Overcoming data deficiency in reptiles

Lucie M. Bland[a, *]

l.bland@unimelb.edu.au

Monika Böhm[b]

[a]Centre of Excellence for Environmental Decisions, School of BioSciences, The University of Melbourne, Parkville 3010, VIC, Australia

[b]Institute of Zoology, Zoological Society of London, Regent's Park, London NW1 4RY, United Kingdom

*Corresponding author.

Abstract

We have no information on the risk of extinction of 21% of reptiles listed as Data Deficient on the Sampled Red List Index (SRLI), an indicator developed to track global change in species status. Data Deficient species are of high research priority, because they contribute to uncertainty in estimates of extinction risk and are neglected by conservation programmes. We review the causes of data deficiency in reptiles; the likely status of Data Deficient reptiles; and possible solutions for their re-assessment. We find that 52% of Data Deficient reptiles lack information on population status and trends, and that few species are only known from type specimens and old records. We build a random forest model for SRLI species of known extinction risk, based on life-history, environmental and threat information. The final model shows perfect classification accuracy (100%) in ten-fold cross validation. We use the model to predict that 56 of 292 Data Deficient reptiles (19%) are at risk of extinction, so the overall proportion of threatened reptiles in the SRLI (19%) remains unchanged. Regions predicted to contain large numbers of threatened Data Deficient reptiles overlap with known centres of threatened species richness. However, the model shows lower accuracy (79%) on 29 species recently re-assessed in the Global Reptile Assessment. Predictive models could be used to prioritize Data Deficient species and reptiles not included in the SRLI, and new reptile assessments could be used to improve model predictions through adaptive learning.

Keywords: Data Deficient; Extinction risk; IUCN Red List; Random forests; Reptiles

1.1 Introduction

Reptiles are one of the most diverse groups of terrestrial vertebrates, with 10,272 species described as of August 2015 (Uetz and Hošek, 2015). Hundreds of reptile species are still being discovered per year (Uetz and Hošek, 2015) on par with rapid species discoveries in amphibians (Köhler et al., 2005). Yet as a group, reptiles are poorly represented on The IUCN Red List of Threatened Species™, with only 44% of described species evaluated (IUCN, 2015). Within those, 1500 were assessed in a representative manner using the Sampled Red List Index (SRLI) method (Baillie et al., 2008), producing the first global analysis of the extinction risk of reptiles. This method selects species randomly from the species list, producing a broadly representative picture of extinction risk status for reptiles at the order and family level (Böhm et al., 2013). The analysis revealed that one in five reptile species is threatened with extinction, with risk highest among freshwater species and in tropical regions (Böhm et al., 2013).

Within the reptile SRLI assessment, 318 species were assessed as Data Deficient (Böhm et al., 2013) due to insufficient information on species' taxonomy, ecology, distribution, population trends, and/or threats (IUCN, 2001). The Data Deficient category does not correspond to a level of extinction risk, but indicates that further information should be collected to assign species to a risk category (IUCN, 2001). High levels of data deficiency within group assessments are problematic for a number of reasons. First, ignoring Data Deficient species (or treating them as threatened or non-threatened) contributes to considerable uncertainty in extinction risk patterns (Bland et al., 2012; Butchart and Bird, 2010; Hoffmann et al., 2010). This uncertainty not only affects the monitoring of progress towards global biodiversity targets (e.g. Convention on Biological Diversity (CBD) Aichi targets), but also conservation priorities which rely on threatened species lists –such as Key Biodiversity Areas, biodiversity hotspots, and many others (Brooks et al., 2006). Data Deficient species are rarely included in national recovery plans, conservation legislation, conservation planning (Sousa-Baena et al., 2013), and investment schemes. For example, only 2% of the awards from the Mohamed Bin Zayed Species Conservation Fund (MBZSC, 2014) and only one project of the World Association of Zoos and Aquaria (WAZA, 2013) exclusively focus on Data Deficient species. All in all, Data Deficient reptiles are offered very little protection and funding due to their uncertain extinction risk.

Reptiles included in the SRLI show intermediate levels of data deficiency (21%; Böhm et al., 2013). Data Deficiency is higher than in birds (0.6%; Butchart and Bird, 2010) and mammals (15%; Schipper et al., 2008), but lower levels than in amphibians (25%; Stuart et al., 2004), Odonata (35%; Clausnitzer et al., 2009), and freshwater crabs (49%; Cumberlidge et al., 2009). Data Deficient reptiles are also better-known than Data Deficient species in other groups (Bland et al., submitted). The reasons for Data Deficient status in reptiles often quote uncertain population status (33%) and uncertain threats (23%). Comparatively few listings are based on severe sources of uncertainty, namely type records (10%), few records (12%), old records (6%), or records of uncertain provenance (4%). These severe sources of uncertainty collectively represent a higher proportion of Data Deficient listings in freshwater crabs (93%), amphibians (43%), and mammals (42%) than in reptiles (32%; Bland et al., submitted). Some reptiles are listed as Data Deficient due to taxonomic uncertainty (6%) and recent discovery within 10 years of the SRLI assessment (6%). Lack of knowledge of reptiles' threat status is not only exemplified by Data Deficient species; indeed more than half of all known reptile species remain Not Evaluated (IUCN, 2015). Addressing data deficiency and broader data gaps are therefore key challenges for reptile assessments.

Data Deficient species have received increased interest from the conservation literature in recent years, with studies investigating the rationale for the use of the category (Butchart and Bird, 2010; Sousa-Baena et al., 2013), the effect of Data Deficient species on conservation priorities (Trindade-Filho et al., 2012), and their potential for informing future biodiversity inventories (Brito, 2010). Considerable progress has been made in predicting the likely status of Data Deficient species based on contextual data (e.g. biology, ecology, phylogeny, environment, and threats). For example, distribution maps are available for most Data Deficient species and can be used to quantify risk factors, such as geographical range size and exposure to anthropogenic threats. These contextual data alone are insufficient for making a decision on formal Red List status, but are available for a wide range of groups and have been used to predict risk in plants (Good et al., 2006; Walker, 2014), mammals (Bland et al., 2015a; Davidson et al., 2009; Jones and Safi, 2011), amphibians (Bland et al., 2015b; Morais et al., 2013), and crayfish (Bland et al., 2015b). Simple rules of thumb based on one or two factors may not capture complex extinction risk dynamics, as strong correlations between factors influencing both knowledge availability and extinction risk may lead to unreliable estimates of risk. For example, Data Deficient reptiles typically possess small geographic ranges (Bland, 2014) and may be more likely to be threatened (Böhm et al., 2015), but Data Deficient reptiles also tend to be small-bodied (Bland, 2014; Vilela et al., 2014), a factor usually associated with low extinction risk (Cardillo and Meijaard, 2012). Predicting risk in Data Deficient must therefore rely on a wide variety of contextual data and robust modelling approaches.

