Supplementary Methods and Results

Assessing classifiers for conservation. RF was applied to develop accurate classifiers from extremely large datasets, resulting in a simple tool to predict the probability that an unassessed species may be under some threat. We included several sampling schemes to overcome potential biases in the data, such as uneven sampling across categories and/or low sample size (Supplementary Tables 1-14). These analyses used two combinations of IUCN category listings as potential classes (i.e., response variables); specifically, we combined classes into 1) those that are of least concern (LC) vs. those that are non-LC, and 2) those that are critically endangered (CR) vs. those that are non-CR. In addition, we used down- and re-sampling approaches to balance the response variables, which has been shown to be particularly useful in biodiversity datasets.

For the 'spatial' datasets, error rates varied across the RF classifiers but the highest accuracy and most balanced per-category error rates were obtained using downsampled data and Red List categories coded as LC vs. non-LC (Supplementary Tables 2-6). This is due to the disparity in the sizes of the categories that occur when using the full category listings (Table 1); unequal representation results in reduced per-category accuracy. Downsampling produces balanced error rates across categories (Supplementary Tables 2-5), and should be applied in RF analyses where datasets have unequal representation across categories. To optimize the accuracy of our spatial classifiers, we ran two additional RF analyses for each continent using LC vs. non-LC as the response variable, with species that were classified incorrectly more than 90% and 80% of the time removed from the classifier. These models had lower overall error rates (80% - 90%; Supplementary Tables 2 & 4).

Quantifying the effect of spatial errors in the predictions

Previous work has investigated and measured the effect of data errors present in GBIF, especially in the framework of biodiversity studies. A particularly relevant work on this is that of Maldonado et al. (1).

There, the authors sought to compare results obtained from a traditional ‘manual’ data collection (visits to herbaria and recording of data directly from curated herbaria sheets) and from a GBIF search. They did this for a specific group of plants (Rubiaceae: Cinchoneae), identifying centers of diversity for the whole group, and estimating range characteristics of each investigated plant species. Unlike some expectations, they show that the most important source of error in the identification of biodiversity patterns using GBIF is the use of observations coming from records that contain no data (NA) for the ‘Locality’ field. These are usually coordinates that are associated to observations for which no precise coordinate exists, and which were assigned to the center of the lowest known political delimitation unit (e.g., city/town, county, province/state, etc.). Their work demonstrated that removing these localities from the GBIF datasets improved the biodiversity results to make them significantly similar to those obtained using a ‘manual’ data collection. Importantly in relation to our study, they could not identify any significant differences in range characteristics (in their case, elevation range), indicating that these errors should not be expected to affect ecological or general spatial characteristics of the taxa considered. In the framework of our study, this suggests that because we are indeed using ecological and range descriptors (climatic variables, range sizes, range spatial characteristics, etc.), it is relatively unlikely that such errors are driving the global and regional patterns we observe. To demonstrate this, we performed a partial and supplementary analysis of our dataset. We first identified the regions harboring the highest and lowest proportions of observations for which the ‘Locality’ GBIF field was empty (=’NA’). When doing so, Central America appeared as the one presenting the lowest (0.05), and Asia as the one presenting the highest (0.29) proportion of data missing values in that field. Using an approach similar to that used by Maldonado et al., we filtered out these localities, we recalculated all variables for each taxon, and we then rebuilt our classifiers. Like with the original dataset, we then calculated the regional per-cell probabilities of belonging to a non-LC category. Finally, and using these new predictions, we compared the results obtained with the original and this newly cleaned dataset. Our results (Supplementary Figures 4 and 5) agree with those obtained by Maldonado et al.: removing these ‘misleading’ points does not affect our predictions in any obvious way. While for the Central America datasets, the results provide probability values that are extremely similar

between the two methods (median of difference in probability: -0.000749), the Asian dataset has values that are larger than the ones obtained for Central America, but which are still close to each other (median of difference in probability: 0.116). Most importantly, however, in both cases, the differences between the datasets are not only small, but they are spatially constant (bottom left panel in Supplementary Figures 4 and 5), which suggests that the error being introduced is not affecting the general conservation need pattern that we observe.

References

1. Maldonado C, et al. (2015) Estimating species diversity and distribution in the era of Big Data: to what extent can we trust public databases? Glob Ecol Biogeogr 24(8):973-984.

Supplementary Table 1. Number of species used to build classifiers for the down-sampled 'spatial' datasets and the 'spatial+morpho' datasets, as well as number of species used to build classifiers where species that were misclassified 90% and 80% of the time were removed.

