



This is a repository copy of *Towards a general framework for predicting threat status of data-deficient species from phylogenetic, spatial and environmental information.*

White Rose Research Online URL for this paper:
<http://eprints.whiterose.ac.uk/88375/>

Version: Accepted Version

Article:

Jetz, W. and Freckleton, R.P. (2015) Towards a general framework for predicting threat status of data-deficient species from phylogenetic, spatial and environmental information. *Philosophical Transactions B: Biological Sciences*, 370 (1662). ISSN 0962-8436

<https://doi.org/10.1098/rstb.2014.0016>

Reuse

Unless indicated otherwise, fulltext items are protected by copyright with all rights reserved. The copyright exception in section 29 of the Copyright, Designs and Patents Act 1988 allows the making of a single copy solely for the purpose of non-commercial research or private study within the limits of fair dealing. The publisher or other rights-holder may allow further reproduction and re-use of this version - refer to the White Rose Research Online record for this item. Where records identify the publisher as the copyright holder, users can verify any specific terms of use on the publisher's website.

Takedown

If you consider content in White Rose Research Online to be in breach of UK law, please notify us by emailing eprints@whiterose.ac.uk including the URL of the record and the reason for the withdrawal request.



eprints@whiterose.ac.uk
<https://eprints.whiterose.ac.uk/>

1
2
3
4 1 **Toward a general framework for predicting threat status of data-**
5
6 2 **deficient species from phylogenetic, spatial and environmental**
7
8
9 3 **information**
10
11
12 4
13
14
15 5
16

17
18 6 Walter Jetz^{1*} & Robert P. Freckleton^{2*}
19
20

21 7
22
23 8 ¹ Department of Ecology and Evolutionary Biology, Yale University, PO Box 208106,
24
25
26 9 New Haven CT 06520-8106, USA.
27
28

29 10
30
31 11 ² Department of Animal & Plant Sciences, University of Sheffield, Sheffield S10 2TN,
32
33
34 12 United Kingdom
35
36

37 13
38
39
40 14 Running head: Predicting species threat status
41
42

43 15
44
45
46 16
47
48
49 17
50
51 18
52
53
54 19
55
56
57 20 * Authors contributed equally
58
59
60

1
2
3 21 **Abstract**
4
5

6 22 In taxon-wide assessments of threat status many species remain not included due to lack
7
8 23 of data. Here we present a novel spatial-phylogenetic statistical framework that uses a
9
10 24 small set of readily available or derivable characteristics, including phylogenetically
11
12 25 imputed body mass and remotely-sensed human encroachment, to provide initial baseline
13
14 26 predictions of threat status for data-deficient species. Applied to assessed mammal
15
16 27 species worldwide the approach effectively identifies threatened species and predicts the
17
18 28 geographic variation in threat. For the 483 data-deficient species the models predict
19
20 29 highly elevated threat, with 69% 'at-risk' species in this set, compared to 22% among
21
22 30 assessed species. This results in 331 additional potentially threatened mammals, with
23
24 31 elevated conservation importance in rodents, bats and shrews, and countries like
25
26 32 Colombia, Sulawesi, and the Philippines. These findings demonstrate the future potential
27
28 33 for combining phylogenies and remotely sensed data with species distributions to identify
29
30 34 species and regions of conservation concern.
31
32
33
34
35

36
37 35
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

36 **Introduction**

37 Human activities continue to cause the loss of many species together with the function
38 and services they provide [1]. In the face of these mounting threats and limited resources
39 to conserve species [2], tools are required to identify those of greatest conservation
40 concern. Global IUCN Red List assessments [3] have provided important knowledge
41 about the state of biodiversity and have helped to identify priority species and regions for
42 conservation [4-7]. On this basis approximately 20% of mammal, bird and amphibian
43 species are currently identified as threatened [3]. In order to minimize potential biases in
44 perceived patterns of biodiversity threat, species should be assessed comprehensively or
45 at least representatively. In addition to undiscovered species [4, 8], species with too little
46 information for threat categorization ('data-deficient species') are thus a major concern.
47 The IUCN assessment process relies on available field-based knowledge of e.g.
48 population size, rate of decline, and range size of each species to assigned threat status
49 [9-12]. Paucity of data, e.g. due to financial or logistical limitations for field studies,
50 makes complete assessments impossible for some species, with little prospects for change
51 in the near future. The number of species lacking data may be substantial, with e.g. 2,436
52 of 11,806 recognized mammal and amphibians species classified as 'data-deficient' in
53 2011 [3], including 834 extant mammals. The potential for data-deficient species to
54 change absolute threat levels of taxa has been acknowledged [4, 7]. In the absence of
55 better knowledge, a risk-averse approach may be to simply assume that all data-deficient
56 species are threatened. But given the sheer number of data-deficient species the
57 implications for conservation prioritization may be substantial and carry a high cost if
58 large numbers are in fact not threatened. At the other extreme data-deficient species may

1
2
3 59 be no more threatened than assessed species, for example as appears to be the case with
4
5 60 data-deficient birds [13]. Model-based initial baseline threat predictions for data-deficient
6
7
8 61 species and a general framework to provide them for groups assessed in the future would
9
10 62 thus hold multiple benefits for conservation practice.
11
12

13 63
14
15
16 64 The relative paucity of data on ecology and threats of many species stands in stark
17
18 65 contrast to our rapidly growing detailed knowledge about species' phylogenetic
19
20 66 relationships and geographic distributions. Technology now allows cost-effective and
21
22 67 rapid generation of phylogenies for thousands of species. While often remaining coarse in
23
24 68 grain [14, 15] or even limited to type specimen and thus for some species a main reason
25
26 69 for data-deficient status, more faceted geographic distribution information is increasingly
27
28 70 becoming available for many species [15; see also mol.org]. Distribution data permits
29
30 71 two types of inference about potential threat status. First, statistical models can
31
32 72 quantitatively capture the association between range size and threat status in assessed
33
34 73 species [16] and then can be applied to data-deficient species [17-19] to account for this
35
36 74 risk component. Second, geographic range information can be intersected with
37
38 75 environmental layers that inform about broad-scale environmental niches and associated
39
40 76 life history signals (e.g. on fecundity, generation time) related to threat status [16, 20].
41
42 77 And, more directly, remotely-sensed layers of land-cover can provide coarse estimates of
43
44 78 potential habitat loss due to human encroachment. Information of this kind has recently
45
46 79 been shown to successfully predict threat status in birds [20] and mammals [21]. Modern
47
48 80 statistical tools allow the development of models of correlates of current threat levels that
49
50 81 incorporate both phylogenetic and spatial data [17, 22-27].
51
52
53
54
55
56
57
58
59
60