Predictive studies of risk in Data Deficient species span a diversity of methods, such as spatial-phylogenetic frameworks, rule-based methods, and machine learning (Bland et al., 2015b; Jones and Safi, 2011; Morais et al., 2013). Machine learning methods are powerful tools for finding patterns in large datasets and rely on few assumptions. This had made machine learning increasingly popular in ecology (Cutler et al., 2007; Prasad et al., 2006) and extinction risk analyses (Bland et al., 2015a; Davidson et al., 2009). A previous study highlighted the usefulness of random forests for predicting extinction risk in different taxonomic groups, including reptiles (Bland et al., 2015b). Random forests are an ensemble method related to classification trees, where many classification trees are constructed and predictions obtained by a majority vote (Breiman, 2001). For each tree, only a randomly chosen subset of the explanatory variables is used at each node, which reduces correlation between trees. Random forests are widely used in ecology due to their high predictive power and their robustness to overfitting and noise (Cutler et al., 2007). In addition, the outputs of random forests are probability estimates of a given outcome, which allow easy interpretation of uncertainty when predicting complex processes such as extinction risk.

Overcoming data deficiency will be a costly exercise (Bland et al., 2015b): research needs to be carried out on the key aspects for which species information is missing (e.g. population status, threats, taxonomy), and this new evidence will need to be synthesized in assessments. Prioritizing Data Deficient species for re-assessment will require information on both their likely risk status and an understanding of the underlying causes of this data deficiency. We predict extinction risk of Data Deficient species in the random representative sample of 1,500 reptiles by Böhm et al. (2013) (hereafter termed 'SRLI assessment') as a first step to reducing uncertainty in extinction risk patterns. We ask the following questions: i) Can a random forest model predict risk in species of known extinction risk (non-Data Deficient)?; ii) What is the predicted level of extinction risk faced by Data Deficient species?; and iii) How can our findings direct re-assessments?

2.2 Methods

2.1 2.1 Dataset

We used a published extinction risk trait database on 1,416 terrestrial reptile species (Appendix A; Bland et al., 2015b) included in the SRLI assessment (Böhm et al., 2013). We did not collect data for non-SRLI species. No species of crocodiles were assessed as Data Deficient so predictions for this taxon cannot be made. We therefore excluded four crocodile species, resulting in 1,412 species in the dataset. Data were compiled from species descriptions, field guides, museum specimens and published life-history studies, and supplemented with data obtained from expert herpetologists during the IUCN Red List assessment process. References and raw data are available in Appendices A and B, respectively. The dataset contained the following life-history and ecological variables: maximum body size (snout-vent length), reproductive mode, habitat mode, trophic level, island presence, and number of IUCN-listed habitats. Information on species' niche and threat exposure was available as mean values from within species' geographic ranges for: mean annual temperature, temperature seasonality (coefficient of variation), annual precipitation, precipitation seasonality, minimum elevation (Hijmans et al., 2005), Human Footprint (CIESIN, 2005a), and mean and minimum human population density for the year 2000 (CIESIN, 2005b). Geographical range size and latitude of

range centroid were computed from the IUCN distribution maps of each species. We did not undertake variable selection as uninformative variables are unlikely to affect predictive performance in analyses with fewer variables than species (low dimension problems; Kuhn, 2008). We defined non-Data Deficient species on the IUCN Red List as threatened (IUCN categories Critically Endangered, Endangered or Vulnerable) or non-threatened (IUCN categories Near Threatened or Least Concern).

2.2.2.2 Modelling

We used random forests to predict the extinction risk (non-threatened or threatened) of data-sufficient species. Random forests cannot currently take into account phylogenetic relatedness among species, so we included taxonomic order, family, and genus as predictor variables to partially account for shared evolutionary history. All numeric variables were centred to a mean of zero and scaled to a standard deviation of one (Kuhn, 2008). All categorical variables were transformed to orthogonal dummy variables. We removed highly correlated ($r > 0.9$) and low variance (frequency ratio > 999 and unique value percentage < 0.0001) variables, which can lead to collinearity and zero variance in cross-validation partitions (Kuhn, 2008). We used ten-fold cross validation to optimize the number of variables chosen randomly at each node for each tree, and grew 500 trees for each random forest iteration. We maximized the area under the receiver operating characteristic curve (AUC) to measure classification performance as AUC is insensitive to unequal numbers of threatened and non-threatened species in training data, and does not require the specification of misclassification costs (Fawcett, 2006). We identified the probability threshold above which a species was identified as threatened by maximizing the Youden index ($Y = $ sensitivity $+$ specificity $- 1$; Youden, 1950), which lends equal weight to detecting threatened and non-threatened species (Bland et al., 2015a). All analyses were conducted in R version 3.2.2 with the *caret* package (Kuhn, 2008). For further details see Appendix A.

Completeness of the predictor variables in our analysis ranged between 96.2 and 100% (Appendix A). Random forests can impute missing data with proximity-weighted averages, but this process may lead to uncertainty in predictions (Breiman, 2003). We therefore conducted the analysis with two datasets (Table 1): i) complete dataset (removing species with missing data); and ii) imputed dataset (imputing missing data). Model performance was identical for the two datasets, so we used the imputed dataset for all further analyses to maximize sample size. Results for the complete dataset are reported in Appendix A.

Table 1 Summary statistics for the Sampled Red List Index assessment, the complete dataset, and the imputed dataset. The percentage of threatened species is calculated for data-sufficient species (non-threatened $+$ threatened).

alt-text: Table 1

	SRLI assessment	Complete dataset	Imputed dataset
Total species	1473	1298	1412
Non-threatened species	939	861	915
Threatened species	223	172	205
Data Deficient species	313	265	292
Percent threatened	19.1	16.7	18.3
Percent Data Deficient	21.2	20.4	20.7

2.3.2.3 Predictions for Data Deficient species

We predicted the status of 292 Data Deficient terrestrial species for which we obtained trait data. We could not obtain data for 21 species. We created threatened species richness maps with a grid of 21,583 hexagons, each with an equal area of 23,529 km^2. The resolution was selected to obtain a reasonable number of species in each cell for congruence and spatial regression analyses, and to reflect the limited accuracy of global range maps (Hurlbert and Jetz, 2007).