	Spatial					Spatial+Morpho	
	All IUCN listing downsampled	LC vs non-LC downsampled	CR vs non-CR downsampled	LC vs non-LC rm 0.9	LC vs non-LC rm 0.8	LC vs non-LC resampled	LC vs non-LC downsampled
Africa	405	1842	162	1994	1944	80	40
Africa endemics	355	1270	142	1434	1386	26	13
Asia	150	526	60	1669	1617	78	39
Asia endemics	120	412	48	620	597	6	3
Australia	25	154	10	589	576	80	40
Australia endemics	20	110	8	189	186	46	23
Central America	145	512	58	652	636	100	50
Central America endemics	130	306	52	368	361	72	36
Europe	145	292	58	956	936	6	3
Europe endemics	10	26	4	121	117	NA	NA
North America	35	122	14	1247	1234	28	14
North America endemics	25	54	10	263	261	NA	NA
South America	420	2482	168	2591	2523	50	25
South America endemics	390	1698	156	2120	2059	32	16
Global species	160	690	64	2571	2514	52	26

Supplementary Table 2. Random Forest out-of-the-bag (OOB) percent error rates for all the tested datasets.

	Spatial								Spatial + Morpho	
Continent	**All IUCN listings**	**LC vs non-LC**	**CR vs non-CR**	**All IUCN listings subsampled**	**LC vs non-LC subsampled**	**CR vs non-CR subsampled**	**LC vs non-LC rm 0.9**	**LC vs non-LC rm 0.8**	**LC vs non-LC subsamplng**	**LC vs non-LC downsamplng**
Africa	37.32	19.5	4.22	61.23	19.83	31.02	18.96	16.67	5.66	17.5
Africa endemics	43.11	21.73	5.12	63.84	22.95	34.72	21.27	18.25	10.14	16.46
Asia	21.72	14.97	1.89	66.89	24.62	33.87	13.66	10.64	9.27	29.98
Asia endemics	42.2	27.87	4.14	76.9	36.1	44.1	27.42	24.79	9.27	29.98
Australia	20.33	13.55	0.83	80.45	27.09	30.39	11.38	8.85	9.12	25.19
Australia endemics	44.1	29.23	2.05	85.87	32.49	54.26	27.51	26.34	9.25	28.72
C. America	40.7	27.53	4.39	66.31	27.31	33.14	25.46	23.27	7.18	24.95
C. America endemics	61.46	39.08	7.55	71.9	40.34	39.42	38.86	36.57	4.32	24.78
Europe	20.31	11.4	3.11	65.33	18.25	23.48	10.56	8.55	14.61	48.24
Europe endemics	25.6	10.4	1.6	95.58	52.03	72.8	7.44	4.27	NA	NA
N. America	6.85	4.17	0.55	70.59	25	28.68	2.41	1.46	8.93	25.06
N. America endemics	14.29	8.79	1.83	76.13	36.88	38.34	4.56	4.6	NA	NA
S. America	44.71	34.69	3.22	63.39	25.2	35.98	23.93	21.96	8.69	31.69
S. America endemics	48.94	25.25	3.66	64.6	27.99	37.67	25.14	22.97	5.45	17.73
Global species	18.13	11	1.3	69.7	21.97	24.76	9.92	7.64	11.35	26.54

Supplementary Table 3. OOB error rates per category for the spatial data sets considering all IUCN categories. Full datasets included all species, without any manipulation to avoid Red List category imbalance. Downsampled datasets randomly sampled the majority class(es) to match the value of the minority class.

Data set	CR	EN	LC	NT	VU
Africa full spatial	1	0.5402	0.1296	0.9934	0.4277
Africa downsampled spatial	0.6529	0.6229	0.3749	0.7389	0.6718
Asia full spatial	0.9666	0.9677	0.0279	1	0.8128
Asia downsampled spatial	0.7042	0.7845	0.4279	0.7187	0.7091
Australia full spatial	1	1	0.0245	0.95	0.9166
Australia downsampled spatial	0.8014	0.8974	0.5609	0.8826	0.8803
C. America full spatial	1	0.7472	0.0911	1	0.6323
C. America downsampled spatial	0.631	0.8107	0.3907	0.6597	0.8232
Europe full spatial	0.9655	0.807	0.0172	0.9545	0.7666
Europe downsampled spatial	0.6844	0.7807	0.3068	0.7858	0.7089
N. America full spatial	0.7142	1	0.0008	1	0.9687
N. America downsampled spatial	0.5743	0.8516	0.45	0.7513	0.9018
S. America full spatial	0.9761	0.8126	0.2162	0.8555	0.2935
S. America downsampled spatial	0.6396	0.7254	0.3933	0.6542	0.7571
Africa full endemics spatial	1	0.5296	0.2073	0.9747	0.3862
Africa downsampled endemic spatial	0.684	0.639	0.429	0.759	0.6811
Asia full endemic spatial	0.9166	0.98	0.0931	1	0.7803
Asia downsampled endemic spatial	0.7318	0.8172	0.7224	0.802	0.7718
Australia full endemic spatial	1	1	0.1217	0.92	0.8125
Australia downsampled endemic spatial	0.8789	0.9146	0.7339	0.855	0.9106
C. America full endemic spatial	1	0.725	0.3716	1	0.5535
C. America downsampled endemic spatial	0.6633	0.8237	0.5525	0.6963	0.8591
Europe full endemic spatial	1	1	0.0329	0.7619	1
Europe downsampled endemic spatial	0.9984	0.933	0.9468	0.9286	0.9718
N. America full endemic spatial	0.6	1	0	0.9333	1
N. America downsampled endemic spatial	0.5916	0.8848	0.6143	0.8429	0.8727
S. America full endemic spatial	0.9871	0.7963	0.3582	0.847	0.2399
S. America downsampled endemic spatial	0.6416	0.7253	0.4484	0.6653	0.7494
Global full spatial	1	0.9537	0.0121	0.9769	0.9024
Global downsampled spatial	0.7443	0.8459	0.4087	0.7032	0.7857
mean all data sets	0.8249	0.8305	0.3155	0.8537	0.7516
mean full data sets	0.9461	0.8392	0.0725	0.9648	0.6883
mean downsampled data sets	0.6697	0.7819	0.4149	0.7416	0.7789
mean full endemics data sets	0.8794	0.8780	0.1748	0.8774	0.7331
mean downsampled endemics data sets	0.7414	0.8197	0.6353	0.7927	0.8309

