
Philip J. Platts¹*, Antje Ahrends², Roy E. Gereau³, Colin J. McClean¹, Jon C. Lovett⁴, Andrew R. Marshall¹,⁵, Petri K. E. Pellikka⁶, Mark Mulligan⁷, Eibleis Fanning⁸ and Rob Marchant¹

ABSTRACT

Aim Data shortages mean that conservation priorities can be highly sensitive to historical patterns of exploration. Here, we investigate the potential of regionally focussed species distribution models to elucidate fine-scale patterns of richness, rarity and endemism.

Location Eastern Arc Mountains, Tanzania and Kenya.

Methods Generalized additive models and land cover data are used to estimate the distributions of 452 forest plant taxa (trees, lianas, shrubs and herbs). Presence records from a newly compiled database are regressed against environmental variables in a stepwise multimodel. Estimates of occurrence in forest patches are collated across target groups and analysed alongside inventory-based estimates of conservation priority.

Results Predicted richness is higher than observed richness, with the biggest disparities in regions that have had the least research. North Pare and Nguu in particular are predicted to be more important than the inventory data suggest. Environmental conditions in parts of Nguru could support as many range-restricted and endemic taxa as Uluguru, although realized niches are subject to unknown colonization histories. Concentrations of rare plants are especially high in the Usambaras, a pattern mediated in models by moisture indices, whilst overall richness is better explained by temperature gradients. Tree data dominate the botanical inventory; we find that priorities based on other growth forms might favour the mountains in a different order.

Main conclusions Distribution models can provide conservation planning with high-resolution estimates of richness in well-researched areas, and predictive estimates of conservation importance elsewhere. Spatial and taxonomic biases in the data are essential considerations, as is the spatial scale used for models. We caution that predictive estimates are most uncertain for the species of highest conservation concern, and advocate using models and targeted field assessments iteratively to refine our understanding of which areas should be prioritised for conservation.

Keywords Biodiversity, conservation planning, endemism, rare species, sampling bias, spatial predictions.
INTRODUCTION

Limited resources for conservation dictate identification of priority regions to achieve effective conservation action (Margules & Pressey, 2000; Myers et al., 2000; Eken et al., 2004; Wilson et al., 2006). A major constraint, particularly at the site scale, is the scarcity of fine-scale data on the distribution of biodiversity (da Fonseca et al., 2000; Käper et al., 2006). Given the urgency of conservation action and the fact that much-needed biodiversity inventories are costly and underfunded (Lawton et al., 1998), the application of distribution models to species occurrence data could provide a practical way forward.

Conservation action is most often driven by decisions at the site scale (Mace et al., 2000; Ferrier, 2002). Such prioritizations can be highly sensitive to the inventory data available at the time, resulting in bias towards sites with a good history of biological exploration (Reddy & Davalos, 2003). Early explorations in the Eastern Arc Mountains (hereafter, EAMs) focused almost exclusively on the Uluguru and Usambara ranges (1880–1980). Over the last 30 years, funding has continued to be spread unevenly, favouring some mountain blocs such as the Usambaras and Udzungwas, whilst others such as North Pare and Ngu remain under-surveyed (A. Ahrends, unpublished data). Recent investment in the Nguru and Rubeho Mountains has resulted in the discovery of new species, altering conservation priorities still further (Doggart et al., 2006; Menegon et al., 2008). Spatially referenced inventory data for regions such as the EAMs have become increasingly accessible in recent years (e.g. http://www.tropicos.org); however, for use in a modelling framework, it is necessary to consider the historical, artifactual and biological processes that underlie them (Graham et al., 2004). For instance, inventory data are often biased not only in geographical space but also towards particular taxonomic groups – in the case of vascular plants, trees tend to be the dominant growth form recorded. Since plant diversity is sometimes employed as an indicator of overall biodiversity value (Bladt et al., 2008; Larsen et al., 2009), it is important to consider whether models predict similar patterns for the different growth forms within this group.

Historical habitat and climate configurations are also important for understanding species distributions, especially for endemic taxa (Jetz et al., 2004; Possingham & Wilson, 2005; Graham et al., 2006). Climatic conditions in the EAMs are thought to have been relatively stable, their proximity to the Indian Ocean providing a buffer against global trends in climate (Lovett, 1990; Marchant et al., 2007). Similar ecoclimatic stability is evident in other regions where highland habitats abut warm tropical oceans, such as the Atlantic rainforests in South America and the Queensland rainforests in Australia (Lovett et al., 2005) and has been suggested as a key driver of endemism in biodiversity hotspots (Fjeldså et al., 1997). Historical and evolutionary processes are particularly pertinent in the EAMs, which are geologically much older than adjacent mountains (Griffiths, 1993; Schlüter, 1997). Recently, however, they have suffered significant deforestation, reducing forest cover by around 70% (Burgess et al., 2007; Hall et al., 2009).

The aims of this article are to investigate the extent to which modelled richness is affected by historical and taxonomic bias in inventory data and to highlight the potential conservation importance of under-researched areas. Present-day climatic conditions, topography and soil parameters are combined with remotely sensed land cover data to estimate the spatial distributions of 452 plant taxa (species, subspecies, varieties), including 71 that are endemic to the EAMs and/or threatened with extinction. Our discussion of results explores the potential of distribution models to help refine conservation priorities in a region where confounding factors are typical of those found in many biodiversity hotspots.

METHODS

Study region

The EAMs are part of the Eastern Afromontane Biodiversity Hotspot (Mittermeier et al., 2004), extending from the Taita Hills in southeast Kenya to the Makambako Gap in south-central Tanzania (Fig. 1 & Table 1). At least 492 vascular plant taxa are endemic (402 species plus 90 subspecies and varieties), including 114 taxa of trees (86 species; R.E.G. unpublished data). Endemism amongst birds is also high (ICBP, 1992;
Stattersfield *et al.*, 1998), and a number of mammals and amphibians are endemic or near-endemic (Burgess *et al.*, 2007; Poynton *et al.*, 2007). Preservation of this region is a priority for biodiversity conservation (Olson & Dinerstein, 1998; Brooks *et al.*, 2002) and crucial to Tanzania’s population, for whom the forests provide a wide variety of ecosystem services including water, electricity, building materials, medicine and tourist revenue (Burgess *et al.*, 2009; Mwakalila *et al.*, 2009).