We compiled species' geographical range maps for three sets of species: i) SRLI species listed as threatened (observed threatened species richness; 218 species); ii) Data Deficient species predicted by the model to be threatened (56 species); and iii) the predicted number of threatened species (274 species), including threatened SRLI species and Data Deficient species predicted to be threatened. For each set, we selected cells with threatened species richness equal to or greater than one to prevent zero inflation in our models. We quantified the spatial correlation (Tjøsthein, 1978) and spatial congruence in centres of observed threatened species richness and predicted threatened richness. We then used spatial autoregressive models of predicted species richness as a function of observed species richness with intercept fixed to zero, and tested for a regression slope different from one.

Some of the species originally included in the SRLI assessment by Böhm et al. (2013) have since been reassessed in expert workshops as part of the Global Reptile Assessment. For our 292 Data Deficient species, we compared our model predictions against changes in Red List status due to reassessment in the Global Reptile Assessment. We computed the number of species retaining Data Deficient status after reassessment. For species which changed status during reassessment, we investigated the congruence between predicted and reassessed status.

3.3 Results

The model achieved perfect predictive performance (AUC: 1; Kappa: 1; accuracy: 1; sensitivity: 1; specificity: 1), indicating perfect discriminability between non-threatened and threatened SRLI reptiles. The predicted probability of risk threshold obtained by maximizing the Youden index (Y: 1) was 0.614. The model predicted 56 Data Deficient species (19.2%) to be threatened with extinction, so the overall percentage of threatened reptiles in the SRLI assessment (19%) remained unchanged. The percentage of threatened species did not change for Sauria and Serpentes, remaining at 21% and 12% respectively. The percentage of threatened species decreased slightly for Testudines (47% threatened compared to 51% in Böhm et al. (2013)), and increased greatly for Amphisbaenians (21% threatened compared to 7% in Böhm et al. (2013)). Families with high numbers of potentially threatened Data Deficient species included Scincidae (12 species), Dipsadidae (7 species), and Amphisbaenidae (5 species). Regions with high numbers of potentially threatened Data Deficient species included Madagascar (9 species), Brazil (7 species), India (6 species), Peru (5 species), Indonesia (4 species), and Venezuela (4 species) (Fig. 1). The proportion of species predicted to be at risk varied among reasons for listing as Data Deficient (Fig. 2), which may reflect genuine differences in risk or uncertainty in contextual data.
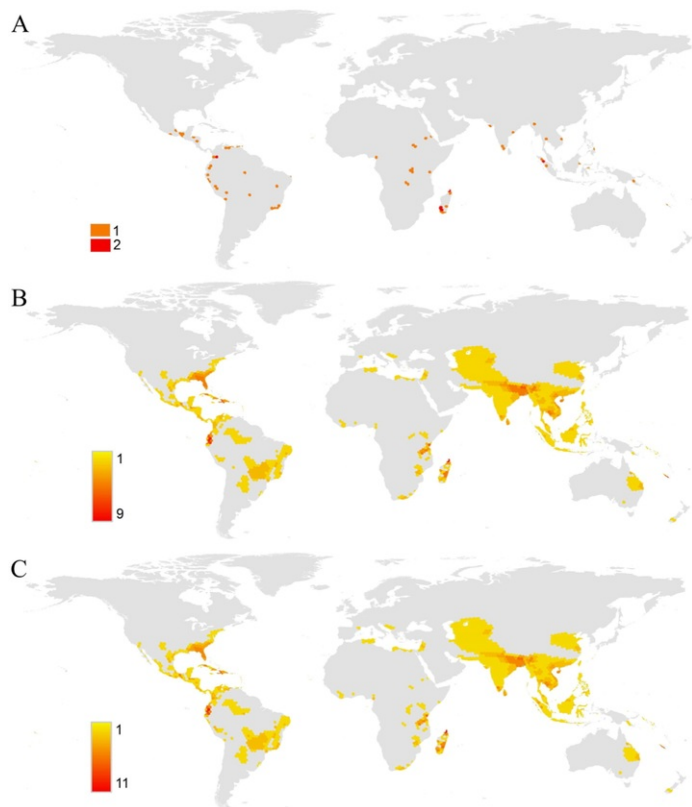

Figure 1. Fig. 1 A) Number of Data Deficient SRLI reptiles predicted to be threatened (n = 56); B) Number of threatened SRLI reptiles (n = 218); C) Number of predicted threatened Data Deficient species and threatened species in the reptile SRLI (n = 274).
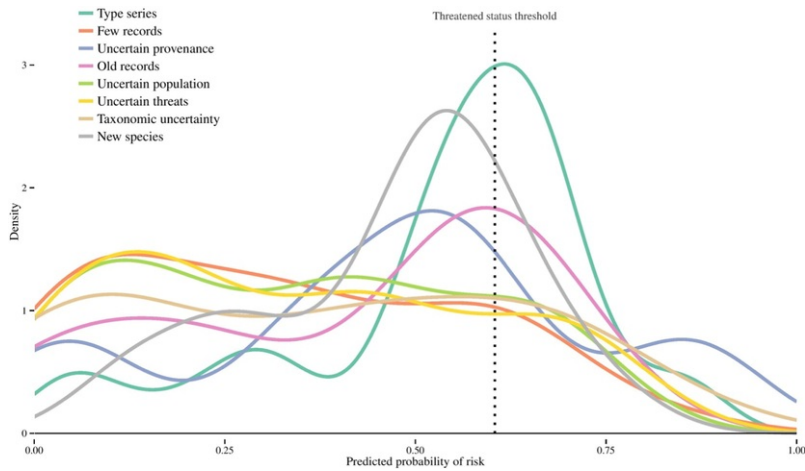
alt-text: Fig. 1

Distribution of predicted probability of risk for 292 Data Deficient reptiles assigned to each justification tag. We assigned eight justifications tags to denote the reason(s) why a species was listed as Data Deficient. The threshold best classifies data-sufficient species as threatened or non-threatened based on the predicted probability of risk. Definitions and sample sizes for each justification are available in Appendix A.

alt-text: Fig. 2

The spatial correlation (Tjøsthein coefficient: 0.219) and the congruence (Fig. S7) between observed and predicted centres of threatened species richness were high. The top 1% of cells containing the highest observed threatened species richness captured 89% of the predicted threatened species richness. The congruence increased to 98% for the top 25% of cells containing the highest observed threatened species richness. Observed levels of risk have not been underestimated according to our regression model of predicted vs. observed threatened species richness (testing for slope $\neq$ 1: slope: 1.003; p = 0.65; $\chi^2_{1, 1534}$ = 0.202).

