

Predicting the conservation status of Data Deficient species

Journal:	<i>Conservation Biology</i>
Manuscript ID:	Draft
Wiley - Manuscript type:	Contributed Paper
Date Submitted by the Author:	n/a
Complete List of Authors:	Bland, Lucie; Zoological Society of London, Institute of Zoology; Imperial College London, Division of Biology Collen, Ben; University College London, Department of Genetics, Evolution and Environment Orme, C. David; Imperial College London, Division of Biology Bielby, Jon; Zoological Society of London, Institute of Zoology
Keywords:	Mammals < Animals, Indicators, Predictive modeling, Red lists, Threatened species, Statistics
Abstract:	We have no appreciation of the level of extinction risk faced by a sixth of the 65,000+ species currently on the IUCN Red List. Determining the status of these Data Deficient (DD) species is essential to developing an accurate picture of global biodiversity and protecting potentially threatened DD species. Using terrestrial mammals as our focal taxon, we compared the outcomes of seven Machine Learning (ML) tools in predicting threat for species of known conservation status using taxonomic, life-history, geographical and threat information. ML tools showed very high species classification accuracy (up to 92%) and ability to correctly identify centres of threatened species richness. Applying the best model to DD species, we predict 313 of 493 DD species (64%) to be at risk, increasing the estimated proportion of threatened terrestrial mammals from 22% to 27%. Regions predicted to contain large numbers of threatened DD species are already conservation priorities, but show considerably higher levels of risk than previously recognized. We conclude that unless directly targeted for monitoring, species classified as DD are likely to slide towards extinction unnoticed. Taking into account information on DD species may therefore help tackle data gaps in biodiversity indicators and conserve the earth's poorly-known biodiversity.

1 **Predicting the conservation status of Data Deficient species**

2 **Running head:** Predicting extinction risk

3 **Keywords:** IUCN Red List, Data Deficient, Extinction risk, Machine Learning

4 **Word count:** 5,950 words

5 Lucie M. Bland^{1,2*}, Ben Collen³, C. David L. Orme², and Jon Bielby¹

6 ¹Institute of Zoology, Zoological Society of London, Regent's Park, London

7 NW1 4RY, UK

8 ²Division of Biology, Imperial College London, Silwood Park, Ascot, SL5 7PY, UK

9 ³Centre for Biodiversity and Environmental Research, University College London, Gower

10 Street, London, WC1 E6BT, UK

11 * e-mail: lucie.bland@ioz.ac.uk

12 **Abstract**

13 We have no appreciation of the level of extinction risk faced by a sixth of the 65,000+
14 species assessed by the IUCN Red List. Determining the status of these Data Deficient (DD)
15 species is essential to developing an accurate picture of global biodiversity and identifying
16 potentially threatened DD species. To address this gap in our knowledge, we used
17 predictive models incorporating species' life-history, geography and threat information to
18 predict the conservation status of DD species within terrestrial mammals. We constructed
19 the models using seven Machine Learning (ML) tools trained on species of known status.
20 The resultant models showed very high species classification accuracy (up to 92%) and
21 ability to correctly identify centres of threatened species richness. Applying the best model
22 to DD species, we predict 313 of 493 DD species (64%) to be at risk, increasing the estimated
23 proportion of threatened terrestrial mammals from 22% to 27%. Regions predicted to
24 contain large numbers of threatened DD species are already conservation priorities, but
25 show considerably higher levels of risk than previously recognized. We conclude that unless
26 directly targeted for monitoring, species classified as DD are likely to slide towards
27 extinction unnoticed. Taking into account information on DD species may therefore help
28 tackle data gaps in biodiversity indicators and conserve the earth's poorly-known
29 biodiversity.

30 Introduction

31 In light of global biodiversity change, the 12th target of the Strategic Plan of the Convention
32 on Biological Diversity (CBD) states that by “2020 the extinction of known threatened
33 species has been prevented” (Convention on Biological Diversity 2010). Understanding the
34 level of extinction risk faced by different species, and why interspecific differences in risk
35 arise are therefore some of the greatest challenges facing conservation biology. Assessment
36 frameworks for threatened species are crucial to identifying risk and monitoring progress
37 towards CBD targets (Jones et al. 2011), and one of the most widely used is the International
38 Union for Conservation of Nature (IUCN) Red List (IUCN 2001; Butchart et al. 2010).

39
40 There has been much improvement in the taxonomic coverage of the Red List over recent
41 years, resulting in a more comprehensive understanding of species’ extinction risk (Collen &
42 Bailie 2010; Böhm et al. 2013). However, a sixth of the 65,000+ species assessed by the IUCN
43 are classified as Data Deficient (DD) due to a lack of information on taxonomy, geographic
44 distribution, population status or threats (IUCN 2010). To date 15% of mammals (Schipper et
45 al. 2008), 25% of amphibians (Stuart et al. 2004), 19% of reptiles (Böhm et al. 2013) and 49%
46 of freshwater crabs (Cumberlidge et al. 2009) are classified as DD. Uncertainty within many
47 groups about the true level of extinction risk of DD species considerably influences our
48 understanding of patterns of threat and risk (Butchart & Bird 2010; Bland et al. 2012), as the
49 distribution of DD species is often taxonomically and spatially biased (Bielby et al. 2006;
50 Bland et al. 2012). For example, 25% of data-sufficient mammals are threatened with
51 extinction, but estimates range from 21% if all DD species were non-threatened to 36% if all
52 DD species were threatened (Hilton-Taylor et al. 2009). In addition, genuinely threatened

53 DD species may be neglected by conservation programmes due to their uncertain extinction
54 risk status.

55

56 Determining the true conservation status of DD species is essential in developing an
57 accurate picture of global biodiversity and enabling the protection of threatened species.

58 Re-assessment of the 10,673 species currently classified as DD to a data-sufficient category
59 could be achieved through focused field surveys, but the prospect of this occurring is
60 unlikely given the monetary and time costs of biodiversity surveys (Balmford & Gaston
61 1999) and current levels of investment in IUCN Red List assessments (Stuart et al. 2010).

62 However, large amounts of life-history, ecological and phylogenetic information are
63 available for DD species. The distribution of many DD species is known, allowing inference of
64 species' geographical range size, environmental niche and exposure to anthropogenic
65 threats. These data alone are insufficient for making a decision on formal Red List status,
66 but could be used to help inform global estimates of risk. Comparative studies of extinction
67 risk based on species trait data have previously yielded insight into the determinants of risk
68 across taxa (Purvis 2008; Cardillo & Meijaard 2012), and could enable the preliminary re-
69 assessment of DD species.

70

71 Comparative datasets frequently contain many variables, with non-linearities, complex
72 interactions and missing values (Cutler et al. 2007), and as such traditional statistical
73 methods may lack predictive ability. Machine Learning (ML) methods, derived from the
74 artificial intelligence literature, are flexible and powerful tools for finding patterns in
75 datasets (Webb 2002; Hastie et al. 2009). They rely on few assumptions and can utilize large
76 amounts of data, which has made them increasingly popular with ecologists (Prasad et al.

77 2006; Ozesmi et al. 2006; Cutler et al. 2007; Olden et al. 2008). A wide range of ML
78 algorithms are available, and their relative predictive performance depends on the study
79 objectives and available data (No Free Lunch Theorem: see Webb 2002 and Hastie et al.
80 2009). A series of comparisons have been made to identify the strengths and weaknesses of
81 different ML algorithms for ecological applications (Elith & Graham 2009; Kampichler et al.
82 2010; Keller et al. 2011), but only tree-based ML methods have been applied to threatened
83 species classification (Jones et al. 2006; Boyer 2008; Davidson et al. 2009, 2012). The
84 outputs of ML algorithms are probability estimates of a given outcome, which allow easy
85 interpretation of levels of certainty in predicting complex processes such as extinction risk.
86 As a result of these properties, ML algorithms represent a robust approach to identifying the
87 complex pathways leading to observed patterns of extinction risk, and deriving rules-of-
88 thumb to predict the true level of risk of DD species.