### Plant inventory data

The plant database (c. 70,000 records) combines our own field data with two large datasets contributed by the Missouri Botanical Garden (http://www.tropicos.org) and Frontier-Tanzania (http://www.frontier.ac.uk). Botanical identifications were verified by herbaria (Royal Botanic Gardens, Kew, Missouri Botanical Garden, and the University of Dar es Salaam); nomenclature was standardized by reference to the African Flowering Plants Database (AFPD, 2009). Threatened and potentially threatened taxa were identified according to an ongoing assessment of the conservation status of the combined EAM and Coastal Forest flora (Gereau *et al.*, 2010). Endemism in the context of this article refers to taxa that have been found only in the EAMs at and above 500 m elevation.

We modelled all taxa with records of occurrence in ten or more distinct 1-km or 2-km grid squares, favouring the higher resolution where specimen locality data allowed (see Appendix S1 in Supporting Information). The modelling subset targets 452 taxa in 90 plant families: 304 trees, 12 lianas, 62 shrubs and 74 herbs. Of these, 319 were modelled at 1-km resolution and 133 at 2-km resolution; 68 are threatened, and 25 are strictly endemic.

### Environmental data

Point patterns observed for our target taxa were regressed against twelve predictor variables, each representing an aspect of the environment thought to directly affect plant distributions in the EAMs (Tables 1 & 2). For temperature, we used interpolated climate surfaces based on records from the period 1950–2000 (WorldClim; Hijmans *et al.*, 2005). These data provide monthly temperature means and extremes at a spatial resolution of 1 km, from which we derived the annual mean and range, potential evapotranspiration (Thorntwaite, 1948) and an associated moisture index (potential evapotranspiration / annual rainfall). Rainfall grids were based on analysis of data from the Tropical Radar Measuring Mission (TRMM 2B31 combined PR, TMI profile): first, mean monthly 1-km gridded atmosphere rainfall was calculated from observations spanning the period 1997–2006 (Mulligan, 2006a); surface-received orographic rainfall was then modelled using wind velocity, slope, aspect and topographic exposure (Mulligan & Burke, 2005). Maximum water deficit represents the length and severity of the dry season and was calculated as the highest cumulative deficit in mean monthly rainfall, where a deficit is <100 mm month$^{-1}$. Estimates of cloud frequency were based on a 1 km climatology derived from the MODIS MOD35 Cloud Mask Product (Mulligan, 2006b).

### Table 1

<table>
<thead>
<tr>
<th>Mountain bloc (north to south)</th>
<th>Forest area (km$^2$)</th>
<th>Number of patches</th>
<th>Altitude (m)</th>
<th>Mean annual temperature ($^\circ$C)</th>
<th>Mean annual rainfall (mm/year)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Taita Hills</td>
<td>7.0</td>
<td>14</td>
<td>*1102–2208 (1585)</td>
<td>16–22 (19)</td>
<td>253–1208 (630)</td>
</tr>
<tr>
<td>North Pare</td>
<td>147.0</td>
<td>2</td>
<td>755–2099 (1274)</td>
<td>16–24 (20)</td>
<td>158–1677 (770)</td>
</tr>
<tr>
<td>South Pare</td>
<td>331.0</td>
<td>6</td>
<td>541–2454 (1384)</td>
<td>13–24 (20)</td>
<td>359–2947 (1100)</td>
</tr>
<tr>
<td>West Usambara</td>
<td>528.8</td>
<td>14</td>
<td>408–2294 (1365)</td>
<td>13–25 (18)</td>
<td>393–3126 (1005)</td>
</tr>
<tr>
<td>East Usambara</td>
<td>391.4</td>
<td>5</td>
<td>124–1484 (628)</td>
<td>17–26 (22)</td>
<td>529–2788 (1176)</td>
</tr>
<tr>
<td>Nguu</td>
<td>416.8</td>
<td>13</td>
<td>709–1998 (1232)</td>
<td>16–23 (20)</td>
<td>333–3543 (1243)</td>
</tr>
<tr>
<td>Nguru</td>
<td>471.8</td>
<td>7</td>
<td>350–2382 (1243)</td>
<td>14–26 (20)</td>
<td>222–3814 (1706)</td>
</tr>
<tr>
<td>Ukaguru</td>
<td>197.3</td>
<td>6</td>
<td>885–2259 (1693)</td>
<td>15–23 (18)</td>
<td>634–2352 (1537)</td>
</tr>
<tr>
<td>Uluguru</td>
<td>308.5</td>
<td>9</td>
<td>255–2636 (1691)</td>
<td>12–27 (18)</td>
<td>579–2352 (1482)</td>
</tr>
<tr>
<td>Malundwe</td>
<td>2.3</td>
<td>1</td>
<td>793–1259 (1054)</td>
<td>20–23 (21)</td>
<td>978–1469 (1132)</td>
</tr>
<tr>
<td>Rubeho</td>
<td>530.7</td>
<td>16</td>
<td>565–2334 (1700)</td>
<td>15–25 (18)</td>
<td>281–1415 (822)</td>
</tr>
<tr>
<td>Udzungwa</td>
<td>1673.2</td>
<td>32</td>
<td>278–2555 (1390)</td>
<td>13–26 (20)</td>
<td>388–2470 (1346)</td>
</tr>
<tr>
<td>Mahenge</td>
<td>70.5</td>
<td>3</td>
<td>347–1478 (749)</td>
<td>18–26 (23)</td>
<td>1100–3238 (1813)</td>
</tr>
<tr>
<td>All Eastern Arc Mountains</td>
<td>5076.4</td>
<td>130</td>
<td>124–2636 (1352)</td>
<td>12–27 (20)</td>
<td>158–3814 (1257)</td>
</tr>
</tbody>
</table>

*Pellikka *et al.* (2009)
slopes the lowest. Soil parameters were obtained from the Soil and Terrain Digital Database (SOTER) and include soil reaction (pH), cation exchange capacity and available water capacity (Batje, 2004).