Sixty Data Deficient species with predicted risk status have already been reassessed and published on the IUCN Red List as part of the Global Reptile Assessment. Thirty one species retained their Data Deficient status — our model predicted the majority of these (23) to be non-threatened. Of the remaining 29 reassessed species, 21 were reassessed as non-threatened and were correctly classified by our model. Eight species were reassessed as threatened, but we only predicted two of these to be at risk. However, seven of these eight species were predicted to have a high probability of risk greater (> 0.41). Our model correctly classified 79.3% of re-assessed Data Deficient species, including 100% of non-threatened species and 25% of threatened species.

4.4 Discussion

Predicting the status of Data Deficient, Not Evaluated and other poorly-known species is a key challenge for conservation planning and global monitoring. We demonstrate that extinction risk can be well predicted in SRLI reptiles based on intrinsic (life-history and ecology) and extrinsic (niche and threat exposure) attributes. This is an encouraging result given the financial and temporal constraints operating on the red listing process (Rondinini et al., 2014). Extinction risk levels estimated with species currently listed as threatened in the SRLI and excluding Data Deficient species (19%) do not differ from those estimated with Data Deficient species predictions (19%). Known centres of threatened species richness are highly congruent with predicted centres (Fig. 1). Our findings suggest that known extinction risk patterns are robust to data deficiency.

Distribution of predicted threatened species reflects the distribution of recognised hotspots of threatened species in need of effective conservation planning (e.g., biodiversity hotspots; Myers et al., 2000), such as the Tropical Andes, Mesoamerica, Sundaland, and Madagascar. We predicted the highest number of threatened Data Deficient reptiles in Madagascar, a country recently highlighted as a centre of high extinction risk for reptiles (39% threatened; Jenkins et al., 2014). Similarly, our model predicted large numbers of threatened Data Deficient species in the highly threatened Atlantic forest and restinga habitats in Brazil (Böhm et al., 2013).

The snake family Dipsadidae contributed one of the largest numbers of predicted threatened species in our analysis. This family was one of the few snake families to not be significantly over-threatened in earlier analyses (Böhm et al., 2013), suggesting that risk levels in snakes may have been underestimated. Estimation of risk levels in amphisbaenians was previously highly uncertain, with 7–54% threatened depending on the treatment of Data Deficient species as either non-threatened or threatened (Böhm et al., 2013). Our model predicts that amphisbaenians are threatened in the same proportion as lizards. Predictive

models may thus be of particular importance for filling data gaps for species that are difficult to observe, such as fossorial and arboreal species.

Our extinction risk model has great potential for conservation applications. It could be used to cost-effectively re-assess Data Deficient reptiles, by prioritizing species with high predicted risk probabilities, low action costs, and high probability of action success (Joseph et al., 2009). Actions required to re-assess a species known from a type specimen collected a hundred years ago will differ from those required for relatively well-known species for which information on threats is uncertain. Prioritizing species for re-assessment with both high predicted risk and large amounts of information already available (e.g. species with only population trends or threat information missing) may show the lowest cost and highest probability of success. Data Deficient reptiles are relatively well-known compared to Data Deficient species from other groups (Bland et al. submitted), so re-assessing reptiles is likely to be less costly, faster, and more successful.

Our extinction risk model compares favourably with models built for other groups. Models for mammals range in sensitivity from 47.7% (Davidson et al., 2009) to 93.5% (Bland et al., 2015a). Similarly, Tingley et al. (2013) obtained R^2 values of 0.54–0.68 in models applied to New Zealand lizards, whilst models applied to other groups tend to show low R^2 values (Cardillo et al., 2006; Cardillo and Meijaard, 2012). Our random forest model performed better than a similar model for reptiles (Bland et al., 2015b), as we did not attempt inter-model comparisons and therefore made more efficient use of training data. Why models vary in predictive performance within and among groups depends on data quality and quantity; the link between the available data and extinction risk; and the modelling method — factors that can be difficult to disentangle. Good performance of our model may be linked to the quality of Red List assessments, as consistent application of the IUCN Red List Categories and Criteria (IUCN, 2001) is likely to help build highly predictive models. Incorrect predictions of risk in mammals have been associated with charismatic and contentious species (e.g. cheetah), and with species that have subsequently changed status due to non-genuine reasons, such as increased information availability (Bland et al., 2015a). A consistent and transparent Red Listing process is therefore likely to improve the usefulness of predictive extinction risk modelling.

Performance may also be linked to the strength of the relationship between predictor variables and extinction risk. Range size is the dominant variable explaining extinction risk in reptiles (Böhm et al., 2015) and for species with small ranges, accessibility of the range and habitat specialization become increasingly important in determining risk (Böhm et al., 2015). This highlights the interaction between multiple factors influencing risk, a process that can be easily captured by tree-based machine learning such as random forests. Geographical range size is used to classify species as threatened in the IUCN Red List, so its use in explanatory models of extinction risk can provide biased parameter estimations. However in a predictive context, geographical range size is a key variable to accurately predict risk, and degrading the quality of species range maps leads to poor model predictions (Bland et al., 2015b).

Applying any model relies on assumptions that may contribute to uncertainty in predictions for Data Deficient species. We assume that the relationship between predictor variables and extinction risk (fitted model) is similar for both data-sufficient and Data Deficient species. We show that the distribution of explanatory variables for data-sufficient species spans that for Data Deficient species (Figs. S1 and S3), but biases in variable distributions may lead to more uncertain predictions for species possessing traits not well represented in the training set, such as smaller bodies (González-Suárez et al., 2012).

The accuracy of data available for Data Deficient species can also be problematic. Limited sampling may cause systematic under-estimation of range size for Data Deficient species, which in turn may lead to over-estimation of risk. The majority of species predicted to be at risk (34 out of 56) were listed as Data Deficient on the basis of being known only from their type specimen(s), and thus having small known distributions. However, it appears that small species distribution will not automatically lead to threatened classifications. The overall level of risk in Data Deficient species (19.2%) is similar to levels found in data-sufficient species (19.1%), suggesting an upper limit to the over-estimation of extinction risk in the dataset. Of the 31 species which remained Data Deficient in the Global Reptile Assessment workshops, our model predicted 23 not to be threatened. Most of these species (12 out of 23) remained Data Deficient as they were only known from type specimens, few specimens, or the type locality.

Whilst our model perfectly predicts the status of data-sufficient species, we misclassify 20.7% of Data Deficient species re-assessed under the Global Reptile Assessment. It is expected that models will perform more poorly on new data compared to validation data. However, because re-assessed species in the Global Reptile Assessment are not a random sample of Data Deficient species, the misclassification rate may not be applicable to other Data Deficient species or non-SRLI species. As information on more (or all) species from the Global Reptile Assessment becomes available, models of extinction risk could be refined through adaptive learning and eventually applied to species not yet assessed by IUCN (e.g. newly discovered species). This highlights the need for true independent data to validate extinction risk models, rather than calibrating models on a single dataset (e.g. SRLI) with cross-validation. Unfortunately, the speed at which IUCN Red List assessments are updated and the inherent biases of many Red List assessments may limit advancements in adaptive learning.