89

90 Here we investigate the performance of ML algorithms in predicting extinction risk and in
91 estimating the prevalence of risk in DD terrestrial mammals. Terrestrial mammals are a well-
92 suited model taxon for the purposes of our study: they contain a high proportion of species
93 of known conservation status (85%) and previous studies (Purvis et al. 2000; Cardillo et al.
94 2005, 2008; Davidson et al. 2009) provide a benchmark against which to measure
95 improvement in predictive accuracy. There is also a high amount of data available on the
96 biology of the clade, even for Data Deficient species. We predict extinction risk from data on
97 a range of intrinsic factors, including species' life history and ecology, and extrinsic factors,
98 including environmental data and measures of threat intensity. Specifically, we address the
99 following questions:

- 100 1) What are the relative powers of seven different ML methods (classification trees,
101 random forests, boosted trees, k-nearest neighbours, support vector machines, neural
102 networks and decision stumps) to predict extinction risk in terrestrial mammals?
- 103 2) How accurately can those methods predict current geographical patterns of extinction
104 risk?
- 105 3) Using the models obtained, what is the predicted level of extinction risk faced by DD
106 species?
- 107 4) How do our findings change current geographical patterns of extinction risk for
108 terrestrial mammals?

109

110 **Methods**

111 **Dataset**

112 We collated a database for 4,461 terrestrial mammal species with threat status classified as
113 non-threatened (LC, NT), threatened (VU, EN, CR) and Data Deficient (DD) (IUCN 2008). We
114 treated species as threatened or non-threatened, as highly imbalanced categories (2,826 LC
115 species versus 157 CR species) are difficult to discriminate using predictive models (Webb
116 2002) and uncertainty around classifications with multiple categories is difficult to interpret
117 and communicate. In contrast, machine learning predictions from our binary classification
118 provide a simple quantification of both the likely probability of threatened status for each
119 species and the level of uncertainty around that prediction.

120

121 For each species, we collated the following life-history traits (IUCN 2008; Jones et al. 2009):
122 body mass, litter size, habitat breadth, trophic level and number of IUCN-listed habitats.

123 Each trait was available for at least 60% of species. Since some ML methods require

124 complete data, missing data was either phylogenetically imputed (Fritz et al. 2009;
125 Bruggeman et al. 2009), or assigned the genus or family median for species missing from the
126 phylogeny. We used species' range maps to determine geographical range size (IUCN 2010),
127 the latitude of range centroid (IUCN 2010), and extract summary statistics within ranges for
128 a range of global variables: annual mean and seasonality of temperature and precipitation
129 (Hijmans et al. 2005); minimum and range of elevation (Hijmans et al. 2005); mean and
130 minimum human population density for the year 2000 (CIESIN 2005a); and averages for
131 each of Net Primary Productivity (NPP) (Imhoff et al. 2004), Human Footprint (CIESIN
132 2005b), GDP for the year 1990 (CIESIN 2002) and human appropriation of NPP (Imhoff et al.
133 2004). Finally, we recorded biogeographical distribution (IUCN 2010), External Threat Index
134 (Cardillo et al. 2004) and habitat suitability (Rondinini et al. 2011) for each species. See
135 Appendix S1 for details. Previous studies have reached inconsistent conclusions about the
136 primary traits explaining variation in extinction risk across species (Cardillo & Meijaard
137 2012). In addition, uninformative explanatory variables are unlikely to affect predictive
138 performance in problems with fewer variables than species (Webb 2002; Kuhn 2008). We
139 therefore do not undertake variable selection, but instead focus on using all available traits
140 implicated in determining extinction risk to make the best predictions.

141

142 **Training of Machine Learning tools**

143 Six ML tools were used to model risk status across all variables: classification trees, random
144 forests, boosted trees, k-nearest neighbours, support vector machines and neural networks.
145 We also computed decision stumps using geographical range size alone to assess the
146 predictive power of that variable. We developed models for all mammals and separately for
147 rodents, bats, primates and carnivores to explore the taxonomic transferability of ML

148 predictive accuracy. ML tools cannot currently take into account phylogenetic relatedness
149 between species, so we included taxonomic order, family and genus in all models to
150 partially account for shared evolutionary history. For each taxonomic dataset, we removed
151 highly correlated ($r=0.9$) and low variance variables, which can lead to colinearity and zero
152 variance in cross-validation partitions. All numeric predictors were centred and scaled to a
153 standard normal distribution before analysis (Kuhn 2008).

154

155 We set aside DD species and, within each taxonomic group, divided the remaining species
156 into a 25% validation set and 75% training set. For each ML method, we used ten-fold cross-
157 validation on the 75% training set to optimize model tuning parameters by maximizing the
158 Area Under the Receiver Operating Characteristic Curve (AUROC), which is insensitive to
159 class imbalance and does not require the specification of misclassification costs (Fawcett
160 2006). The best ML tool for each dataset for predicting threatened and non-threatened
161 status was then found by comparing AUROC values of various tuned models on the 25%
162 validation set.

163

164 In all models, we used Youden's index (Youden 1950) to identify a probability threshold
165 above which species are identified as threatened. This lends equal weight to detecting
166 threatened and non-threatened species, which does not reflect the true prevalence of
167 threat but is reasonable given the importance of identifying threatened species (Vié et al.
168 2009). All analyses were conducted in R version 2.14.1, using the *caret* package (Kuhn 2008)
169 to optimize model parameters. For further details see Appendix S1.

170

171 **Spatial analysis of predictions**

172 Using species' range maps (IUCN 2010), we then computed the observed and predicted
173 proportion of threatened species from the 991 species in the 25% validation set across a
174 global grid of 4,505 equal-area hexagons. We fitted a linear regression across cells of
175 observed threat as a function of predicted threat, cell species richness and average range
176 size of species, excluding cells with fewer than 10 species. We also fitted simultaneous
177 autoregressive models to account for spatial autocorrelation (Appendix S1). We produced
178 maps in ArcGIS 9.3 and conducted all analyses in R version 2.14.1.

179

180 **Predictions for Data Deficient species**

181 We predicted the status of 493 DD species from the best performing global model, using the
182 same threshold as for the validation dataset (Appendix S2) and tabulated the number of DD
183 species predicted to be threatened and non-threatened in 6,593 hexagons. We then
184 compared the proportion of threatened species in cells with and without incorporating our
185 predictions for DD species. Finally, we used linear regression and spatial autoregressive
186 models of observed threat as a function of predicted threat to test for a regression slope
187 different from one.

188

189 **Results**

190 **Comparison of Machine Learning tools and taxonomic levels**

191 Area Under Receiver Operator Characteristic Curve (AUROC) for best models ranged
192 between 0.873 and 0.961 (Table 1), indicating that ML tools calibrated on species-specific
193 information can accurately predict species threat. The best model for the global dataset
194 identified correctly 93.5% of threatened species and 88.7% of non-threatened species

195 (Appendix S1). There were significant differences in performance across tools (Friedman
196 test, $\chi^2=18.3$, $p=0.005$, $df=6$). *Post hoc* symmetry tests showed that this difference was
197 caused by the lack of power of decision stumps based on geographical range size alone,
198 compared to boosted trees ($p=0.05$, $df=1$), neural networks ($p=0.05$, $df=1$) and support
199 vector machines ($p=0.05$, $df=1$). Predictions from the global model for individual orders
200 achieved higher AUROC than predictions from the order-specific models (Appendix S1),
201 indicating that predictions are more reliable when information from all mammals is taken
202 into account.

203

204 **Spatial predictions**

205 Observed and predicted proportions of threatened species in assemblages of the validation
206 set were broadly consistent (Fig. 1), indicating that ML tools can correctly predict
207 macroecological patterns of extinction risk. In both ordinary least squares (OLS) and spatial
208 regression (SAR) models, we found a strong positive association between predicted
209 assemblage threat on observed assemblage threat (OLS: slope=0.592, $p<0.0001$, $t_{1,4501} =$
210 79.03, AIC= -18182; SAR: slope= 0.596, $p<0.0001$, $t_{1,4499}=5.457$, AIC= -19050). The
211 relationship is mediated by a significant interaction with assemblage species richness in
212 both OLS and SAR models (OLS: slope=0.066, $p\text{-value}<0.001$, $t_{1,4501} = 3.865$; SAR:
213 slope=0.096, $p\text{-value}<0.0001$, $t_{1,4499} = 5.448$), with model fit improving with larger
214 assemblage size (Appendix S1). Mean assemblage risk was globally over-predicted
215 (observed: 36.8%, predicted: 46.7%), mirroring over-predictions at the species level
216 (observed: 22.1%, predicted: 26.7%).

