Using coarse-scale species distribution data to predict extinction risk in plants

Sarah E. Darrah 1*, Lucie M. Bland 2, Steven P. Bachman 3,4, Colin P. Clubbe 3 and Anna Trias-Blasi 3

1 United Nations Environment Programme World Conservation Monitoring Centre (UNEP-WCMC), Cambridge CB3 0DL, UK, 2 School of BioSciences, The University of Melbourne, Parkville, VIC 3010, Australia, 3 Royal Botanic Gardens, Kew, Richmond, Surrey TW9 3AB, UK, 4 School of Geography, University of Nottingham, Nottingham NG7 2RD, UK

ABSTRACT

Aim Less than 6% of the worlds described plant species have been assessed on the IUCN Red List, leaving many species invisible to conservation prioritization. Large-scale Red List assessment of plant species is a challenge, as most species’ ranges have only been resolved to a coarse scale. As geographic distribution is a key assessment criterion on the IUCN Red List, we evaluate the use of coarse-scale distribution data in predictive models to assess the global scale and drivers of extinction risk in an economically important plant group, the bulbous monocotyledons.

Location Global.

Methods Using coarse-scale species distribution data, we train a machine learning model on biological and environmental variables for 148 species assessed on the IUCN Red List in order to identify correlates of extinction risk. We predict the extinction risk of 6439 'bulbous monocot' species with the best of 13 models and map our predictions to identify potential hotspots of threat.

Results Our model achieved 91% classification accuracy, with 88% of threatened species and 93% of non-threatened species accurately predicted. The model predicted 35% of bulbous monocots presently 'Not Evaluated' under IUCN criteria to be threatened and human impacts were a key correlate of threat. Spatial analysis identified some hotspots of threat where no bulbous monocots are yet on the IUCN Red List, for example central Chile.

Main conclusions This is the first time a machine learning model has been used to determine extinction risk at a global scale in a species-rich plant group. As coarse-scale distribution data exist for many plant groups, our methods can be replicated to provide extinction risk predictions across the plant kingdom. Our approach can be used as a low-cost prioritization tool for targeting field-based assessments.

Keywords bulbous monocotyledons, extinction risk, IUCN Red List, machine learning, plants, predictive modelling.

INTRODUCTION

Plant diversity across the globe is at risk from anthropogenic changes such as habitat loss, degradation and overexploitation (Corlett, 2016), yet the extent and drivers of species extinction, as well as the strategies to counteract them, remain poorly known. These challenges are compounded by the sheer size of the plant kingdom, with 350,699 accepted species names on The Plant List to date (The Plant List, 2013). The size of the group, compared to comprehensively assessed vertebrate groups like birds (10,425 species) and mammals (5513 species), is one of the reasons plants are underrepresented on the International Union for the Conservation of Nature (IUCN) Red List of Threatened Species™. The IUCN Red List contains 21,898 plant assessments (IUCN, 2016) which equates to 6% of described plants, although as much as a third of these assessments are considered out of date as they are either more than 10 years old or...
use an earlier version of the criteria (IUCN Standards And Petitions Subcommittee, 2014). Of those plants assessed, 53% are listed as threatened (Critically Endangered, Endangered, Vulnerable) (IUCN, 2016), but risk is likely to be overestimated as assessment efforts have been preferentially directed towards the most threatened species and areas (Rodrigues et al., 2006).

To better understand the global threat status of plants, the IUCN Sampled Red List Index (SRLI) for plants assessed the status of a random sample of plant species to give a representative view of plant extinction risk globally (Brummitt et al., 2015). The results indicated a much lower proportion of species are threatened (21%) compared to species published on the current version of the Red List (53%), which nevertheless implies that as many as 74,000–84,000 plants could be at risk. Although Red List assessments in isolation are not an appropriate device for conservation prioritization (Possingham et al., 2002), they are an important component for a range of prioritization and funding schemes such as Mohamed Bin Zayed Conservation Trust (www.speciesconser vation.org), Save our Species (SOS; www.sospecies.org) and the Critical Ecosystem Partnership Fund (CEPF; www.cepf.net). The absence of a strategic prioritization system for plant Red List assessments means that many species could be missing out on conservation funding (Bland et al., 2015b).

Given the large number of plants remaining to be assessed on the IUCN Red List, devising approaches to facilitate their assessment or to include them in conservation prioritization without formal assessments are of the highest priority (Callmander et al., 2005; Royal Botanic Gardens, Kew, 2010). We therefore must refine our understanding of what factors contribute to high extinction risk and which plant species are likely to be at highest risk (Duffy et al., 2009; Brummitt et al., 2015). Predictive modelling of extinction risk is a widely used tool to quantify threat levels across taxa and prioritize conservation research (Purvis et al., 2000; Cardillo et al., 2008; Di Marco et al., 2014; Jetz & Frecleton, 2015). Machine learning models such as random forests are popular due to their ability to find patterns in large and complex datasets (Di Marco et al., 2014; Bland et al., 2015a), and there is clear potential to apply these models to plants (Duffy et al., 2009). Existing studies of extinction risk in plants focus on national or regional scales, for example Amazonian plants (Feeley & Silman, 2009). Global extinction risk studies are more challenging as they rely on the availability of systematic, global data for the large majority of species under study. In particular, lack of high-resolution, high-quality occurrence data for most plant species presents limitations in scaling up the previous models (Bland et al., 2015b).

The World Checklist of Selected Plant Families (WCSP) is a coarse-scale geographic dataset of species presence according to a hierarchical geographic coding system derived from the Taxonomic Databases Working Group (TDWG; Brummitt et al., 2001). Data are complete for a large number of plant families and comprehensive for broad taxonomic groups such as the monocotyledons (WCSP, 2014). The coding at level one is continental, level two is regional and level three is broadly equivalent to small countries and islands (Brummitt et al., 2001). This coarse-scale geographic coding of plant distributions has not yet been explored as a predictive tool for extinction risk assessment. In this paper, we quantify the effect of high-quality range data on variable importance by comparing models using both coarse-scale and fine-scale species distribution data.