**Model calibration**

Spatial data were projected to UTM zone 37S and resampled to 1 or 2 km, depending on the taxon. Observed distributions were related to environmental predictors using generalized additive models (GAMs), calibrated using logit link functions and binomial error terms and allowing between one and four degrees of freedom for smoothers (Yee & Mitchell, 1991). For statistical calculations and the manipulation of map layers, we used R 2.10.0 (R Development Core Team, 2009) and GRASS GIS 6.3.0 (GRASS Development Team, 2009).

**Background data**

As is often the case when working with plot and herbarium data, ground-truthed absences were not available. Instead, we generated pseudo-absence (background) data to constrain the models. Because presence localities were spatially biased, it was appropriate to impose similar bias on the background data (Phillips et al., 2009). In a previous application of this approach, we targeted pseudo-absences for EAM tree species towards locations known to have been surveyed using similar methods (Platts et al., 2008). Here, we extend this methodology to consider separately the four different growth forms of plants – tree data are more plentiful than herb data, for example, not because tree species are necessarily more abundant but because vegetation plot assessments (c. 70% of our data) often target plants of a minimum size (e.g. ≥10 cm diameter at reference height c. 1.3 m). Thus, background data were placed only in locations where a matching growth form of plant has been sampled in the past (excluding presence sites for that taxon), using a ratio of five absence points for every presence point (Appendix S2).

**Predictor selection**

Two pairs of predictors were strongly collinear: mean annual temperature vs. potential evapotranspiration, and aspect north vs. aspect south-east (Table 2). These were reduced prior to modelling by constructing additive models separately for each taxon-predictor pair and retaining whichever yielded the strongest prediction. Minimal predictor sets were then identified using forward-backward selection, beginning with a null model and adding or removing terms iteratively according to Akaike Information Criterion. Next, we sought alternative solutions using backward-forward selection, beginning with a full model and removing or adding terms according to Bayesian Information Criterion. In each case, the most powerful predictive model was selected by cross-validating the area under the receiver-operating characteristic curve (AUC) – a threshold independent measure that incorporates
both type I and type II error rates (Green & Swets, 1974). We used a fivefold cross-validation procedure (80:20 training-testing split) stratified with respect to prevalence and averaged over ten independent runs (Parker et al., 2007). These ‘best-model’ solutions were combined in performance-weighted averages to give multimodel estimates of occurrence.

Spatial autocorrelation

A common problem with using regression techniques in ecology is that environmental variables are rarely sufficient to explain fully spatial dependence in species data (Dormann et al., 2007; Miller et al., 2007). Consequently, model residuals exhibit spatial structure, violating the statistical assumption that they are independent and identically distributed. Spatial autocorrelation in model predictions was parameterized by appending autocovariate terms to the GAM formulae (Augustin et al., 1996):

$$\log\left(\frac{P_i}{1 - P_i}\right) = \alpha + \beta_1 \text{cov}_1 + \ldots + \beta_n \text{cov}_n + \beta_n \text{autocov}_i$$

where $P_i$ is the probability of occurrence in focal cell $i$, and autocov$_i$ is a distance-weighted average of occurrence probabilities in surrounding grid cells (neighbourhood size = 10 km). There is a risk, however, that autocovariate models may underestimate environmental controls on species distributions, resulting in less stable predictions (Dormann, 2007; Platts et al., 2008). Autocovariate terms were therefore retained if and only if they improved predictive performance on unseen data (fivefold AUC).

Testing and validation

In addition to the measures of model performance employed during calibration, final model predictions were further validated using a fully independent test set. These presence data were omitted from calibration because of low or uncertain spatial accuracy but remained useful for gauging the sensitivity of predictions, and in particular the ability of models to predict occurrence in novel mountain blocs, i.e. those within a plant’s documented range but that were not represented in the presence data for that taxon. Test data accurate to c. 2 km were available for 286 taxa (1956 records); data with lower spatial accuracy were available for 341 (1578 records). The pattern was similar across growth forms, but only significant for trees. Tree models were particularly stable, retaining significantly more of the AUC under cross-validation than models for lianas, shrubs or herbs (Appendix S3).

The extent to which sampling distributions captured the range of environmental conditions in EAM forests was investigated using envelope uncertainty maps – spatial representations of where and to what extent particular models were extrapolated beyond the niche-breadth of the training data (Appendix S3; Platts et al., 2008).

Richness estimates

Plant richness was calculated by summing maps of estimated presence-absence over all taxa in a target group (e.g. trees or endemics). Distribution models predicted occurrence on a continuous scale, from 0 to 1; these predictions were dichotomised using taxon-specific occurrence thresholds, chosen by maximizing the sum of sensitivity and specificity (Appendix S2).

Because of uncertain colonization histories, we produced three versions of each richness map. First, model predictions were extrapolated to all forested grid squares, regardless of location. Richness maps derived from these estimates are tentative predictions, because they assume no historical barriers to dispersal. Second, models were extrapolated only to mountain blocs within a plant’s documented range. Derived richness is less speculative but biased by the level of research. Third, we map the disparity between predicted and confirmed richness, giving an indication of which areas should be prioritised for future exploration.

RESULTS

Model performance

According to validation statistics, models performed well and were rarely forced to extrapolate far beyond the niche-breadth used for calibration (Table 3 & Appendix S3). The balance of errors favoured correctly predicted presences (higher sensitivity), which is preferable because presence locations have been ground-truthed whereas background data are likely to contain genuine misclassifications. Even so, fully independent tests revealed that models for endemic taxa often failed to predict known occurrences accurately (median error = 4.24 km), especially in blocs beyond the spatial range of training data (Table 3). The sensitivity of novel-bloc predictions was also comparatively low for threatened taxa.

When training data were reused for testing, models calibrated at 2-km resolution outperformed those calibrated at 1-km resolution, but for unseen data 1-km models were significantly better (fivefold AUC, $P < 0.001$; Wilcoxon rank sum). The pattern was similar across growth forms, but only significant for trees. Tree models were particularly stable, retaining significantly more of the AUC under cross-validation than models for lianas, shrubs or herbs (Appendix S3).