217

218 **Predictions for Data Deficient species**

219 Our model outputs predict 313 of 493 DD species to be threatened with extinction, implying
220 that underlying risk levels are much greater in DD species (63.5%) than data-sufficient
221 species (22.1%) (Appendix S2). The spatial congruence between threat hotspots identified
222 using only data-sufficient species and hotspots incorporating our DD species predictions was
223 very high (Spearman rank correlation= 0.987, $p < 0.001$; Fig. 2 and 3). Additionally, the levels
224 of threat in centres of threatened species richness may previously have been
225 underestimated according to our regression model of observed vs. predicted threat (testing
226 for slope \neq 1: OLS: slope=1.036, $p < 0.0001$, $F_{1,6591}=242.96$; SAR: slope= 1.043, $p < 0.0001$,
227 $\chi^2_{1,6589}=214.15$).

228

229 Discussion

230 We have no appreciation of the true level of extinction risk faced by one in six species on
231 the IUCN Red List. These Data Deficient species are of great conservation concern, as they
232 contribute to considerable uncertainty in estimates of risk (Butchart & Bird 2010; Bland et
233 al. 2012) and are neglected by conservation programmes due to their uncertain status.
234 Accurate predictive models of risk based on species traits could therefore enhance our
235 understanding of risk patterns, and enable the proactive conservation of threatened Data
236 Deficient species.

237 Predictions for Data Deficient species

238 We predict 313 of 493 (63.5%) DD species are threatened with extinction (Appendix S2). A
239 previous random forests model (Davidson et al. 2009) predicted only 28 of 341 (8.2%) DD
240 terrestrial mammals to be at risk, perhaps reflecting the low sensitivity of the model to
241 detection of threatened species (sensitivity of 47.7% compared to 93.5% in our best model).

242 A recently published prediction of species extinction risk using eigenvector methods
243 predicted 35% of 481 DD species to be at risk (Jones & Safi 2011), but the ability of the
244 method to integrate phylogenetic signal has been questioned (Freckleton et al. 2011). Our
245 estimates are considerably larger, increasing the estimated proportion of threatened
246 terrestrial mammals from 22% to 27% globally.

247

248 Despite this apparent increase in risk, spatial distribution of predicted risk suggests that
249 global spatial prioritization based on current knowledge is robust to uncertainty. Large
250 model residuals (Fig. 2) were caused by the predicted threatened status of a few wide-
251 ranging DD species, such as the northwestern Australian marsupial mole *Notoryctes*
252 *caurinus*. Our findings echo those of Joppa *et al.* (Joppa et al. 2011), who found that regions
253 predicted to contain large numbers of undiscovered plant species are already conservation
254 priorities, but show considerably higher levels of species risk than previously acknowledged.
255 Additionally, areas containing DD species have been shown to contain more recently
256 described amphibian species than expected by chance (Brito 2010), suggesting that these
257 sites might hold many undescribed species (Bini et al. 2006). A better understanding of the
258 likely status of DD species may therefore provide an efficient method for targeting surveys,
259 as well as incorporating the world's poorly-known and undescribed species in conservation
260 planning.

261

262 Our results suggest that DD species are of great conservation concern. DD species have
263 smaller ranges (median=9,891 km²) than their data-sufficient counterparts (median=
264 1,666,107 km²), which contributes to their high extinction risk. Maps of DD species ranges
265 may be uncertain and underestimated when collection effort is low. Nonetheless, the data

266 suggest that many DD species are likely to be range-restricted and that geographical
267 measures derived from the species' range maps are broadly representative of the species'
268 environment. We make the best use of the information available for each species, and note
269 that risk predictions for individual DD species should be interpreted in the context of their
270 IUCN Red List documentation. Since 2008, two DD mammal species (pale fox *Vulpes pallida*
271 and long-nosed mosaic-tailed rat *Paramelomys levipes*) have been re-assigned as least
272 concern; both were predicted not to be at risk by our model. These cases, along with the
273 high consistency between predicted probability of threat and Red List category in our
274 validation set (Appendix S1), indicate that DD species that are assigned a high probability of
275 threat are likely to be at imminent risk of extinction.

276

277 Many Data Deficient mammals are nocturnal, and most are bats and rodents (75%), which
278 are difficult to observe and identify in the field without expert knowledge. Worryingly,
279 nearly 40% of DD species are only known from few specimens, old records or from unknown
280 provenance (Appendix S1), indicating a severe lack of knowledge of mammalian diversity.
281 Predicted threat levels in those very-poorly known species are particularly high (79.6%),
282 compared to species classified as DD due to unknown population trends and threats (51.2%)
283 or uncertain taxonomic status and new discoveries (61.7%). High rate of species
284 rediscoveries indicate that many species missing for long periods of time remain extant
285 (particularly those that are only known from type specimens (Scheffers et al. 2011)), but
286 show considerably higher levels of threat than other species (Scheffers et al. 2011). We may
287 therefore expect very poorly-known DD species to be extant, but on the brink of extinction.

288

289 Ninety-one species listed as DD in the 1996 IUCN Red List assessment were assigned to a
290 data-sufficient category in 2008 (Collen et al. 2011), including 31 (34%) as threatened. We
291 predict 53 out of 90 species (59%) listed as DD in both the 1996 and 2008 IUCN Red Lists to
292 be at risk of extinction. This suggests that species already re-assigned to a data-sufficient
293 category are more abundant and widespread than species still listed as DD on the 2008 Red
294 List. Hence, we expect threatened DD species to be the last species to be assigned their true
295 conservation status in future iterations of the Red List. This finding highlights the
296 importance of prioritizing potentially threatened DD species for monitoring and re-
297 assessment. Collection of life-history and distribution information is especially urgent for
298 the 174 DD species excluded from our analysis due to insufficient data.

299

300 **Comparison of Machine Learning tools and taxonomic levels**

301 For all mammals and within the orders analysed, ML tools achieved very clear discrimination
302 between threatened and non-threatened species in the independent validation sets.
303 Classification trees and k-nearest neighbours are conceptually simpler and computationally less
304 intensive than other tools, and never achieved highest classification performance. Random forests,
305 boosted trees, support vector machines and neural networks performed particularly well,
306 and we recommend them as powerful methods for predicting species extinction risk. Why
307 tools differ in predictive performance depends on the link between the algorithm, fitted
308 functions and data distribution, which can be investigated by simulating data (see Elith &
309 Graham (2009) for an example in species distribution modelling). In addition, studies
310 focusing on explaining the role of underlying risk drivers rather than risk prediction could
311 undertake variable selection and model simplification.

312

313 Whether one or all of the recommended methods should be applied to a given situation of
314 extinction risk prediction depends on available computational resources. We believe that
315 even small increases in performance achieved by using multiple techniques justify their
316 combined use, given the importance of accurately predicting species conservation status.
317 Geographical range size alone provided reasonable discriminatory power in decision
318 stumps, as expected from its role in categorising species under IUCN criterion B (Purvis et al.
319 2000). However, the high AUROC observed in models with all explanatory variables
320 included indicates that these extra data are necessary to identify species not listed under
321 criterion B, and to achieve suitable performance for use in conservation decision-making.

322

323 Although comparative studies of extinction risk have been criticized for not providing
324 findings that are applicable across taxa (Cardillo & Meijaard 2012), our results suggest that,
325 at least in mammals, information obtained from a wider range of species improves
326 extinction risk prediction. The additional power provided by including all terrestrial mammal
327 species has important implications for the development of predictive systems for
328 conservation. Transferability of predictive power across taxa, and the trade-off between
329 amount of contextual information and predictive ability should be the focus of future
330 research.

331

332 **Limitations**

333 Although our models achieved high discrimination between threatened and non-threatened
334 species, a number of factors may have negatively affected predictive performance.

335 Discarding species due to the absence of a range map and setting aside 25% of the data as
336 validation reduced the sample size. Our study also lacked a phylogenetic framework, though

337 we took into account taxonomy in our models by including taxonomic levels (order, family
338 and genus) and building four order-level models. However, order-level models achieved
339 lower predictive performance than order-level predictions from the global model (Appendix
340 S1), indicating a modest role of order-specific processes in determining extinction risk.