Modelling extinction risk relies on identifying appropriate variables which correlate with extinction risk in species of known conservation status. Correlates broadly fall into intrinsic (life history traits, e.g. habit and dispersal mode, and ecological traits, e.g. range size) and extrinsic variables (environmental and anthropogenic) and the most informative models combine both types of data (e.g. Davies et al., 2011). For vertebrates, studies often focus on intrinsic traits such as body mass, fecundity and niche characteristics (Purvis et al., 2005; Di Marco et al., 2014). Intrinsic correlates of extinction risk are less researched for plants at higher taxonomic levels, including pollination syndrome, height, habit, sexual system and dispersal mode (e.g. Laliberté, 2016), but there have been some recent advancements in understanding plant functional traits (e.g. Bullock et al., 2002; Pywell et al., 2003; Cornwell et al., 2014; and Diaz et al., 2016). Extrinsic variables are known to be good predictors of extinction risk (Lee & Jetz, 2010; Murray et al., 2014; Di Marco & Santini, 2015). The SRLI for plants found human impacts to be the greatest cause of threat to plants, particularly the conversion of natural habitats to agricultural land (Brummitt et al., 2008, 2015) and global data on human threats are increasingly available for use in models (e.g. CIESIN & CIAT, 2005; Hansen et al., 2013).

We bridge this gap by focusing on bulbous monocotyledons, a relatively well-known plant group. Bulbous monocot is an informal term that refers to all Monocotyledons in the orders Liliales and Asparagales with a geophytic life-form and petaloid flowers (excluding Orchidaceae) (A. Trias-Blasi, Royal Botanic Gardens, Kew, London, pers. comm.). There are approximately 7000 bulbous monocot species in eight different plant families (Amaryllidaceae, Asparagaceae (subfamilies Scilloideae and Brodiaeaeoideae), Colchicaceae, Iridaceae, Ixoliriaceae, Liliaceae, Melanthiaceae and Tecophilaeaceae) (WCSP, 2014). Many bulbous monocot taxa are economically important due to their horticultural, medicinal and nutritional value (Marshall, 1993). Bulbous monocots, such as snowdrops, have been greatly affected by illegal collecting activities linked to international trade, which has led to their inclusion on the checklist of the Convention on Trade in Endangered Species of Wild Fauna and Flora (CITES) (Davis, 1999; Yüzbaşıoğlu, 2008; Newton et al., 2014). Small range size, extractive activities and habitat loss from agricultural development, grazing, urban expansion, road building and tourism in littoral/montane sites also threaten species survival across global ranges and are likely to be good predictors of extinction risk (Özhatay et al., 2013; IUCN, 2014). To date, the extinction risk of only c. 2% (148 species) of bulbous monocots has been assessed against the IUCN Red List criteria (IUCN, 2014).
so the majority of species are categorized as Not Evaluated (NE). As bulbous monocots have complete geographic range data on the WCSP (WCSP, 2014), they are an excellent case study to test the applicability of extinction risk models, not only for monocotyledons but also for plants in general.

Using bulbous monocots as a study group, we build models to discriminate threatened and non-threatened species based on species-level data. Using species assessments from the IUCN Red List and Sampled Red List Index assessments, we build models to predict the threat status of bulbous monocots that are yet to be assessed. Our aims are to: (1) test the utility of coarse-scale distribution data in extinction risk models compared to fine-scale data; (2) identify correlates of extinction risk in bulbous monocots; and (3) predict levels of extinction risk in non-assessed species. If models perform well using coarse-scale distribution data, they will provide a much needed tool to understand the drivers of threat in bulbous monocots and to prioritize conservation efforts at a global scale. It could also provide opportunities for application to other plant groups and be upscaled to the whole plant kingdom.

**METHODS**

Predicting species extinction risk and prioritizing conservation action can be achieved in five steps: data collection, model validation, predictive modelling, prioritization and review of the process (Fig. 1).

**Distribution data**

We obtained a list of all bulbous monocot species and associated Taxonomic Databases Working Group (TDWG) level 3 distributions from the World Checklist of Selected Plant Families (WCSP, 2014). We partitioned the list into two groups: *assessed* and *non-assessed*, based on IUCN Red List and Sampled Red List Index assessments. Species assessed as Critically Endangered (CR), Endangered (EN), Vulnerable (VU), Near Threatened (NT) and Least Concern (LC) formed the assessed group, comprising 148 species, and Data Deficient (DD) and Not Evaluated (NE) species formed the non-assessed group, comprising 6439 species (Table 1). We excluded species classified as Extinct (EX) or Extinct in the Wild (EW), five species with missing distribution data, and families with either no assessed species (Ixioliriaceae and Tecophilaeaceae) or fewer than eight assessed species (Melanthiaceae).

**Table 1** IUCN Red List risk categories of bulbous monocotyledon species used in the model.

| IUCN Red list category | Number of species |
|------------------------|------------------|
| Critically endangered (CR) | 7 |
| Endangered (EN) | 24 |
| Vulnerable (VU) | 28 |
| Near threatened (NT) | 6 |
| Least concern (LC) | 83 |
| **Total assessed** | **148** |
| Data deficient (DD) | 56 |
| Not evaluated (NE) | 6383 |
| **Total non-assessed** | **6439** |
| **Total species** | **6587** |

Figure 1 Framework of steps to predict species extinction risk and prioritize conservation action. This analysis starts with the first three steps: data collection, model validation and modelling predicted extinction risk. To ensure effective conservation on the ground, we recommend the results of these steps inform conservation prioritization and that the process is reviewed and refined as data sources and model techniques are improved.
Correlates of extinction risk

We selected predictor variables based on expected correlates of extinction risk and data availability. We used seven variables grouped into taxonomy, geographic distribution, human impacts and conservation action (Table 2).

