What determines plant species diversity in Central Africa?

Andreas S.J. van Proosdij

What determines plant species diversity in Central Africa?

Andreas S.J. van Proosdij

Thesis committee

Promotor
Prof. Dr M.S.M. Sosef
Professor of Biosystematics
Wageningen University & Research

Co-promotors
Dr N. Raes
Postdoctorate Fellow Biodiversity Dynamics
Naturalis Biodiversity Center, Leiden

Dr J.J. Wieringa
Researcher / Assistant professor, Biosystematics
Naturalis Biodiversity Center, Leiden / Wageningen University & Research

Other members
Prof. Dr L. Poorter, Wageningen University & Research
Prof. Dr P. Linder, Zurich University, Switzerland
Prof. Dr O. Hardy, Université Libre de Bruxelles, Belgium
Prof. Dr H. ter Steege, Naturalis Biodiversity Center, Leiden & VU Amsterdam

This research was conducted under the auspices of the Graduate School for Production Ecology and Resource Conservation

What determines plant species diversity in Central Africa?

Andreas S.J. van Proosdij

Thesis
submitted in fulfilment of the requirements for the degree of doctor
at Wageningen University
by the authority of the Rector Magnificus,
Prof. Dr A.P.J. Mol,
in the presence of the
Thesis Committee appointed by the Academic Board
to be defended in public
on Friday 6 October 2017
at 1.30 p.m. in the Aula.

Andreas S.J. van Proosdij
What determines plant species diversity in Central Africa?,
161 pages.

PhD thesis, Wageningen University, Wageningen, the Netherlands
(2017)
With references, with summary in English

ISBN 978-94-6343-661-8
DOI: https://doi.org/10.18174/420928

Contents

Chapter 1

General introduction

A mega diverse world

Planet Earth hosts an overwhelming biological diversity. The term 'biodiversity' may refer to genetic diversity, species diversity, ecosystem diversity, and even more, but commonly species richness is used as a representation of biodiversity (Gaston, 1996a). The exact number of species occurring on our planet remains unknown, but estimates range from 5 to 11 million eukaryotic species, including 400,000-450,000 species of plants (Costello *et al.*, 2013, Joppa *et al.*, 2011, Mora *et al.*, 2011, Pimm *et al.*, 2014).

Ever since Europeans set out to explore the world, they noticed the overwhelming species richness of the Tropics. Throughout the past three centuries, living and dried plants, seeds, bulbs and tubers were sent to European scientists at universities and botanic gardens. This enabled the founders of modern taxonomy, including e.g. Linnaeus, to study, describe and name these species. In the 19th century, many of the great explorers and describers of natural science, such as Joseph Banks, Charles Darwin, Joseph Dalton Hooker, Alexander von Humboldt, and Alfred Russel Wallace, discussed the reason behind the exceptionally higher biodiversity in the Tropics compared to temperate regions (Wallace, 1876). This phenomenon is known as the latitudinal diversity gradient and has been documented for many taxonomic groups of plants and animals (Rosenzweig, 1995, Willig *et al.*, 2003). In terms of vascular plants, the Neotropics harbour approx. 90,000 species (Paton, 2013, Thomas, 1999), 41,000-42,500 species occur in the Malesian region (Malaysia, Singapore, Indonesia, Brunei, the Philippines and Papua New Guinea) (Paton, 2013, Roos, 1993) and 30,000-35,000 species of flowering plants in tropical Africa excluding Madagascar (Klopper *et al.*, 2007). In contrast, only 11,500 vascular plant species occur in Europe (Tutin, 1964-1980) and 20,500 species in North America (1993+, Paton, 2013). The latitudinal diversity gradient appears to be mainly driven by higher energy levels in the Tropics (Brown, 2014), although other factors may play a role too, such as climatic stability, larger habitat heterogeneity, larger areas for each climate zone near the Equator, as well as the central position of the Tropics on the globe (Gaston, 2000, Lomolino *et al.*, 2010). In addition to this unequal distribution of species richness on Earth, large differences exist between the range size or prevalence of individual species.

Rarity of species

In 1859, Charles Darwin noted "*Who can explain why one species ranges widely and is very numerous, and why another allied species has a narrow range and is rare? Yet these relations are of the highest importance, for they determine the present welfare and, as I believe, the future success* *and modification of every inhabitant of this world*". Since then, many researchers tried to answer this question for a wide array of taxa. Rarity can be defined in multiple ways, best summarized by Rabinowitz's (1981) 'seven forms of rarity', which are based on the following three criteria: a) species can be restricted to a small geographic area, b) species can have a narrow ecological niche, thus being restricted to rare biotopes, or c) species can have a low abundance, although being present in a large area. One of the most simple definitions of a geographic rare species is that of an 'endemic', a species that is restricted to a small area (Gaston, 1994). However, endemism is often defined using political boundaries, which often have limited ecological meaning. Moreover, the use of endemism is criticised as it is a qualitative classification: a species is either restricted to a specific area or not, hence its rarity is not quantified. Quantitative methods for describing the rarity of a species include applying discrete rarity classes (Marshall *et al.*, 2016) or applying a continuous system of weighted endemism (Crisp *et al.*, 2001).

The geographic rarity of a species is often quantified by its range size. Out of several methods to measure the geographic range size of a species, two are widely applied, e.g. for the IUCN Red List assessment of species (IUCN, 2001, IUCN, 2014). The Extent Of Occurrence (EOO) is the area within the shortest polygon encompassing all known occurrences. The Area Of Occupancy (AOO), in contrast, is the area that is actually occupied by the species, for which commonly a standard buffer area is applied around each known occurrence (Gaston, 1991, Gaston & Fuller, 2009, Lomolino *et al.*, 2010). Instead of known occurrences, predicted occurrences based on models can also be used to assess the range size (Syfert *et al.*, 2014).

Most species on Earth are rare and few are abundant, concluding from data based on e.g. plants (ter Steege *et al.*, 2013), birds (Gaston, 1996b), mammals (Schipper *et al.*, 2008), and fish (Magurran & Henderson, 2003). In other words, when the number of species is plotted against the range size, this range size frequency distribution (RSFD) is generally strongly right-skewed, thus towards narrow-ranged species. The same skew is observed when the number of species is plotted against the abundance of individual species.

African rainforests

The tropical rainforests (TRFs) of the Neotropics (South America), Africa and South-East Asia are the most species-rich terrestrial biomes on Earth (Gentry, 1992). They fulfil crucial ecosystems services including storage of carbon and production of oxygen, as well as the provision of food, timber, construction materials, medicines, and other non-timber forest products (Hassan *et al.*, 2005, Schulze & Mooney, 1994).

African rainforests are restricted to four phytogeographical regions: Upper Guinea (Liberia to Ghana), Lower Guinea (Nigeria to northern Angola), Congolia (the Congo Basin), and the Eastern Arc and coastal forests of Kenya and Tanzania (Linder, 2001, Sosef *et al.*, 2017, White, 1979). The Dahomey Gap separates the Upper Guinean from the Lower Guinean forests and is mainly covered by savannah. The Sangha river interval consists of extensive swamp forests that separate Lower Guinea from Congolia. Africa has been termed the "Odd man out" as African rainforests contain far less species than South American and Asian rainforests (Richards, 1973). Exact numbers are lacking, but recent estimations based on tropical trees indicate a total of 4,500-6,000 tree species for Africa versus 19,000-25,000 tree species for each of both other regions (Slik *et al.*, 2015). Based on plot data, tree alpha-diversity in Amazonian forests is twice as high compared to African forests, except for the driest and coolest African forests (Parmentier *et al.*, 2007). From the many hypothesized causes of this difference, higher speciation rates in the Neotropics and South-East Asia seem a more likely explanation than higher extinction rates in Africa (reviewed in Couvreur, 2015). In addition, African rainforests are overall drier and have probably experienced a much stronger decrease in range size during Pleistocene glacial periods; the latter is explained below.