341

342 Missing and inexact explanatory variables and incomplete characterization of the
343 threatening processes may also have caused misclassifications. For example, Purvis et al.
344 (2000) identified population density as a significant predictor of elevated extinction risk in
345 primates, but were unable to use this variable due to its poor coverage across terrestrial
346 mammals. Analyses based on species' geographic range maps have been criticized as
347 species are not evenly distributed across their range, and because some habitats may be
348 unsuitable or inaccessible for species (Rondinini et al. 2006). Making use of more refined
349 maps of species range, such as those derived from habitat suitability modelling (Rondinini et
350 al. 2011), may shed light on how higher resolution range data inform extinction risk
351 prediction. Anthropogenic threat impacts included in the model were mainly based on
352 properties of the human population in the area, e.g. human population density and gross
353 domestic product. Due to the limited characterization of threatening processes, our models
354 are less likely to identify species threatened by over-exploitation and invasive species than
355 those affected by habitat loss.

356

357 Finally, model misclassifications may indicate latent potential for recovery or threat and may
358 be used to inform future species assessments. Three of the 15 species incorrectly classified
359 as non-threatened by our models (*Proechimys roberti*, *Reithrodontomys microdon* and
360 *Scotonycteris ophiodon*) were down-listed to a non-threatened category in 2010.

361

362 **Conclusions**

363 Data Deficient species should be of high conservation interest: they bias our understanding
364 of patterns of extinction risk (Butchart & Bird 2010; Bland et al. 2012) and are neglected by
365 conservation programmes due to their uncertain status. Resolution of taxonomic
366 uncertainty and extensive field surveys are unlikely prospects for all 10,673 species currently
367 listed as DD on the IUCN Red List, given monetary and time costs of surveys (Balmford &
368 Gaston 1999) and risk assessments (Stuart et al. 2010). Predicting species extinction risk
369 from contextual information could be a rapid and inexpensive approach for prioritizing taxa
370 and geographical regions under limited knowledge. ML methods are extremely powerful
371 tools for statistical pattern recognition, which can readily incorporate decision-makers' risk
372 attitudes and quantify prediction uncertainty. As such, they show great potential for
373 predictive conservation science under increasing availability of biodiversity data. The seven
374 ML tools used across two taxonomic levels of terrestrial mammals accurately predicted
375 species extinction risk and centres of threatened species richness. Data Deficient mammal
376 species are likely to be disproportionately at risk, and unless directly targeted for
377 conservation action may slide towards extinction unnoticed. Although our study leaves
378 global mammalian conservation priorities generally unaffected, we conclude risk levels in
379 terrestrial mammals are likely to have been considerably underestimated. Predicting the
380 conservation status of DD species can reduce uncertainty in global patterns of threat, and
381 enable the transparent prioritization for field surveys of potentially threatened DD species.
382 Such an approach could be particularly cost-effective for taxa containing large numbers of
383 DD species, such as invertebrates (Samways & Böhm 2010). Finally, DD species may be
384 indicative of spatial knowledge deficiency and could inform species inventories. Taking into

385 account information on DD species may therefore help tackle data gaps in biodiversity
386 indicators, as well as conserve the earth's poorly-known biodiversity.

387 **Acknowledgements**

388 We thank J. Bruggeman for his assistance with the PhyloPARS programme. We thank E.J.
389 Milner-Gulland, R. Ewers, G. Mace, E. Nicholson, M. McCarthy, and S. Canessa for thoughtful
390 comments on the manuscript. B.C. is partly supported by a grant from the Rufford
391 Foundation.

392 **Supporting Information**

393 Supplementary methods, tables and figures (Appendix S1) and predicted conservation status of
394 Data Deficient terrestrial mammals (Appendix S2) are available online. The authors are solely
395 responsible for the content and functionality of these materials. Queries (other than absence of
396 the material) should be directed to the corresponding author.

397

398 **Literature cited**

- 399 Balmford, A., and K. J. Gaston. 1999. Why biodiversity surveys are good value. *Nature*
400 **398**:204–205. Macmillan Magazines Ltd.
- 401 Bielby, J., A. A. Cunningham, and A. Purvis. 2006. Taxonomic selectivity in amphibians:
402 ignorance, geography or biology? *Animal Conservation* **9**:135–143. Blackwell Publishing
403 Ltd.
- 404 Bini, L. M., J. A. F. Diniz-Filho, T. F. L. V. B. Rangel, R. P. Bastos, and M. P. Pinto. 2006.
405 Challenging Wallacean and Linnean shortfalls: knowledge gradients and conservation
406 planning in a biodiversity hotspot. *Diversity and Distributions* **12**:475–482.
- 407 Bland, L. M., B. Collen, C. D. L. Orme, and J. Bielby. 2012. Data uncertainty and the selectivity
408 of extinction risk in freshwater invertebrates. *Diversity and Distributions* **18**:1211–
409 1220.

- 410 Böhm, M. et al. 2013. The conservation status of the world's reptiles. *Biological*
411 *Conservation* **157**:372–385.
- 412 Boyer, A. G. 2008. Extinction patterns in the avifauna of the Hawaiian islands. *Diversity and*
413 *Distributions* **14**:509–517.
- 414 Brito, D. 2010. Overcoming the Linnean shortfall: Data deficiency and biological survey
415 priorities. *Basic and Applied Ecology* **11**:709–713.
- 416 Bruggeman, J., J. Heringa, and B. W. Brandt. 2009. PhyloPars: estimation of missing
417 parameter values using phylogeny. *Nucleic Acids Research* **37**:179–184.
- 418 Brummitt, N., S. P. Bachman, and J. Moat. 2008. Applications of the IUCN Red List: towards a
419 global barometer for plant diversity. *Endangered Species Research* **6**:127–135.
- 420 Butchart, S. H. M., and J. P. Bird. 2010. Data Deficient birds on the IUCN Red List: What don't
421 we know and why does it matter? *Biological Conservation* **143**:239–247.
- 422 Butchart, S. H. M., A. J. Stattersfield, J. Baillie, L. A. Bennun, S. N. Stuart, H. R. Akçakaya, C.
423 Hilton-Taylor, and G. M. Mace. 2005. Using Red List Indices to measure progress
424 towards the 2010 target and beyond. *Philosophical Transactions of the Royal Society B:*
425 *Biological Sciences* **360**:255–268.
- 426 Cardillo, M., G. M. Mace, J. L. Gittleman, K. E. Jones, J. Bielby, and A. Purvis. 2008. The
427 predictability of extinction: biological and external correlates of decline in mammals.
428 *Proceedings of the Royal Society B: Biological Sciences* **275**:1441–8.
- 429 Cardillo, M., G. M. Mace, K. E. Jones, J. Bieby, O. R. P. Bininda-Emonds, W. Sechrest, C. D.
430 L. Orme, and A. Purvis. 2005. Multiple causes of high extinction risk in large mammal
431 species. *Science* **309**:1239–1241.
- 432 Cardillo, M., and E. Meijaard. 2012. Are comparative studies of extinction risk useful for
433 conservation? *Trends in ecology & evolution* **27**:167–171. Elsevier Ltd.
- 434 Cardillo, M., A. Purvis, W. Sechrest, J. L. Gittleman, J. Bieby, and G. M. Mace. 2004. Human
435 population density and extinction risk in the world's carnivores. *PLoS Biology* **2**:909–
436 914.
- 437 CIESIN. 2002. Country-level Population and Downscaled Projections based on the B2
438 Scenario (1990). Palisades, NY. Retrieved from
439 <http://www.ciesin.columbia.edu/datasets/downscaled>.
- 440 CIESIN. 2005a. Gridded Population of the World (2000), Version 3 (GPWv3). Socioeconomic
441 Data and Applications Center (SEDAC), Columbia University, Palisades, NY. Retrieved
442 from <http://sedac.ciesin.columbia.edu/gpw>.
- 443 CIESIN. 2005b. Last of the Wild Data Version 2 (LWP-2): Global Human Footprint dataset
444 (HF).