We used the count of TDWG level 3 regions (representing small countries or islands) within a species range as an indicator of range size, with a score of one representing endemic species. We also included an indicator of isolation, measured as percentage of islands across a species distribution. Data on life history traits are sparse and incomplete across plant groups. For the majority of plant groups, only life-form data (habit) are consistently available, which are redundant for this analysis as all bulbous monocots are by definition geophytes. To indicate differences in life history traits that could mediate extinction risk we used taxonomic family to account for shared evolutionary history. We use taxonomy as a surrogate for phylogeny due to data availability. Although there are some phylogenetic studies of bulbous monocot families (Seberg et al., 2012; Chen et al., 2013; García et al., 2014), it is currently difficult to incorporate phylogenetic relatedness in machine learning when there is no consolidated phylogeny at the group level. In addition, whilst phylogeny can be informative, it is currently not possible to include phylogeny in a computationally appropriate way in machine learning models, and data on life history traits is preferable for both predictive power and interpretability (Bland et al., 2015a). Within each TDWG level 3 region, we queried GIS layers of human impacts and conservation action to calculate average values across all the regions within each species range. Human impact variables included the following: Human Footprint Index (WCS, 2005), human population density (CIESIN & CIAT, 2005) and global forest loss (Hansen et al., 2013). We used forest loss as a surrogate for habitat loss, as it has been shown to have adverse effects on a diverse range of habitats, for example grassland and savannas (Boakes et al., 2010). We measured conservation action as the percentage area formally protected by intersecting TDWG level 3 regions with protected area polygons (IUCN & UNEP-WCMC, 2014). We then calculated average percentage area across each species distribution in an equal area cylindrical projection.

As TDWG level 3 regions are not equal in terms of area coverage, each region was normalized by the area of the region before averaging across the species range. We base predictions on current levels of threat and do not attempt to predict future risk levels under different scenarios, so climate change variables were not included.

### Extinction risk models and spatial analysis

We normalized predictor variables and split the data to create a training set of assessed species and a prediction set of non-assessed species. Multiple methods are available to model and predict extinction risk (Di Marco et al., 2014; Bland et al., 2015a; Luiz et al., 2016). Phylogenetic least squares is a popular method to explain extinction risk (Purvis et al., 2000; Cardillo et al., 2008), but is unsuitable for groups with no consolidated global phylogeny and shows

| Variable group | Variable | Description | Reference |
|----------------|----------|-------------|-----------|
| Extinction risk | Threat status – Red List assessed | Species assessed through IUCN Red List. Threat categories: CR, EN, CN, LC, NT, DD. Assessment dates range from 2003 to 2013 | IUCN (2014) |
|                 | Threat status – SRLI assessed | Species assessed in Sampled Red List Index project with IUCN threat categories | Royal Botanic Gardens, Kew (2010) |
| Taxonomy        | Plant family | Bulbous monocotyledon plant family group | WCSP (2014) |
| Geographic distribution | Range size | Number of Taxonomic Databases Working Group level 3 regions across a species range | |
|                 | Isolation | Percentage islands across species range | |
| Human impacts   | Human Footprint Index (1995–2004) | Raster 1-km resolution dataset created from nine global data layers covering human population pressure, human land use, infrastructure and human access | WCS (2005) |
|                 | Human population density (2000) | Raster 2.5-arc minute resolution dataset, each grid cell represents population density | CIESIN & CIAT (2005) |
|                 | Global forest loss (2000–2012) | Raster 30-m-resolution dataset, each grid cell represents forest loss encoded as 1 (loss) and 0 (no loss) | Hansen et al. (2013) |
| Conservation action | % Area protected | Percentage area covered by protected areas within species range. Calculated by intersecting World Database on Protected Areas with Taxonomic Databases Working Group level 3 regions in each species range | IUCN & UNEP-WCMC (2014) |
limited predictive power (Bland, 2015a). Because of the absence of consolidated global phylogenies for bulbous monocots (but see species available on Open Tree of Life) and our focus on predicting the status of non-assessed species, we use the machine learning model ‘random forests’, an ensemble of decision trees that repeatedly split predictors into increasingly homogenous groups. Random forests show high predictive power in extinction risk analyses compared to other machine learning tools (Bland, 2015a). These models were trained with assessed species and associated predictor variables to model the probability of threat in non-assessed species (Kuhn, 2008; R Core Development Team, 2014). Within the training set, we categorized threat status as ‘threatened’ (Critically Endangered, Endangered, Vulnerable) and ‘non-threatened’ (Near Threatened and Least Concern) due to difficulty in discriminating five imbalanced categories (e.g. 83 Least Concern and 7 Critically Endangered species; see Appendix S1 in Supporting Information, Table S1.2) (Hand, 2012; Luiz et al., 2016). To test for variation in model results due to sampling bias, we first ran models separately for IUCN Red List and SRLI species. To maximize sample size, we then grouped all species to test likely correlates of extinction risk.

To reliably predict the status of non-assessed and assessed species, it is imperative to assess model performance, that is, the ability of the model to accurately predict the extinction risk of species of known conservation status. We set model parameters to run 500 trees to prevent overfitting and assessed predictive accuracy using 10-fold cross-validation repeated five times on the training set. We used area under the receiver operating characteristic curve (AUC) as the most appropriate measure of model performance, as it assigns equal weight to both sensitivity and specificity whilst accounting for imbalanced threat categories in our training set (see Appendix S1, Table S1.2). We selected the optimal probability threshold by maximizing the Youden Index (sensitivity + specificity − 1; Perkins & Schisterman, 2006). We then undertook variable selection to understand if drivers identified by the SRLI project are also correlates of extinction risk in bulbous monocots, which were underrepresented in the original assessment. We measured variable importance with the mean decrease in Gini Index; a large decrease indicates high levels of statistical dispersion (Breiman et al., 1984).