In the past 2.5 million years, 21 cycles of global glaciations took place: 9 major and 12 minor. From fossil pollen records and carbon isotope data of vegetation and soils it is concluded that, in response to these climate changes, rainforests contracted into forest refugia during drier and cooler glacials and expanded again during wetter and warmer interglacials (Maley, 1996, Pietsch & Gautam, 2013, Plana, 2004). Where contraction of the rainforests resulted in a reduction of surface by 54% for Amazonian rainforests, African rainforests were reduced by as much as 84% (Anhuf *et al.*, 2006). The location of the forest refugia has been identified based on the composition of pollen in lake sediments (Maley, 1996), carbon isotope data (Pietsch & Gautam, 2013), and the presence of endemic species (White, 1979), or species that are known to be bad dispersers, such as e.g. *Begonia* species (Sosef, 1994), Caesalpinioid legume trees with ballistic seeds (Leal, 2004, Wieringa, 1999) and some Rubiaceae taxa (Robbrecht, 1996). The role of gallery forests as smaller but more widespread riverine forest refugia in addition to the larger montane forest refugia is illustrated by the restricted distribution of some of these plant species to gallery and montane forests (Leal, 2001, Robbrecht, 1996, Wieringa, 1999), as well as by the distribution of allelic endemism (Hardy *et al.*, 2013). The distribution of primates also suggests the survival of patches of lowland rainforest during glacials (Colyn *et al.*, 1991). During the last glacial period, the African rainforest reached its smallest size at the last glacial maximum (LGM) around 18,000 years BP. Around 12,000 years BP, the forest rapidly expanded until it reached its maximum size between 9,000 and 5,000 years BP, at that time even bridging the Dahomey Gap. Between 2,800 and 2,500 years BP, the Lower Guinean forest experienced a short but strong contraction due to increased precipitation seasonality and hence an extension of the dry season. Soon after, when precipitation increased again, the rainforest recovered and expanded to its current size that is slightly smaller than its previous maximum extent (Maley, 1996, Maley, 2002). Although some studies suggest an important role for humans in the Holocene rainforest contraction, following and reinforcing the climate-induced forest contraction around 2,800 years BP (Oslisly *et al.*, 2013), other studies did not find proof for this (Ngomanda *et al.*, 2009). Data on the distribution of forest pioneer species such as oil palm and Okoumé trees as well as comparison of air photos and satellite data show that the expansion of Central African rainforest to the costs of savannahs continues until today (Leal, 2004, Maley, 2002).

This thesis focusses on the central African country of Gabon. It has a surface of 268,000 km^2, is situated in the centre of the Lower Guinean phytogeographical region, and is a Central African biodiversity hotspot (Barthlott *et al.*, 2007, Kier *et al.*, 2005, Linder, 2014) hosting an estimated 7000-7500 species of vascular plants (Küper *et al.*, 2004, Sosef *et al.*, 2017, Sosef *et al.*, 2006). The country has protected over 10% of its surface in 13 National Parks, and has an active conservation policy, supported by international organisations such as the Wildlife Conservation Society, the World Wildlife Fund and coordinated by the Agence Nationale des Parcs Nationaux, a national authority. Some 80% of the surface of Gabon is covered with what is believed to be the most species-rich lowland rainforest in Africa (Sayer *et al.*, 1992). This high level of plant species richness and endemism has been, at least in part, attributed to the history of the region functioning as LGM forest refugia characterized by climatic stability and high levels of habitat heterogeneity (Maley, 1996, Sosef, 1996). The high level of species richness and endemism, the postulated past dynamics in rainforest coverage, and the availability of an excellent dataset on plant species distributions render Gabon an excellent study area to assess the driving forces of species richness and endemism in a tropical rainforest.

Global change

Biodiversity is affected by human-induced changes that act on a global scale, including climate change, loss of habitat, the introduction of invasive species and diseases, as well as acidification of oceans, fresh water bodies and rain. Of these, loss of habitat, climate change, and invasive species form the largest threats to biodiversity and are driving a biodiversity tragedy, particularly in the Tropics (Bradshaw *et al.*, 2009). The effect of climate change is likely to be particularly severe in tropical ecosystems, as these contain by far the largest number of species and individual tropical species have much narrower thermal tolerances compared to temperate species (Deutsch *et al.*, 2008). In addition to the direct effects, the extinction of species itself appears to accelerate changes in ecosystem processes leading to further loss of species and ecosystem functions (Hooper *et al.*, 2012). Moreover, synergies in these extinction drivers further increase the extinction risk (Brook *et al.*, 2008, Gilman *et al.*, 2010). The predicted effects of global change on biodiversity are severe, but estimates vary and depend on the taxonomic group, spatial scale, temporal window, and applied biodiversity loss metrics (Bellard *et al.*, 2012). For example, 81-97% of sub-Saharan species is estimated to loose parts or all of their suitable habitat by 2085 due to climate change (McClean *et al.*, 2005). The effect of habitat loss on species extinction has been estimated e.g. for the Amazon region, ranging from 5-9% of plant species going extinct by 2050 (Feeley & Silman, 2009) to an extinction of 20-33% of all tree species by 2020 (Hubbell *et al.*, 2008). Effects on local scales are much stronger, with up to 76,5% species loss for the worst-affected habitats (Newbold *et al.*, 2015). On a positive side, some studies showed that change of land use offers opportunities for species to expand their ranges when given the possibility to migrate in time (Feeley & Silman, 2010a). Individual species can respond to climate change either by colonizing newly suitable habitats, or trough changes in their phenology, morphology and genetic structure (Bellard *et al.*, 2012, Parmesan, 2006). However, for many species, the speed of these global changes may be too high, driving them to extinction and leading the world towards the 7th mass extinction.

Conservation priorities

Limited availability of resources and the scale at which global change affects biodiversity forces governments and NGOs to set priorities in conservation (Margules & Pressey, 2000, Moilanen & Arponen, 2011). This is particularly important for species in tropical forests, where the risk of extinction escalates (Ricketts *et al.*, 2005, Vamosi & Vamosi, 2008). Both species richness and endemism have been recommended and used as key criteria to set priorities in conservation (Brooks *et al.*, 2006, Huang *et al.*, 2016, Myers *et al.*, 2000). Consequently, the availability of accurate and sufficient information on the spatial distribution of species is crucial for making informed conservation decisions (Boitani *et al.*, 2011, Whittaker *et al.*, 2005). Similar, knowing which species are threatened and where these species occur now as well as in the future is crucial in order to spend conservation resources efficiently and effectively (Brooks *et al.*, 2006). Unfortunately, precisely this knowledge on the distribution, rarity and threat status is lacking for most species (Pimm *et al.*, 2014). This knowledge gap is worrying, because half of the world's plant species may well be qualified as threatened with extinction under the classification of the International Union for Conservation of Nature (IUCN, 2001, Pitman & Jorgensen, 2002), although sufficient data for a formal classification of these species are still lacking.

Data availability

The lack of detailed and often even basic distributional information for species is termed the Wallacean shortfall (Lomolino, 2004) and applies to most species worldwide, but particularly to species in the Tropics (Collen *et al.*, 2008). The Linnaean shortfall (Brown & Lomolino, 1998), the difference between the number of species existing and the number actually described, causes an additional data gap. Fortunately, the availability of species occurrence data is increasing rapidly, notably due to the ongoing digitalisation of natural history collections, inventory data, and other observations (Graham *et al.*, 2004) and online publishing of these in e.g. the Global Biodiversity Information Facility (www.gbif.org) and targeted projects such as the African RAINBIO project (Dauby *et al.*, 2016). These data are typically presence-only data, referring to the observed presence of a species at a particular time and location. Absence data are usually not available. Notwithstanding the increasing availability of these big data and the opportunities this offers to bridge the Wallacean shortfall, many species and habitats remain highly underrepresented (Feeley, 2015, Küper *et al.*, 2006, Sosef *et al.*, 2017, Stropp *et al.*, 2016). These shortfalls, in addition to biases and uncertainties in the data, continue to hamper biodiversity research and conservation efforts (Meyer *et al.*, 2016). The low availability of data on African TRFs is illustrated by Fig. 1 showing the number of vascular plant species known from Gabon plotted against the number of available records from either Gabon alone or from all African countries. Using only records from Gabon, out of the 5,323 species documented from Gabon, 2512 are known from 5 or fewer records and 980 are known from only a single collection. When records from other African countries are included, 2047 species are known from 5 or fewer records and 688 are known from only one collection. This low availability of distributional data is typical for tropical ecosystems.