1
2
3 82
4
5
6

7 83 In previous work, modeled threat predictions for data-deficient species have been made
8
9 84 without environmental or phylogenetic information [19], or without habitat encroachment
10
11 85 information and using eigenvectors [17, 18] which are highly constrained in their ability
12
13 86 to appropriately represent both phylogenetic and spatial signals [28, 29]. A general
14
15 87 framework that readily capitalizes on the ever increasing availability of species
16
17 88 distribution and remote sensing data, and rigorously incorporates phylogenetic and
18
19 89 geographic information is thus still missing. In this study we build on our earlier work
20
21 90 linking spatial and phylogenetic models [27] and predicting threat data with GIS-derived
22
23 91 habitat information [20] to develop such a framework. We demonstrate the approach
24
25 92 applied to mammals by parameterizing models of threat status based on readily available
26
27 93 variables capturing key aspects of life history, rarity and range loss (body mass,
28
29 94 geographic range size, human encroachment on species' ranges) together with spatial and
30
31 95 phylogenetic dependency for 3,703 mammal species across 16 orders with sufficient
32
33 96 information to be assessed by the Red List. We then apply these models to 483 species
34
35 97 classified as data-deficient species. We show, that the presented framework may offer a
36
37 98 cost-effective way for initial baseline threat evaluation of many understudied (and
38
39 99 potentially at-risk) species.
40
41
42
43
44
45
46

47 100
4849
50 **101 Methods**

51
52 102 **Data.** We analyzed data on 4,186 terrestrial mammal species from 16 orders in the IUCN
53
54 103 Red List [30] that could also be placed in the mammalian super tree phylogeny [31] (with
55
56 104 recent updates). Of these, 3703 species had been assessed (with 812 deemed threatened,
57
58
59
60

1
2
3 105 i.e. categories “Vulnerable”, “Endangered” and “Critically Endangered”) and 483
4
5
6 106 recognized but not assessed (category “Data-Deficient”) by IUCN. We gathered
7
8 107 information on mammal body masses from [32] and from F. Smith (pers. comm.). One
9
10 108 order (Perissodactyla) contained no data-deficient species. We selected native and
11
12 109 reintroduced resident and breeding ranges that were extant or probably extant from the
13
14 110 IUCN expert range maps [30] which we extracted over a 110x110km grid in Equal Area
15
16 111 Cylindrical projection. We overlaid each species range map with information on
17
18 112 transformed habitats owing to anthropogenic activities. Specifically, we estimated
19
20 113 ‘Encroachment’ as the proportion of expert range transformed by past human activities
21
22 114 (i.e. cultivated or managed, mosaics, including cropland and urban areas) according to the
23
24 115 Global LandCover 2000 land-cover classification [33]. At 1km native resolution this
25
26 116 information is collected at much finer scale than expert range maps and analysis grid
27
28 117 [14], but used as range summary measure it offers a concrete first-order estimate of
29
30 118 overall range encroachment, and has recently been shown to be a strong correlate of
31
32 119 expert-assessed IUCN threat status in birds [20]. We note that other high-resolution
33
34 120 global land cover classifications exist and that all suffer from remaining classification
35
36 121 errors [34]. As additional metric we also calculated the average Human Influence Index
37
38 122 (‘HII’) value [35, 36] over the species ranges.
39
40
41
42
43
44
45
46
47
48

49 124 **Summary of approach**

50
51
52 125 To summarize our approach, we first imputed the body masses of species for which data
53
54 126 are missing and then used Generalized Linear Models that include phylogenetic and
55
56 127 spatial dependence to predict IUCN status. We account for statistical uncertainty in our
57
58
59
60

1
2
3 128 estimates of body mass by using Multiple Imputation. In order to incorporate uncertainty
4
5 129 in our overall predictions, we express the model outputs as threat probabilities; i.e. given
6
7
8 130 the predictions of the model and the statistical uncertainty in these, what is the probability
9
10 131 that each species is threatened (i.e. IUCN categories Vulnerable, Endangered or Critically
11
12
13 132 Endangered) or not?
14
15
16 133

18 134 **Statistical modelling framework**

20
21 135 The starting point for our analyses is a linear statistical model relating the values of a trait
22
23 136 of interest to a set of predictors [24, 26]. The errors are assumed to have a multivariate
24
25 137 normal distribution with mean 0 and a variance-covariance matrix that is defined by the
26
27 138 phylogeny [23, 24, 26] and spatial distances [27]. Predictions from our models were
28
29 139 generated by using the fitted parameter values together with the degree of phylogenetic
30
31 140 and spatial similarity of species using the approach described in [26]. Our predictions
32
33 141 therefore account for the phylogenetic /spatial structure in the data, i.e. they have the
34
35 142 property that closely related, or species that live in the same place, should be similar to
36
37 143 each other. We calculated variances for predicted values using the formulae in [24].
38
39 144 These variances are used to calculate the variance in estimates of body mass and IUCN
40
41 145 status (below).
42
43
44
45
46
47
48 146

50 147 **Phylogenetic and spatial models for trait covariances**

52
53 148 We use the generalized least squares (GLS) approach described in Freckleton & Jetz [27]
54
55 149 to account for both spatial and phylogenetic effects. A parameter ϕ is included in the
56
57
58
59
60

1
2
3 150 model to account for the influence of space. According to this model, of the total
4
5
6 151 variance, a proportion ϕ is attributed to spatial variance, $(1 - \phi)$ is due to the non-spatial
7
8 152 component. We also used the λ transformation suggested by Pagel [22, 37]. In the
9
10 153 context of modelling spatial and phylogenetic effects simultaneously, the λ statistic
11
12 154 allows us to include trait variation independent of both phylogeny and space in our
13
14 155 analysis: a proportion $\gamma = (1 - \phi) (1 - \lambda)$ of the trait variation is independent of phylogeny
15
16 156 or space [27]. This approach is akin to including a ‘nugget’ in a spatial model [38]. We
17
18 157 estimated ϕ and λ by maximum likelihood [39].
19
20
21
22
23
24
25

26 158
27 159 The spatial matrix was calculated and tested using the approach described in Freckleton
28
29 160 & Jetz [27]. The spatial matrix reduces the spatial configuration of the data to a series of
30
31 161 pairwise distances that measure the distance between each species. Following Freckleton
32
33 162 & Jetz [27] we did this by calculating the distances between the centroids of the ranges of
34
35 163 each pair of species. The assumption is, therefore, that the variance between species’
36
37 164 traits grows linearly with spatial distances. As we showed before, this assumption can be
38
39 165 tested graphically and in the analyses reported here, as well as in Freckleton & Jetz [27],
40
41 166 this assumption was found to be adequate. Following Freckleton & Jetz [27], in order to
42
43 167 aid interpretation of the model we define λ' as the relative contribution of phylogeny (λ'
44
45 168 $=\lambda (1 - \phi)$) once the effects of space have been accounted for. This parameterisation
46
47 169 allows a simple interpretation of the joint estimates of ϕ and λ because, as shown in
48
49 170 Freckleton & Jetz [27], the sum of γ , λ' , and ϕ is always 1. These parameters can be
50
51 171 interpreted as the individual proportional contributions to variance of the different
52
53
54
55
56
57
58
59
60