Some variables are likely to correlate with extinction risk at subnational and local levels. Therefore, we also tested models based on fine-scale range size data, using a subset of assessed species for which these data were available. We used IUCN polygon range maps for IUCN Red List species. Because range maps are not available for SRLI species, we constructed maps based on occurrence points. We buffered each occurrence point with a 10-km radius and created a minimum convex polygon around those points. The 10-km radius was chosen to represent the area over which a population is expected to experience threats. We calculated range area in km² and extracted mean values of variables across species ranges. These models quantify the effect of high-quality range data on variable importance, but cannot be used to predict extinction risk in non-assessed species for which range size data are not available.

We predicted the status of 6439 species with the best model and mapped the spatial distribution of predicted threatened species to identify potential hotspots of threat. These maps were compared with the distribution of threatened bulbous monocot species that are currently on the IUCN Red List to identify priority areas where predicted threat levels are high, but few or no species have been assessed. Such gaps can be used to prioritize future conservation assessments.

**RESULTS**

**Assessing the ability of our model to predict extinction risk in species of known conservation status**

The model results demonstrate good predictive performance with a high area under the receiver operating characteristic curve (AUC = 0.98) indicating a good fit of the model to the data (Pearce & Ferrier, 2000) and high classification accuracy (91% of species were correctly classified during cross-validation). The model accurately classified 88% of threatened species (sensitivity) and 93% of non-threatened species (specificity; Table 3).

There was good discrimination between the threatened and non-threatened classes with particularly strong discrimination between Critically Endangered species and non-threatened species (Near Threatened and Least Concern) (Fig. 2). All Critically Endangered species were correctly predicted, indicating strong predictive capacity for species with high extinction risk. Near Threatened species, which are on the threshold between threatened and non-threatened, had the lowest proportion correctly classified. Identification of individual species misclassified revealed that some species (almost all of which are endemic to a single TDWG level 3 unit) were consistently misclassified across models. We found

**Table 3 Performance statistics of random forest model for assessed bulbous monocotyledons.**

| Parameter                      | Value        |
|--------------------------------|--------------|
| No. species                    | 148          |
| No. variables                  | 7            |
| Accuracy                       | 0.91         |
| 95% CI                         | 0.86, 0.95   |
| Sensitivity                    | 0.88         |
| Specificity                    | 0.93         |
| AUC                            | 0.98         |
| Threshold probability          | 0.46         |
| Youden Index                   | 0.82         |
| Kappa                          | 0.82         |
| % Predicted threatened         | 35           |
that classification accuracy for endemic species was lower than the average for the model (86% compared to 91%). Of particular note is the higher misclassification rate of species endemic to Cape Provinces (5 of 26 endemics were misclassified, 81% accuracy), including the two outliers in Fig. 2 (Crennum variable – misclassified as threatened and Romulea aquatic – misclassified as not threatened).

**Identifying correlates of high extinction risk**

Analysis of variable importance revealed that human impact variables (Human Footprint Index, human population density and forest loss) and the conservation action variable (percentage area under formal protection) were strong correlates of extinction risk (Fig. 3). The surrogate for range size calculated from TDWG regions was also important with the second greatest contribution to the predictive ability of the model. Plant family and isolation were not strong correlates of threat but still contributed to overall predictive performance, possibly because of interaction effects with other variables.

The model based on fine-scale range data derived from IUCN polygons and SRLI point data again showed good discrimination between threatened and non-threatened classes (see Appendix S1, Table S1.4), with all but one species correctly classified. Species range size had a much larger contribution to model performance when we used fine-scale range data than in models using TDWG distributions (see Appendix S1, Fig. S1.2).

**Predicting the status of non-assessed species**

When we applied the model with the highest area under the receiver operating characteristic curve (AUC) to predict extinction risk in bulbous monocots, we found that 35% (2237 of 6439) of species were predicted to be threatened (see Appendix S2, Table S2). Sequential removal of variables with the lowest contribution to the model reduced model performance, so the final model contained all seven predictor variables (Table 2). Running a separate model on IUCN Red List species resulted in reduced model performance (78% of species correctly classified; see Appendix S1, Table S1.3), whilst the model on SRLI species resulted in high model performance (91% correctly classified) but is based on a sample of only 38 assessed species. There was no substantial difference in variable importance between the two models based on IUCN Red List and SRLI species, so all species were pooled to fit the final model.

Analysis of model predictions by geographic region reveals a strong geographic clustering of extinction risk predictions. ‘Hotspots’ of global threat include the following: Turkey and Cape Provinces (> 153 species predicted to be threatened), Iran, central Chile, Greece, Spain, Lebanon, Syria, Northern Provinces and Transcaucasia (> 90 species predicted to be threatened; Fig. 4c). Regions with a high density of species assessed as threatened were reflected with a high density of predicted threatened species, for example, in Greece, Spain and Cape Provinces. In contrast, some regions, for example Ecuador, with a high density of assessed threatened species were not predicted to have a high density of threatened species (Fig. 4b,c).

**DISCUSSION**

In light of global environmental change, it is paramount to identify which species are at high risk of extinction and to diagnose drivers of risk in order to reverse biodiversity declines. Acquiring this information is particularly challenging for species-rich and poorly studied groups such as plants (Brummitt et al., 2015). Bulbous monocots are a good study group as they are a relatively well-known group of plants.
Figure 4 Global geographic distribution of bulbous monocotyledon species occurrence: (a) species richness, (b) number of assessed threatened species (Sampled Red List Index and IUCN Red List; Royal Botanic Gardens, Kew, 2010; IUCN, 2014), (c) predicted number of threatened species, (d) predicted proportion of threatened species. All maps are based on Taxonomic Databases Working Group level 3 regions in Mollweide projection.
comprising horticulturally important plants which are widely traded (Davis, 1999; Yüzübaşioğlu, 2008; Newton et al., 2014), and threatened in the wild (Özhatay et al., 2013; IUCN, 2014). We present the first global analysis of bulbous monocots, giving much needed indications of threat levels and correlates of risk across this economically important plant group. This information is crucial in order to prioritize geographic regions and individual species for future conservation interventions, including IUCN Red List assessments and trade controls.