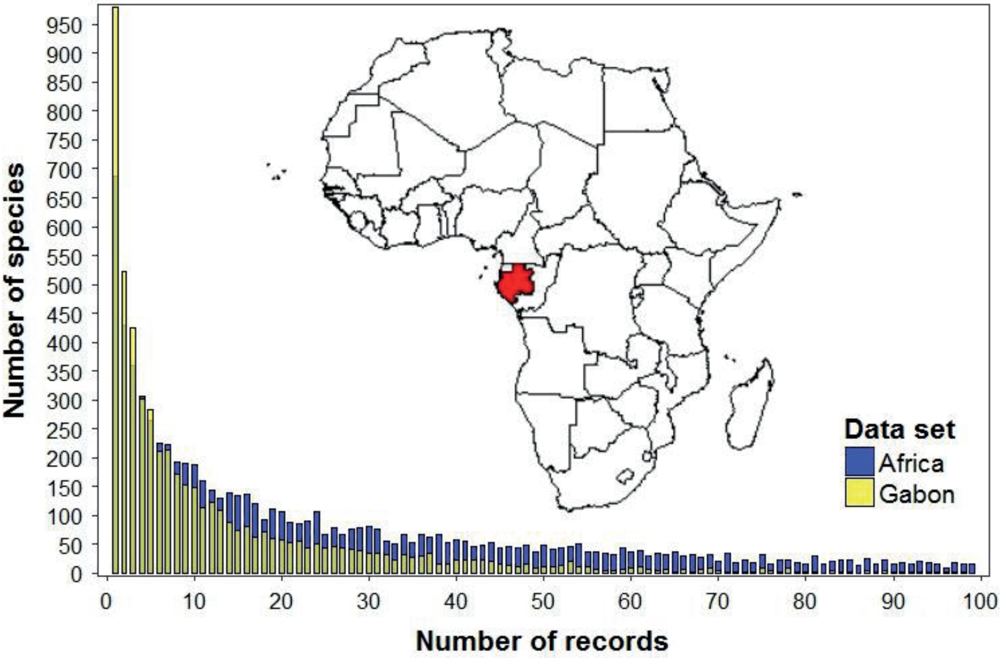


Figure 1: Data availability of Gabonese vascular plant species. *The number of species is plotted against the number of records representing the species when using only records from Gabon (yellow) as well as when including records from other African countries (blue). To increase readability, only the species with 100 or fewer records are shown.*

Species Distribution Models

Species Distribution Models (SDMs) offer opportunities to infer the spatial distribution of species using a limited number of observation data. SDMs are correlative models that link known localities of species to high resolution environmental data, typically climate, soil and altitudinal data (Elith & Leathwick, 2009). The currently available suite of methods forms a pipeline (Fig. 2) through which significant SDMs can be generated rapidly and patterns of species richness and weighted endemism using these SDMs can be inferred for a large number of species including plants, animals, fungi and microbes. At the start of this pipeline, species occurrence data are retrieved, usually from online available resources such as GBIF <www.gbif.org> and Specieslink <http://splink.cria.org.br>. These are then combined with high resolution environmental data on climate, soil and altitude, recently also including remotely sensed climate data (Bookhagen & Strecker, 2008, Duan & Bastiaanssen, 2013, Platts *et al.*, 2015). The pipeline returns spatially projected models visualising the predicted habitat suitability of each species in a specific study area under past, current or future climatic conditions. By superimposing individual SDMs, patterns of species richness and weighted endemism can be inferred.

The development of SDM techniques has resulted in a tremendous increase in knowledge on the distribution of individual species as well as an improvement of the quality of macroecological, biogeographical and conservation research (Franklin, 2009). As such, these techniques help to bridge the Wallacean shortfall described above (Anderson, 2012, Guisan & Thuiller, 2005). Nowadays, SDMs are widely used to study past, present and future distributions of individual species (Franklin, 2009, Wieringa *et al.*, 2013), to assess the risk of invasive species (Jiménez-Valverde *et al.*, 2011), to infer patterns of species richness, identify biodiversity hotspots and assess the factors causing these patterns (Amaral *et al.*, 2017, Raes *et al.*, 2009, Zhang *et al.*, 2016). In addition, the impact of climate change (Schweiger *et al.*, 2012), and the effectiveness of protected areas in the light of conservation planning has been assessed using SDMs (Guisan *et al.*, 2013, Zhang *et al.*, 2012).

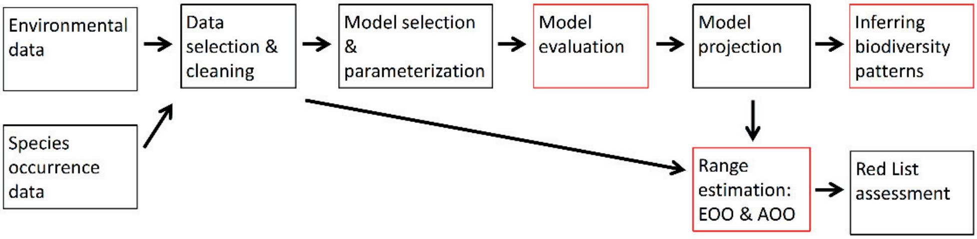

Environmental data
Species occurrence data
Data selection & cleaning
Model selection & parameterization
Model evaluation
Model projection
Inferring biodiversity patterns
Range estimation: EOO & AOO
Red List assessment

***Figure 2: Methodological pipeline to generate SDMs and infer biodiversity patterns** such as of species richness and weighted endemism using SDMs. Parts of the pipeline addressed specifically in this thesis are marked red.*

Limitations of SDMs

Notwithstanding the advantages of SDMs, each step in the pipeline has its own assumptions, limitations and uncertainties, including errors in specimen identification and geographic locality of occurrence records, inaccuracies of environmental data, stochastic effects, as well as the effect of model parameterization, model selection, model evaluation and variable selection methods (Araújo & Guisan, 2006, Araújo & Peterson, 2012, Zimmermann *et al.*, 2010), to name but a few. The exact effect of these matters on SDMs usually remains unknown, particularly when taken in concert, but warrant consideration, before jumping to conclusions based on the outcomes. The effect of some limitations has been addressed by others, e.g. the effect of the extent of study area in relation to the species' range size (Barve *et al.*, 2011, McPherson *et al.*, 2004, VanDerWal *et al.*, 2009), the effect of multicollinearity of environmental variables (Dormann *et al.*, 2013), the effect of using remotely sensed climate data with higher accuracy over data interpolated from scarce weather stations (Deblauwe *et al.*, 2016), the effect of incomplete sampling of the species' niche (Raes, 2012), the effect of using big data on species occurrence through massive digitalisation projects over smaller, expert-verified data sets (Beck *et al.*, 2013), the effect of selection of modelling algorithm (Elith *et al.*, 2006), the use of null models to assess model accuracy (Deblauwe *et al.*, 2016, Raes & ter Steege, 2007), as well as the effect of using different methods to infer species richness patterns from SDMs of individual species (Calabrese *et al.*, 2014). Transferability of the model in place and time, especially to non-analogue environments is crucial, meaning that the model can be projected successfully to combinations of environmental variables not present in the training area (Randin *et al.*, 2006). In addition, model performance can be seriously improved when data on dispersal limitations and biotic interactions are incorporated. Unfortunately, such data are available for only very few species (Anderson, 2017, Franklin, 2010). Another crucial, but often neglected matter is the dependency of model accuracy on the number of records used to train the model (van Proosdij *et al.*, 2016b, Wisz *et al.*, 2008). For rare species, which are most in need of analysis of their distribution and threat using predictive distribution modelling, the availability of species occurrence data is low, and hence their model accuracy (or reliability) is low, a contrast known as the "rare species modelling paradox" (Lomba *et al.*, 2010).

Collecting bias

The vast majority of species occurrence data is not collected in a randomized way and hence contain taxonomic, geographical and temporal biases of different, often unknown sizes (Meyer *et al.*, 2016). Such biases can result in less accurate SDMs as well as incorrect estimations of species richness (Schmidt-Lebuhn *et al.*, 2012). Even smaller, unbiased data sets may generate more accurate models than larger, biased data sets (Beck *et al.*, 2014). If bias is based on the preference of the collector, this collectors' bias of the data is towards the species of interest. Botanical collectors, for example, are known for their effort to never collect the same species twice in the region they visit (ter Steege *et al.*, 2011). In addition, many species may remain undetected in the field due to a variety of reasons including life form (geophytic, nocturnal, migratory), size, biotope, or training-level of the observer (Chen *et al.*, 2013). Geographical and temporal biases originate from the inaccessibility of areas due to logistic or political reasons. Finally, differences in the effort spent to digitize and publish herbarium specimens data results in differences in data accessibility. In general, species occurrence data are biased towards easy observable and collectable species and towards easy accessible areas such as near cities as well as along roads and rivers (Hortal *et al.*, 2007, Reddy & Davalos, 2003). A geographical bias may result in an environmental bias with some environments in the study area being underrepresented by collections (Kadmon *et al.*, 2004, Loiselle *et al.*, 2008), although this is not necessarily always the case (McCarthy *et al.*, 2012, Tessarolo *et al.*, 2014). Many regions of the world are severely undersampled, including several large parts of Africa (Meyer *et al.*, 2015, Meyer *et al.*, 2016, Sosef *et al.*, 2017, Stropp *et al.*, 2016).