1
2
3 172 variance components.
4
5

6 173
7
8

9 174 **Imputation of mammalian body mass**
10

11 175 We used estimates of mammalian body masses of 3462 species in the 16 analyzed orders
12
13 176 to predict the values for the 723 species without body mass data. For each order we used
14
15 177 the GLS approach described above to predict body mass based on the species with body
16
17 178 mass data along with phylogenetic and spatial information. We conducted this analysis at
18
19 179 the level of orders as previous analysis has shown that the Brownian model, modified to
20
21 180 allow for varying degrees of phylogenetic dependence, provides an adequate description
22
23 181 of body mass variation within orders [40]. Body mass was log-transformed prior to
24
25 182 analysis.
26
27
28
29
30

31 183
32
33

34 184 For species missing body mass we used the predicted values predicted as estimates of log
35
36 185 mass in the modeling of IUCN threat status. A problem with using single imputation of
37
38 186 this sort is that although parameter estimates should be unbiased [41], there is a
39
40 187 possibility of under-estimation of variances for parameters using this method. We
41
42 188 therefore conducted significance tests for our models using multiple imputation. For this
43
44 189 we calculated for all species lacking body mass data predicted values using the above
45
46 190 GLS model, along with a variance for each prediction (using the method in [24], see
47
48 191 above). These estimates formed the basis for the multiple imputations [for further
49
50 192 background on the method see 42, 43, 44; for specific implementation here see also
51
52 193 Nakagawa & Freckleton 2008]. We used 10 imputations, and the statistical tests reported
53
54
55
56
57
58
59
60

1
2
3 194 in Table S1 are the outcome of this analysis. We found that in practice the variance
4
5 195 across the imputations was very small indeed so that this step was not vital in this case,
6
7
8 196 although this need not always be true.
9

10
11 197
12
13 198 In order to evaluate the accuracy of the predictions of body mass we used a simple
14
15
16 199 randomization. Estimates of λ and ϕ from the best fitting model for each order were used
17
18
19 200 to construct a variance-covariance matrix. This variance matrix formed the basis for
20
21 201 generating randomized multivariate normally distributed data (using the `rmvnorm` in the
22
23 202 R `mvtnorm` package). Species originally missing data were then removed and their values
24
25
26 203 imputed. The correlation of these imputed values with the true values was then
27
28 204 calculated. Note that because this analysis is conducted on randomly generated data, this
29
30 205 is different from a cross-validation which is based on removal of data from the original
31
32 206 data and would not normally be conducted using single-species removals. This was
33
34
35 207 repeated 1000 times per missing species per order. The results of this analysis are
36
37 208 summarized in Table S2.
38

39
40 209

41 42 43 210 **Application to IUCN categories**

44
45
46 211 The IUCN categories were treated as a five point ordinal scale ranging from “Least
47
48 212 Concern”, 1, to “Critically Endangered”, 5. Although the response variable is a discrete
49
50 213 ordinal variable, the models described observed threat levels well, offering explanatory
51
52 214 power equal to, or better than that found in previous studies (Table S1, Fig. S1). This
53
54
55 215 same approach has been taken in other recent analyses of threat status [45]. We compared
56
57
58
59
60

1
2
3 216 our results with those of generalized linear models in which responses are treated as
4
5 217 multinomial or ordered logistic responses, which yielded very similar results, but are
6
7
8 218 unable to address the spatial and phylogenetic covariance (see Figs S2, S3 and below).
9
10 219 The main problem in generating an output from the model is that a fitted / predicted value
11
12 220 is a point estimate and does not account for the statistical uncertainty in our estimates. To
13
14 221 incorporate uncertainty, after the analysis we converted our predictions of IUCN status
15
16 222 into probabilities of threat. Previous analyses have taken a similar approach in the
17
18 223 analysis of threat status, but instead converting the threat to a binary variable before the
19
20 224 regression analysis [17]. This has the disadvantage that information on the ordinal nature
21
22 225 of the IUCN scale is ignored. Our analysis, however, retained the continuous information
23
24 226 in the model fitting: for example we account for the fact that a species classified as
25
26 227 category 5 (Critically Endangered) is more at risk than a species in category 3
27
28 228 (Vulnerable).

29
30
31
32
33
34 229
35
36
37 230 To produce these threat probabilities we calculated the probability that each species was
38
39 231 threatened or not from the predictions of IUCN status. This was simply done by
40
41 232 calculating:

42
43
44
45 233
$$P_i^{threatened} = Z\left(\frac{y_i^{pred} - 2.5}{\sigma_i}\right) \quad (1)$$

46
47
48
49 234 where $Z()$ is the cumulative z (standardized normal) distribution, y^{pred} is the predicted
50
51 235 value and σ is its standard deviation. This is the probability that the predicted value of
52
53 236 species i is *greater* than 2.5 (see also [17, 20]). The choice of threshold in equation (1) is
54
55 237 dependent on the interpretation of the categories and how these relate to continuous
56
57
58
59
60

1
2
3 238 model predictions. With eqn. (1), a species with an IUCN status predicted to be 2.5 (i.e.
4
5 239 in between “near threatened” and “vulnerable”) will have a threat probability of 0.5. We
6
7
8 240 repeated the analysis using a threshold of 2 which yielded a visually clearer
9
10 241 discrimination between the higher IUCN categories, but did essentially not affect the
11
12 242 results of Fig. 1(Fig. S4), because the probabilities are simply rescaled such that the mean
13
14 243 probability is 0.5 at a predicted value of 2 rather than 2.5. The results in Fig 2 are also
15
16 244 extremely similar (Fig. S5), because the estimates of the proportions of species to be
17
18 245 threatened or not are set by a threshold estimated from the data by receiver operator
19
20 246 characteristic analysis (below). Thus, results were broadly invariant to the choice of
21
22 247 threshold in equation (1).
23
24
25
26
27
28
29

30 249 We used the full models in Table S1 for making predictions and did not attempt model
31
32 250 reduction. There were several reasons for this. First, model reduction by elimination of
33
34 251 variables (e.g. based on statistical significance) has undesirable consequences, such as
35
36 252 degenerate sampling distributions and model selection bias [46]. Second, examination of
37
38 253 the coefficients for the predictors indicated that, independent of statistical significance,
39
40 254 the directions of effects were usually quite consistent between orders. For example 15 out
41
42 255 of 16 coefficients for the effect of body mass are positive even if all are not statistically
43
44 256 significant (Table S1); 12 of 16 coefficients for the encroachment variable are positive
45
46 257 (Table S1). Finally, we checked predictions with and without the least significant
47
48 258 variables and confirmed that the R^2 values were not unduly inflated and giving a false
49
50 259 impression of good fit. In order to test the predictive ability of the threat probabilities we
51
52 260 assessed how well the fitted threat probabilities predicted for assessed species were able
53
54
55
56
57
58
59
60