Misclassifications can provide clues for improving the model. All six species re-assessed as threatened under the Global Reptile Assessment but not predicted to be at risk by our model were re-assessed based on small range size and continuing decline in extent and quality of habitat, processes that our model may not capture adequately. Our anthropogenic threat variables (e.g. Human Population Density and Human Footprint) may be poor proxies of threat, or may not be of sufficient spatial resolution to capture changes in reptile habitats since many species primarily rely on microhabitats or seasonal habitats (Michael et al., 2015). This highlights the need for future work to focus on the appropriateness of emerging data on spatial habitat change (e.g. forest loss; Hansen et al., 2013), including spatial data on climate change (Böhm et al.

this issue) and invasive species-. Threat-specific data may allow improved prediction of extinction risk, for example in island species which are disproportionately affected by invasive species (30% are affected by those vs. 12% on continents). At the same time, there is a need to regularly update existing spatial data layers to allow effective assessment of threats over time (Joppa et al., 2016).

Our study did not consider freshwater and marine reptiles, which globally exhibit higher levels of risk but lower levels of data deficiency than terrestrial reptiles (Böhm et al., 2013). However, some marine reptiles show high levels of data deficiency (e.g. 34% in marine elapid snakes; Elfes et al., 2013) and levels of data deficiency in the entire Red List are twice as high in marine groups than non-marine groups (Webb and Mindel, 2015). Developing predictive models of risk for marine Data Deficient species is a priority, and will rely on the quantification of anthropogenic threats relevant to marine systems (Halpern et al., 2008).

Extinction risk models based on two categories (threatened vs. non-threatened) can be highly predictive, yet distinct IUCN Red List categories are needed for calculating the Red List Index, monitoring progress towards Aichi targets (Convention on Biological Diversity, 2010), and for finer scale conservation prioritization. This can be achieved by selecting probability thresholds representing category boundaries (Fig. S2 and S4), but a key challenge is to directly model five risk categories when these are highly imbalanced (e.g. 842 Least Concern vs. 22 Critically Endangered species in our dataset). This could be achieved through improved machine learning (e.g. resampling or anomaly detection) or alternative modelling methods (e.g. cumulative link models; Luiz et al., 2016).

This study provides an encouraging step forward for assessing reptile biodiversity. We find that extinction risk is highly predictable in SRLI reptiles, and that known patterns of risk are robust to data deficiency. The real challenges lie in assessing all reptile species globally, and fully understanding and mitigating drivers of risk. The results of this study should be interpreted in the context of the SRLI, which is based on a limited sample of reptile biodiversity. Whilst the representativeness of the Sampled Red List approach has been demonstrated at the taxonomic level (Baillie et al., 2008), the representativeness of spatial patterns derived from the approach remains unproven. Assessing all reptiles under the Global Reptile Assessment will improve our understanding of drivers and spatial patterns of extinction risk, and enhance our capacity to predict risk across reptile biodiversity.

Acknowledgements

Appendix A. Appendix A. Supplementary data

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.biocon.2016.05.018.

References

Baillie J.E.M., Collen B., Amin R., Akcakaya H.R., Butchart S.H.M., Brummitt N., Meagher T.R., Ram M., Hilton-Taylor C. and Mace G.M., Toward monitoring global biodiversity, Conservation LettersConserv. Lett. **1**, 2008, 18–26.

Bland L.M., Resolving the eEffects of Data Deficient sSpecies on the eEstimation of eExtinction rRisk, 2014, Imperial College London; London, UK, 158.

Bland L.M., Collen B., Orme C.D.L. and Bielby J., Data uncertainty and the selectivity of extinction risk in freshwater invertebrates, Diversity and DistributionsDivers. Distrib. **18**, 2012, 1211–1220.

Bland L.M., Collen B., Orme C.D.L. and Bielby J., Predicting the Cconservation Sstatus of Ddata-Ddeficient Sspecies, Conservation BiologyConserv. Biol. **29**, 2015a, 250–259.

Bland L.M., Orme C.D.L., Bielby J., Collen B., Nicholson E. and McCarthy M.A., Cost-effective assessment of extinction risk with limited information, Journal of Applied EcologyJ. Appl. Ecol. **52**, 2015b, 861–870.

Bland L.M., Bielby J., Kearney S., Orme C.D.L., Watson J.E.M. and Collen B., Towards reassessing Data Deficient species, Conserv. Biol. 2016, (Submitted).

Böhm M., Collen B., Baillie J.E.M., Bowles P., Chanson J., Cox N., Hammerson G., Hoffmann M., Livingstone S.R., Ram M., Rhodin A.G.J., Stuart S.N., van Dijk P.P., Young B.E., Afuang L.E., Aghasyan A., García A., Aguilar C., Ajtic R., Akarsu F., Alencar L.R.V., Allison A., Ananjeva N., Anderson S., Andrén C., Ariano-Sánchez D., Arredondo J.C., Auliya M., Austin C.C., Avci A., Baker P.J., Barreto-Lima A.F., Barrio-Amorós C.L., Basu D., Bates M.F., Batistella A., Bauer A., Bennett D., Böhme W., Broadley D., Brown R., Burgess J., Captain A., Carreira S., Castañeda M.D.R., Castro F., Catenazzi A., Cedeño-Vázquez J.R., Chapple D.G., Cheylan M., Cisneros-Heredia D.F., Cogalniceanu D., Cogger H., Corti C., Costa G.C., Couper P.J., Courtney T., Crnobrnja-Isailovic J., Crochet P.-A., Crother B., Cruz F., Daltry J.C., Daniels R.J.R., Das I., de Silva A., Diesmos A.C., Dirksen L., Doan T.M., Dodd C.K., Doody J.S., Dorcas M.E., Duarte de Barros Filho J., Egan V.T., El Mouden E.H., Embert D., Espinoza R.E., Fallabrino A., Feng X., Feng Z.-J., Fitzgerald L., Flores-Villela O., França F.G.R., Frost D., Gadsden H., Gamble T., Ganesh S.R., Garcia M.a., García-Pérez J.E., Gatus J., Gaulke M., Geniez P., Georges A., Gerlach J., Goldberg S., Gonzalez J.-C.T., Gower D.J., Grant T., Greenbaum E.,