A few methods have been developed to mitigate the negative effect of collecting bias on model accuracy. For example, spatial thinning or geographic filtering can be applied by excluding records from areas with high sampling density (Aiello-Lammens *et al.*, 2015). Similarly, environmental filtering, applied in parameter space rather than geographic space, reduces environmental bias in data sets even better (Varela *et al.*, 2014). However, for many species, the number of available occurrence records is simply too low to enable filtering. Alternatively target-group background sampling can be used, which applies the same bias present in the species occurrence records to the background data that are used to train the model with. Background records are then drawn from known collecting localities of a specified target-group, often localities where collections were made, but where the species itself was never collected (Phillips *et al.*, 2009, Syfert *et al.*, 2013). Similarly, bias-corrected null models can be used to evaluate models by using null models that are trained on such target-group background localities instead of random localities (Raes & ter Steege, 2007). Finally, some recent studies on species occupancy models report to account for imperfect detection of the species (Dorazio, 2014, Guillera-Arroita, 2017).

Biotic interactions

The interaction with other species, either competitors, food sources, predators, hosts, pollinators, dispersers, or pests influences the distribution of species (Godsoe *et al.*, 2015, Soberón & Peterson, 2005). These biotic interactions are affected by climate change, increasing their importance for the accurate prediction of species distributions under global climate change (Blois *et al.*, 2013). Therefore, including biotic interactions in some stage of the modelling process is recommended (Anderson, 2017, Kissling *et al.*, 2012). Biotic interactions have been shown to play an important role at high spatial resolutions, but their imprint rapidly vanishes with decreasing spatial resolution and is largely absent at resolutions typically used for species distribution modelling (Thuiller *et al.*, 2015), although others found evidence for substantial impact at regional, continental and even global extents too (Wisz *et al.*, 2013). Recently, new methods have been explored to incorporate species interactions directly in the modelling process (Boulangeat *et al.*, 2012, Giannini *et al.*, 2013) or to model the species interactions themselves (Kissling *et al.*, 2012). However, for most species, interactions between species and their importance in shaping species distributions remain largely unknown (Godsoe *et al.*, 2015, Wisz *et al.*, 2013), leaving biotic interactions a challenging aspect of species distribution modelling.

Model evaluation and the use of the AUC

Testing the accuracy of SDMs or model evaluation is challenging, in particular when SDMs are trained on presence-only data. Commonly, a confusion matrix is used (Table 1), showing the numbers of correctly and incorrectly predicted presences and absences (Fielding & Bell, 1997, Metz, 1978). Two types of errors are included: a commission error (or false positive, FP) when a real absence is predicted to be a presence and an omission error (or false negative, FN) when a real presence is predicted to be an absence. Sensitivity of a model or its power to correctly predict all presences, is computed by dividing the number of true positives (TP) by the total number of real presences (TP + FN). Model specificity is defined as its power to correctly discriminate between presences and absences and is computed by dividing the number of true negatives (TN) by the total number of real absences (TN + FP). Generally, higher sensitivity will come at the price of lower specificity and vice versa. A good model combines high sensitivity with high specificity.

Table 1: Confusion matrix *showing the number of correctly and incorrectly predicted presences and absences, by comparing the predictions with the actual situation.*

		Actual		
		Presence	Absence	Total
Predicted	Presence	*True Positives*	*False Positives (Commission)*	*Predicted Presences*
	Absence	*False Negatives (Omission)*	*True Negatives*	*Predicted absences*
	Total	*Observed Presences*	*Observed Absences*	*All observations*

Model evaluation criteria derived from the confusion matrix include sensitivity, specificity, the True Skill Statistics (TSS), defined as sensitivity + specificity – 1 (Allouche *et al.*, 2006), and Cohen's kappa which corrects for chance and accounts both omission and commission errors in one performance indicator value (Cohen, 1960). Each of these indicators apply a fixed threshold to convert the model's continuous probability of occurrence into presences and absences. The level of this threshold strongly influences the balance between the number of presences and absences and hence between sensitivity and specificity.

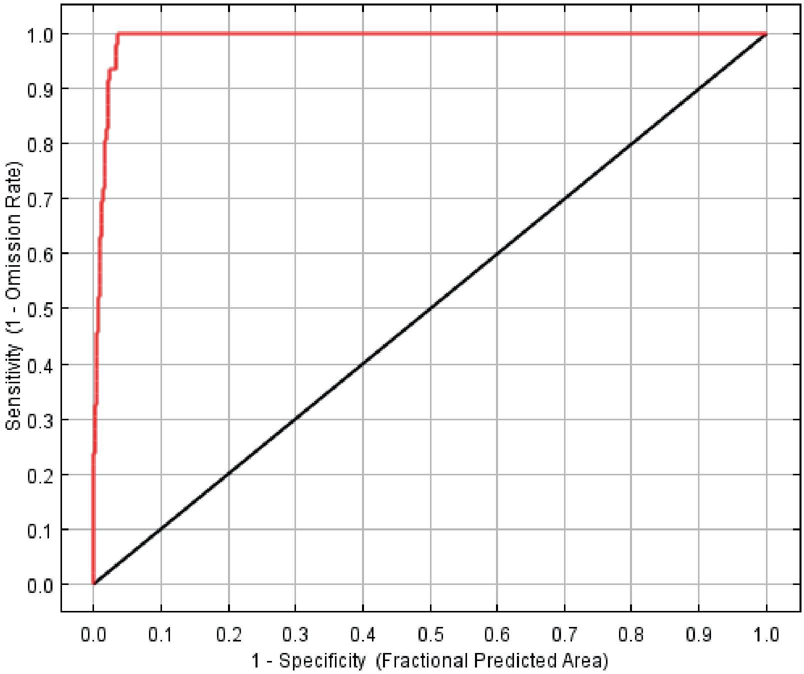


***Figure 3: ROC curve of the SDM** of the endemic legume species* Gabonius ngouniensis*, based on 46 training records showing sensitivity as a function of 1 – specificity, with the model AUC (0.991, red line) and the AUC of random chance (0.500, black line).*

The most widely used indicator of model performance in species distribution modelling is the Area Under the Curve (AUC) of the Receiver Operator Characteristic (ROC) (Metz, 1978). In the ROC curve, sensitivity is plotted as a function of commission error (1 – specificity) for every possible threshold value, which makes this indicator threshold-independent (Fig. 3). Low values on the x-axis indicate low commission error rates and high values on the y-axis indicate low omission error rates.

When the curve approaches the upper left corner of the graph, the level of both omission and commission errors is minimised. The area under this ROC curve (AUC) is a measure of the discriminative power of the model and codes for the probability that a randomly taken presence has a higher predicted value than a randomly selected absence point (Jiménez-Valverde, 2012). An AUC value of 0.5 indicates random model performance, a value of 1.0 perfect model fit. AUC values above 0.7 are generally regarded to represent a fit that is useful, while a value above 0.9 represents high model accuracy (Pearce & Ferrier, 2000, Swets, 1988). In addition to being threshold-independent, the AUC is the only indicator of model performance largely independent of sampling prevalence (McPherson *et al.*, 2004), making it the only useful indicator for SDMs that are trained on small presence-only data sets.

To compute omission and commission rates, independent evaluation data are required. However, these data are typically not available and to obtain them through collecting in the field is very time-consuming and costly. To overcome this problem, the available data can be partitioned into a training and test dataset, although at the cost of reducing the number of records used for model training.

The use of AUC as indicator of model performance for models trained on presence-only data has been criticized (Jiménez-Valverde, 2012, Lobo *et al.*, 2008, Raes & ter Steege, 2007). The problem is that, when using presence-only data, pseudo-absences are sampled from the study area. As the true distribution of the species is unknown, some pseudo-absences will be sampled from localities where the species actually is present. This results in lower values of the AUC. When pseudo-absences are sampled randomly, the maximum possible value of the AUC is $1 - a/2$, where a represents the fraction of localities where the species is present (Phillips *et al.*, 2006). Models for widespread species therefore will have a lower maximum achievable AUC than narrow-ranged species. In addition, the impact of bias in species records is usually ignored and rarely tested (but see Smith, 2013). When sampling of test sites is biased towards areas with higher prediction values, more test presence sites have a higher prediction than random absence sites, resulting in the maximum value of the AUC $> 1 - a/2$. On the other hand, if test sites have lower prediction values, the maximum achievable AUC is lower than $1 - a/2$ (Smith, 2013). As the species' distribution is typically unknown, it's unclear if a higher AUC value is caused by an increased actual model performance or by a bias in sampling. In short, this makes the AUC an inappropriate indicator if used on presence-only data in a direct way. To overcome this problem, a null-model test has been developed to test if the SDM deviates significantly from random chance (Raes & ter Steege, 2007). In this null-model test, an SDM performs significantly better than chance ($p < 0.05$) if its AUC value ranks > 95 when grouped with the 99 AUC values derived from the null models. The bias in species occurrence data discussed above can be mitigated by using a bias-corrected null-model, where null-model records are drawn from a pool of target-group presence localities.