The two alternative stepwise models frequently returned different solutions (21% agreement), but predictive performance was similar. On average, forward–backward models were smaller than backward–forward models (mean number of predictors = 3 and 4, respectively) and so were preferred for inferring causal relationships (Table 2). Temperature variables were the most often selected, reflecting the importance of altitude in determining species distributions in mountainous regions. Predictors of moisture availability, including cloud frequency, were also important, as were slope orientation and cation exchange capacity. The least selected predictor was soil acidity, although it contributed highly when included (Table 2). Response shapes for soil variables were not always sensible, indicating that they captured broad geographical patterns rather than functional relationships (see also Appendix S5).
Spatial autocovariates were retained in 30% of cases, more often in larger (backward–forward selection) and more stable (1 km) models. The median increase in explained deviance was only 6%, so environmental constraints were well-represented alongside spatial terms.

**Sampling bias**

Bias in exploration history was quantified by survey intensity, which we calculated at bloc level using all available data. The East Usambaras and Udzungwas are by far the best researched blocs, each with 20,000–30,000 data points. There is a steep drop to the Ulugurus and West Usambaras (6000–8000), followed by Nguru and Rubeho (3000–4000), South Pare, Mahenge then Ukaguru (1000–3000). The Taita Hills, North Pare, Nguu and Malundwe have fewer than 500 records amongst them. Tree species dominate, accounting for over 80% of specimens in most blocs (60% in Taita and South Pare); the remainder are mainly shrub and herb records, with lianas accounting for <5%.

The relationship between the number of modelled taxa observed in each mountain bloc and the number predicted to have potential niche-space was highly significant (Fig. 2a), reflecting both genuine biogeographical patterns and spatial bias in exploration history. Survey intensity explained 89% of the deviance in observed plant richness (log-linear relationship). The fit was lower for predicted richness (66%) with a shallower gradient, but still highly significant; Malundwe Mountain was an outlier with models predicting fewer taxa than expected (Cook’s distance = 1.2).

For species of conservation concern, the fit was stronger for predicted richness than for observed richness, and the gradient of the slope remained comparatively steep (Fig. 2c). This may be a consequence of non-climatic factors such as isolation: survey intensity and environmental correlates predict similar richness in Rubeho and South Pare, yet observed richness is very different. Combined with lower performance in independent tests (Table 3), we find that endemic and narrow-ranged taxa may be particularly sensitive to sampling bias.

**Richness**

**Confirmed at bloc level**

Extrapolating predictions within but not between mountain blocs, Fig. 3a shows a clear bias towards better studied regions, especially the East Usambaras and Udzungwas. Localized richness was also high in parts of South Pare, Uluguru and Rubeho. Average richness across grid cells in West Usambaras was comparatively low given that it ranked second at the bloc resolution (modelling subset, Table 4). Fig. 4 shows that many taxa in this bloc were not predicted to be widespread in larger forests, suggesting high species turnover. The same may be true of Nguru, which is also ranked higher than the 1-km map suggests (cf. Fig. 3a & Table 4). In South Pare, richness was concentrated mainly in Chome Forest Reserve, reflecting a bias in collection localities.

Endemic and threatened taxa were most prevalent across grid cells in the Uluguru and Usambara Mountains (Fig. 3b), with the South Pares and parts of Udzungwa also important. Compared with overall richness, relative concentrations were higher in Nguru and Ukaguru, and lower in Rubeho and Udzungwa, although the bloc total for Udzungwa was still high (ranked fourth in Table 4). In Table 5, we provide details of 18 taxa that are both endemic to the EAMs and threatened with extinction, including area-based recommendations for the IUCN Red List.

**Predictive estimates**

Predicted richness was greater than observed richness in all cases, with the size of the disparity showing a negative

---

**Table 3** Summary of model performance. Explained deviance, area under the receiver-operator characteristic curve (AUC) including a fivefold cross-validation, and the proportion of presences (sensitivity) and pseudo-absences (specificity) classified correctly. Figures shown are median values because of negative skew. Using high-resolution independent test data (c. 2 km accuracy), we present the median distance to the nearest predicted occurrence. Using all available test data (bloc-level accuracy), we assess the ability of models to predict occurrence successfully in novel mountain blocs (those with no presence points in the training data): mean sensitivity ± one standard deviation (medians = 1). See also Appendix S3.

<table>
<thead>
<tr>
<th></th>
<th>N</th>
<th>Explained deviance</th>
<th>AUC*</th>
<th>5-fold AUC</th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>Distance to nearest occurrence (km)</th>
<th>Novel-bloc sensitivity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Trees</td>
<td>302</td>
<td>0.66</td>
<td>0.95</td>
<td>0.87</td>
<td>0.94</td>
<td>0.87</td>
<td>1.00</td>
<td>0.91 ± 0.25</td>
</tr>
<tr>
<td>Lianas</td>
<td>12</td>
<td>0.60</td>
<td>0.94</td>
<td>0.82</td>
<td>0.90</td>
<td>0.91</td>
<td>0.00</td>
<td>1.00 ± 0.00</td>
</tr>
<tr>
<td>Shrubs</td>
<td>62</td>
<td>0.67</td>
<td>0.95</td>
<td>0.83</td>
<td>0.94</td>
<td>0.92</td>
<td>0.00</td>
<td>0.92 ± 0.25</td>
</tr>
<tr>
<td>Herbs</td>
<td>74</td>
<td>0.62</td>
<td>0.93</td>
<td>0.79</td>
<td>0.93</td>
<td>0.89</td>
<td>0.00</td>
<td>0.94 ± 0.20</td>
</tr>
<tr>
<td>Endemic</td>
<td>25</td>
<td>0.73</td>
<td>0.98</td>
<td>0.89</td>
<td>1.00</td>
<td>0.89</td>
<td>4.24</td>
<td>0.87 ± 0.26</td>
</tr>
<tr>
<td>Threatened</td>
<td>68</td>
<td>0.71</td>
<td>0.96</td>
<td>0.88</td>
<td>0.99</td>
<td>0.91</td>
<td>1.00</td>
<td>0.85 ± 0.30</td>
</tr>
<tr>
<td>All species</td>
<td>452</td>
<td>0.65</td>
<td>0.95</td>
<td>0.86</td>
<td>0.94</td>
<td>0.89</td>
<td>1.00</td>
<td>0.92 ± 0.24</td>
</tr>
</tbody>
</table>