- 445 Collen, B., and J. M. Bailie. 2010. The barometer of life: sampling. *Science* **329**:140.
- 446 Collen, B., S. T. Turvey, C. Waterman, H. M. R. Meredith, T. S. Kuhn, J. E. M. Baillie, and N. J.
447 B. Insaac. 2011. Investing in evolutionary history: implementing a phylogenetic
448 approach for mammal conservation. *Philosophical Transactions of the Royal Society B:*
449 *Biological Sciences* **366**:2611–2622.
- 450 Convention on Biological Diversity. 2010. TARGET 12 - Technical Rationale. COP 10 Decisions
451 Tenth meeting of the Conference of the Parties to the Convention on Biological
452 Diversity. CBD, Nagoya, Japan.
- 453 Cumberlidge, N., P. K. L. Ng, D. C. J. Yeo, C. Magalhães, M. R. Campos, F. Alvarez, T. Naruse,
454 S. R. Daniels, L. J. Esser, and F. Y. K. Attipoe. 2009. Freshwater crabs and the
455 biodiversity crisis: importance, threats, status, and conservation challenges. *Biological*
456 *Conservation* **142**:1665–1673.
- 457 Cutler, R. D., T. C. Edwards, K. H. Beard, A. Cutler, K. T. Hess, J. Gibson, and J. J. Lawler. 2007.
458 Random forests for classification in ecology. *Ecology* **88**:2783–92.
- 459 Davidson, A. D., A. G. Boyer, H. Kim, S. Pompa-Mansilla, M. J. Hamilton, D. P. Costa, G.
460 Ceballos, and J. H. Brown. 2012. Drivers and hotspots of extinction risk in marine
461 mammals. *Proceedings of the National Academy of Sciences* **109**:3395–400.
- 462 Davidson, A. D., M. J. Hamilton, A. G. Boyer, J. H. Borwn, and G. Ceballos. 2009. Multiple
463 ecological pathways to extinction in mammals. *Proceedings of the National Academy of*
464 *Sciences* **106**:10702–10705.
- 465 De'ath, G., and K. E. Fabricius. 2000. Classification and regression trees: a powerful yet
466 simple technique for ecological data analysis. *Ecology* **81**:3178–3192.
- 467 Duda, R. O., P. E. Hart, and D. G. Stork. 2001. *Pattern Classification*. Page 654 p. Wiley, USA.
- 468 Elith, J., and C. H. Graham. 2009. Do they? How do they? WHY do they differ? On finding
469 reasons for differing performances of species distribution models. *Ecography* **32**:66–77.
- 470 Fawcett, T. 2006. An introduction to ROC analysis. *Pattern Recognition Letters* **27**:861–874.
- 471 Freckleton, R. P., N. Cooper, and W. Jetz. 2011. Comparative methods as a statistical fix: the
472 dangers of ignoring an evolutionary model. *The American Naturalist* **178**:E10–7.
- 473 Fritz, S. A., O. R. P. Bininda-Emonds, and A. Purvis. 2009. Geographical variation in predictors
474 of mammalian extinction risk: big is bad, but only in the tropics. *Ecology letters* **12**:538–
475 49.
- 476 Hastie, T., R. Tibshirani, and J. Friedman. 2009. *The Elements of Statistical Learning*. Page
477 746 p. Springer, NY, USA.

- 478 Hijmans, S. E., J. L. Cameron, P. G. Parra, A. Jones, and R. J. Jarvis. 2005. Very high resolution
479 interpolated climate surfaces for global land areas. *International Journal of Climatology*
480 **25**:1965–1978.
- 481 Hilton-Taylor, C., C. M. Pollock, J. S. Chanson, S. H. M. Butchart, T. E. E. Oldfield, and V.
482 Katariya. 2009. State of the world's species. Pages 15–41 *Wildlife in a changing world.*
483 An analysis of the 2008 IUCN Red List of Threatened Species. IUCN, Gland, Switzerland.
- 484 Imhoff, M. L., L. Bounoua, T. Ricketts, C. Loucks, R. Harriss, and W. T. Lawrence. 2004. Global
485 patterns in human consumption of net primary production. *Nature* **429**:870–873.
- 486 IUCN. 2001. IUCN Red List Categories and Criteria: version 3.1. Gland, Switzerland and
487 Cambridge, UK.
- 488 IUCN. 2008. 2008 IUCN Red List of Threatened Species. Gland, Switzerland. Retrieved
489 October 10, 2011, from www.iucnredlist.org.
- 490 IUCN. 2010. 2010 IUCN Red List of threatened species. Version 2010.3. Retrieved October
491 10, 2011, from www.iucnredlist.org.
- 492 Jones, J. P. G. et al. 2011. The Why, What, and How of Global Biodiversity Indicators Beyond
493 the 2010 Target. *Conservation Biology* **25**:450–457. Blackwell Publishing Inc.
- 494 Jones, K. E. et al. 2009. PanTHERIA: a species-level database of life history, ecology, and
495 geography of extant and recently extinct mammals. *Ecology* **90**:2648–2648.
- 496 Jones, K. E., and K. Safi. 2011. Ecology and evolution of mammalian biodiversity.
497 *Philosophical Transactions of the Royal Society B: Biological Sciences* **366**:2451–2461.
- 498 Jones, M. J., A. Fielding, and M. Sullivan. 2006. Analysing extinction risk in parrots using
499 decision trees. *Biodiversity and Conservation* **15**:1993–2007.
- 500 Joppa, L. N., D. L. Roberts, N. Myers, and S. L. Pimm. 2011. Biodiversity hotspots house most
501 undiscovered plant species. *Proceedings of the National Academy of Sciences*
502 **108**:13171–6.
- 503 Kampichler, C., R. Wieland, S. Calmé, H. Weissenberger, and S. Arriaga-Weiss. 2010.
504 Classification in conservation biology: A comparison of five machine-learning methods.
505 *Ecological Informatics* **5**:441–450. Elsevier B.V.
- 506 Keller, R. P., D. Kocev, and S. Džeroski. 2011. Trait-based risk assessment for invasive
507 species: high performance across diverse taxonomic groups, geographic ranges and
508 machine learning/statistical tools. *Diversity and Distributions* **17**:451–461.
- 509 Kuhn, M. 2008. Building predictive models in R using the caret package. *Journal of Statistical*
510 *Software* **28**:1–26.

- 511 Olden, J. D., J. J. Lawler, and N. L. Poff. 2008. Machine learning methods without tears: a
512 primer for ecologists. *The Quarterly review of biology* **83**:171–93.
- 513 Ozesmi, S., C. Tan, and U. Ozesmi. 2006. Methodological issues in building, training, and
514 testing artificial neural networks in ecological applications. *Ecological Modelling*
515 **195**:83–93.
- 516 Prasad, A. M., L. R. Iverson, and A. Liaw. 2006. Newer classification and regression tree
517 techniques: bagging and random forests for ecological prediction. *Ecosystems* **9**:181–
518 199.
- 519 Purvis, A. 2008. Phylogenetic approaches to the study of extinction. *Annual Review of*
520 *Ecology, Evolution and Systematics* **39**:301–319.
- 521 Purvis, A., J. L. Gittleman, G. Cowlishaw, and G. M. Mace. 2000. Predicting extinction risk in
522 declining species. *Proceedings of the Royal Society of London Series B: Biological*
523 *Sciences* **267**:1947–1952.
- 524 Rondinini, C. et al. 2011. Global habitat suitability models of terrestrial mammals.
525 *Philosophical Transactions of the Royal Society B: Biological Sciences* **366**:2633–41.
- 526 Rondinini, C., K. A. Wilson, L. Boitani, H. Grantham, and H. P. Possingham. 2006. Trade offs
527 of different types of species occurrence data for use in systematic conservation
528 planning. *Ecology letters* **9**:1136–45.
- 529 Samways, M., and M. Böhm. 2010. Invertebrata. Are vertebrates representative of animal
530 biodiversity as a whole? Pages 55–61 in J. E. M. Bailie, J. Griffiths, S. T. Turvey, J. Loh,
531 and B. Collen, editors. *Zoological Society of London, London, UK.*
- 532 Scheffers, B. R., D. L. Yong, J. B. C. Harris, X. Giam, and N. S. Sodhi. 2011. The world's
533 rediscovered species: back from the brink? *PloS One* **6**:e22531.
- 534 Schipper, J. et al. 2008. The status of the world's land and marine mammals: diversity,
535 threat, and knowledge. *Science* **322**:225–30.
- 536 Stuart, S. N., J. S. Chanson, N. A. Cox, B. E. Young, A. S. L. Rodrigues, D. L. Fischman, and R.
537 W. Waller. 2004. Status and trends of amphibian declines and extinctions worldwide.
538 *Science* **306**:1783–1786.
- 539 Stuart, S. N., E. O. Wilson, J. A. McNeely, R. A. Mittermeier, and J. P. Rodríguez. 2010. The
540 barometer of life. *Science* **328**:177. *American Association for the Advancement of*
541 *Science.*
- 542 Vié, J.-C., C. Hilton-Taylor, C. M. Pollock, J. S. Ragle, J. Smart, S. S. Stuart, and R. Tong. 2009.
543 The IUCN Red List: a key conservation tool. Pages 1–13 in J.-C. Vié, C. Hilton-Taylor, and
544 S. N. Stuart, editors. *Wildlife in a changing world. An analysis of the 2008 IUCN Red List*
545 *of Threatened Species. IUCN, Gland, Switzerland.*

546 Webb, A. 2002. Statistical Pattern Recognition. Page 496 p. Wiley, Chichester, UK.

547 Youden, W. J. 1950. An index for rating diagnostic tests. *Cancer* **3**:32–35.

548

For review only

549 **Tables**

550 Table 1. Number of data-sufficient species, proportion of threatened species, number of
 551 Data Deficient species and number of explanatory variables used in the models across
 552 datasets.