Supplementary Table 4. OOB error rates per category for the spatial data sets using LC and non-LC as the response categories. Full datasets included all species, without any manipulation to avoid Red List category imbalance. Downsampled datasets randomly sampled the majority class(es) to match the value of the minority class.

Data set	**LC**	**noLC**
Africa full spatial	0.1873	0.2041
Africa downsampled spatial	0.2136	0.183
Africa full spatial rm at 0.9	0.1798	0.201
Africa full spatial rm at 0.8	0.16	0.1745
Asia full spatial	0.0357	0.768
Asia downsampled spatial	0.2781	0.2143
Asia full spatial rm at 0.9	0.0329	0.7479
Asia full spatial rm at 0.8	0.0252	0.7046
Australia full spatial	0.0397	0.7922
Australia downsampled spatial	0.2653	0.2765
Australia full spatial rm at 0.9	0.0379	0.7704
Australia full spatial rm at 0.8	0.0284	0.75
C. America full spatial	0.1851	0.4179
C. America downsampled spatial	0.2992	0.247
C. America full spatial rm at 0.9	0.1766	0.38
C. America full spatial rm at 0.8	0.1641	0.3458
Europe full spatial	0.0354	0.5547
Europe downsampled spatial	0.2053	0.1598
Europe full spatial rm at 0.9	0.0366	0.5182
Europe full spatial rm at 0.8	0.0293	0.4705
N. America full spatial	0.0041	0.7868
N. America downsampled spatial	0.2496	0.2504
N. America full spatial rm at 0.9	0.0041	0.6578
N. America full spatial rm at 0.8	0.0033	0.56
S. America full spatial	0.2933	0.2048
S. America downsampled spatial	0.2707	0.2332
S. America full spatial rm at 0.9	0.284	0.1986
S. America full spatial rm at 0.8	0.2613	0.1849
Global full spatial	0.0189	0.7072
Global downsampled spatial	0.2407	0.1986
Global full spatial rm at 0.9	0.0181	0.6925
Global full spatial rm at 0.8	0.0163	0.6054
mean all data sets	0.1337	0.4425
mean full data sets	0.1115	0.5326
mean downsampled data sets	0.2545	0.2235
mean rm at 0.9 data sets	0.1074	0.4963
mean rm at 0.8 data sets	0.0959	0.4558

Supplementary Table 4 (cont).