* AUC: 0.5–0.7, better than chance; 0.7–0.9, good performance; 0.9–1.0, excellent performance (Swets, 1988)
correlation with survey intensity (Fig. 2b–c). Unconfirmed but potentially suitable habitat was therefore most common in Taita, North Pare, Nguu and Malundwe. Environmental conditions in Nguru, Ukaguru, Rubeho and Mahenge suggest these areas could also support higher concentrations of species than currently documented (Figs 3 & 4). Compared with observed richness at the bloc level, predicted richness ranked the Nguru and Rubeho Mountains above South Pare (Table 4). Also, North Pare and Nguu were ranked above Ukaguru and Mahenge despite sampling bias in favour of the latter. Predictive rankings for endemic and threatened taxa followed a similar pattern, except that the Ulugurus were ranked slightly lower, and the importance of Mahenge is predicted to be higher than inventory data suggest (Table 4 & Appendix S4).

Growth form

Figure 5 shows patterns of richness to be similar across growth forms, with the notable exception that tree richness is highest in the two most researched mountain blocs (East Usambara and Udzungwa), whereas lianas, shrubs and herbs have equally high (confirmed) or higher (predicted) richness in other areas, particularly the West Usambara and Rubeho Mountains.

DISCUSSION

The prioritization of areas for conservation within the EAMs has tended to change with the availability of new field data. First, the Usambaras and Ulugurus were ranked most important; subsequently, the importance of Udzungwa was recognized, followed by Nguru and now Rubeho (CEPF, 2003; Doggart et al., 2006). This reshuffling of conservation priorities is symptomatic of a paucity of survey data common to many high biodiversity regions and highlights the need for strategically targeted field sampling. Distribution models are an appealing tool for obtaining high-resolution estimates of richness in well-researched areas, and tentative estimates of conservation importance elsewhere. Alongside other considerations such as threats to habitat, richness in other taxonomic groups and ecosystem value (e.g. carbon stocks, hydrology, natural resources, ecotourism; Naidoo et al., 2008), they could form part of a more consistent approach to conservation priority setting and strategic planning of surveys.

In many cases, the data available for modelling are biased both in geographical space and towards particular groups of organisms. Here, tree data were the most abundant and tree models the most stable. Our results suggest that if the bias were towards lianas, shrubs or herbs, instead of trees, then we might favour the mountain blocs in a slightly different order. Faced with insufficient data, conservation planners must determine the degree to which different taxonomic groups and growth forms can serve as surrogates for each other in the prioritization of areas for conservation (Burgess et al., 2006). We find that even within the group of vascular plants, it is preferable to consider all growth forms in the analysis of conservation.
priority. Low levels of congruence have also been reported for vertebrates (Grenyer et al., 2006) and when comparing patterns of endemism across a range of taxonomic groups (Kremen et al., 2008).

Because of broad-scale geographical bias in the occurrence data, coupled with uncertain colonization histories, we have been careful to distinguish between those mountain blocs where a taxon is known to occur and those where it is to-date undocumented. When dispersal limitations are not considered, models predict that richness could be more evenly distributed across the mountains than is currently documented (Fig. 3). In the 2003 Ecosystem Profile of the EAMs and Coastal Forests (CEPF, 2003), the Usambaras, Ulugurus and Udzungwas were identified as being the most species-rich blocs. Predictive

Figure 3 Spatial estimates of plant richness calculated across (a) all taxa and (b) taxa of conservation concern. Scale bars show the number of taxa predicted to have potential niche-space in 1-km grid squares. In the left panel, modelled distributions are extrapolated to all forest patches with suitable environmental conditions. In the centre panel, predictions are restricted to just those mountain blocs where the respective taxa have been confirmed present. The right panel shows predictions of occurrence in unconfirmed blocs (left panel minus centre panel) – we suggest this map can be helpful in selecting future sites for exploration.
estimates largely confirm this ranking, whilst indicating that the importance of Nguru and Rubeho may still be underestimated, particularly for rare species (see also Doggart et al., 2006). Lesser researched blocs, especially North Pare and Nguu, could also be important, following higher rankings despite low survey intensity. Predictions such as these could be verified and subsequently refined by ongoing and targeted field assessments (Guisan et al., 2006).

Using independent test data, we found that models were generally quite successful at predicting occurrence in novel mountain blocs. These validations were, however, limited to bloc-level sensitivity, so the extent of over-prediction remains uncertain. Models for threatened and endemic taxa were most likely to under-predict when extrapolated into novel blocs, indicating gaps in the documented environmental niche. This could be a problem for wider-ranging taxa too, for it is difficult to know whether or not the complete range of conditions under which a taxon exists has been sampled. Further, we suspect that in some cases the soil predictors, which vary broadly by mountain bloc, simply identified spatial biases in the sampling distribution, rather than truly casual factors. Given the sensitivity of predictions to survey intensity and the

Table 4  Conservation priorities, based on the number of plant taxa confirmed or predicted in each mountain bloc. These rankings are not corrected for forest area, and therefore favour larger mountain blocs such as Udzungwa. See Figs 3 & 4 for fine-scale richness estimates.