1
2
3 261 to distinguish threatened from non-threatened species using the Area Under the Curve
4
5 262 (AUC) in the Receiver Operating Characteristic (ROC) curve [47]. AUC varies between
6
7
8 263 0.5, which indicates that the predictions are no better than random, and 1, which is
9
10 264 perfect agreement between observed and predicted. As threshold for assigning
11
12 265 probabilities into binary categories of threatened and non-threatened, we used the value at
13
14
15 266 which sensitivity equaled specificity in a given order.
16
17
18 267
19
20
21 268

268 **Model Approach and Limitations**

269 The methodology we have used is based on currently available tools and will be
270 improved by future developments that include techniques such as logistic and
271 multinomial generalized linear mixed models that could account for phylogenetic and
272 spatial dependence and would enable to better model the discrete ordinal state variable
273 [48, 49]. However such tools require very large datasets: logistic regression requires large
274 amounts of data because binary observations contain relatively little information.
275 Multinomial or ordered responses are an extension of logistic regression and as the
276 number of states increases the data requirements increase. Given this, the approach taken
277 here to treat the data as continuous is unlikely to seriously compromise the results (see
278 also supplementary results). Moreover existing methods for such responses do not
279 combine spatial and phylogenetic signals, and can be very difficult to implement and
280 tune. In the future, faster methods for fitting phylogenetic models are under development
281 and these should facilitate further methodological advances [50]. We have assumed that
282 the variance scales linearly with both phylogenetic and geographic distances. This is
283 supported by diagnostics (e.g. see [27] for a worked examples). The assumption of

1
2
3 284 linearity is not terribly critical so long as variance increases with distance. In previous
4
5 285 work we suggested how the assumption could be varied (Table 1 in [27]). However it
6
7
8 286 should be noted that nonlinear transformations of variance matrices are potentially
9
10
11 287 difficult to work with. For example we have recently shown that a commonly used
12
13 288 transformation (the Ornstein Uhlenbeck) is severely biased under most circumstances for
14
15 289 even large datasets (Thomas et al. in review).

16
17
18 290 The models we developed are strongly dependent on range size as a predictor of IUCN
19
20 291 status, which reflects the importance of range size in the formal assessment process. It is
21
22 292 important to note that our predictive models are not aimed at *testing* the relevance of this
23
24 293 variable (which would require variable elimination to avoid circularity), but to use this
25
26 294 formally recognized association for prediction. In other words, we use A (assessed
27
28 295 species) modeled by B (novel framework and independent variables) to predict C (not yet
29
30 296 assessed species), not to make inference about A.
31
32
33
34

35 297

36 37 38 298 **Results**

39 40 41 299 **Assessed species**

42
43
44 300 For the 16 mammal orders analyzed the threat probabilities (whether a species is non-
45
46 301 threatened or threatened) predicted by the models successfully explain observed variation
47
48 302 in assessed threat score (R^2 values were typically ~40% or greater; Table S1) and
49
50 303 effectively predict species threat category (Fig. 1A). Range size and body mass were
51
52 304 generally strong correlates of threat status, with smaller ranging and larger species
53
54
55 305 typically being subject to greater threat (Table S1). Given the inherent role of range size
56
57
58
59
60

1
2
3 306 in the IUCN assessment process [11] these associations are not altogether unexpected and
4
5 307 confirm previous findings [19, 20, 51, 52]. Less consistently than recently observed in all
6
7
8 308 terrestrial birds [20], land-cover encroachment and human influence measures are
9
10 309 strongly positively correlated with IUCN threat category in several orders. This
11
12 310 contributes to the overall predictive ability of the models and confirms the relevance of
13
14
15 311 such variables for threat predictions (Table S1).
16
17
18 312

19
20 313 In addition to the strong phylogenetic dependence of body mass (Table S2), 9 of the
21
22 314 orders showed phylogenetic or spatial dependence in the residuals of the models for
23
24
25 315 IUCN threat. The degree of net phylogenetic signal in the residuals of the final models is
26
27 316 generally low, with the phylogenetic effect estimated as zero for seven and very low (0.1
28
29 317 or less) for five orders. Notably higher estimates are obtained for primates (0.66). Six
30
31 318 orders showed strong spatial signals, with estimates of the spatial coefficient, ϕ , as high
32
33 319 as 0.6-1.0 (Table S1). The threat probabilities resulting from our models are the
34
35 320 probabilities that each species is in one of the threatened states rather than not threatened,
36
37 321 given the mean and variance predicted by the model (see methods). The Receiver
38
39 322 Operating Characteristic plot (Fig 1B) indicates a very strong discrimination of
40
41 323 threatened from non-threatened species with an AUC of 0.90 for the whole dataset and a
42
43 324 median of 0.91 for all orders. These were associated with high degrees of sensitivity and
44
45 325 specificity (typically ca. 0.8 - 0.9, Table S1). Predicted threat probabilities are remarkably
46
47 326 effective in delimiting threat status, as especially illustrated by the most and least
48
49 327 threatened IUCN classes: only 4% of species assessed to be of 'least concern' were
50
51 328 predicted to have a threat probability of 0.5 or greater (Fig. 1C; see Fig. S1 for order-
52
53
54
55
56
57
58
59
60

1
2
3 329 level plots) and only 11% of species assessed to be ‘critically endangered’ were assigned
4
5 330 a threat probability lower than 0.5 (Fig. 1C). Across all threat categories, 61% of species
6
7
8 331 assessed as being under some degree of threat had estimated threat probabilities greater
9
10 332 than 0.6 and (with nearly 31% greater than 0.8) (Fig. 1C). Overall, our predicted threat
11
12 333 probabilities are a strong discriminator of threat status with particularly high values
13
14
15 334 (>0.8) extremely unlikely for species that are not actually threatened.
16
17

18 335
19
20 336 The relative richness of species assessed as being threatened is geographically very
21
22 337 uneven (Fig. 3A). Applied to assessed species, our model predicts this observed pattern
23
24 338 very well (Fig. 3B). Overall, however, there is a strong association between the predicted
25
26 339 average probability or predicted proportion of species threatened and the observed
27
28 340 proportion of species assessed threatened ($r = 0.74$ and $r = 0.68$, respectively; Fig. 4A, B)
29
30 341 as well as, expectedly, between predicted and observed threatened richness ($r = 0.82$, Fig.
31
32 342 4C). This suggests that our models successfully capture the biogeography of assessed
33
34 343 threatened species.
35
36
37
38
39

40 344

41 42 43 345 **Data-deficient species**

44
45 346 Data-deficient species are predicted to be substantially more likely to be threatened than
46
47 347 assessed species (Fig. 1D), with an average predicted threat probability of 0.40 compared
48
49 348 to 0.21 in assessed species. For data-deficient species 28% of threat probability estimates
50
51 349 were greater than 0.6, whereas for assessed species it was only 11%. Overall threat
52
53 350 probabilities were higher for data-deficient species in 10 orders, and statistically
54
55
56
57
58
59
60