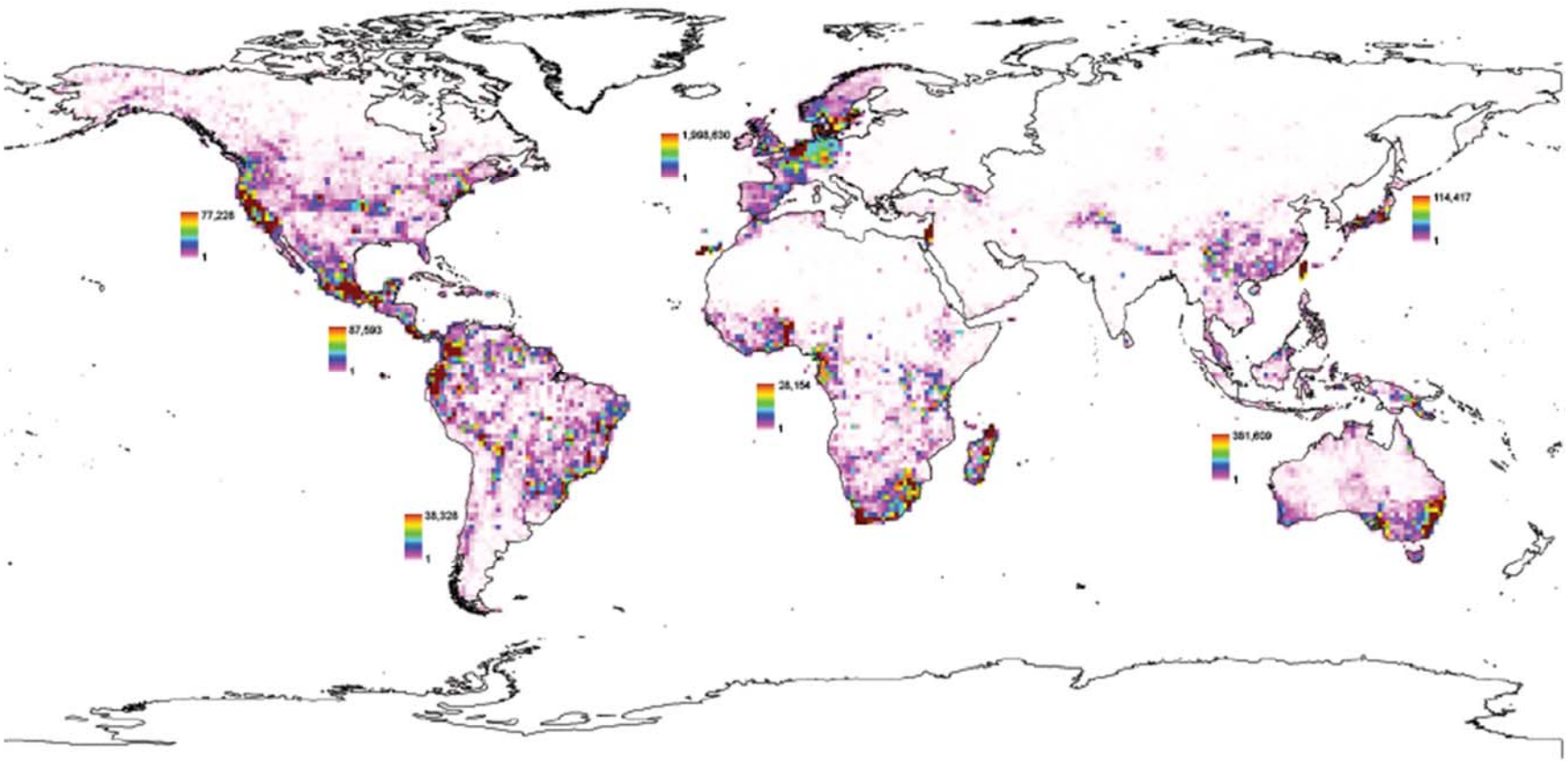
Data set	LC	noLC
Africa full endemic spatial	0.2897	0.1604
Africa downsampled endemic spatial	0.2444	0.2146
Africa full endemic spatial rm at 0.9	0.2798	0.1602
Africa full endemic spatial rm at 0.8	0.2495	0.1324
Asia full endemic spatial	0.1066	0.631
Asia downsampled endemic spatial	0.3435	0.3785
Asia full endemic spatial rm at 0.9	0.1066	0.6313
Asia full endemic spatial rm at 0.8	0.095	0.6136
Australia full endemic spatial	0.1357	0.6909
Australia downsampled endemic spatial	0.3135	0.3363
Australia full endemic spatial rm at 0.9	0.1428	0.653
Australia full endemic spatial rm at 0.8	0.1357	0.6527
C. America full endemic spatial	0.562	0.2706
C. America downsampled endemic spatial	0.3856	0.4213
C. America full endemic spatial rm at 0.9	0.5733	0.2614
C. America full endemic spatial rm at 0.8	0.5208	0.2626
Europe full endemic spatial	0	1
Europe downsampled endemic spatial	0.5057	0.5348
Europe full endemic spatial rm at 0.9	0	1
Europe full endemic spatial rm at 0.8	0	1
N. America full endemic spatial	0.0203	0.7037
N. America downsampled endemic spatial	0.3331	0.4044
N. America full endemic spatial rm at 0.9	0.0121	0.5294
N. America full endemic spatial rm at 0.8	0.0081	0.6666
S. America full endemic spatial	0.3981	0.156
S. America downsampled endemic spatial	0.285	0.2749
S. America full endemic spatial rm at 0.9	0.398	0.1553
S. America full endemic spatial rm at 0.8	0.3661	0.1444
mean all data sets	0.2433	0.4657
mean full data sets	0.2161	0.5161
mean downsampled data sets	0.3444	0.3664
mean rm at 0.9 data sets	0.2161	0.4844
mean rm at 0.8 data sets	0.1965	0.4960

Supplementary Table 5. OOB error rates per category for the spatial data sets using CR and non-CR as the response categories. Full datasets included all species, without any manipulation to avoid Red List category imbalance. Down-sampled datasets randomly sampled the majority class(es) to match the value of the minority class.