IUCN Red List assessments and indices underpin a number of the Aichi Biodiversity Targets of the Strategic Plan for Biodiversity 2011–2020 as well as the cross-cutting Global Strategy for Plant Conservation. Achieving Target 2 of the Global Strategy for Plant Conservation ‘an assessment of the conservation status of all known plant species, as far as possible, to guide conservation action’ by 2020 will require rapid advances to increase the rate and reduce the high cost of species assessments (Baillie et al., 2008; Juffe-Bignoli et al., 2016). In addition, with limited funding there will undoubtedly be trade-offs between increasing the number of plant assessments and keeping current assessments up to date (Rondinini et al., 2014). A recent review of progress in achieving the Global Strategy for Plant Conservation indicated that Target 2 is unlikely to be met by 2020 (Sharrock et al., 2014), highlighting the urgent need for rapid and cost-effective assessments (Bland et al., 2015b, 2016).

Our model predicts that 35% of non-assessed bulbous monocots are currently at risk of extinction, resulting in a potential 2254 additional threatened species to the 59 already assessed as threatened on the IUCN Red List (IUCN, 2015). This estimate is higher than the global average for plants based on the SRLI (20%) (Brummitt et al., 2015). However, these predictions are much closer to extinction levels in another horticulturally important group, the cacti (Cactaceae), for which 31% of species are at risk (Goettsch et al., 2015).

Coarse-scale global analyses can be used to predict hotspots of threat and systematically prioritize species assessments using local data (Smith et al., 2009). In addition to predicting known hotspots of threat, for example Turkey and Cape Provinces, our model predicted hotspots where no bulbous monocot species have yet been assessed, for example central Chile (72% of species predicted to be threatened), and some hotspots where no species have yet been assessed as threatened, for example Iran, Lebanon and Syria (Fig. 4b, c). Many species-poor regions exhibit high levels of predicted extinction risk (Fig. 4a,c), for example, more than half of India’s 50 species are predicted to be threatened, including 25 endemic species. On the other hand, some species-rich areas are predicted to have few or no threatened species, for example Namibia where none of its 204 species are predicted to be threatened. This global view can be used to prioritize international conservation efforts to target the most vulnerable regions, thereby increasing cost-effectiveness and demonstrating value for money.

The drivers of extinction risk in bulbous monocots have received limited research at the group level, yet trade in bulbs has been increasing (CITES Trade Database, 2016) and the value of international floriculture exports, including bulbs, increased from US $8.5 billion in 2011 to US $20.6 in 2013 (UN Comtrade, 2014). Our model highlights common patterns in the human impact and biological variables used to predict extinction risk. As also highlighted by other taxon-wide studies (Cardillo et al., 2008; Murray et al., 2014; Di Marco & Santini, 2015) and the SRLI project, human impacts were key in distinguishing between threatened and non-threatened species. The Human Footprint Index and population density were strong correlates of extinction risk (Fig. 3). This is likely due to close correlation with habitat degradation through urbanization, the second most common threat for bulbous monocots on the IUCN Red List (IUCN, 2014).

Range size was also an important correlate, reflecting the inherent role of range size in IUCN criteria (IUCN, 2001). The strongest effects were seen in species with narrow ranges (fewer than three Taxonomic Databases Working Group level 3 regions; see Appendix S1, Fig. S1.3a). The relationship between percentage area protected and likelihood of extinction is less clear, perhaps reflecting poor data coverage or because protected areas are not usually created and managed for plants. Whilst the World Database on Protected Areas is the best available global dataset for protected areas, it is compiled from the submission of national data and, as such, the data are incomplete and outdated for many countries (UNEP-WCMC, 2015). Issues of data coverage are particularly prevalent in Middle Eastern countries, for example Turkey, where bulbous monocot species richness is highest. Consistent with previous studies, plant family does not appear to be a good correlate of extinction risk (Fréville et al., 2007). Isolation was also a poor correlate but the indicator could be strengthened if population data were collected more widely. Understanding how drivers of threat in bulbous monocots compare to plants as a whole is an important step in ensuring plant conservation strategies are suitable at the group level.

To be confident in predictions of extinction risk in non-assessed species, it is paramount to assess the capacity of the model to accurately predict risk in species of known conservation status. Our model showed high classification accuracy (91%) in line with extinction risk models for other taxa, for example 90% for mammals in Di Marco et al. (2014) and 95% for birds in Machado & Loyola (2013). However, the accuracy of our model in predicting threatened species (88% model sensitivity) is an improvement on previous models, for example 68% in Di Marco et al. (2014) and 24% in Machado & Loyola (2013). The ability to accurately identify threatened species is appropriate for conservation problems as it is more risk averse to predict species as threatened when they are not (false positives) than to predict species as not threatened when they are (false negatives). Model performance is particularly high for Critically Endangered species, whereas Near Threatened species had the lowest classification
accuracy. This may reflect the lack of quantitative criteria to define the Near Threatened threshold (IUCN, 2001), meaning that the boundary for this category is more ‘fuzzy’ than for threatened categories.

Although model performance was good, poor data availability imposes limitations on the inferences we can make from the results (Table 4 shows limitations and recommendations). Model validation was based on a sample of 148 assessed species which were used to predict the extinction risk of 6439 non-assessed species. Jetz & Freckleton (2015) consider predictions to be sufficiently reliable when 60% of species have been assessed and are used to train predictions for non-assessed and/or Data Deficient species. To predict extinction risk in bulbous monocots through modelling alone, this would equate to random assessment of 3952 species before then predicting the remaining 40%. We therefore recommend first validating this approach using random samples at national and subnational levels, focusing on areas for which there is an abundance of plant information and assessments. For example, the South African National Biodiversity Institute (SANBI, 2014) has nationally assessed 1987 bulbous monocots and is a good candidate for further model validation. However, obtaining suitable re-assessment data to validate extinction risk models remains a challenge (Bland et al., 2016) and such validation is in itself biased and may not estimate the true accuracy of the model (Hastie et al., 2009).