Dataset	Number of data-sufficient species	Proportion of threatened species	Number of Data Deficient species	Number of explanatory variables
Global	3967	22.1%	493	35
Bats	828	17%	108	36
Carnivores	188	23.2%	14	36
Primates	304	56.7%	12	32
Rodents	1666	17%	263	29

553

554 Table 2. Area Under the Receiver Operator Characteristic Curve (AUROC) for each
 555 combination of tool and dataset on the validation sets.

	CT	RF	BT	KNN	SVM	NNET	DS
Global	0.895	0.944	0.935	0.906	0.932	0.922	0.75
Bats	0.872	0.894	0.897	0.858	0.871	0.891	0.727
Carnivores	0.896	0.901	0.919	0.849	0.922	0.961	0.736
Primates	0.803	0.854	0.866	0.788	0.873	0.857	0.738
Rodents	0.871	0.951	0.933	0.925	0.949	0.935	0.792

556 * CT: Classification Tree, RF: Random Forests, BT: Boosted Trees, KNN: K-Nearest
 557 Neighbours, SVM: Support Vector Machine, NNET: Neural Networks, DS: Decision Stump.

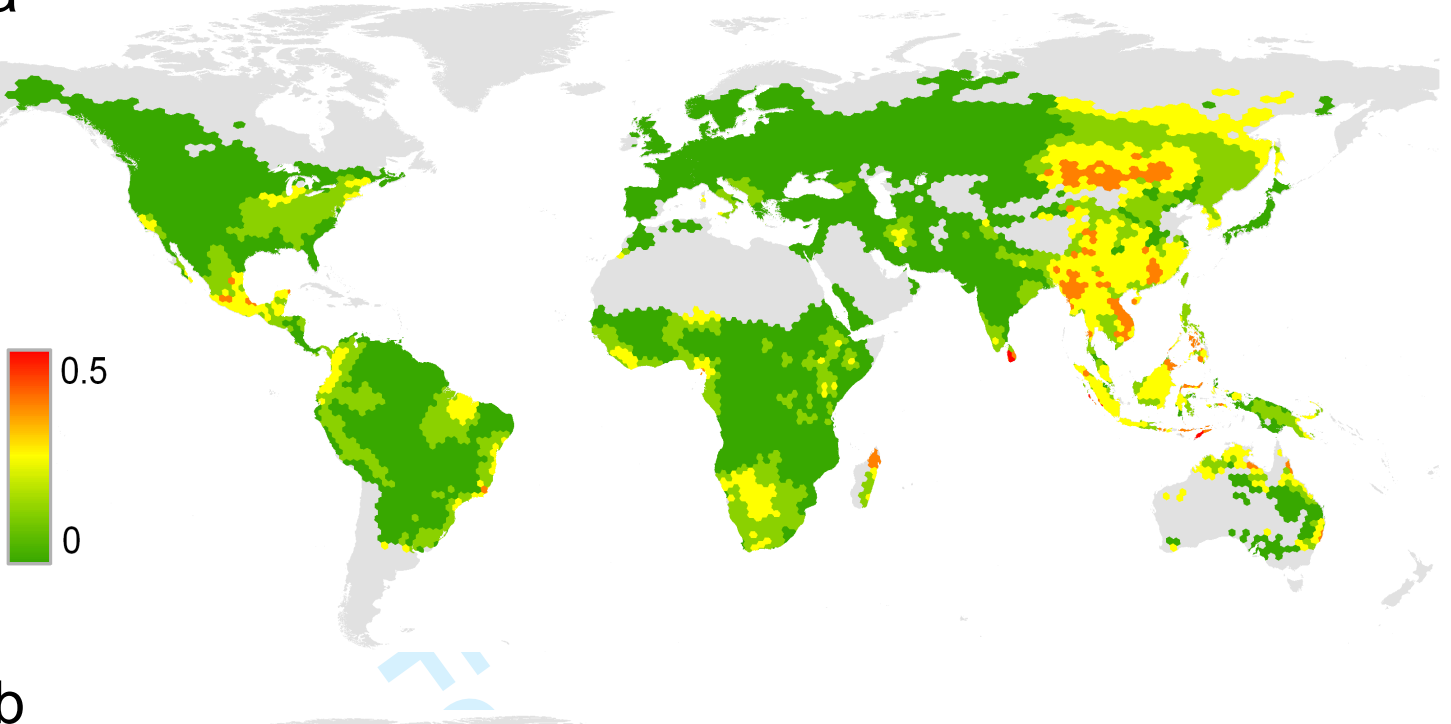
558 **Figure Legends**

559 Figure 1. Global geographic distribution of terrestrial mammal extinction risk in the
560 validation set. Observed (a) and predicted (b) proportion of threatened species and
561 standardized model residuals (c).

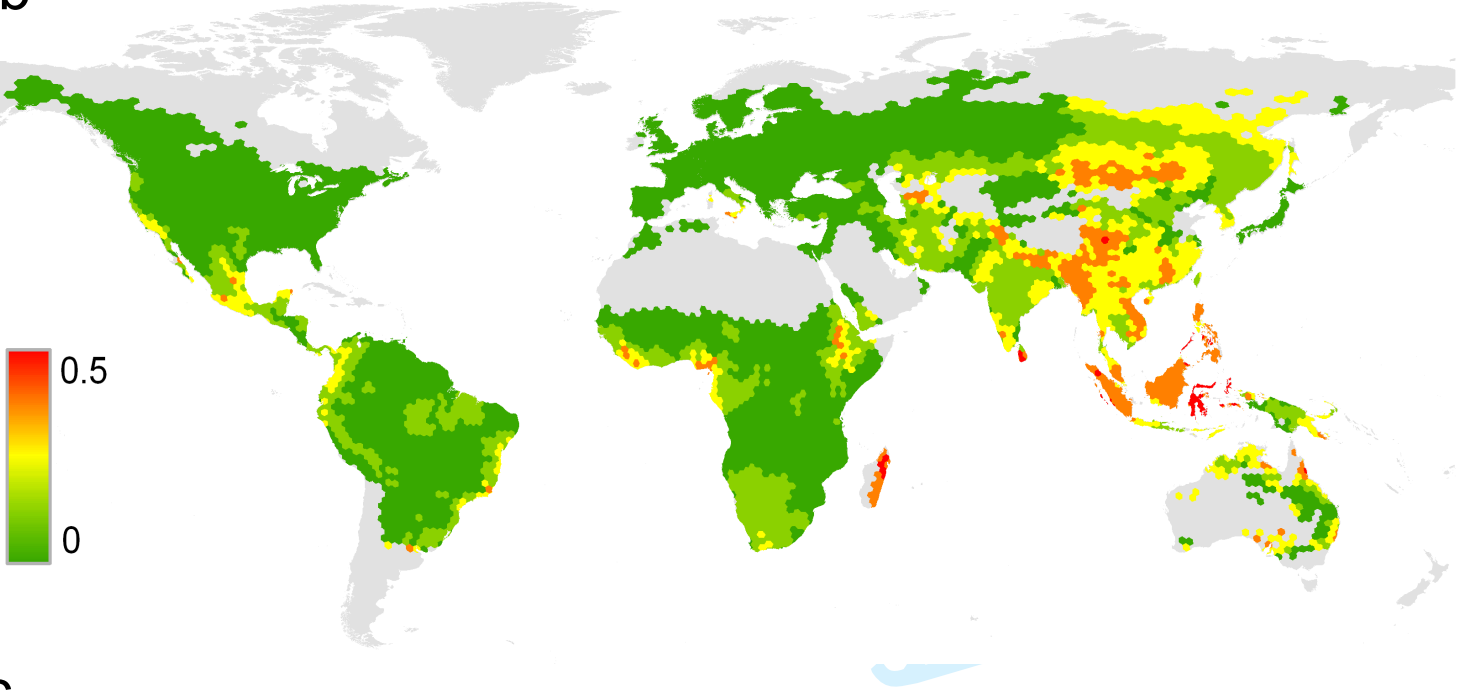
562 Figure 2. Global geographic distribution of terrestrial mammal extinction risk. Proportion of
563 threatened species when Data Deficient species are excluded (a), when Data Deficient
564 species model predictions are included (b) and standardized model residuals (c).