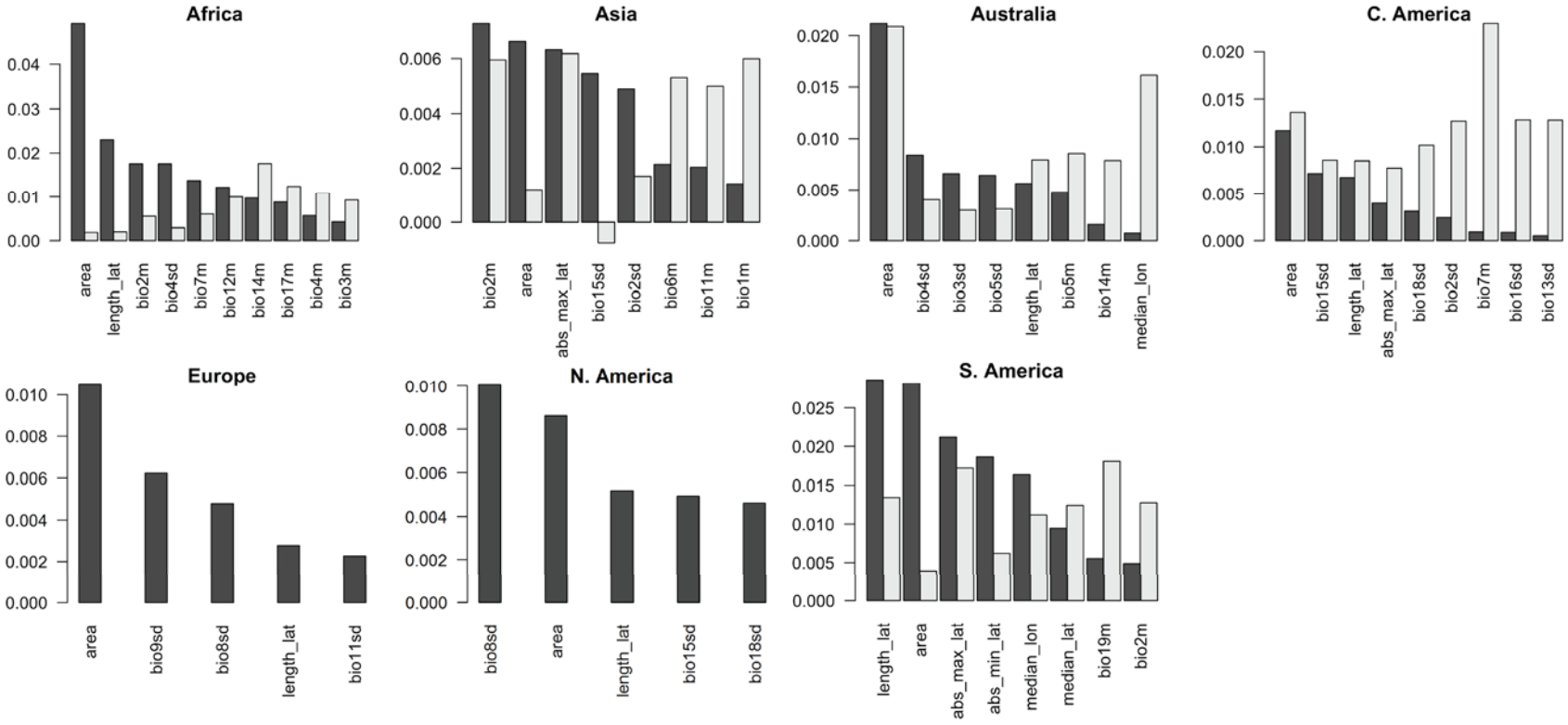
Data set	**CR**	**noCR**
Africa full spatial	1	0.002
Africa downsampled spatial	0.3078	0.3126
Asia full spatial	1	0.0012
Asia downsampled spatial	0.3342	0.3431
Australia full spatial	1	0
Australia downsampled spatial	0.2595	0.3483
C. America full spatial	1	0
C. America downsampled spatial	0.3111	0.3518
Europe full spatial	1	0.001
Europe downsampled spatial	0.2397	0.2299
N. America full spatial	1	0
N. America downsampled spatial	0.2883	0.2852
S. America full spatial	1	0
S. America downsampled spatial	0.3595	0.3601
Africa full endemics spatial	1	0.0021
Africa downsampled endemic spatial	0.3445	0.35
Asia full endemic spatial	0.9583	0.0049
Asia downsampled endemic spatial	0.4385	0.4434
Australia full endemic spatial	1	0
Australia downsampled endemic spatial	0.5067	0.5784
C. America full endemic spatial	1	0.0057
C. America downsampled endemic spatial	0.3685	0.4183
Europe full endemic spatial	1	0
Europe downsampled endemic spatial	0.8076	0.6486
N. America full endemic spatial	1	0
N. America downsampled endemic spatial	0.3817	0.3848
S. America full endemic spatial	1	0
S. America downsampled endemic spatial	0.3827	0.3707
Global full spatial	1	0.0007
Global downsampled spatial	0.2561	0.2391
mean all data sets	0.6848	0.1894
mean full data sets	0.8750	0.0005
mean downsampled data sets	0.3000	0.3187
mean full endemics data sets	0.9940	0.0018
mean downsampled endemics data sets	0.4615	0.4563