Our model assumes the relationships between correlates and assessed species in the training set are transferrable to non-assessed species. This assumption is substantial considering the apparent bias towards threatened species assessments in IUCN Red List. Nevertheless, models run separately for IUCN Red List and randomly sampled SRLI species did not show substantial differences in variable importance. The most important predictor variables were indicators of habitat degradation, a non-species-specific threat, which the SRLI found to hold true across plant groups under random sampling (Brummitt et al., 2015). Most predictor variables used in the model show strong correlations with extinction risk; however, increasing the number and range of variables is likely to better capture variations in extinction risk globally. For example, forest loss data, mainly based on forest cover derived from satellite imagery, may not adequately represent the nuances of habitat degradation such as selective logging (Burovalova et al., 2015). We recommend that land cover/land use data are included alongside forest loss in future models.

As our approach is restricted to systematic, globally available datasets, the range of species-specific threat data that could be included in the model was limited. Inclusion of such data, for example overharvesting of wild populations for trade, is likely to improve the accuracy of predictions yet are currently only available for a small number of species (see Table 1 in Smith et al., 2011).

This study highlights that models based on coarse-scale species distribution data can provide rapid and low-cost options for preliminary assessments of extinction risk and conservation prioritization. As the Taxonomic Databases

| Data category                  | Limitations                                                                 | Recommendations                                                                 |
|-------------------------------|-----------------------------------------------------------------------------|---------------------------------------------------------------------------------|
| **Data availability and quality** |                                                                             |                                                                                 |
| Species assessments            | Small training set – 148 species                                             | Increase sample size                                                            |
|                               | Possible inconsistency of application of the Near Threatened IUCN category    |                                                                                 |
| Species-specific data          | No species-specific predictor variables, for example phylogeny, trade data, population trends | Increase data collection efforts                                                 |
| Species range data             | Taxonomic Databases Working Group regions used as surrogates for species distributions, likely overestimating range size | Match resolution of range size data to that of environmental and socioeconomic variables |
|                               | Coarse-scale range data may mask the effects of predictor variables acting at subnational levels | Investigate use of data aggregators e.g. GBIF as a source for finer resolution geographic range data |
| Other predictor variables      | Forest loss used as a surrogate for habitat degradation                      | Included land cover and land use data within the model                          |
|                               | Climate data not included                                                    | Include climate data                                                            |
| **Model limitations**          |                                                                             |                                                                                 |
| Model selection                | Only one type of model was tested                                           | Re-run analysis using a variety of model types, for example classification trees, boosted trees and GLMs |
| Assumptions                    | Relationships in the training set are assumed to apply to non-assessed species | Further test the modelled relationships by repeating the analysis with species-specific data |
| Assessment bias                | Training set includes both non-random (Red List) and random (Sampled Red List Index) assessments | Training set should comprise only randomly sampled/SRLI species                 |

Table 4 Key limitations and recommendations on the data and model used in the analysis.
Working Group system for the World Checklist of Selected Plant Families represents the most comprehensively compiled distribution data presently available for plants (WCSP, 2014), the high performance of these models has implications for future application to predict extinction risk for all plants when the checklist is complete. In line with previous studies, our findings indicate that human impacts and species range size are key drivers of extinction risk in bulbous monocots, providing much needed insight into correlates of threat in the group. Our model predicts that 35% of bulbous monocots may be threatened with extinction and highlights predicted hotspots of threat for systematic prioritization of species assessments. In an era of budget constraints for biodiversity assessments, our approach provides a low-cost option to achieving ambitious conservation targets based on limited information and financial investment.

ACKNOWLEDGEMENTS

We would like to thank E.J. Milner-Gulland (Imperial College London) and Walter Jetz (Yale University) for comments on the project proposal. We acknowledge the work of Masato Matoki and William Baker on modelling extinction risk in palms (Arecaceae) that inspired this study, and all contributors to open access data used in this research, without which the analysis would not have been possible, namely, the International Union for Conservation of Nature, Royal Botanic Gardens, Kew, Wildlife Conservation Society, Centre for International Earth Science Information Network, Columbia University, International Center for Tropical Agriculture, United Nations Environment Programme World Conservation Monitoring Centre and Hansen et al. (2013).