1
2
3 351 significantly so for 7 orders (Fig. 2). Classifying data deficient species into binary threat
4
5 352 categories using a standardized order-specific threshold (the value at which sensitivity
6
7
8 353 equals specificity) results in a total of 331 of 483 species predicted threatened, i.e. 69% of
9
10 354 species threatened compared to 29% among assessed species. This difference is repeated
11
12 355 among almost all orders, with a total of 298 potentially threatened data-deficient species
13
14 356 identified among the Chiroptera, Rodentia and Eulipotyphla alone (for species-level
15
16 357 results see Table S3).
17
18
19
20
21 358

22
23 359 Geographically, data-deficient species are predicted to exhibit substantially higher
24
25 360 average probabilities and proportions of species threatened (grid cell assemblage values
26
27 361 of 0.12 and 0.17, respectively) than assessed species (0.06 and 0.06, respectively). At the
28
29 362 grid cell level, the predicted average threat probabilities or proportion of data-deficient
30
31 363 species shows barely any relationship with the proportion of species assessed to be
32
33 364 threatened (Spearman rank correlations: $r = 0.30$ and $r = 0.29$, respectively; Figs. 4D,E).
34
35 365 Equally, the richness of data-deficient species predicted to be threatened is only weakly
36
37 366 correlated with that of species assessed to be threatened ($r = 0.30$, Fig. 4F). The
38
39 367 discordance in geographic ‘hotspots’ of predicted assessed and data-deficient threat is
40
41 368 apparent when comparing the maps of their predicted threat probabilities and species
42
43 369 richness in Figure 3. Threat levels predicted for data-deficient species substantially
44
45 370 exceed those of assessed species in many locations (note different color scales). Data-
46
47 371 deficient species hold much higher predicted threat levels than assessed species in
48
49 372 Colombia and Central America, Southern South America, and parts of Southeast Asia. In
50
51 373 terms of species richness (Fig 3F, 4F), data-deficient species are predicted to strongly
52
53
54
55
56
57
58
59
60

1
2
3 374 increase the number of at-risk species in Sumatra, New Guinea, Colombia, and especially
4
5
6 375 Sulawesi, where in grid cell 10 likely threatened mammal species add to the known 16.
7
8 376 This suggests that these regions are even more important for conservation than previous
9
10 377 global conservation prioritization analyses may have suggested [53, 54]. In contrast, data-
11
12 378 deficient species predicted *not* to be threatened occur both outside (e.g. Southern South
13
14 379 America) and inside (e.g. Borneo, Central and West African forests) some main areas of
15
16 380 known (assessed) high prevalence of threatened species (Fig 3D, E).
17
18
19
20
21
22
23
24
25

381

382

383 **Discussion**

26
27
28
29 384 In this study we have shown that data-deficient species are much more likely to be under
30
31 385 threat than those that have already been assessed and that the geographic distribution of
32
33 386 data-deficient species that are likely threatened is different to that of assessed threatened
34
35 387 species. This may have important implications for global mammal conservation strategies
36
37 388 [55]. According to our analysis it is extremely likely that well over three hundred
38
39 389 additional mammal species (69% of those data-deficient) are threatened, many of them
40
41 390 potentially severely so. This is over an order of magnitude more than suggested by
42
43 391 Davidson et al [19] which identified 28 data-deficient mammal species as potentially
44
45 392 threatened, but did not use environmental, spatial or phylogenetic information. Using
46
47 393 eigenvectors, no encroachment data and model validation with only bats, Jones and Safi
48
49 394 [18] estimated 35% of 481 data-deficient mammal species to be potentially threatened.
50
51 395 Our statistically more robust approach [28, 29] that additionally uses remotely sensed
52
53 396 encroachment information thus suggests much greater levels of threat in data-deficient
54
55
56
57
58
59
60

1
2
3 397 species than previously thought. The relatively low degree of phylogenetic signal of
4
5 398 IUCN status we found here contrasts with previous related results in carnivores [17]. This
6
7
8 399 difference has two sources: firstly from the inclusion of species' body masses in our
9
10 400 analyses, and secondly from the inclusion of spatial effects, which also has phylogenetic
11
12 401 signal. In particular mean mass is both strongly phylogenetically determined in all orders
13
14 402 and strongly related to IUCN status (Table S1). Accounting for body mass thus decreases
15
16 403 the detectable phylogenetic signal.
17
18
19
20 404

21
22
23 405 Our findings suggest that data-deficient species cannot be ignored in conservation threat
24
25 406 assessments and in interpretation of threat status for policy setting. In mammals data-
26
27 407 deficient species are clearly more likely than non data-deficient species to be under
28
29 408 significant threat. The association between threat status and data deficiency arises,
30
31 409 because narrow-ranged (and thus often scarce), large-bodied (that thus often low-density,
32
33 410 long generation time) species, are also very likely to be those for which little data exist
34
35 411 (Table S1, [see also 45]). There are notable exceptions: for instance, the threat probability
36
37 412 of data-deficient primates is no higher than that of those that have been assessed, likely
38
39 413 reflecting the relatively higher research effort directed at primates in the past. In contrast,
40
41 414 rodents are much more difficult to study (they are small, live in inaccessible habitat and
42
43 415 are frequently nocturnal) and for them over half of data-deficient species are predicted to
44
45 416 be threatened whereas only 16% have been assessed as threatened (Fig 2). Our findings
46
47 417 contrast with recent results for birds, where just 0.6% of species are data-deficient and
48
49 418 where species that were recently moved from this category were found to be less
50
51 419 threatened than non-data-deficient species [13]. However, these only recently assessed
52
53
54
55
56
57
58
59
60

1
2
3 420 bird species are likely not a representative sample of data-deficient species as whole, as
4
5 421 data-deficient species assessed first will likely be ones that are more easily studied (and
6
7
8 422 thus face different, potentially lesser threats) than those assessed last. The statistical
9
10 423 results gathered from all species may offer more reliable guidance.
11
12

13 424

16 425 Our general aim was to demonstrate how readily-available information can be used to
17
18 426 make initial predictions about the likely conservation status of species for which a formal
19
20 427 assessment has not yet been possible. If a similar proportion of data-deficient yet
21
22 428 threatened mammal species (69%) was to be found amongst data-deficient amphibians
23
24 429 (1,600 out of 6,312 species are data-deficient, [3]), it would represent a very large
25
26 430 increase of amphibian species at-risk. Such a scenario would add many new species to
27
28 431 the threatened categories in the Red List with strong potential consequences for geographic
29
30 432 conservation prioritization. The transferability of mammal-based estimates to other taxa
31
32 433 is of course unclear, but this realization highlights the importance of expanding
33
34 434 assessment work and seizing the increasing opportunities for rigorous statistical inference
35
36 435 of threat status.
37
38
39
40
41
42