Supplementary Table 6. OOB error rates per category for the spatial+morpho data sets using LC and non-LC as response categories. Downsampled datasets randomly sampled the majority class(es) to match the value of the minority class. Resampled datasets randomly sampled the majority class(es) to double that of the minority, in order to increase the minority class count.

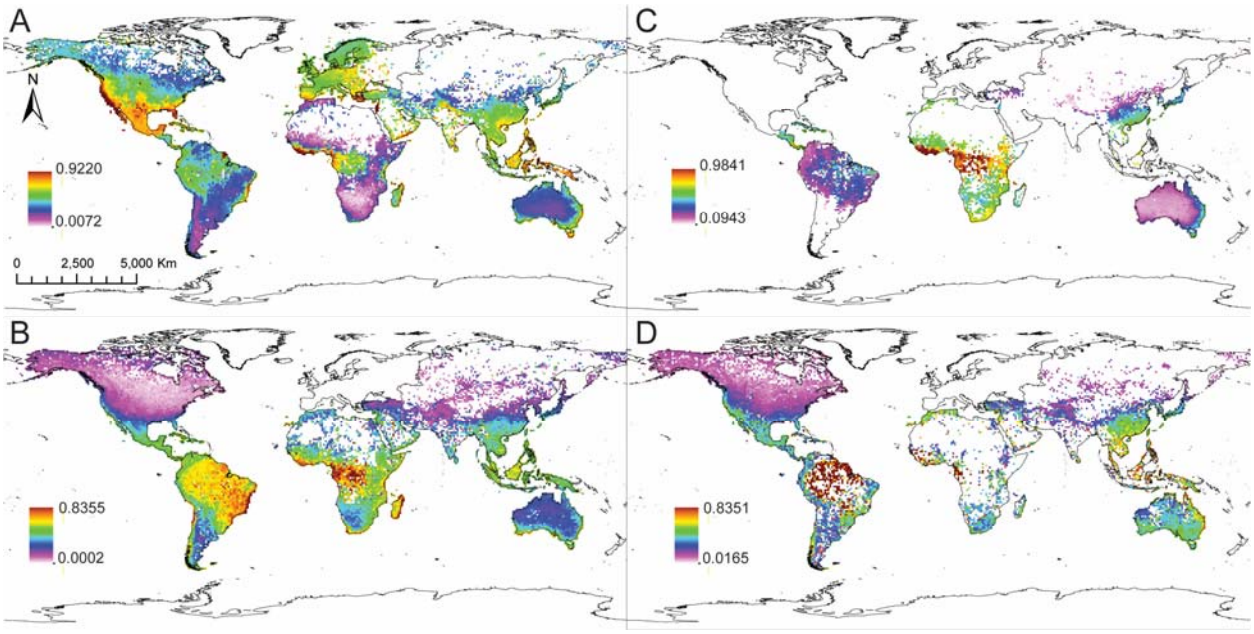
Data set	**LC**	**noLC**
Global downsampled	26.66	26.43
Global resampled	16.14	6.56
Africa downsampled	16.54	18.2
Africa subssampled	7.72	3.61
Africa downsampled endemic	30.59	9.89
Africa resampled endemic	7.98	12.3
Asia downsampled	22.99	36.96
Asia subssampled	12.01	6.52
Asia downsampled endemic	22.99	36.96
Asia resampled endemic	12.01	6.52
Australia downsampled	23.16	27.22
Australia subssampled	12.91	5.33
Australia downsampled endemic	26.38	31.07
Australia resampled endemic	12.11	6.39
C. America downsampled	22.51	27.39
C. America resampled	8.81	5.55
C. America full endemic	17.17	32.39
C. America resampled endemic	3.28	5.36
Europe downsampled	44.72	51.74
Europe subssampled	21.34	7.88
N. America downsampled	23.54	26.57
N. America resampled	14.33	52.11
S. America downsampled	29.28	34.1
S. America resampled	11.03	6.34
S. America downsampled endemic	14.2	23.24
S. America resampled endemic	6.49	4.4
mean all data sets	17.95731	19.655
mean downsampled data sets	26.55143	31.72
mean resampled data sets	11.245	5.22375
mean downsampled endemic data sets	22.266	26.71
mean resampled endemic data sets	8.374	6.994