Virtual ecologist approach

To overcome the problem of not knowing the true distribution of species hampering quantitative assessments of different aspects of SDMs, simulated species can be used. The use of such simulations has been advocated as it offers opportunities to systematically assess the effect of individual aspects in a complex world (Saupe *et al.*, 2012), a methodology termed 'virtual ecologist approach (Zurell *et al.*, 2010). Recently, software to generate simulated species distributions have become available (Duan *et al.*, 2015, Leroy *et al.*, 2016), although both not offering the wide variety of species definitions available in the method presented in Chapter 2 (van Proosdij *et al.*, 2016b). By simulating a species, its characteristics are fully controlled and its spatial distribution is known by definition. This enables an assessment of the role and importance of crucial aspects of species distribution modelling including data sampling strategy, model selection, model evaluation as well as biases and specific ecological characteristics of species (Austin *et al.*, 2006, Hirzel *et al.*, 2001, Meynard & Kaplan, 2012, Meynard & Kaplan, 2013).

Outline of this thesis

Back in 2005, 'What determines species diversity?' was discussed in Science as one of the 25 most important, but still unanswered questions in science (Kennedy & Norman, 2005, Pennisi, 2005). This question has been puzzling scientists for centuries. In this thesis, I address this question in the context of Gabon, Central Africa: "What determines botanical diversity in Gabon, Central Africa?". The choice for Gabon is driven by its botanical richness, harbouring one of the most species-rich lowland rainforests in Africa, whereas at the same time, African TRFs are considerably more species-poor than their Neotropical and Asian counterparts. Therefore, understanding patterns of species richness and endemism in Gabon will contribute to a better understanding of the factors that determine species richness. I approach this question by using species distribution models. In particular, I investigate the effects on global climate change on the future of Gabonese plant species richness. In the process of doing so, I encountered two methodological and one biological question that are addressed first.

In Chapter 2 I use simulated species to identify the minimum number of species occurrence records that is required to generate accurate SDMs (van Proosdij *et al.*, 2016b). I show that this lower limit of sample size differs for species of different range sizes. The results show that larger sample sizes are required for more widespread species. The method developed and presented here can be applied to any species group and study area. In addition, the behaviour of the AUC based on presence-only is compared with the behaviour of the AUC based on presence-absence data.

In Chapter 3 I assess the accuracy of 11 methods, including a novel one presented in this chapter, to estimate the range size or prevalence of a species. The species' range size or prevalence, defined as the fraction of raster cells in a study area where the species is (predicted to be) present, is an important characteristic of a species and one of the key variables used in IUCN Red List assessments. In addition, as is shown in Chapter 2, species' prevalence influences the minimum required sample size for species distribution modelling. In Chapter 3, I apply the method to simulate species developed in Chapter 2.

For Chapter 4 I created a large dataset with all available species occurrence data of Gabonese vascular plant species, supplemented with available data related to these species from other African countries. I use stacked SDMs of large numbers of species to compute for the first time patterns of plant species richness and of weighted endemism in Gabon. The methods developed and presented in Chapters 2 and 3 are used to generate the best possible SDMs, meaning to select those SDMs that meet the desired accuracy for species of each particular prevalence class. I then identify biodiversity hotspots and areas of exceptional levels of endemism. Specifically, I quantify the contribution of widespread and of narrow-ranged species to patterns of species richness and weighted endemism and show that these groups of species contribute differently to these patterns. The question which subsets of species, narrow-ranged, widespread or randomly selected species, best represents patterns of species richness and of weighted endemism is addressed here (van Proosdij *et al.*, 2016a).

In Chapter 5 I again use SDMs based on real data, but now project the generated models to future climate scenarios for 2085. I infer future patterns of predicted species richness and describe the expected changes of the species richness patterns in Gabon in terms of species gain, loss and turnover. Second, I quantify the additional effect of dispersal limitations. Finally, I identify for each climate anomaly its unique explanatory power to species gain, loss and turnover.

In the synthesis in Chapter 6 I bring together the results of the previous chapters and place these in a wider scientific and societal perspective. Future lines of research are discussed, specifically those that build on the here applied virtual ecologist approach.

Chapter 2

Minimum required number of specimen records to develop accurate species distribution models

André S.J. van Proosdij[1,2], Marc S.M. Sosef[1,3], Jan J. Wieringa[1,2], Niels Raes[2]

Published in Ecography, 2016, 39 (6): 542-552

DOI: 10.1111/ecog.01509

1 Biosystematics Group, Wageningen University, Droevendaalsesteeg 1, 6708 PB Wageningen, the Netherlands

2 Naturalis Biodiversity Center (Botany section), Darwinweg 2, 2333 CR Leiden, the Netherlands

3 Botanic Garden Meise, Nieuwelaan 38, 1860 Meise, Belgium

Abstract

Species Distribution Models (SDMs) are widely used to predict the occurrence of species. Because SDMs generally use presence-only data, validation of the predicted distribution and assessing model accuracy is challenging. Model performance depends on both sample size and species' prevalence, being the fraction of the study area occupied by the species. Here, we present a novel method using simulated species to identify the minimum number of records required to generate accurate SDMs for taxa of different pre-defined prevalence classes. We quantified model performance as a function of sample size and prevalence and found model performance to increase with increasing sample size under constant prevalence, and to decrease with increasing prevalence under constant sample size. The Area Under the Curve (AUC) is commonly used as a measure of model performance. However, when applied to presence-only data it is prevalence-dependent and hence not an accurate performance index. Testing the AUC of an SDM for significant deviation from random performance provides a good alternative. We assessed the minimum number of records required to obtain good model performance for species of different prevalence classes in a virtual study area and in a real African study area. The lower limit depends on the species' prevalence with absolute minimum sample sizes as low as 3 for narrow-ranged and 13 for widespread species for our virtual study area which represents an ideal, balanced, orthogonal world. The lower limit of 3, however, is flawed by statistical artefacts related to modelling species with a prevalence below 0.1. In our African study area lower limits are higher, ranging from 14 for narrow-ranged to 25 for widespread species. We advocate identifying the minimum sample size for any species distribution modelling by applying the novel method presented here, which is applicable to any taxonomic clade or group, study area or climate scenario.

Key words

Simulated species, prevalence, AUC, minimum number of records, model performance, null model, Species Distribution Model

Introduction

Despite globally increasing investments in biodiversity research, our knowledge of the biodiversity on our planet is still limited, especially for data-sparse areas like the tropics (Costello *et al.*, 2013, Whittaker *et al.*, 2005). We can only guess at the total number of extant species, let alone that we know their spatial distribution (Mora *et al.*, 2011). Rare species, those with either a small range or a low abundance (Rabinowitz, 1981), represent the vast majority of species (Longino *et al.*, 2002, ter Steege *et al.*, 2013) and are consequently represented by few samples in natural history collections, our primary source of distributional data. Typically, these collections have a long-tailed relative abundance distribution as illustrated by ter Steege et al. (2011) for the Guianas. Species distribution models (SDMs) have been developed to overcome this lack of information (Araújo & Peterson, 2012, Guisan & Zimmermann, 2000) as they are able to predict the probability of occurrence of species for non-sampled areas too. SDMs relate recorded species presences to abiotic factors that are thought to determine the species' distribution (Araújo & Peterson, 2012). SDMs are thereby built on the assumption that the sample data cover the species' full ecological range (Raes, 2012, Sánchez-Fernández *et al.*, 2011).