Grieco C., Guo P., Hamilton A.M., Hare K., Hedges S.B., Heideman N., Hilton-Taylor C., Hitchmough R., Hollingsworth B., Hutchinson M., Ineich I., Iverson J., Jaksic F.M., Jenkins R., Joger U., Jose R., Kaska Y., Kaya U., Keogh J.S., Köhler G., Kuchling G., Kumlutaş Y., Kwet A., La Marca E., Lamar W., Lane A., Lardner B., Latta C., Latta G., Lau M., Lavin P., Lawson D., LeBreton M., Lehr E., Limpus D., Lipczynski N., Lobo A.S., López-Luna M.a., Luiselli L., Lukoschek V., Lundberg M., Lymberakis P., Macey R., Magnusson W.E., Mahler D.L., Malhotra A., Mariaux J., Maritz B., Marques O.a.V., Márquez R., Martins M., Masterson G., Mateo J.a., Mathew R., Mathews N., Mayer G., McCranie J.R., Measey G.J., Mendoza-Quijano F., Menegon M., Métrailler S., Milton D.a., Montgomery C., Morato S.a.a., Mott T., Muñoz-Alonso A., Murphy J., Nguyen T.Q., Nilson G., Nogueira C., Núñez H., Orlov N., Ota H., Ottenwalder J., Papenfuss T., Pasachnik S., Passos P., Pauwels O.S.G., Pérez-Buitrago N., Pérez-Mellado V., Pianka E.R., Pleguezuelos J., Pollock C., Ponce-Campos P., Powell R., Pupin F., Quintero Díaz G.E., Radder R., Ramer J., Rasmussen A.R., Raxworthy C., Reynolds R., Richman N., Rico E.L., Riservato E., Rivas G., da Rocha P.L.B., Rödel M.-O., Rodríguez Schettino L., Roosenburg W.M., Ross J.P., Sadek R., Sanders K., Santos-Barrera G., Schleich H.H., Schmidt B.R., Schmitz A., Sharifi M., Shea G., Shi H.-T., Shine R., Sindaco R., Slimani T., Somaweera R., Spawls S., Stafford P., Stuebing R., Sweet S., Sy E., Temple H.J., Tognelli M.F., Tolley K., Tolson P.J., Tuniyev B., Tuniyev S., Üzüm N., van Buurt G., Van Sluys M., Velasco A., Vences M., Veselý M., Vinke S., Vinke T., Vogel G., Vogrin M., Vogt R.C., Wearn O.R., Werner Y.L., Whiting M.J., Wiewandt T., Wilkinson J., Wilson B., Wren S., Zamin T., Zhou K. and Zug G., The conservation status of the world's reptiles, *Biol. Conserv.* **157**, 2013, 372–385.

Böhm M., Williams R., Bramhall H., McMillan K., Davidson A., Garcia-Aguayo A., Bland L., Bielby J. and Collen B., Correlates of extinction risk in squamate reptiles: the relative importance of biology, geography, threat and range size, *Glob. Ecol. Biogeogr.* **25**, 2015, 391–405.

Breiman L., Random forests, *Mach. Learn.* **45**, 2001, 5–32.

Breiman L., Manual setting up, using and understanding random forests v4. 0, http://oz.berkeley.edu/users/breiman.Using random forestsv4.0.pdf 2003.

Brooks T.M., Mittermeier R.a., da Fonseca G.a.B., Gerlach J., Hoffmann M., Lamoreux J.F., Mittermeier C.G., Pilgrim J.D. and Rodrigues a.S.L., Global biodiversity conservation priorities, *Science* **313**, 2006, 58–61.

Butchart S.H.M. and Bird J.P., Data Deficient birds on the IUCN Red List: what don't we know and why does it matter?, *Biol. Conserv.* **143**, 2010, 239–247.

Cardillo M. and Meijaard E., Are comparative studies of extinction risk useful for conservation?, *Trends Ecol. Evol.* **27**, 2012, 167–171.

Cardillo M., Mace G.M., Gittleman J.L. and Purvis A., Latent extinction risk and the future battlegrounds of mammal conservation, *Proc. Natl. Acad. Sci. U. S. A.* **103**, 2006, 4157–4161.

Clausnitzer V., Kalkman V.J., Ram M., Collen B., Bailie J.M., Bedjanic M., Darwall W.R.T., Dijkstra K.-D.B., Dow R., Hawking J., Karube H., Malikova E., Paulson D., Schutte K., Suhling F., Villanueva R.J., von Ellenrieder N. and Wilson K., Odonata enter the biodiversity crisis debate: the first global assessment of an insect group, *Biol. Conserv.* **142**, 2009, 1864–1869.

Convention on Biological Diversity, Aichi Targets, Nagoya, Japan, 2010.

Cumberlidge N., Ng P.K.L., Yeo D.C.J., Magalhães C., Campos M.R., Alvarez F., Naruse T., Daniels S.R., Esser L.J. and Attipoe F.Y.K., Freshwater crabs and the biodiversity crisis: importance, threats, status, and conservation challenges, *Biol. Conserv.* **142**, 2009, 1665–1673.

Cutler R.D., Edwards T.C., Beard K.H., Cutler A., Hess K.T., Gibson J. and Lawler J.J., Random forests for classification in ecology, *Ecology* **88**, 2007, 2783–2792.

Davidson A.D., Hamilton M.J., Boyer A.G., Brown J.H. and Ceballos G., Multiple ecological pathways to extinction in mammals, *Proc. Natl. Acad. Sci.* **106**, 2009, 10702–10705.

Elfes C.T., Livingstone S.R., Lane A., Lukoschek V., Sanders K.L., Courtney A.J., Gatus J.L., Guinea M., Lobo A.S. and Milton D., Fascinating and forgotten: the conservation status of the world's sea snakes, *Herpetol. Conserv. Biol.* **8**, 2013, 37–52.

Fawcett T., An introduction to ROC analysis, *Pattern Recogn. Lett.* **27**, 2006, 861–874.

González-Suárez M., Lucas P.M. and Revilla E., Biases in comparative analyses of extinction risk: mind the gap, *J. Anim. Ecol.* **81**, 2012, 1211–1222.

Good T.C., Zjhra M.L. and Kremen C., Addressing data deficiency in classifying extinction risk: a case study of a radiation of Bignoniaceae from Madagascar, *Conserv. Biol.* **20**, 2006, 1099–1110.

Halpern B.S., Walbridge S., Selkoe K.A., Kappel C.V., Micheli F., D'Agrosa C., Bruno J.F., Casey K.S., Ebert C. and Fox H.E., A global map of human impact on marine ecosystems, *Science* **319**, 2008, 948–952.

Hansen M.C., Potapov P.V., Moore R., Hancher M., Turubanova S., Tyukavina A., Thau D., Stehman S., Goetz S. and Loveland T., High-resolution global maps of 21st-century forest cover change, *Science* **342**, 2013, 850–853.