43 436

45 437 The strong importance of select life history traits and range size for predicting threat
46
47 438 status has previously been illustrated [16]. Recently the complementary power of
48
49 439 remotely-sensed measures of human land encroachment to predict threat status has also
50
51 440 been demonstrated for birds [20]. Combined with an increasingly thorough understanding
52
53 441 of the spatial context of species [15] and ever-improving data on the phylogenetic signal,
54
55 442 a general predictive framework is emerging that may be instrumental for statistically
56
57
58
59
60

1
2
3 443 assessing the thousands of species for which an individual evaluation is time- or cost-
4
5 444 prohibitive. By identifying already assessed species with highly over- or under-predicted
6
7
8 445 threat status for further scrutiny, it may also someday help improve the Red Listing
9
10 446 process which is not without human error. Clearly, the presented framework is no silver
11
12 447 bullet to replace the need for expert assessment based on field ecological data. We expect
13
14 448 that assessment data for at least 50% of species, depending on representativeness, is
15
16 449 needed to provide reasonably reliable threat predictions, but this will vary by group and
17
18 450 likely often be higher. But this does potentially free up resources and lower completion
19
20 451 thresholds [56, 57] that would benefit the assessment of neglected taxa such as
21
22 452 invertebrates and plants. More generally, a complementary approach to traditional expert-
23
24 453 based assessment may emerge that combines available phylogenetic/biological data with
25
26 454 improved species distribution knowledge linked to a remotely-sensed monitoring of land
27
28 455 cover [15] – all facilitating a dynamic and continuous baseline assessment of the state of
29
30 456 species.
31
32
33
34
35
36
37
38
39
40
41
42

459 **Acknowledgements**

460 We thank Arne Mooers, Tien Ming Lee, and members of the Jetz Lab for feedback on the
461 manuscript. RPF was funded by a Royal Society University Research Fellowship for this
462 project. WJ acknowledges support from NSF grants DBI 0960550 and DEB 1026764 and
463 NASA Biodiversity Grant NNX11AP72G. We are grateful to Felisa Smith and the
464 NESCent body size group for sharing mammal body mass data.

465

466

References

1. Pereira H.M., Leadley P.W., Proença V., Alkemade R., Scharlemann J.P.W., Fernandez-Manjarrés J.F., Araújo M.B., Balvanera P., Biggs R., Cheung W.W.L., et al. 2010 Scenarios for Global Biodiversity in the 21st Century. *Science* **330**(6010), 1496-1501. (doi:10.1126/science.1196624).
2. Wilson K.A., McBride M.F., Bode M., Possingham H.P. 2006 Prioritizing global conservation efforts. *Nature* **440**, 337-340.
3. IUCN. 2011 IUCN Red List of Threatened Species 2011.1 - <http://www.iucnredlist.org>. Downloaded on 29 Oct 2011. (Gland, Switzerland, IUCN - <http://www.iucnredlist.org>).
4. Schipper J., Chanson J.S., Chiozza F., Cox N.A., Hoffmann M., Katariya V., Lamoreux J., Rodrigues A.S.L., Stuart S.N., Temple H.J., et al. 2008 The Status of the World's Land and Marine Mammals: Diversity, Threat, and Knowledge. *Science* **322**(5899), 225-230. (doi:10.1126/science.1165115).
5. Stattersfield A.J., Capper D.R., Dutson G.C.L., BirdLife International, IUCN. 2000 *Threatened birds of the world : the official source for birds on the IUCN Red List*. Cambridge, Barcelona, BirdLife International; Lynx Edicions; xii, 852 p.
6. Stuart S.N., Chanson J.S., Cox N.A., Young B.E., Rodrigues A.S.L., Fischman D.L., Waller R.W. 2004 Status and Trends of Amphibian Declines and Extinctions Worldwide. *Science* **306**(5702), 1783-1786. (doi:10.1126/science.1103538).
7. Hoffmann M., Hilton-Taylor C., Angulo A., Böhm M., Brooks T.M., Butchart S.H.M., Carpenter K.E., Chanson J., Collen B., Cox N.A., et al. 2010 The Impact of Conservation on the Status of the World's Vertebrates. *Science* **330**(6010), 1503-1509. (doi:10.1126/science.1194442).
8. Ceballos G., Ehrlich P.R. 2009 Discoveries of new mammal species and their

- 1
2
3 implications for conservation and ecosystem services. *Proceedings of the National*
4 *Academy of Sciences* **106**(10), 3841-3846. (doi:10.1073/pnas.0812419106).
- 5
6
7
8 9. IUCN. 2001 IUCN Red List Categories & Criteria (version 3.1). (p. 30. Gland,
9 Switzerland, IUCN.
- 10
11 10. IUCN. 2006 IUCN Standards and Petitions Working Group: Guidelines for Using
12 the IUCN Red List Categories and Criteria. Version 6.2. Prepared by the Standards
13 and Petitions Working Group of the IUCN SSC Biodiversity Assessments Sub-
14 Committee in December 2006. Downloadable from
15 <http://app.iucn.org/webfiles/doc/SSC/RedList/RedListGuidelines.pdf>. (
16
17
18
19
20
21 11. Mace G., Collar N., Gaston K., Hilton-Taylor C., Akcakaya H., Leader-Williams
22 N., Milner-Gulland E., Stuart S. 2008 Quantification of extinction risk: IUCN's system
23 for classifying threatened species. *Conserv Biol* **22**(6), 1424-1442.
- 24
25
26
27 12. Mace G.M., Lande R. 1991 Assessing Extinction Threats: Toward a Reevaluation
28 of IUCN Threatened Species Categories. *Conserv Biol* **5**(2), 148-157.
- 29
30
31 13. Butchart S.H.M., Bird J.P. 2010 Data Deficient birds on the IUCN Red List: What
32 don't we know and why does it matter? *Biol Conserv* **143**(1), 239-247.
33 (doi:10.1016/j.biocon.2009.10.008).
- 34
35
36
37 14. Hurlbert A.H., Jetz W. 2007 Species richness, hotspots, and the scale dependence
38 of range maps in ecology and conservation. *PNAS* **104**, 13384-13389.
39 (doi:10.1073/pnas.0704469104).
- 40
41
42
43 15. Jetz W., McPherson J.M., Guralnick R.P. 2012 Integrating biodiversity
44 distribution knowledge: toward a global map of life. *Trends in Ecology and Evolution*
45 **27**(3), 151-159. (doi:10.1016/j.tree.2011.09.007).
- 46
47
48
49 16. Cardillo M., Mace G.M., Jones K.E., Bielby J., Bininda-Emonds O.R.P., Sechrest
50 W., Orme C.D.L., Purvis A. 2005 Multiple Causes of High Extinction Risk in Large
51 Mammal Species. *Science* **309**(5738), 1239-1241.
- 52
53
54
55 17. Safi K., Pettorelli N. 2010 Phylogenetic, spatial and environmental components of
56 extinction risk in carnivores. *Glob Ecol Biogeogr* **19**(3), 352-362.
- 57
58
59
60