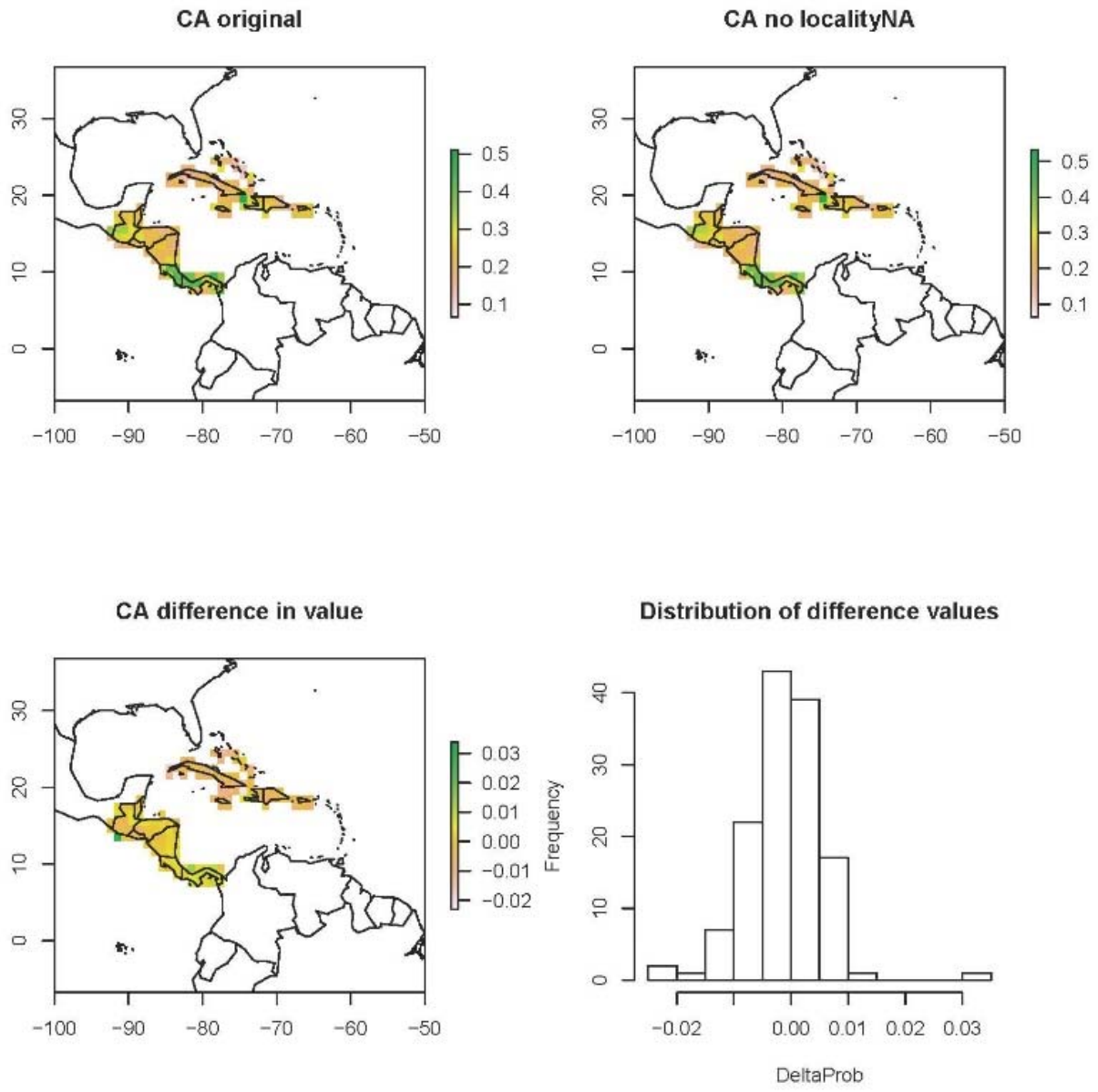
Supplementary Figure 1. Map of all georeferenced localities downloaded from GBIF. The number of localities per grid cell are on a log-scale and are separate for each continent. Only unique localities from each species were counted.