The effect of sample size on model accuracy is an aspect that is often neglected (Mateo *et al.*, 2010, Wisz *et al.*, 2008). However, ignoring this effect results in increased levels of error in distribution models for species represented by (too) few records, which are mostly rare species. In addition to sample size, the species' prevalence has a strong impact on model performance; species' prevalence is defined as the fraction of the study area occupied by a species (McPherson *et al.*, 2004). Model performance for ecologically and geographically narrow-ranged species is significantly better compared to widespread species found in a wider range of habitats (Hernandez *et al.*, 2006, Lobo & Tognelli, 2011, Mateo *et al.*, 2010, Tessarolo *et al.*, 2014). Identifying the lower limit of the number of records that is required to develop accurate SDMs in relation to the species' prevalence is therefore highly topical. Despite the large number of studies using SDMs and the results from recent studies on the negative effects of small sample sizes on SDM performance (Loiselle *et al.*, 2008, Mateo *et al.*, 2010, Tessarolo *et al.*, 2014, Wisz *et al.*, 2008), few studies actually address the minimum number of unique records required to generate an accurate SDM. Model performance is known to rapidly decrease for sample sizes smaller than 20 (Stockwell & Peterson, 2002) or 15 (Papeş & Gaubert, 2007), and is dramatically poor for samples sizes smaller than 5 records (Pearson *et al.*, 2007). Contrary to this, high model accuracy was observed using several modelling techniques for models based on samples as small as 5, 10 and 25 compared to models based on 100 samples (Hernandez *et al.*, 2006). Given that the true distribution of a species is unknown, model evaluation in these studies is based on in-sample test data and not on independent test data.

Standard evaluation parameters are based on a confusion matrix measuring both sensitivity (correctly predicted presences) and specificity (correctly predicted absences) (Fielding & Bell, 1997). Consequently, these metrics require absence data, which are typically not available when presence-only data from herbaria or zoological collections are used. To remedy the lack of absence data, random background data or pseudo-absences are used instead (Phillips *et al.*, 2009). Commonly, the number of background points is high compared to the number of presences, resulting in a low sampling prevalence; where sampling prevalence is defined as the number of presences relative to the entire sample. From the suite of model accuracy measures, the Area Under the Curve (AUC) of the Receiver Operator Characteristic (ROC) is the only one shown to be largely independent of sampling prevalence when applied to presence-absence data (McPherson *et al.*, 2004). This renders the AUC the only useful indicator of model accuracy applicable to SDMs based on low-prevalence data, typical for presence-only data samples (Fielding & Bell, 1997, Metz, 1978). The AUC value translates to the chance that a randomly chosen presence has a higher predicted probability of occurrence than a randomly chosen absence. However, when applied to presence-only data, the use of AUC values is strongly criticized for the above mentioned imbalance between presences and absences, where including more absences that are environmentally more distant from the species' presences increases the fraction of correctly predicted absences (specificity), resulting in higher AUC values (Jiménez-Valverde, 2012, Lobo *et al.*, 2008). In addition, when used on unbiased presence-only data, the maximum achievable value of the AUC is not 1, but 1 - *a*/2, where *a* represents the fraction of the area covered by the species' true distribution, which is typically not known (Jiménez-Valverde, 2012, Phillips *et al.*, 2006, Raes & ter Steege, 2007, Smith, 2013). Hence, when applied to presence-only data, the maximum AUC value *is* species' prevalence sensitive after all, and the commonly applied AUC value of 0.7 indicative for an SDM with acceptable accuracy is flawed (Raes & ter Steege, 2007). To overcome this problem, Raes and ter Steege (2007) developed a null-model test to assess whether the AUC value of an SDM deviates significantly from random expectation. However, a null-model test neither assesses how accurate the species' real distribution is modelled, nor how many records are required to obtain high model accuracy. Here, we introduce the use of simulated species with defined occurrence probability to rigorously assess how many records are required to develop accurate SDMs.

In a virtual environment using simulated species, the species' response to environmental variables is fully controlled and thereby its reciprocal spatial distribution is defined and known (Austin *et al.*, 2006, Duan *et al.*, 2015, Hirzel *et al.*, 2001, Zurell *et al.*, 2010). The use of simulated species has been advocated to systematically evaluate how specific aspects of data, sampling strategy, model building and model evaluation affect SDMs (Miller, 2014, Saupe *et al.*, 2012). Studies using simulated species offer unique opportunities to assess SDM accuracy for different sample sizes and species' prevalence classes (Jiménez-Valverde *et al.*, 2009, Meynard & Quinn, 2007). AUC values can be calculated on defined presence and absence data derived from a probability of occurrence distribution and compared with AUC values of SDMs based on presence-only and background data. In addition, the defined probability of occurrence distribution can be compared with the predicted probability of occurrence distribution using Schoener's *D* and Hellinger distance *I* metrics (Schoener, 1970, Warren *et al.*, 2008), which are widely used to measure niche overlap (Rödder & Engler, 2011). Once tested in a fully controlled virtual environment, the same method can be applied to a real environment. As a pilot we selected tropical Africa as the real environment, focusing on the country of Gabon, as we prepare a botanical diversity assessment using SDMs for this country.

Specifically, we assess the effects of sample size and species' prevalence on SDM accuracy using simulated species in a virtual as well as an African study area. We present a novel method to rigorously identify the minimum number of records relative to the species' prevalence that is required to generate SDMs with high model accuracy.

Material & methods

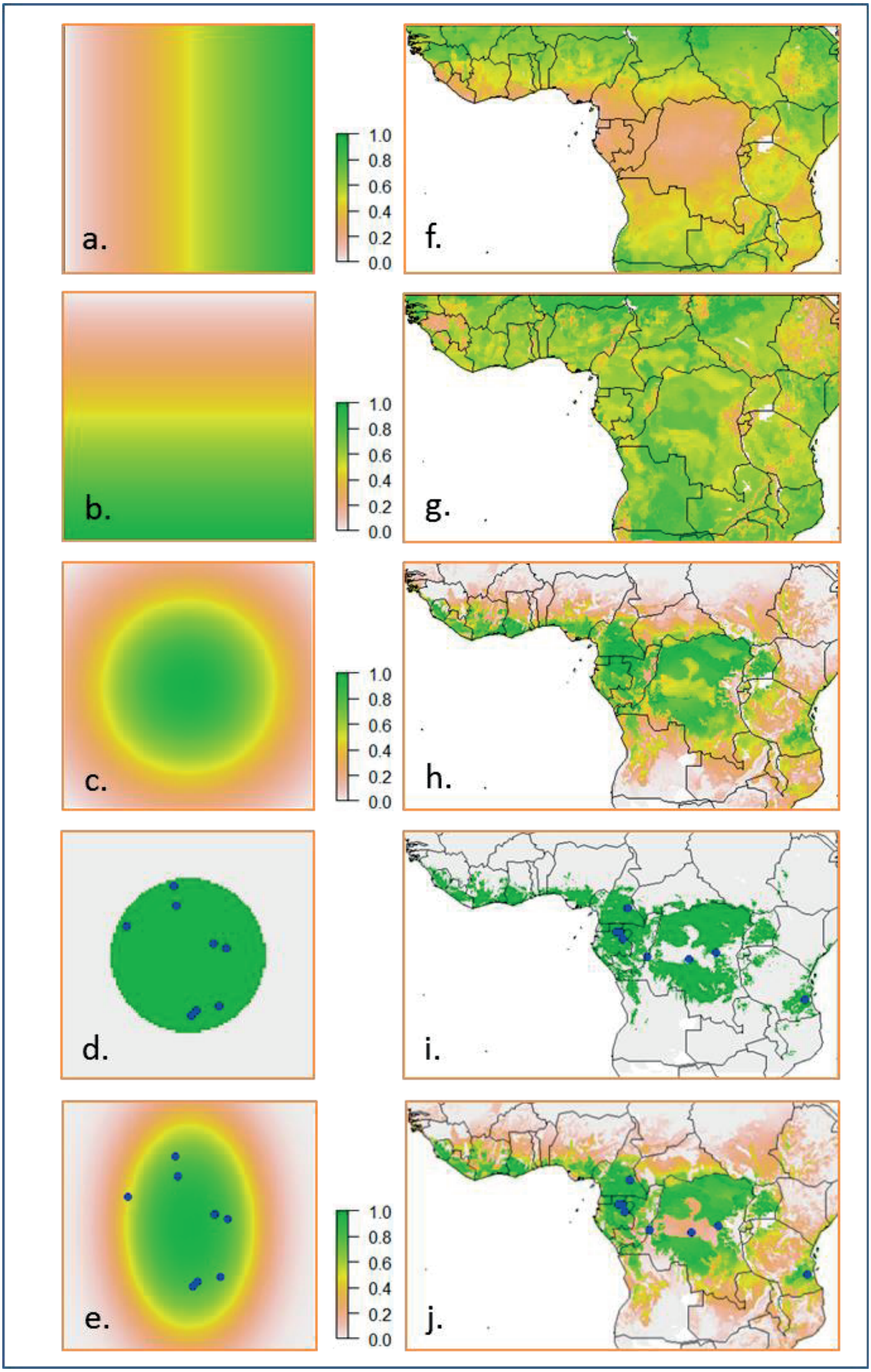
We used the following procedures to define simulated species for different prevalence classes in a virtual as well as an African study area. To increase readability, 'simulated species' are referred to as 'species', unless stated otherwise. All analyses were performed in R (R Core Team, 2016) using functions described below and provided in six separate R scripts (Supplementary material Appendices 4-9) A brief manual explaining the application of the method presented here is provided (Supplementary material Appendix 3).