<table>
<thead>
<tr>
<th>Rank</th>
<th>All taxa</th>
<th>Modelled taxa (confirmed)</th>
<th>Modelled taxa (predicted)</th>
<th>Endemic and/or threatened taxa</th>
<th>Modelled taxa (confirmed)</th>
<th>Modelled taxa (predicted)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Udzungwa 2546</td>
<td>382</td>
<td>425</td>
<td>Udzungwa 319</td>
<td>51</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>E. Usambara 1108</td>
<td>337</td>
<td>417</td>
<td>E. Usambara 233</td>
<td>43</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>South Pare 894</td>
<td>302</td>
<td>404</td>
<td>South Pare 75</td>
<td>22</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>Uluguru 835</td>
<td>255</td>
<td>398</td>
<td>Uluguru 58</td>
<td>15</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>W. Usambara 713</td>
<td>246</td>
<td>389</td>
<td>W. Usambara 162</td>
<td>41</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>Rubeho 665</td>
<td>206</td>
<td>383</td>
<td>Rubeho 58</td>
<td>10</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>Nguru 658</td>
<td>203</td>
<td>358</td>
<td>Nguru 311</td>
<td>8</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>Mahenge 583</td>
<td>125</td>
<td>350</td>
<td>Mahenge 58</td>
<td>44</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>North Pare 108</td>
<td>59</td>
<td>311</td>
<td>North Pare 13</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>Ukaguru 103</td>
<td>59</td>
<td>306</td>
<td>Ukaguru 23</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>Taita Hills 57</td>
<td>27</td>
<td>299</td>
<td>Taita Hills 23</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>Malundwe 31</td>
<td>11</td>
<td>283</td>
<td>Taita Hills 23</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>Nguu 29</td>
<td>5</td>
<td>167</td>
<td>Nguu 3</td>
<td>0</td>
<td></td>
</tr>
</tbody>
</table>

Figure 4  Histograms showing patterns of within-bloc patch occupancy. Horizontal axes represent the largest contiguous area of forest providing environmentally suitable conditions for a particular taxon. Vertical axes show the number of taxa in each 30-km² patch size interval. Filled bars relate to confirmed occurrence at the bloc level; open bars relate to predictions of occurrence in novel mountain blocs. Total forest area in each bloc shown in parentheses (km²). Patterns for endemic/threatened taxa are presented in Appendix S4.
Table 5 Model estimates of the habitat available for 18 plant taxa endemic to the Eastern Arc Mountains, presented in descending order of rarity. Also shown are the current IUCN Red List designations (no designation for eight taxa; IUCN, 2009), the proposed Red List status of each taxon in an ongoing assessment of plant conservation in East Africa (Gereau et al., 2010), and recommendations based solely on environmentally suitable habitat in mountain blocs where taxa are known to occur: critically endangered (CR), area of occupancy < 10 km²; endangered (EN), area of occupancy < 500 km²; vulnerable (VU), area of occupancy < 2000 km² (cf. Hall et al., 2009). Note that areas of occupancy are not the only consideration in determining the level of threat: Englerodendron usambarense has a very narrow range but is proposed as Not Threatened because it is well protected within Amani Nature Reserve. For full details of Red List categories and criteria visit http://www.iucnredlist.org/.

<table>
<thead>
<tr>
<th>Endemic species or infra-specific taxon</th>
<th>Growth form</th>
<th>Mountain bloc coverage</th>
<th>Suitable habitat (km²)</th>
<th>Current IUCN listing (version 2.3 or 3.1)</th>
<th>Proposed threat status</th>
<th>Area-based Recommendation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cynometra longipedicellata</td>
<td>Tree</td>
<td>eU</td>
<td>132</td>
<td>VU [B1 + 2b], ver. 2.3</td>
<td>CR</td>
<td>EN</td>
</tr>
<tr>
<td>Englerodendron usambarense</td>
<td>Tree</td>
<td>eU</td>
<td>156</td>
<td>VU [B1 + 2c], ver. 2.3</td>
<td>NT</td>
<td>EN</td>
</tr>
<tr>
<td>Mammeea usambarensis</td>
<td>Tree</td>
<td>sP, wU</td>
<td>157</td>
<td>VU [B1 + 2b], ver. 2.3</td>
<td>VU</td>
<td>EN</td>
</tr>
<tr>
<td>Allophylus mollidiorus</td>
<td>Tree</td>
<td>wU, eU, Nr</td>
<td>214</td>
<td>VU [B1 + 2b]</td>
<td>PT</td>
<td>EN</td>
</tr>
<tr>
<td>Eugenia tosananatolica</td>
<td>Tree</td>
<td>sP, wU, Mh</td>
<td>233</td>
<td>VU [B1 + 2b], ver. 2.3</td>
<td>PT</td>
<td>EN</td>
</tr>
<tr>
<td>Cola usambarensis</td>
<td>Tree</td>
<td>eU</td>
<td>243</td>
<td>VU [B1 + 2b], ver. 2.3</td>
<td>PT</td>
<td>EN</td>
</tr>
<tr>
<td>Musaenda microdonta subsp. microdonta</td>
<td>Tree</td>
<td>wU, Nr, Ul</td>
<td>295</td>
<td>VU [B1 + 2b], ver. 2.3</td>
<td>PT</td>
<td>EN</td>
</tr>
<tr>
<td>Memecylon cogniauxii</td>
<td>Shrub</td>
<td>sP, wU, eU, Nr, Ul</td>
<td>302</td>
<td>PT [B1 + 2b], ver. 2.3</td>
<td>EN</td>
<td>EN</td>
</tr>
<tr>
<td>Casearia engleri</td>
<td>Tree</td>
<td>sP, wU</td>
<td>328</td>
<td>VU [B1 + 2b], ver. 2.3</td>
<td>EN</td>
<td>EN</td>
</tr>
<tr>
<td>Syzygium micklethwaitii</td>
<td>Tree</td>
<td>sP, wU, Nr, Uk, Ul</td>
<td>468</td>
<td>VU [B1 + 2b], ver. 2.3</td>
<td>EN</td>
<td>EN</td>
</tr>
<tr>
<td>Coffea mangensis</td>
<td>Tree</td>
<td>wU, eU, Nr, Ud</td>
<td>535</td>
<td>VU [B1 + 2b], ver. 2.3</td>
<td>LC</td>
<td>VU</td>
</tr>
<tr>
<td>Impatiens pallidereosa</td>
<td>Herb</td>
<td>Uk, Ul, Ru</td>
<td>543</td>
<td>VU [B1 + 2b], ver. 2.3</td>
<td>VU</td>
<td>VU</td>
</tr>
<tr>
<td>Craterispermum longipedunculatum</td>
<td>Tree</td>
<td>Ul, Ud</td>
<td>712</td>
<td>VU [B1 + 2b], ver. 2.3</td>
<td>PT</td>
<td>VU</td>
</tr>
<tr>
<td>Lasiandathus pedunculatus</td>
<td>Tree</td>
<td>Nr, Uk, Ul, Ud</td>
<td>867</td>
<td>VU [B1 + 2b], ver. 2.3</td>
<td>PT</td>
<td>VU</td>
</tr>
<tr>
<td>Zerokaella capparidacea</td>
<td>Tree</td>
<td>wU, eU, Nr, Ul</td>
<td>872</td>
<td>VU [B1 + 2b], ver. 2.3</td>
<td>EN</td>
<td>VU</td>
</tr>
<tr>
<td>Polyscias stuhlmannii</td>
<td>Tree</td>
<td>sP, wU, Uk, Ul</td>
<td>933</td>
<td>EN B2ab(iii), ver. 3.1</td>
<td>EN</td>
<td>VU</td>
</tr>
<tr>
<td>Dicranolepis usambarica</td>
<td>Tree</td>
<td>Ta, sP, wU, eU, Nr, Ul</td>
<td>996</td>
<td>PT [B1 + 2b], ver. 2.3</td>
<td>PT</td>
<td>VU</td>
</tr>
<tr>
<td>Allamblackia ulugurensis</td>
<td>Tree</td>
<td>Nr, Uk, Ul, Ud</td>
<td>1133</td>
<td>VU [B1 + 2c], ver. 2.3</td>
<td>VU</td>
<td>VU</td>
</tr>
</tbody>
</table>