- 1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
18. Jones K.E., Safi K. 2011 Ecology and evolution of mammalian biodiversity. *Philosophical Transactions of the Royal Society B: Biological Sciences* **366**(1577), 2451-2461. (doi:10.1098/rstb.2011.0090).
 19. Davidson A.D., Hamilton M.J., Boyer A.G., Brown J.H., Ceballos G. 2009 Multiple ecological pathways to extinction in mammals. *Proceedings of the National Academy of Sciences* **106**(26), 10702-10705. (doi:10.1073/pnas.0901956106).
 20. Lee T.M., Jetz W. 2011 Unravelling the structure of species extinction risk for predictive conservation science. *Proceedings of the Royal Society B: Biological Sciences* **278**(1710), 1329-1338. (doi:10.1098/rspb.2010.1877).
 21. Cardillo M., Mace G.M., Gittleman J.L., Jones K.E., Bielby J., Purvis A. 2008 The predictability of extinction: biological and external correlates of decline in mammals. *Proceedings of the Royal Society B: Biological Sciences*.
 22. Pagel M. 1997 Inferring evolutionary processes from phylogenies. *Zoologica Scripta* **26**, 331-348.
 23. Garland T., Midford P.E., Ives A.R. 1999 An introduction to phylogenetically-based statistical methods with a new method for confidence intervals on ancestral values. *American Zoologist* **39**, 374-388.
 24. Garland T.J., Ives A.R. 2000 Using the past to predict the present: confidence intervals for regression equations in phylogenetic comparative methods. *American Naturalist* **155**, 346-364.
 25. Peres-Neto P.R. 2006 A unified strategy for estimating and controlling spatial, temporal and phylogenetic autocorrelation in ecological models. *Oecologia Brasiliensis* **10**, 105-119.
 26. Martins E.P., Hansen T.F. 1997 Phylogenies and the comparative method: A general approach to incorporating phylogenetic information into the analysis of interspecific data. *American Naturalist* **149**(4), 646-667.
 27. Freckleton R.P., Jetz W. 2009 Space versus phylogeny: disentangling phylogenetic and spatial signals in comparative data. *Proceedings of the Royal Society B: Biological Sciences* **276**(1654), 21-30.

- 1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
28. Freckleton R.P., Cooper N., Jetz W. 2011 Comparative Methods as a Statistical Fix: The Dangers of Ignoring an Evolutionary Model. *American Naturalist* **178**(1), E10-E17. (doi:10.1086/660272).
29. Beale C.M., Lennon J.J., Yearsley J.M., Brewer M.J., Elston D.A. 2010 Regression analysis of spatial data. *Ecol Lett* **13**(2), 246-264.
30. IUCN. 2009 IUCN Red List of Threatened Species. Version 2009, Version 2010.4. <http://www.iucnredlist.org>. Downloaded on 27 October 2010. (
31. Bininda-Emonds O.R.P., Cardillo M., Jones K.E., MacPhee R.D.E., Beck R.M.D., Grenyer R., Price S.A., Vos R.A., Gittleman J.L., Purvis A. 2007 The delayed rise of present-day mammals. *Nature* **446**, 507-512.
32. Jones K.E., Bielby J., Cardillo M., Fritz S.A., O'Dell J., Orme C.D.L., Safi K., Sechrest W., Boakes E.H., Carbone C. 2009 PanTHERIA: a species-level database of life history, ecology, and geography of extant and recently extinct mammals. *Ecology* **90**(9), 2648-2648.
33. Bartholomé E., Belward A. 2005 GLC2000: a new approach to global land cover mapping from Earth observation data. *International Journal of Remote Sensing* **26**(9), 1959-1977.
34. Herold M., Mayaux P., Woodcock C.E., Baccini A., Schmullius C. 2008 Some challenges in global land cover mapping: An assessment of agreement and accuracy in existing 1 km datasets. *Remote Sensing of Environment* **112**(5), 2538-2556. (doi:DOI: 10.1016/j.rse.2007.11.013).
35. Sanderson E., Jaiteh M., Levy M., Redford K., Wannebo A., Woolmer G. 2002 The human footprint and the last of the wild. *Bioscience* **52**(10), 891-904.
36. WCS. 2005 Last of the Wild Project, Version 2, 2005 (LWP-2): Global Human Influence Index (HII) Dataset (Geographic). Palisades, NY: NASA Socioeconomic Data and Applications Center (SEDAC). . (Wildlife Conservation Society (WCS); Center for International Earth Science Information Network (CIESIN), Columbia University. 2005. .
37. Pagel M. 1999 Inferring the historical patterns of biological evolution. *Nature*

- 1
2
3 401, 877-884.
4
5
6 38. Haining R. 1990 *Spatial data analysis in the social and environmental sciences*.
7 Cambridge, Cambridge University Press.
8
9
10 39. Freckleton R.P., Harvey P.H., Pagel M. 2002 Phylogenetic dependence and
11 ecological data: a test and review of evidence. *American Naturalist* **160**, 716-726.
12
13
14 40. Cooper N., Purvis A. 2010 Body Size Evolution in Mammals: Complexity in
15 Tempo and Mode. *The American Naturalist* **175**(6), 727-738.
16
17 (doi:doi:10.1086/652466).
18
19
20 41. Nakagawa S., Freckleton R.P. 2008 Missing inaction: the dangers of ignoring
21 missing data. *Trends in Ecology and Evolution* **23**, 592-596.
22
23
24 42. Rubin D.B. 1987 *Multiple imputation for nonresponse in surveys*. New York,
25 John Wiley & Sons.
26
27
28 43. Schafer J.L. 1997 *Analysis of incomplete multivariate data*. Chapman & Hall.
29
30
31 44. McKnight P.E., Mcknight K.M., Sidani S., Figueredo A.J. 2007 *Missing data: a*
32 *gentle introduction*. New York, Guilford Press.
33
34
35 45. González-Suárez M., Lucas P.M., Revilla E. 2012 Biases in comparative analyses
36 of extinction risk: mind the gap. *J Anim Ecol* **81**(6), 1211–1222. (doi:10.1111/j.1365-
37 2656.2012.01999.x).
38
39
40 46. Whittingham M.J., Stephens P.A., Bradbury R.B., Freckleton R.P. 2006 Why do
41 we still use stepwise modelling in ecology and behaviour? *J Anim Ecol* **75**(5), 1182-
42 1189.
43
44
45 47. Hanley J.A., McNeil B.J. 1982 The Meaning and Use of the Area under a
46 Receiver Operating Characteristic (ROC) Curve. *Radiology* **143**(1), 29-36.
47
48
49 48. Hadfield J.D., Nakagawa S. 2010 General quantitative genetic methods for
50 comparative biology: phylogenies, taxonomies and multi-trait models for continuous
51 and categorical characters. *J Evolution Biol* **23**(3), 494-508. (doi:10.1111/j.1420-
52 9101.2009.01915.x).
53
54
55 49. Ives A.R., Helmus M.R. 2011 Generalized linear mixed models for phylogenetic
56
57
58
59
60

