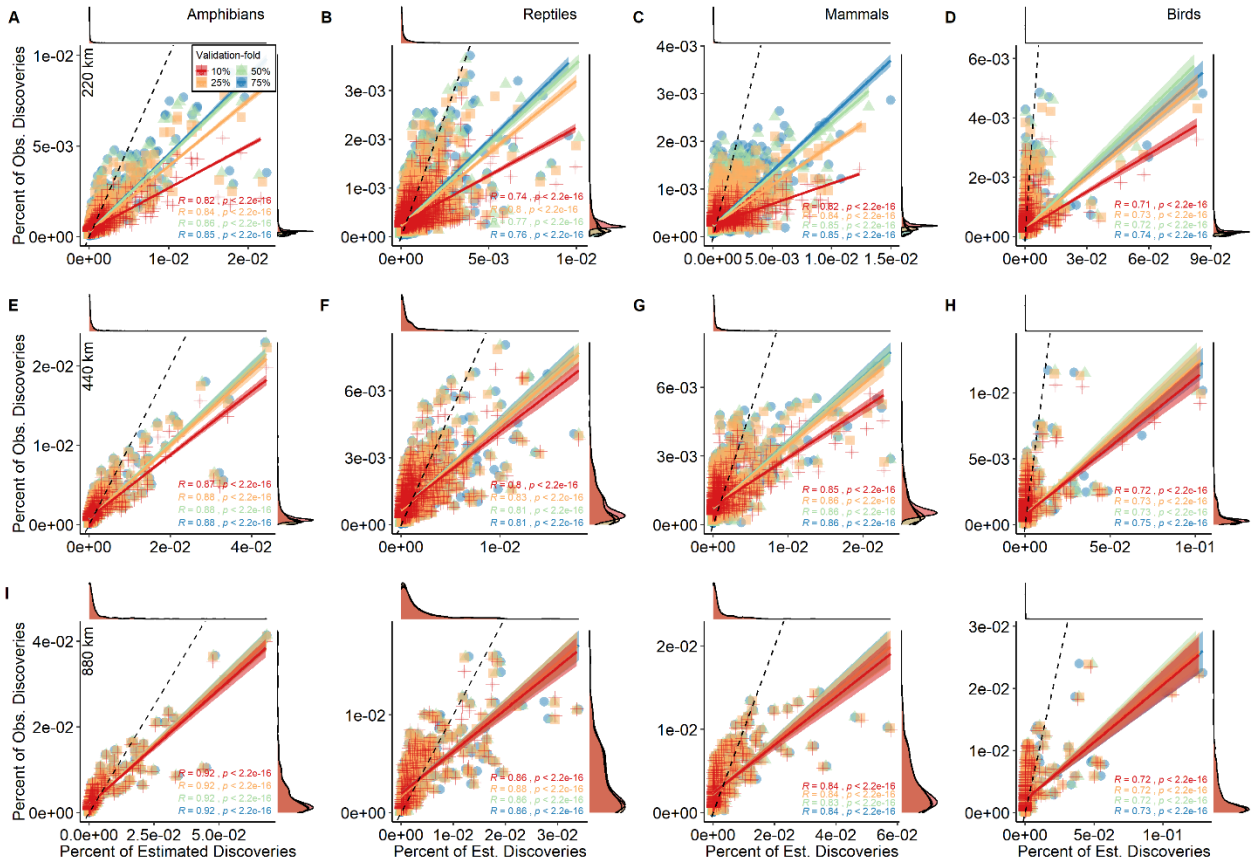
Supplementary Figure 13.

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



Supplementary Figure 14.

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



Supplementary Figure 15.

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

SUPPLEMENTARY DATA

Supplementary Data 1. (separate file) File (TaxonLevelEstimates.zip) includes:

TaxonLevelEstimates_HigherLevelGrouping.csv

TaxonLevelEstimates_TaxonomicFamily.csv

Supplementary Data 2. (separate file) File (AssemblageLevelEstimates.zip) includes:

AssemblageLevelEstimates_220km.csv

AssemblageLevelEstimates_440km.csv

AssemblageLevelEstimates_880km.csv

Supplementary Data 3. (separate file) File (BioregionLevelEstimates.zip) includes:

BioregionLevelEstimates_220km.csv

BioregionLevelEstimates_440km.csv

BioregionLevelEstimates_880km.csv

Supplementary Data 4. (separate file) File (CountryLevelEstimates.zip) includes:

CountryLevelEstimates_220km.csv

CountryLevelEstimates_440km.csv

CountryLevelEstimates_880km.csv