CR, critically endangered; EN, endangered; VU, vulnerable; NT, near threatened; LC, least concern; PT, potentially threatened but not yet evaluated; B1, restricted extent of occurrence; 2b, area of occupancy continuing to decline; 2c, extent and/or quality of habitat declining; Ta, Taita; nP, North Pare; sP, South Pare; wU, West Usambara; eU, East Usambara; Nu, Nguu; Nr, Nguru; Uk, Ukaguru; Ul, Uluguru; MI, Malundwe; Ru, Rubeho; Ud, Udzungwa; Mh, Mahenge.

fact that realized distributions of EAM endemics are highly dependent on past connectivity, we caution that it is for the taxa of highest conservation concern that predictive estimates are most uncertain.

Restricting analyses to confirmed blocs only, we find that environmental conditions across most forests in Udzungwa have potential to support large numbers of plant taxa; concentrations of rare and endemic taxa, meanwhile, are predicted to be lower than in the Usambaras and Ulugurus – possibly a real pattern given the close proximity of non-EAM habitats. Mahenge is predicted to be suitable for many of the rare plants modelled here, but occurrence is unconfirmed in most cases. The Usambaras and Ulugurus are better known centres of endemism (Iversen, 1991; Temu & Andrew, 2008), promoted by geographical isolation and exposure to rain-bearing ocean winds. High levels of endemism have also been recorded in the Taita Hills (Beentje, 1988, 1994); however, this bloc is not well represented in our database, leading models to under-estimate its importance. Forests in Taita are of particular conservation concern, having been reduced to just a few remnant patches (Rogo & Oguge, 2000; Pellikka et al., 2009).

Human activity has resulted in widespread fragmentation and degradation of many tropical forests, yet modelled estimates of diversity often do not consider the minimum forest area required for species persistence, nor the vulnerability of small fragments to degradation. Here, we map forest cover using remotely sensed land cover data. Whilst these estimates are not without error, they can at least be indicative of potential threats. We show that many taxa, especially those predicted to occur in blocs beyond their documented range, have suitable conditions only in relatively small forest patches (Fig. 4). Species across many taxonomic groups are less likely to persist in smaller and more isolated habitats, even if environmental conditions are favourable (MacArthur & Wilson, 1967; Lomolino, 2000; Marshall et al., 2010). Around one-fifth of the forests we identify from the land cover map are both smaller than 1 km² and more than 1 km from another patch. Much of this fragmentation is relatively recent, so in
many cases the extinction debt has yet to be realized. In less isolated fragments, long-term persistence might be possible via seed recruitment from neighbouring populations (Lehouck et al., 2009) – it is therefore imperative to conserve forests of all sizes to maintain connectivity (Fjeldså & Lovett, 1997a). Although not considered here, there is scope to address such patch dynamics post hoc by linking predicted distributions with spatially explicit population models (Keith et al., 2008).

Exacerbated by forest loss, the extinction risk for narrow-range endemics is considerable (Brooks et al., 2001). The tree *Platypterocarpus tanganyikensis* Dunkley & Brenan was once found in the West Usambara Mountains, but collections show no record of its presence since 1953, even before high levels of forest clearance in the 1960s. Distributions models for rare species require particular scrutiny, but as part of a wider assessment they can be useful for indicating the appropriate level of threat on the IUCN Red List (Table 5). One of the rarest endemics modelled here is the tree *Cynometra longipedicellata* Harms, known only from the East Usambaras. Models identify potential niche-space in Mahenge, but this species is more likely endemic to north-eastern Tanzania. We estimate the area of occupancy to be c. 132 km², probably less given competition for niche-space and other factors beyond the scope of our models (Pulliam, 2000). Based on the tree’s observed altitudinal range, Hall et al. (2009) estimate that *C. longipedicellata* may have only 70 km² of habitat remaining, a decrease of over 70% since 1955. This species is currently assessed as Vulnerable (IUCN, 2009); we recommend elevating the threat status to Endangered, EN B1ab(iii) + B2ab(iii) (extent of occurrence < 5000 km², area of occupancy < 500 km², extent and/or quality of habitat declining) or Critically Endangered, CR B1ab(iii) (extent of occurrence < 100 km²).