Virtual study area and simulated species

The virtual study area is defined as a square of 100 by 100 raster cells in which two orthogonal gradients of equal length shaped the ecological landscape: the first linearly increasing from west to east, the second linearly increasing from north to south with the mean value for both located in the center of the study area (0, 0) (Fig. 1a & b). We defined species' prevalence as the fraction of raster cells where the species is present and used six prevalence classes: 0.05, 0.1, 0.2, 0.3, 0.4, and 0.5.

Recent studies on trees (Boucher-Lalonde *et al.*, 2012) and on birds and mammals (Boucher-Lalonde et al. 2014) using temperature and precipitation, showed that in general a species' niche can be described as a Gaussian function to these environmental variables. Based on these findings, we used a bivariate normal function as did others for the same reasons (Broennimann *et al.*, 2012, Duan *et al.*, 2015, Varela *et al.*, 2014). We defined our simulated species by computing their habitat suitability or probability of occurrence as a bivariate normal response to the two orthogonal gradients using the *dmvnorm* function of the R-library *mvtnorm* (Gentz *et al.*, 2014). We assumed that the virtual species' distribution is shaped by these two orthogonal factors only and that niche filling is complete. We defined the different prevalence classes by increasing the standard deviation (*SD*) of the bivariate normal response (Fig. 1c). This procedure resulted in a defined habitat suitability score for each raster cell for each species. In our virtual study area, the optimum of the ecological niche of each species was set at the center (0, 0). We defined a species to be present in raster cells whose environmental bivariate variables are within the central region of the bivariate normal density that has probability 68%. Here, this region is represented by a circle as the two axes represent fully orthogonal, normalized variables with the same variance. Hence, the 68% circle cuts the axes at the points (*optimum* – 1 *SD*, 0), (0, *optimum* + 1 *SD*), (*optimum* + 1 *SD*, 0), and (0, *optimum* – 1 *SD*) (Fig. 1d). For each prevalence class, a small initial *SD* was iteratively increased until the desired prevalence value was approximated by less than 1% difference (for details see Supplementary material Appendix 5).

Figure 1 (next page). Methodological steps illustrated by examples of simulated species *with prevalence 0.2 in the virtual (a-e) and in the real African study area (f-j): 2 orthogonal variables shaping the study area (a & b); defined habitat suitability (c); defined presence (green areas) and absence (white areas) of simulated species and sampled locations (blue dots, sample size 8) (d); predicted habitat suitability and sampled locations (e); 2 orthogonal variables (PCA1 & PCA2) shaping the study area (f & g); defined habitat suitability (h); defined presence (green areas) and absence (white areas) of simulated species and sampled locations (blue dots, sample size 8) (i); predicted habitat suitability and sampled locations (j).*


a.
b.
c.
d.
e.
f.
g.
h.
i.
j.
1.0
0.8
0.6
0.4
0.2
0.0

African study area and simulated species

Our real world study area encompassed most of tropical Africa ranging from 15°N to 19°S and from 18°W to 43°E. The African study area covered 179,994 raster cells with environmental data at 5 arc-minutes spatial resolution, excluding oceans and other large water bodies. Similar to our virtual study area following Broennimannn & al. (2012), we used two orthogonal gradients, that were constructed by means of a principal components analysis (PCA) on fifteen selected environmental variables. These included bioclimatic variables (worldclim.org) (Hijmans *et al.*, 2005), soil variables (Harmonized World Soil Database) (FAO/IIASA/ISRIC/ISSCAS/JRC, 2012), and 90 m resolution elevation data (srtm.csi.cgiar.org) (Supplementary material Appendix 1). From 39 original variables we selected fifteen variables that had a Spearman's |*rho*| < 0.7 (Dormann *et al.*, 2013) (for details see Supplementary material Appendix 4). We used the first two standardized PCA axes that together explained 43% of the variance in multivariate environmental parameter space (Figs. 1f & g). In our African study area, the niche optimum was different for each simulated species, reflecting that each species has a unique ecological niche (Aguirre-Gutiérrez *et al.*, 2014). For each species, the species' optimum was defined by randomly selecting one raster cell from the area delineating Gabon extended by a 5 degree buffer zone. The values of the two PCA-based predictors at this randomly selected location were used as the means of the species' bivariate normal response curve, thus defining the species' optimum (Fig. 1h, for details see Supplementary material Appendices 7). Again, we defined a species to be present in raster cells whose environmental bivariate variables are within the central region of the bivariate normal density that has probability 68% (Fig. 1i), following the same procedure as in the virtual study area. Species' prevalence classes in the African study area were the same as in the virtual study area.

Sampling and replications

For both study areas and for each prevalence class we defined twenty-four sample sizes: 3-20, 25, 30, 35, 40, 45 and 50. For each study area, species and sample size, presences were drawn from the defined presence cells. Sampling probability was equal to the defined habitat suitability score, reflecting higher abundance and therefore higher detectability of species in areas with optimal environmental values (Lomolino *et al.*, 2010). In our virtual study area, where the optimum of every species was (0, 0), we created six species (one for each of 6 prevalence classes). These species were sampled, with each sample size replicated 100 times (6 prevalence classes, 24 sample sizes, 100 replications each), summing to a total of 14,400 species samples from the virtual study area (Fig. 1d). For the African study area, where each species has a unique optimum, for each of the prevalence classes the species definitions were replicated 100 times, resulting in 600 species (6 prevalence classes, 100 replications each). Subsequently, we sampled each species for the 24 different sample sizes (Fig. 1i), also resulting in 14,400 species samples from the African study area.

Species Distribution Modelling

All SDMs were developed with MaxEnt (Phillips *et al.*, 2006). MaxEnt estimates the species potential geographic distribution by finding the distribution of maximum entropy (closest to uniform) subject to the constraint that the expected value of each feature under this estimated distribution matches its empirical average. MaxEnt was developed to use presence-only data and has shown to outperform other algorithms, including when applied to small data sets (Aguirre-Gutiérrez *et al.*, 2013, Elith *et al.*, 2006, Hernandez *et al.*, 2006). Default MaxEnt settings were adjusted to include linear and quadratic features for all sample sizes, while hinge, product and threshold features were excluded to prevent over-parameterization of the models (Merow *et al.*, 2013). Restricting MaxEnt to only use linear features for sample sizes smaller than 10, disables MaxEnt to fit a model on data that demonstrate other responses such as the bivariate normal response of our simulated species. This illustrates the need to adjust default MaxEnt settings based on biologically motivated modelling decisions (Merow *et al.*, 2013). All above mentioned samples were subsequently modelled resulting in 28,800 SDMs (Figs. 1e & j, for details see Supplementary material Appendices 6 & 7).

Testing model accuracy

For each SDM, we calculated the real AUC value (real AUC) by cross-validating the predicted MaxEnt habitat suitability scores with our defined presences and absences using the *evaluate* function of the R-library *dismo* (Hijmans *et al.*, 2013). Due to computational limitations of R, the real AUC for the African study area was calculated using a 10% random subsample of the presences and absences. Second, for each SDM we obtained the internal AUC value calculated by MaxEnt (MaxEnt AUC), which is based on the predicted habitat suitability scores of the sampled presences and background sites. Third, for each sample size, null models were generated by randomly selecting the same number of background sites as sample size records from the entire study area, replicated 99 times. These 99 sets of random points were treated as presences and modelled similarly as the species, resulting in 99 AUC values of MaxEnt models based on randomly drawn points for each sample size. The species' SDM is regarded significantly better than random expectation if its AUC value exceeds rank number 95, when ranked with the 99 null model AUC values, corresponding to a one-sided significance level of 0.05 (Raes & ter Steege, 2007). Both the real AUC and MaxEnt AUC values of each SDM were tested against their corresponding 95th AUC value of the null-distribution (Supplementary material Appendix 6 & 7). Fourth, we calculated the Spearman rank correlation rho values between the defined and predicted habitat suitability based on a cell-by-cell comparison. Finally, for reasons of comparison, we included an analysis of Schoener's *D* and Hellinger distance I (Schoener, 1970, Warren *et al.*, 2008). Here, high overlap indicates that SDMs produce an accurate prediction of our defined species distribution. Both *D* and I were calculated with the *niche.overlap* function of the R-library *phyloclim* (Heibl & Calenge, 2013) applied to the defined and predicted habitat suitability for each sample.