REFERENCES

Baillie, J.E.M., Collen, B., Amin, R., Akcakaya, R.H., Butchart, S.H.M., Brummitt, N., Meagher, T.R., Ram, M., Hilton-Taylor, C. & Mace, G.M. (2008) Towards monitoring global biodiversity. Conservation Letters, 1, 18–26.
Bland, L.M., Collen, B., Orme, C.D.L. & Bielby, J. (2015a) Predicting the conservation status of data-deficient species. Conservation Biology, 29, 250–259.
Bland, L.M., Orme, C.D.L., Bielby, J., Collen, B., Nicholson, E. & McCarthy, M.A. (2015b) Cost-effective assessment of extinction risk with limited information. Journal of Applied Ecology, 52, 861–870.
Bland, L.M., Bielby, J., Kearney, S., Orme, C.D.L., Watson, J.E.M. & Collen, B. (2016) Toward reassessing data-deficient species. Conservation Biology. doi:10.1111/cobi.12850.
Boakes, E.H., McGowan, P.J.K., Fuller, R.A., Chang-qing, D., Clark, N.E., O’Connor, K. & Mace, G.M. (2010) Distorted views of biodiversity: spatial and temporal bias in species occurrence data. PLoS Biology, 8, e1000385. doi:10.1371/journal.pbio.1000385.
Breiman, L., Friedman, J., Stone, C. & Olshen, R. (1984) Classification and regression trees. Chapman & Hall/CRC, London, UK.
Bruggeman, J., Herina, J. & Brandt, B.W. (2009) PhyloPars: estimation of missing parameter values using phylogeny. Nucleic Acids Research, 37, W179–W184.
Brummitt, R.K., Pando, F. & Hollis, S. (2001) Plant taxonomic database standards no. 2. World geographical scheme for recording plant distributions. Hunt Institute for Botanical Documentation, Pittsburgh.
Brummitt, N., Bachman, S. & Moat, J. (2008) Applications of the IUCN Red List: towards a global barometer for plant diversity. Endangered Species Research, 6, 127–135.
Brummitt, N.A., Bachman, S.P., Griffiths-Lee, J. et al. (2015) Green plants in the Red: a baseline global assessment for the IUCN Sampled Red List Index for plants. PLoS ONE, 10, e0135152. doi:10.1371/journal.pone.0135152.
Bullock, J.M., Moy, I.L., Pywell, R.F., Coulson, S.J., Nolan, A.J. & Caswell, H. (2002) Plant dispersal and colonisation process at local and landscape scales. Dispersal ecology (ed. J.M. Bullock, R.E. Kenward and R.S. Hails), pp. 279–302. Blackwell Publishing, Oxford, UK.
Buroválová, A., Bauert, M.R., Hassold, S., Fatrouxandjianfi- nonjasolomiovazo, N.T. & Pin Koh, L. (2015) Relevance of global forest change data set to local conservation: case study of forest degradation in Masoala National Park, Madagascar. Biotropica, 47, 267–274.
Callmander, M.W., Schatz, G.E. & Lowry, P.P. (2005) IUCN Red List assessment and the global strategy for plant conservation: taxonomists must act now. Taxon, 54, 1047–1050.
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, 275, 1441–1448.
Chen, S., Kim, D., Chase, M.W. & Kim, J. (2013) Networks in a large-scale phylogenetic analysis: reconstructing evolutionary history of Asparagales (Lilianae) based on four plastid genes. PLoS ONE, 8, e59472. doi:10.1371/journal.pone.0059472.
CIESIN & CIAT (2005) Gridded population of the world, version 3 (GPWv3): population density grid. NASA Socioeconomic Data and Applications Center (SEDAC). Available at: http://sedac.ciesin.columbia.edu/data/set/gpw-v3-population-density. (accessed 24 April 2014).
CITES Trade Database (2016) CITES trade statistics derived from the CITES Trade Database, UNEP World Conservation Monitoring Centre, Cambridge, UK. Available at: https://trade.cites.org/.
Corlett, R.T. (2016) Plant diversity in a changing world: status, trends, and conservation needs. Plant Diversity, 38, 10–16.
Cormwell, W.K., Westoby, M., Falster, D.S. et al. (2014) Functional distinctiveness of major plant lineages. Journal of Ecology, 102, 345–356.
Davies, T.J., Smith, G.F., Bellstedt, D.U., Boatwright, J.S., Bytebier, B., Cowling, R.M., Forest, F., Harmon, L.J.,
Predicting extinction risk in plants

IUCN (2014) IUCN Red List of threatened species. Version 2014-2. Available at: www.iucnredlist.org. (accessed 28 July 2014).

IUCN (2015) IUCN Red List of threatened species. Version 2015-4. Available at: www.iucnredlist.org. (accessed 12 November 2015).

IUCN (2016) IUCN Red List of threatened species. Version 2016-2. Available at: www.iucnredlist.org. (accessed 25 October 2016).

IUCN Standards and Petitions Subcommittee (2014) Guidelines for using the IUCN Red List Categories and Criteria. Version 11 (February 2014). Prepared by the Standards and Petitions Working Group of the IUCN SSC Biodiversity Assessments SubCommittee in August 2008, 1, 86.

IUCN and UNEP-WCMC (2014) The world database on protected areas (WDPA). Available at: www.protectedplanet.net. UNEP-WCMC, Cambridge, UK. (accessed 24 April 2014).

Jetz, W. & Freckleton, R.P. (2015) Towards a general framework for predicting threat status of data-deficient species from phylogenetic, spatial and environmental information. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, 370, 20140016. DOI: 10.1098/rstb.2014.0016.

Juffe-Bignoli, D., Brooks, T.M., Butchart, S.H.M. et al. (2016) Assessing the cost of global biodiversity and conservation knowledge. *PLoS ONES*, 11, e0160640. doi:10.1371/journal.pone.0160640.

Kuhn, M. (2008) Building predictive models in R using the caret package. *Journal of Statistical Software*, 28, 1–26.

Laliberté, E. (2016) Below-ground frontiers in trait-based plant ecology. *New Phytologist*, doi:10.1111/nph.14247.

Lee, T.M. & Jetz, W. (2010) Unraveling the structure of species extinction risk for predictive conservation science. Proceedings of the Royal Society B, Biological Sciences. doi:10.1098/rspb.2010.1877.

Luiz, O.J., Woods, R.M., Madin, E.M.P. & Madin, J.S. (2016) Predicting IUCN extinction risk categories for the world’s data deficient groupers (Teleostei: Epinephelidae). *Conservation Letters*, 9, 342–350.

Machado, N. & Loyola, R.D. (2013) A comprehensive quantitative assessment of bird extinction risk in Brazil. *PLoS ONE*, 8, e72283. doi:10.1371/journal.pone.0072283.

Marshall, N.T. (1993) The gardener’s guide to plant conservation. World Wildlife Fund-US, Washington DC, USA.

Murray, K.A., Verde Arregoitia, L.D., Davidson, A., Di Marco, M. & Di Fonso, M.M.I. (2014) Threat to the point: improving the value of comparative extinction risk analysis for conservation action. *Global Change Biology*, 20, 483–494.

Newton, R., Davis, A., Rutherford, C., Trias-Blasi, A., Trivedi, C. & Wilford, R. (2014) Conservation and sustainable use of snowdrops. RHS Daffodil, Snowdrop and Tulip Yearbook.

Özhatay, N., Koçyiğit, M., Yüzbasioğlu, S. & Gürdal, B. (2013) Mediterranean flora and its conservation in Turkey: with special reference to Monocot geophytes. *Flora Mediterranea*, 23, 195–208.
S. E. Darrah et al.

Pearce, J. & Ferrier, S. (2000) Evaluating the predictive performance of habitat models developed using logistic regression. Ecological Modelling, 133, 225–245.