Patterns of endemism are often complex (Jetz et al., 2004). Our perceptions of these patterns and our ability to identify causal factors are likely to be influenced by the spatial resolution used for modelling (Whittaker et al., 2001; Rahbek, 2005). We find that higher resolution models are more stable,

---

**Figure 5** Box plots detailing how plant richness varies according to growth form. In the left panel, modelled distributions are extrapolated to all forest patches with suitable environmental conditions. In the right panel, predictions are restricted to just those mountain blocs where the respective taxa have been confirmed present. Box widths are proportional to the area of forest remaining in each mountain bloc.
presumably because micro-climatic conditions are better represented. High levels of endemism in the EAMs have been attributed to historical isolation coupled with long-term climatic stability, with persistent orographic rainfall and mist having minimized climatically linked extinctions (Fjeldså & Lovett, 1997b; Fjeldså et al., 1997). Recent pollen analyses confirm that whilst there were shifts in abundance, few if any plant taxa were lost during the last glacial maximum (Mumbi et al., 2008; Finch et al., 2009). Analysis of model predictions also suggests that moisture is a key driver for concentrations of endemism, with the annual moisture index explaining 31% of deviance across forested grid squares (Appendix S5). Similarly, other studies have found contemporary rainfall to be a good predictor of endemism in the EAMs (Fjeldså & Lovett, 1997b) and of range-size rarity in West Africa (Holmgren & Poorter, 2007). Cloud cover explains little of the spatial variation in endemism but was an important predictor for some of the rarest plants (e.g. C. longipedicellata). The correlation between cloud frequency and overall richness was higher (13% explained deviance), with frequencies over 50% promoting climatic suitability for the most taxa (Appendix S5). Annual temperature range was the best climatic predictor of modelled richness (24%), with lower seasonality correlating with higher diversity. Given the importance of the moisture index, these results suggest that measures of seasonal constancy in the water balance might be worth including in future studies.

CONCLUSIONS

The application of distribution models to plant inventory data can provide useful indications of which areas may be important for biodiversity conservation, and offers a means to estimate the niche-space available for species of conservation concern. Whilst models are highly sensitive to spatial bias in the inventory data, especially for rare species, we suggest that predictive definitions of conservation priority could be systematically improved by targeting field sampling towards locations with large discrepancies between observed and predicted diversity. As improvements in data quality cease to increase model stability, the limits of environmental controls on species’ distributions will become clearer, providing a baseline by which to quantify the roles of historical and non-climatic factors in shaping contemporary patterns of biodiversity. Our results indicate that it is necessary to consider all growth forms of plants in the prioritization of sites for conservation, and so we draw attention to the sometimes excessive dominance of tree species in botanical inventories.

ACKNOWLEDGEMENTS

Many thanks to all who contributed to the species database, including field researchers at Frontier-Tanzania, data providers of the Missouri Botanical Garden’s TROPICOS database and John Hall (University of Wales, UK). Thanks also to Ruth Swetnam and Jon Green (University of Cambridge, UK), Neil Burgess and Jon Fjeldså (University of Copenhagen, Denmark), Boniface Mbilyini (Sokoine University of Agriculture, Tanzania) and Barnaby Clark (University of Helsinki, Finland) for their work on land cover classifications. Logistical support and permissions for fieldwork in Tanzania were provided by the Tanzanian Commission for Science and Technology, the Forestry and Beekeeping Division of the Ministry of Natural Resources and Tourism, the University of Dar es Salaam and WWF Tanzania. This research was funded by the Marie-Curie programme of the European 6th Framework under contract MEXT-CT-2004–517098 to R. Marchant, and contributes to the Global Land Project (http://www.globallandproject.org) and to the Valuing the Arc Programme (http://www.valuingthe-arc.org). The latter is funded by the Leverhulme Trust with additional support from the Royal Society. Marshall was funded by Natural Environment Research Council grant NER/S/A/2002/11177 and by the National Geographic Society and Margot Marsh Biodiversity Foundation. Funding from the Academy of Finland for the Taita Project (http://www.helsinki.fi/science/taita/) is also gratefully acknowledged. Finally, we thank Janet Franklin (Associate Editor), Alistair Jump (University of Stirling, UK), Daniel Kissling (Aarhus University, Denmark) and two anonymous reviewers, whose comments greatly improved this manuscript.

REFERENCES


Burgess, N.D., Hales, J.D., Ricketts, T.H. & Dinerstein, E. (2006) Factoring species, non-species values and threats into


MNRT (1997) *National reconnaissance level land use and natural resources mapping project*. Ministry of Natural Resources and Tourism, The United Republic of Tanzania.


Platts, P.J., McClean, C.J., Lovett, J.C. 


Possingham, H.P. 


Poynton, J.C., Loader, S.P., Sherratt, E. 


Reddy, S. 


Rogo, L. 


Stattersfield, A.J., Crosby, M.J., Long, A.J. 


Temu, R.P.C. 


Whittaker, R.J., Willis, K.J. 


Wilson, K.A., McBride, M.F., Bode, M. 


Yee, T.W. 


**SUPPORTING INFORMATION**

Additional supporting information may be found in the online version of this article:

**Appendix S1** Details of plant taxa modelled.

**Appendix S2** Occurrence thresholds and sensitivity to prevalence.

**Appendix S3** Analysis of model performance.

**Appendix S4** Patch occupancy for endemic/threatened taxa.

**Appendix S5** Drivers of richness and endemism.

As a service to our authors and readers, this journal provides supporting information supplied by the authors. Such materials are peer-reviewed and may be re-organized for online delivery but are not copy-edited or typeset. Technical support issues arising from supporting information (other than missing files) should be addressed to the authors.

**BIOSKETCH**

Philip J. Platts is a graduate student at the York Institute for Tropical Ecosystem Dynamics (KITE), based in the Environment Department at the University of York. KITE is a Marie-Curie funded Excellence Centre that explores the relationship between ecosystem dynamics, climate change and human impacts in Africa. See our website for more information (http://www.york.ac.uk/res/kite/).

Author contributions: This study was conceived by P.J.P., who programmed and implemented the modelling experiments, analysed the results and prepared the manuscript. R.E.G. provided details of endemism and proposed Red List designations. A.A., E.B., R.E.G., J.C.L and A.R.M. contributed to the botanical database, which was compiled and cleaned by A.A. and R.E.G. Estimates of forest cover in the Taita Hills were provided by P.K.E.P. Rainfall and cloud data were provided by M.M. The study was supervised by R.M. and C.J.M. All authors discussed the results and commented on the manuscript.

Editor: Janet Franklin