565 Figure 3. Extent of congruence between hotspots of proportion of threatened species under
566 two scenarios, shown across a range of hotspot definitions. The two scenarios are: 1)
567 exclusion of Data Deficient species and 2) inclusion of Data Deficient species model
568 predictions. Horizontal line shows expectation under full congruence; vertical arrow shows
569 2.5% hotspot definition.

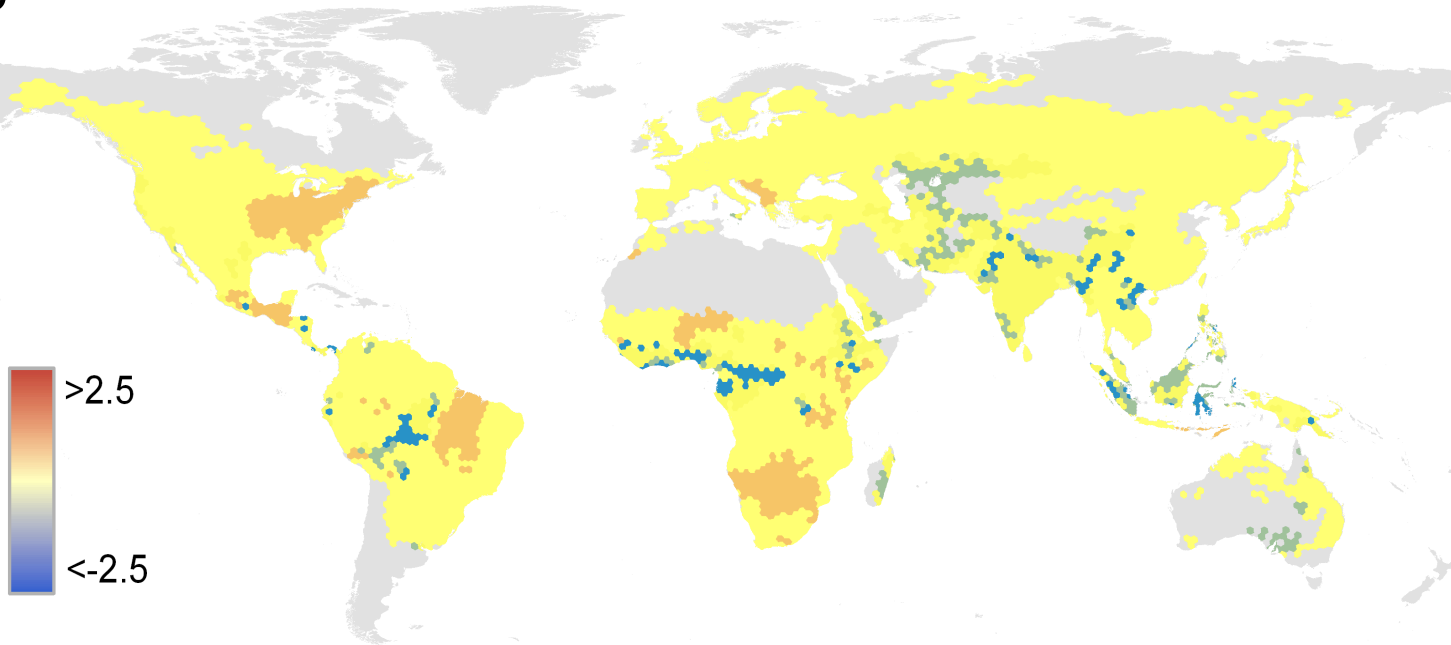
a



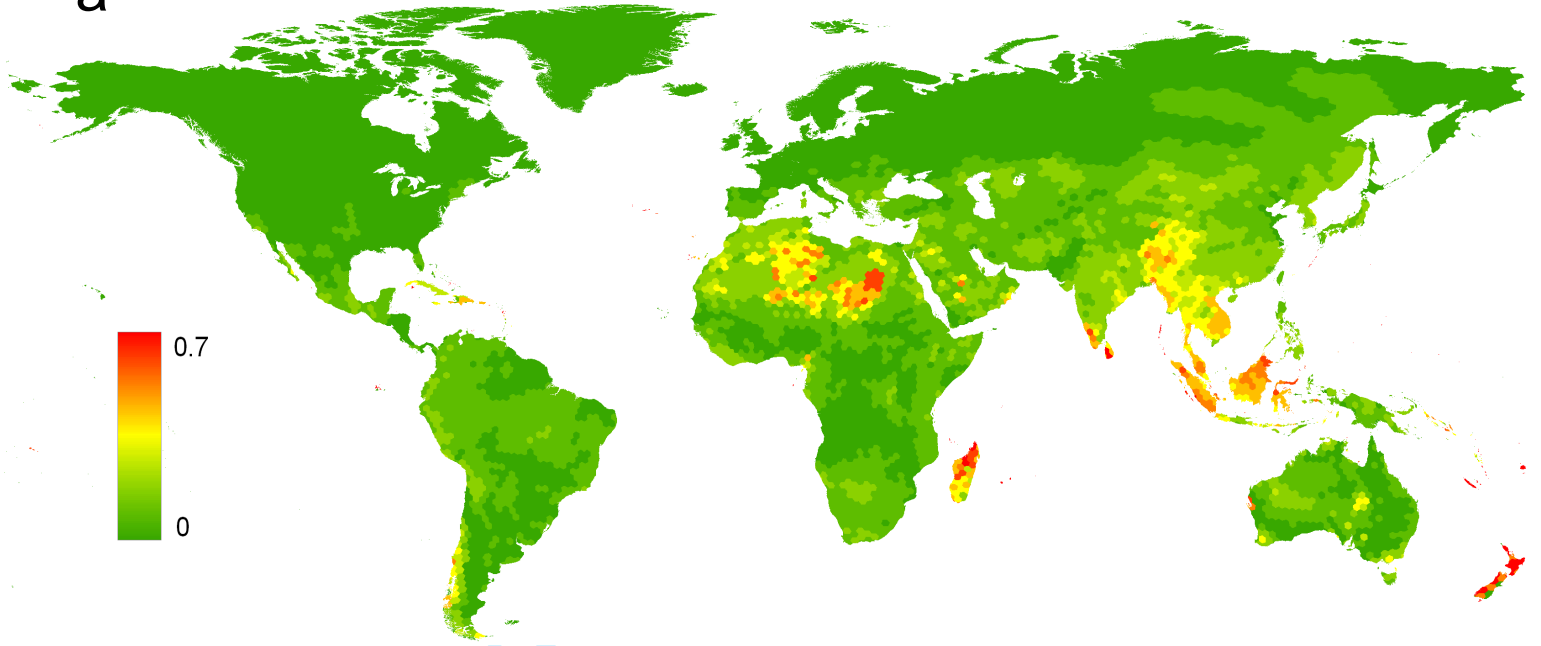
b



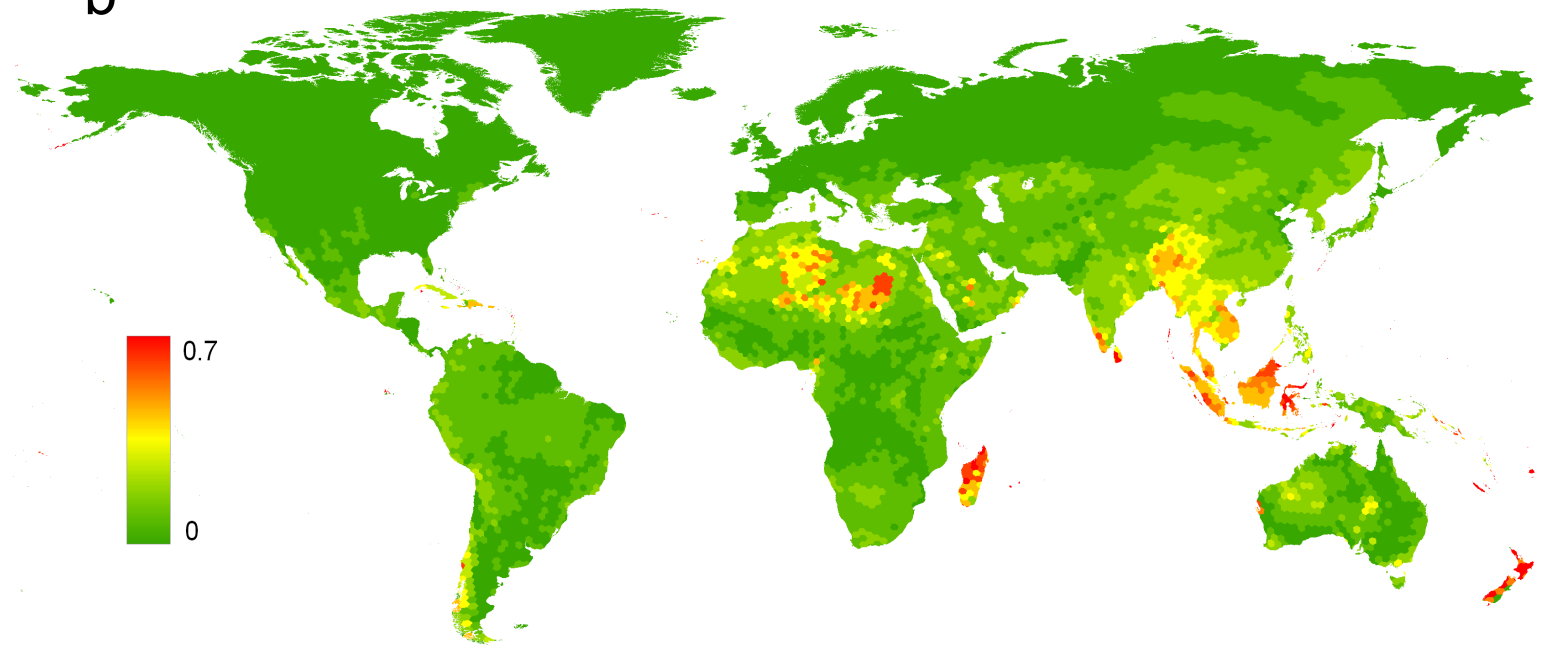