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., Darwall W.R.T., Dulvy N.K., Harrison L.R., Katariya V., Pollock C.M., Quader S., Richman N.I., Rodrigues A.S.L., Tognelli M.F., Vié J.-c., Aguiar J.M., Allen D.J., Allen G.R., Amori G., Ananjeva N.B., Andreone F., Andrew P., Luz A., Ortiz A., Baillie J.E.M., Baldi R., Bell B.D., Biju S.D., Bird J.P., Black-decima P., Blanc J.J., Bolaños F., Bolivar-g W., Burfield I.J., Burton J.A., Capper D.R., Castro F., Catullo G., Cavanagh R.D., Channing A., Chao N.L., Chenery A.M., Chiozza F., Clausnitzer V., Collar N.J., Collett L.C., Collette B.B., Fernandez C.F.C., Craig M.T., Crosby M.J., Cumberlidge N., Cuttelod A., Derocher A.E., Diesmos A.C., Donaldson J.S., Duckworth J.W., Dutson G., Dutta S.K., Emslie R.H., Farjon A., Fowler S., Garshelis D.L., Gerlach J., Gower D.J., Grant T.D., Hammerson G.A., Harris R.B., Heaney L.R., Hedges S.B., Kottelat M., Kovacs K.M., Kuzmin S.L., Marca E.L., Lewison R.L., Lichtenstein G., Livingstone S.R., Lukoschek V., Mallon D.P., McGowan P.J.K., McIvor A., Moehlman P.D., Molur S., Alonso A.M., Musick J.A., Nowell K., Nussbaum R.A., Olech W., Orlov N.L., Papenfuss T.J., Parra-olea G., Perrin W.F., Polidoro B.A., Pourkazemi M., Racey P.A., Ragle J.S., Ram M., Rathbun G., Reynolds R.P., Rhodin A.G.J., Richards S.J., Rodríguez L.O., Ron S.R., Rondinini C., Rylands A.B., Mitcheson Y.S.D., Sanciangco J.C., Sanders K.L., Santos barrera G., Schipper J., Self-sullivan C., Shi Y., Shoemaker A., Short F.T., Sillero-zubiri C., Silvano D.L., Smith K.G., Smith A.T., Snoeks J., Stattersfield A.J., Symes A.J., Taber A.B., Talukdar B.K., Temple H.J., Timmins R., Tobias J.A., Tsytsulina K., Tweddle D., Ubeda C., Valenti S.V., Dijk P.P.V., Veiga L.M., Veloso A., Wege D.C., Wilkinson M., Williamson E.A., Xie F., Young B.E., Akçakaya H.R., Bennun L., Blackburn T.M., Boitani L., Dublin H.T., Fonseca G.A.B., Gascon C., Lacher T.E., Jr., Mace G.M., Mainka S.A., McNeely J.A., Mittermeier R.A., Reid G.M., Rodriguez J.P., Rosenberg A.A., Samways M.J., Smart J., Stein B.A. and Stuart S.N., The ~~I~~impact of ~~C~~conservation on the ~~S~~status of the ~~W~~world~~'~~'s ~~V~~vertebrates, *Science* **330**, 2010, 1503–1509.

Hurlbert A.H. and Jetz W., Species richness, hotspots, and the scale dependence of range maps in ecology and conservation, ~~*Proceedings of the National Academy of Sciences of the United States of America*~~*Proc. Natl. Acad. Sci. U. S. A.* **104**, 2007, 13384–13389.

IUCN, IUCN Red List Categories and Criteria: ~~v~~Version 3.1, pp. 30, Gland, Switzerland and Cambridge, UK, 2001.

IUCN, The IUCN Red List of Threatened Species. Version 2015-3, 2015.

Jenkins R.K., Tognelli M.F., ~~Bowles P.~~, Cox N., Brown J.L., Chan L., Andreone F., Andriamazava A., Andriantsimanarilafy R.R. and Anjeriniaina M., Extinction ~~R~~risks and the ~~C~~conservation of Madagascar's ~~R~~reptiles, ~~*PLoS ONE*~~*PLoS One* **9** (8), 2014, e100173.

Jones K.E. and Safi K., Ecology and evolution of mammalian biodiversity, *Philos. Trans. R. Soc. B* **366**, 2011, 2451–2461.

Joppa L., O'Connor B., Visconti P., Smith C., Geldmann J., Hoffmann M., Watson J., Butchart S., Virah-Sawmy M. and Halpern B., Filling in biodiversity threat gaps, *Science* **352**, 2016, 416–418.

Joseph L.N., Maloney R.F. and Possingham H.P., Optimal allocation of resources among threatened species: a project prioritization protocol, ~~*Conservation Biology*~~*Conserv. Biol.* **23**, 2009, 328–338.

Köhler J., Rodriguez Vieites D., Vences M., Bonett R.M., Hita Garcia F., Glaw F. and Steinke D., New ~~A~~amphibians and ~~G~~global ~~C~~conservation: ~~A~~a ~~B~~boost in ~~S~~species ~~D~~discoveries in a ~~H~~highly ~~E~~endangered ~~V~~vertebrate ~~G~~group, ~~*BioScience*~~*Bioscience* **55**, 2005, 693.

Kuhn M., Building predictive models in R using the caret package, ~~*Journal of Statistical Software*~~*J. Stat. Softw.* **28**, 2008, 1–26.

Luiz O.J., Woods R.M., Madin E.M. and Madin J.S., Predicting IUCN ~~E~~extinction ~~R~~risk ~~C~~categories for the World's Data Deficient Groupers (Teleostei: Epinephelidae), *Conserv. Lett.* 2016.

Michael D.R., Kay G.M., Crane M., Florance D., MacGregor C., Okada S., McBurney L., Blair D. and Lindenmayer D.B., Ecological niche breadth and microhabitat guild structure in temperate Australian reptiles: implications for natural resource management in endangered grassy woodland ecosystems, *Austral Ecol.* **40**, 2015, 651–660.

Mohamed bin Zayed Species Conservation, F., The Mohamed ~~b~~Bin Zayed Species Conservation Fund, 2014.

Morais A.R., Siqueira M.N., Lemes P., Maciel N.M., De Marco P. and Brito D., Unraveling the conservation status of Data Deficient species, ~~*Biological Conservation*~~*Biol. Conserv.* **166**, 2013, 98–102.

Myers N., Mittermeier R.A., Mittermeier C.G., da Fonseca G.A. and Kent J., Biodiversity hotspots for conservation priorities, *Nature* **403**, 2000, 853–858.

Prasad A.M., Iverson L.R. and Liaw A., Newer classification and regression tree techniques: bagging and random forests for ecological prediction, *Ecosystems* **9**, 2006, 181–199.