Perkins, N. & Schisterman, E. (2006) The inconsistency of “optimal” cutpoints obtained using two criteria based on the receiver operating characteristic curve. American Journal of Epidemiology, 163, 670–675.

Possingham, H.P., Andelman, S.J., Burgman, M.A., Medellin, R.A., Master, L.L. & Keith, D.A. (2002) Limits to the use of threatened species lists. Trends in Ecology & Evolution, 17, 503–507.

Purvis, A., Gittleman, J.L., Cowlishaw, G. & Mace, G.M. (2000) Predicting extinction risk in declining species. Proceedings of the Royal Society B: Biological Sciences, 267, 1947–1952.

Purvis, A., Cardillo, M., Grenyer, R. & Collen, B. (2005) Correlates of extinction risk: phylogeny, biology, threat and scale. Phyllogeography and conservation (ed.by A. Purvis, T.M. Brooks and J.L. Gittleman), pp. 295–316. Cambridge University Press, Cambridge, UK.

Pywell, R.F., Bullock, J.M., Roy, D.B., Warman, L.I.Z., Walker, K.J. & Rothery, P. (2003) Journal of Applied Ecology, 40, 65–77.

R Core Development Team (2014) Version 3.0.2. Available at: http://www.r-project.org/. (accessed 30 April 2014).

Rodrigues, A.S., Pilgrim, J.D., Lamoreux, J.F., Hoffmann, M. & Brooks, T.M. (2006) The value of the IUCN Red List for conservation. Trends in Ecology & Evolution, 21, 71–76.

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

Royal Botanic Gardens, Kew (2010) Plants under pressure—a global assessment: the first report of the IUCN sampled red list index for plants. Royal Botanic Gardens, Kew, UK.

SANBI (2014) Red List of South African Plants. Version 2014.1. Available at: redlist.sanbi.org. (accessed 14 July 2014).

Seberg, O., Petersen, G., Davies, J.L., Pires, J.C., Stevenson, D.W., Chase, M.W., Fay, M.F., Devey, D.S., Jørgensen, T., Sytsma, K.J. & Pillon, Y. (2012) Phylogeny of the Asparagales based on three plastid and two mitochondrial genes. American Journal of Botany, 99, 875–889.

Sharrock, S., Oldfield, S. & Wilson, O. (2014) Plant Conservation Report 2014: a review of progress in implementation of the global strategy for plant conservation 2011–2020. Secretariat of the Convention on Biological Diversity, Montréal, Canada and Botanic Gardens Conservation International, Richmond, UK. Technical Series No. 81, 56 pages. Available at: https://www.cbd.int/doc/publications/cbd-ts-81-en.pdf. (accessed 7 April 2016).

Smith, R.J., Verissimo, D., Leader-Williams, N., Cowling, R.M. & Knight, A.T. (2009) Let the locals lead. Nature, 462, 280–281.

Smith, M.J., Benítez-Díaz, H., Africa, Clemente-Muñoz M., Donaldson, J., Hutton, J.M., Noel, McGough H., Medellin, R.A., Morgan, D.H.W., O’Criodain, C., Oldfield, T.E.E., Schippman, W. & Williams, R.J. (2011) Assessing the impacts of international trade on CITES-listed species: current practises and opportunities for scientific research. Biological Conservation, 144, 82–91.

The Plant List (2013) Version 1.1. Available at: http://www.thewildlife.org/. (accessed 18 December 2015).

UN Comtrade (2014) United Nations Commodity Trade Statistics Database. Available at: http://comtrade.un.org/db/default.aspx

UNEP-WCMC (2015) World database on protected areas user manual 1.0. Cambridge, UK, UNEP-WCMC.

WCBS (2005) Last of the Wild Project, Version 2, 2005, (LWP-2): Global Human Footprint Dataset (Geographic). Palisades, New York: NASA Socioeconomic Data and Applications Center (SEDAC). Available at: https://doi.org/10.7927/H4M61H5F. (accessed 26 April 2014).

WCSP (2014) World checklist of selected plant families. Royal Botanic Gardens, Kew, UK.

Yüzbasioğlu, S. (2008) The development of non-detriment findings for Galanthus elwesii Hook. F., in Turkey. International Expert Workshop on CITES Non-Detriment Findings, Cancun. Available at: http://www.conabio.gob.mx/institucion/cooperacion_internacional/TallerNDF/Links-Documentos/WG-CS/WG4-GeophytesandEpiphytes/WG4-CS7%20Galanthusel2/WG4-CS7.pdf.

DATA ACCESSIBILITY

R script for this analysis is available on GitHub: https://github.com/sdarrah/extinctionrisk-plants.git.

SUPPORTING INFORMATION

Additional Supporting Information may be found in the online version of this article:

Appendix S1 Supplementary tables and figures.

Appendix S2 Online database of model results.

Figure S1.1 Random forest predictions for validation set.

Figure S1.2 Variable importance for model based on fine-scale range species distribution data.

Figure S1.3 Partial dependence plots showing likelihood of extinction risk for predictor variables.

Table S1.1 Number of bulbous monocotyledon species used in the analysis by plant family.

Table S1.2 Proportion of assessed bulbous monocotyledon species correctly classified by the model.

Table S1.3 Performance statistics for models based on IUCN Red List and SRLI assessed bulbous monocotyledon species.

Table S1.4 Performance statistics for model based on species with fine-scale species distribution data.
Table S2 Online database of model results showing predicted extinction risk for assessed and non-assessed bulbous monocotyledon species.

BIOSKETCH

Sarah Darrah is interested in plant biodiversity indicators and the application of extinction risk modelling to conservation prioritization schemes.

Author contributions: All authors contributed to research design and writing of the manuscript. A.T.-B. and S.B. jointly conceived the main objectives. Data were supplied from multiple sources as outlined in the text and references. L.B. provided coding script and S.D. carried out the modelling and analysis.

Editor: Kenneth Feeley