- 1
2
3 analyses of community structure. *Ecological Monographs* **81**(3), 511-525.
4 (doi:10.1890/10-1264.1).
5
6
7
8 50. Freckleton R.P. 2012 Fast likelihood calculations for comparative analyses.
9 *Methods in Ecology and Evolution* **3**(5), 940-947. (doi:10.1111/j.2041-
10 210X.2012.00220.x).
11
12
13 51. Purvis A., Gittleman J.L., Cowlishaw G., Mace G.M. 2000 Predicting extinction
14 risk in declining species. *Proc R Soc Lond Ser B-Biol Sci* **267**(1456), 1947-1952.
15
16
17 52. Cooper N., Bielby J., Thomas G.H., Purvis A. 2008 Macroecology and extinction
18 risk correlates of frogs. *Glob Ecol Biogeogr* **17**(2), 211-221.
19
20
21 53. Ceballos G., Ehrlich P.R. 2006 Global mammal distributions, biodiversity
22 hotspots, and conservation. *PNAS* **103**(51), 19374-19379.
23 (doi:10.1073/pnas.0609334103).
24
25
26
27 54. Grenyer R., Orme C.D.L., Jackson S.F., Thomas G.H., Davies R.G., Davies T.J.,
28 Jones K.E., Olson V.A., Ridgely R.S., Rasmussen P.C., et al. 2006 Global distribution
29 and conservation of rare and threatened vertebrates. *Nature* **444**(7115), 93-96.
30
31
32
33 55. Rondinini C., Rodrigues A.S.L., Boitani L. 2011 The key elements of a
34 comprehensive global mammal conservation strategy. *Philosophical Transactions of*
35 *the Royal Society B: Biological Sciences* **366**(1578), 2591-2597.
36 (doi:10.1098/rstb.2011.0111).
37
38
39
40 56. Stuart S.N., Wilson E.O., McNeely J.A., Mittermeier R.A., Rodriguez J.P. 2010
41 The Barometer of Life. *Science* **328**(5975), 177-. (doi:10.1126/science.1188606).
42
43
44
45 57. Baillie J.E.M., Collen B., Amin R., Akcakaya H.R., Butchart S.H.M., Brummitt
46 N., Meagher T.R., Ram M., Hilton-Taylor C., Mace G.M. 2008 Toward monitoring
47 global biodiversity. *Conservation Letters* **1**(1), 18-26.
48
49
50
51 58. Fielding A.H., Bell J.F. 1997 A review of methods for the assessment of
52 prediction errors in conservation presence/absence models. *Environ Conserv* **24**(1),
53 38-49.
54
55
56
57
58
59
60

Figure Legends

Fig. 1. Explanatory and discriminative power of the fitted models of threat status for assessed species. (A), the relationship between fitted threat probability and IUCN status for assessed species. Threat probability is the probability a species is in one of the ‘threatened’ categories according to our spatial-phylogenetic multi-predictor model. (B), Receiver Operator Characteristic (ROC) curve, showing the relationship between true positive (sensitivity) and false positive (1 minus specificity) rate. The dashed line is the expected pattern if the threat probabilities were no better than random at discriminating threatened species. The AUC, which varies between 0.5 and 1, is the area highlighted in grey and is a measure of explanatory power. (C), the frequency distribution of fitted probabilities for species of contrasting conservation status. The green bars refer to species of ‘least concern’ (IUCN status 1 in *A*), whilst the red bars refer to species which are ‘critically endangered’ (IUCN status 5 in *A*). (D) Fitted / predicted threat probabilities shown separately for assessed species (grey) and data-deficient species (red).

Fig. 2. Prediction of threats for individual mammal orders. For each order the average fitted threat probabilities for assessed species (black points) and predicted threat probabilities for data-deficient species (red points) are shown (\pm standard errors). F-ratios and p-values refer to tests of differences between the mean fitted threat probabilities of assessed and data-deficient species. The numbers of assessed species is given for each order, together with the number of data-deficient and threatened (i.e. not ‘least concern’) species. The grey vertical bars show the threshold threat probability for each order (see Table S1), which is used to denote which data-deficient species are predicted to be threatened. The threshold is the point at which sensitivity = specificity (where

1
2
3 threatened and non-threatened status have equal chance of being correctly predicted [58]). Based
4
5 on this probability, the final column gives the number of data-deficient species which are predicted
6
7 to be threatened. Note that there are no data-deficient species in Perrissodactyla and the orders is
8
9 thus not included here.
10
11
12
13
14
15

16 **Fig. 3. The geography of observed and predicted mammal threat levels and richness.** Panels
17
18 illustrate the observed and predicted grid-cell proportions of all species assessed by IUCN to be
19
20 threatened and analyzed here (A,B, 3,703 species, for model details see Table S1), and the
21
22 predicted proportion of data-deficient species threatened (C, 483 species, for details see Fig 2).
23
24 (D-F) show observed and predicted richness patterns: the richness of observed (assessed)
25
26 threatened species (D, 812 of 3,703 species in analysis, i.e. all those assessed “vulnerable”,
27
28 “endangered” or “critically endangered”), those data-deficient species predicted by the combined
29
30 spatial, phylogenetic and environmental model to be non-threatened (E, N = 152 of 483 species),
31
32 and those predicted to be threatened (F, 331 of 483 species). Quantile classification of values
33
34 across 110km equal area grid cells in Behrman projection. Note that color scales vary to emphasize
35
36 geographic differences.
37
38
39
40
41
42
43
44
45

46 **Fig. 4. Relationships between observed and predicted threat levels of grid cell assemblages.**

47
48 The model-based predictions of average probability (A) and total proportion (B) of species
49
50 threatened successfully captures the observed variation in proportion species threatened (A,
51
52 Spearman $r = 0.72$; B, $r = 0.66$; 3,703 species; cf. Fig 3A, B). Observed and predicted richness of
53
54 threatened assessed species is tightly associated (C, $r = 0.77$, cf. Fig 3D). In contrast, the predicted
55
56 average threat levels and proportions of data-deficient species (cf. Fig 3C) show only very weak
57
58
59
60

1
2
3 association with the proportional threat patterns of assessed species (D, $r = 0.33$, E, $r = 0.23$; 843
4 data-deficient species). Equally, the areas with high richness of data-deficient species predicted to
5 be threatened shows little covariance with those of high assessed threatened richness (F, $r = 0.31$,
6 cf. Fig 3F). Darker gray represent higher density of points, line indicates 1:1 relationship. A total
7 of 11,331 110km equal area grid cells that had $\geq 50\%$ dry land or were oceanic islands and had \geq
8 2 assessed species were analyzed.
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

For Review Only

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47

Figure 1