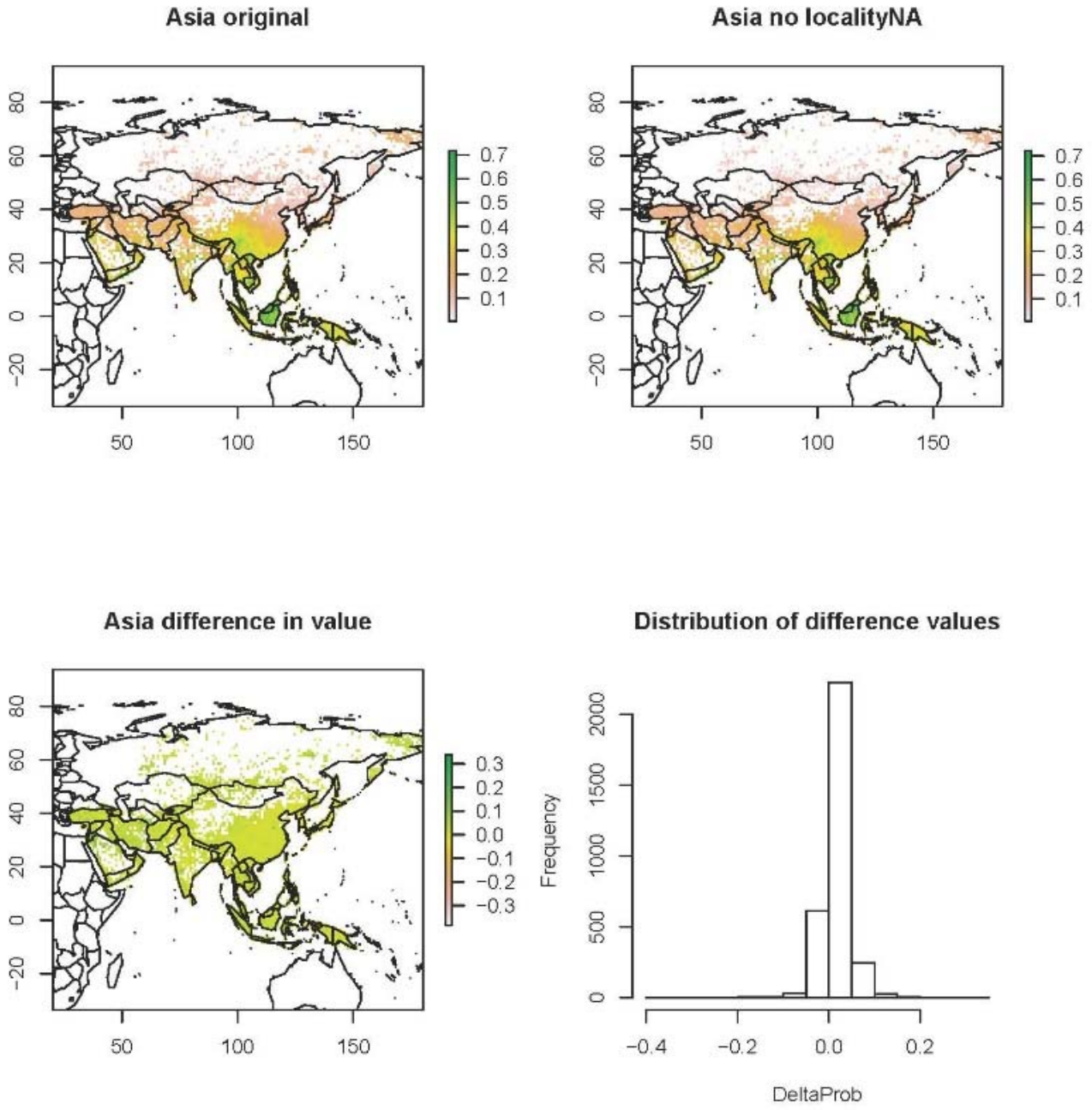
Supplementary Figure 2. Variable importance for the continental endemic datasets ranked by the mean decrease in accuracy. Black bars: 'spatial' dataset. Grey bars: 'spatial+morpho' dataset. Only the top five predictor variables for each model are included for simplicity, and are ordered according to the 'spatial' data.



Supplementary Figure 3. Average per grid-cell probability of being listed as non-LC calculated by the RF classifier using the 'spatial' endemic (A), 'spatial+morpho' endemic (B), 'spatial' global (C), and 'spatial+morpho' global (D) datasets. See scales for values.



Supplementary Figure 4 – Comparison between predictions obtained using the original GBIF dataset and the dataset with all observations without "LOCALITY" data excluded, for the Central American dataset. Top left: original predictions; top right: predictions with the filtered dataset; bottom left: difference in the prediction value per grid cell; bottom right: distribution of difference values. Values indicate per-cell probability of harboring non-LC taxa.



Supplementary Figure 5 – Comparison between predictions obtained using the original GBIF dataset and the dataset with all observations without "LOCALITY" data excluded, for the Asian dataset. Top left: original predictions; top right: predictions with the filtered dataset; bottom left: difference in the prediction value per grid cell; bottom right: distribution of difference values. Values indicate per-cell probability of harboring non-LC taxa.