Rondinini C., Di Marco M., Visconti P., Butchart S.H.M. and Boitani L., Update or outdate: long-term viability of the IUCN Red List, ~~Conservation Letters~~Conserv. Lett. **7**, 2014, 126–130.

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., Baillie J., Boitani L., Lacher T.E., Mittermeier R.A., Smith A.T., Absolon D., Aguiar J.M., Amori G., Bakkour N., Baldi R., Berridge R.J., Bielby J., Black P.A., Blanc J.J., Brooks T.M., Burton J.A., Butynski T.M., Catullo G., Chapman R., Cokeliss Z., Collen B., Conroy J., Cooke J.G., da Fonseca G.A.B., Derocher A.E., Dublin H.T., Duckworth J.W., Emmons L., Emslie R.H., Festa-Bianchet M., Foster M., Foster S., Garshelis D.L., Gates C., Gimenez-Dixon M., Gonzalez S., Gonzalez-Maya J.F., Good T.C., Hammerson G., Hammond P.S., Happold D., Happold M., Hare J., Harris R.B., Hawkins C.E., Haywood M., Heaney L.R., Hedges S., Helgen K.M., Hilton-Taylor C., Hussain S.A., Ishii N., Jefferson T.A., Jenkins R.K.B., Johnston C.H., Keith M., Kingdon J., Knox D.H., Kovacs K.M., Langhammer P., Leus K., Lewison R., Lichtenstein G., Lowry L.F., Macavoy Z., Mace G.M., Mallon D.P., Masi M., McKnight M.W., Medellín R.A., Medici P., Mills G., Moehlman P.D., Molur S., Mora A., Nowell K., Oates J.F., Olech W., Oliver W.R.L., Oprea M., Patterson B.D., Perrin W.F., Polidoro B.A., Pollock C., Powel A., Protas Y., Racey P., Ragle J., Ramani P., Rathbun G., Reeves R.R., Reilly S.B., Reynolds J.E., Rondinini C., Rosell-Ambal R.G., Rulli M., Rylands A.B., Savini S., Schank C.J., Sechrest W., Self-Sullivan C., Shoemaker A., Sillero-Zubiri C., De Silva N., Smith D.E., Srinivasulu C., Stephenson P.J., van Strien N., Talukdar B.K., Taylor B.L., Timmins R., Tirira D.G., Tognelli M.F., Tsytsulina K., Veiga L.M., Vié J.-C., Williamson E.A., Wyatt S.A., Xie Y. and Young B.E., The status of the world's land and marine mammals: diversity, threat, and knowledge, *Science* **322**, 2008, 225–230.

Sousa-Baena M.S., Garcia L.C. and Peterson T.A., Knowledge behind conservation status decisions: data basis for "Data Deficient" Brazilian plant species, *Biol. Conserv.* 2013.

Stuart S.N., Chanson J.S., Cox N.A., Young B.E., Rodrigues A.S.L., Fischman D.L. and Waller R.W., Status and trends of amphibian declines and extinctions worldwide, *Science* **306**, 2004, 1783–1786.

Tingley R., Hitchmough R.A. and Chapple D.G., Life-history traits and extrinsic threats determine extinction risk in New Zealand lizards, ~~Biological Conservation~~Biol. Conserv. **165**, 2013, 62–68.

Tjøsthein D., A measure of association for spatial variables, *Biometrika* **65**, 1978, 109–114.

Uetz P. and Hošek J., The Reptile Database, 2015.

Vilela B., Villalobos F., Rodríguez M.Á. and Terribile L.C., Body ~~S~~size, ~~E~~extinction ~~R~~risk and ~~K~~knowledge ~~B~~bias in New World Snakes, ~~PLoS ONE~~PLoS One **9**, 2014, e113429.

Walker S., Applying ~~e~~Extinction ~~r~~Risk ~~m~~Modelling to ~~d~~Develop ~~g~~Global ~~c~~Conservation ~~p~~Priorities for ~~b~~Bulbous ~~m~~Monocots, 2014, Imperial College London; London, UK, 65.

Webb T.J. and Mindel B.L., Global patterns of extinction risk in marine and non-marine systems, ~~Current Biology~~Curr. Biol. **25**, 2015, 506–511.

World Association of, Z., Aquaria, World Association of Zoos and Aquaria, 2013.

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

~~Appendix A.~~Appendix A. Supplementary data

[Multimedia Component 1](#)

Supplementary material 1

alt-text: Image 1

[Multimedia Component 2](#)

Supplementary material 2

alt-text: Image 2

Highlights

- We use random forests to predict the extinction risk of Data Deficient reptiles.

- We find that 19% of Data Deficient reptiles are likely to be threatened.

- Global patterns of threatened species richness are robust to data deficiency.

- New information from the Global Reptile Assessment will improve risk models.

Queries and Answers

Query:

Your article is registered as belonging to the Special Issue/Collection entitled "Reptile Conservation". If this is NOT correct and your article is a regular item or belongs to a different Special Issue please contact v.sethuraman@elsevier.com immediately prior to returning your corrections.

Answer: Yes

Query:

The author names have been tagged as given names and surnames (surnames are highlighted in teal color). Please confirm if they have been identified correctly.

Answer: Yes

Query:

Refs. "Trindade-Filho et al. 2012", "Brito 2010" are cited in text but not provided in the reference list. Please provide it in the reference list or delete the citation from the text.

Answer: Trindade-Filho, J., et al. (2012). "How does the inclusion of Data Deficient species change conservation priorities for amphibians in the Atlantic Forest?" Biodiversity and Conservation 21(10): 2709-2718

Brito, D. (2010). "Overcoming the Linnean shortfall: Data deficiency and biological survey priorities." Basic and Applied Ecology 11(8): 709-713.

Query:

Refs. "Hijmans et al. 2005", "CIESIN 2005a", "CIESIN 2005b" are cited in text but not provided in the reference list. Please provide it in the reference list or delete the citation from the text.

Answer: Hijmans, S. E., et al. (2005). "Very high resolution interpolated climate surfaces for global land areas." International Journal of Climatology 25: 1965-1978.

Ciesin (2005b). Gridded Population of the World (2000), Version 3 (GPWv3). Palisades, NY, Socioeconomic Data and Applications Center (SEDAC), Columbia University.

Ciesin (2005a). Last of the Wild Data Version 2 (LWP-2): Global Human Footprint dataset (HF).

Query:

Supplementary caption was not provided. Please check the suggested data if appropriate, and correct if necessary.

Answer: Appendix A

Supplementary methods and results, including Tables S1-S5, Figures S1-S8, and references for data collection.