c



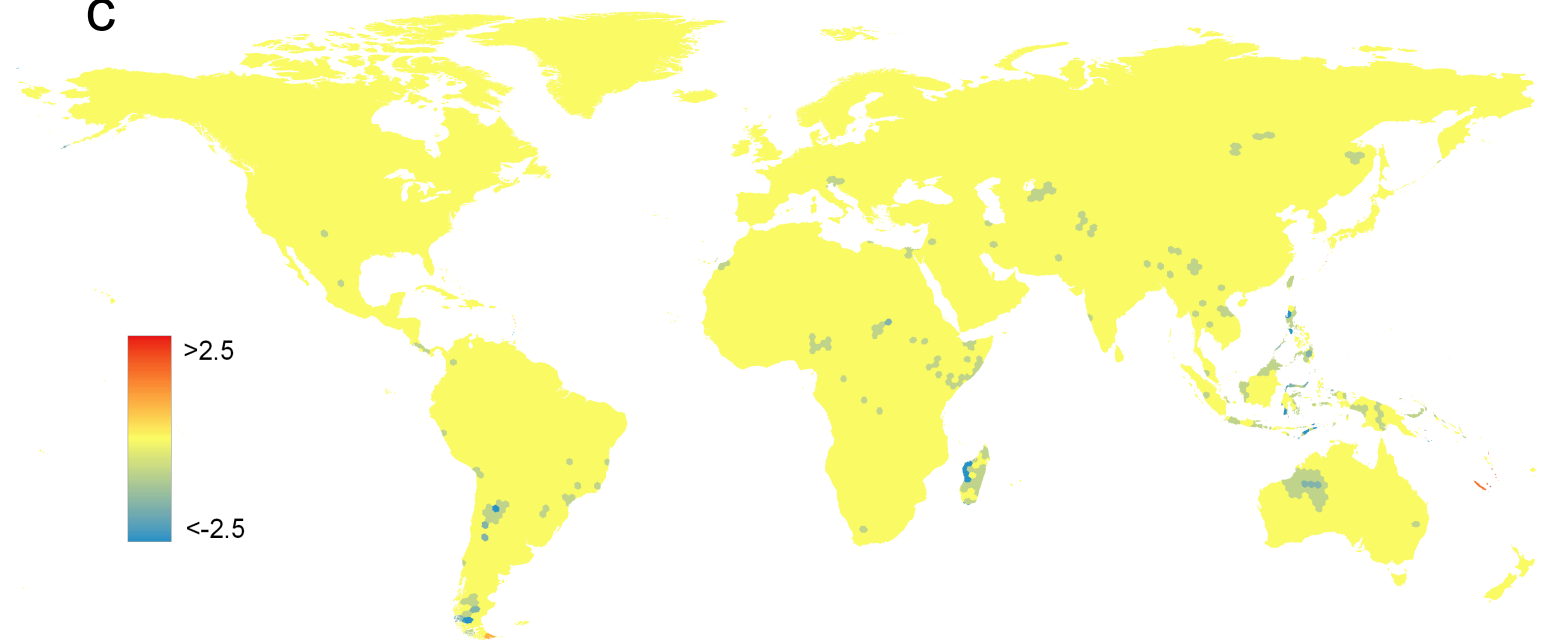
a



b



c



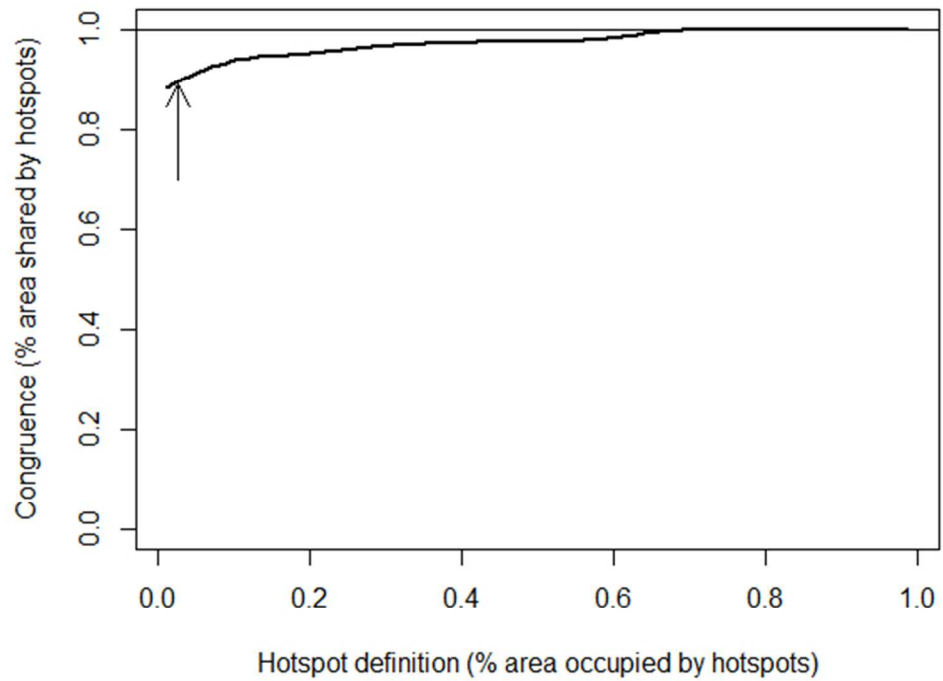


Figure 3. Extent of congruence between hotspots of proportion of threatened species under two scenarios, shown across a range of hotspot definitions. The two scenarios are: 1) exclusion of Data Deficient species and 2) inclusion of Data Deficient species model predictions. Horizontal line shows expectation under full congruence; vertical arrow shows 2.5% hotspot definition.

158x117mm (96 x 96 DPI)