Identifying the required minimum number of records

We used the upper 95% range values from the 100 replications for each combination of prevalence class and sample size of the real AUC, MaxEnt AUC, real AUC rank, MaxEnt AUC rank, Spearman rank correlation, Schoener's *D* and Hellinger distance I values. This effectively excludes the 5% worst performing SDMs. To mask out stochastic effects, the lower and upper limits of this upper 95% range values were smoothed by applying the *loess* function of the R-library *stats* (R Core Team, 2016) with default settings.

To identify for each prevalence class the minimum number of records that is required to accurately model a species' distribution, we evaluated model performance using three decision rules applied to the smoothed lower range limit values for both study areas separately. First, for each prevalence class, we identified the sample size for which the lower range limit of the SDM's real AUC values exceeds 0.9. An AUC value of 0.9 is commonly used as indicative for a 'very good model' performance (Manel *et al.*, 2001, Pearce & Ferrier, 2000), although the original author did not explicitly state so (Swets, 1988). Second, for each prevalence class, we identified the sample size for which the lower range limit of the SDM's real AUC rank values exceeds 95, corresponding to a performance significantly better than random expectation based on a significance level of $p < 0.05$. Finally, for each prevalence class, we identified the sample size for which the lower range limit of the SDM's Spearman rank correlation values exceeds 0.9, indicating strong correlation between defined and modelled species distributions. The behavior of the MaxEnt AUC, Schoener's *D*, and Hellinger distance I values as a function of sample size and species' prevalence are discussed in the context of identifying the required minimum sample sizes.

Results

For both the virtual and the real African study area and for each prevalence class, model performance defined as the lower range limit of the upper 95% range values of the real AUC, MaxEnt AUC, real AUC rank, MaxEnt AUC rank, Spearman rank correlation, Schoener's *D*, and Hellinger distance *I* increased with increasing sample sizes (Figs. 2-5). As expected, the mean and maximum MaxEnt AUC values of our simulated species decreased with increasing sample size, which is in line with the observations of others using real species (Merckx *et al.*, 2011, Raes & ter Steege, 2007).

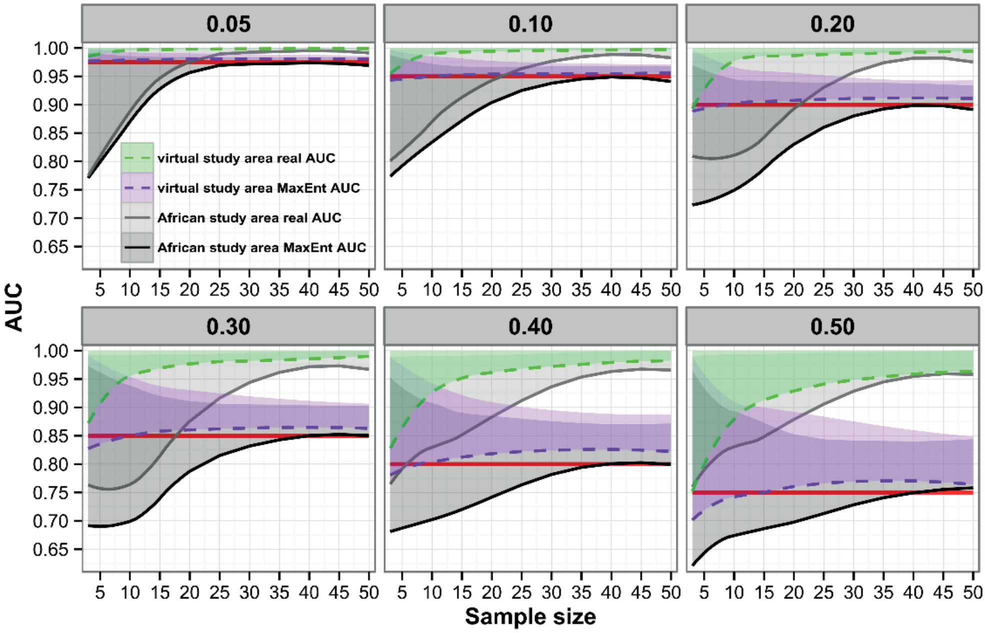


Figure 2. Model performance as a function of species' prevalence class and sample size in a virtual and in an African study area based on AUC values *with separate panels for each prevalence class. Ribbons show the upper one-sided 95% range of the observed values, excluding the 5% worst performing models; darker lines show the lower range limit; dashed lines for the virtual study area; solid lines for the African study area. Red, horizontal lines show the AUC values of 1 – a/2 (where a is the species' prevalence).*

Our results show a strong effect of species' prevalence on model performance: SDMs for species with a small prevalence perform better than SDMs for species with a large prevalence when using the same number of records to train the model (Figs. 2-4), which is in line with results reported by others (Lobo & Tognelli, 2011, Manel *et al.*, 2001, Mateo *et al.*, 2010, McPherson & Jetz, 2007). In contrast, values for Schoener's *D* and Hellinger distance *I* increase with increasing prevalence using the same number of records (Fig. 5).

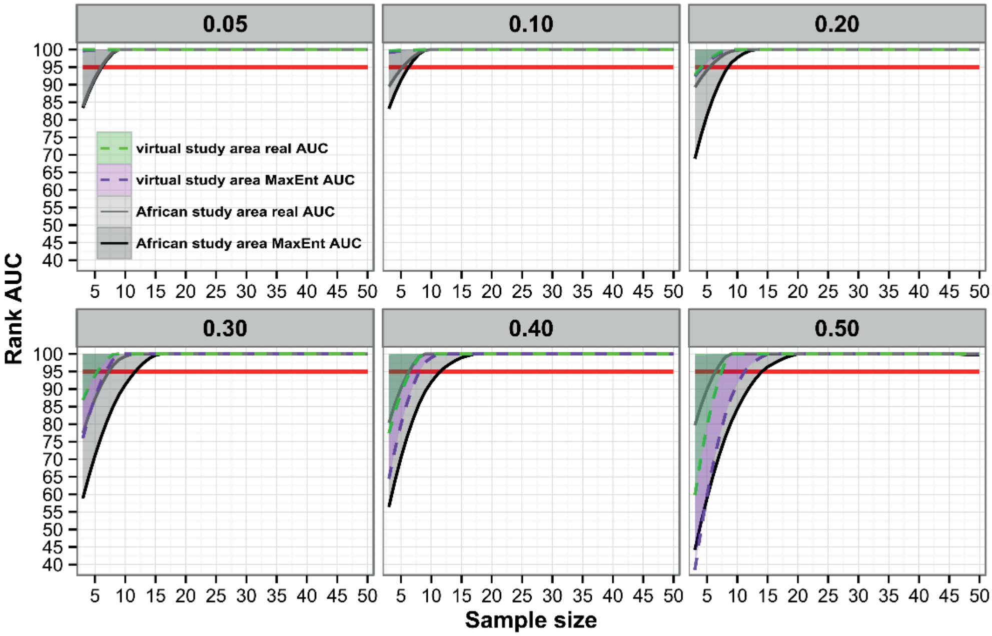


Figure 3. Model performance as a function of species' prevalence class and sample size in a virtual and in an African study area based on rank numbers of AUC values *with separate panels for each prevalence class. Ribbons show the upper one-sided 95% range of the observed values, excluding the 5% worst performing models; darker lines show the lower range limit; dashed lines for the virtual study area; solid lines for the African study area; red, horizontal lines show the critical AUC rank value of 95.*

Figure 2 shows that the MaxEnt AUC values approach an asymptote with increasing sample sizes for each prevalence class. The value to which this asymptote converges strongly decreases with prevalence (Fig. 2). This difference in maximum possible MaxEnt AUC value underlines the importance of being cautious when using the AUC value based on presence-only data (Jiménez-Valverde, 2012, Lobo *et al.*, 2008, Phillips *et al.*, 2006, Raes & ter Steege, 2007). In our results the MaxEnt AUC values slightly exceed the expected value of 1 - *a*/2, where *a* represents the species' prevalence (Fig. 2: red horizontal line), which is caused by our sampling strategy favoring locations with a higher given habitat suitability (Jiménez-Valverde, 2012, Smith, 2013).

The large spread in AUC values at small sample sizes (Figs. 2 & 3) illustrates the low accuracy of SDMs based on small sample sizes. The spread in AUC values, rank AUC values and Spearman rank correlation values decrease with decreasing prevalence and with increasing sample sizes (Figs. 2-4). The spread in observed values was largest at small sample sizes and for widespread species, which illustrates the low accuracy of SDMs based on sample sizes smaller than the required minimum number of records to obtain an accurate SDM.

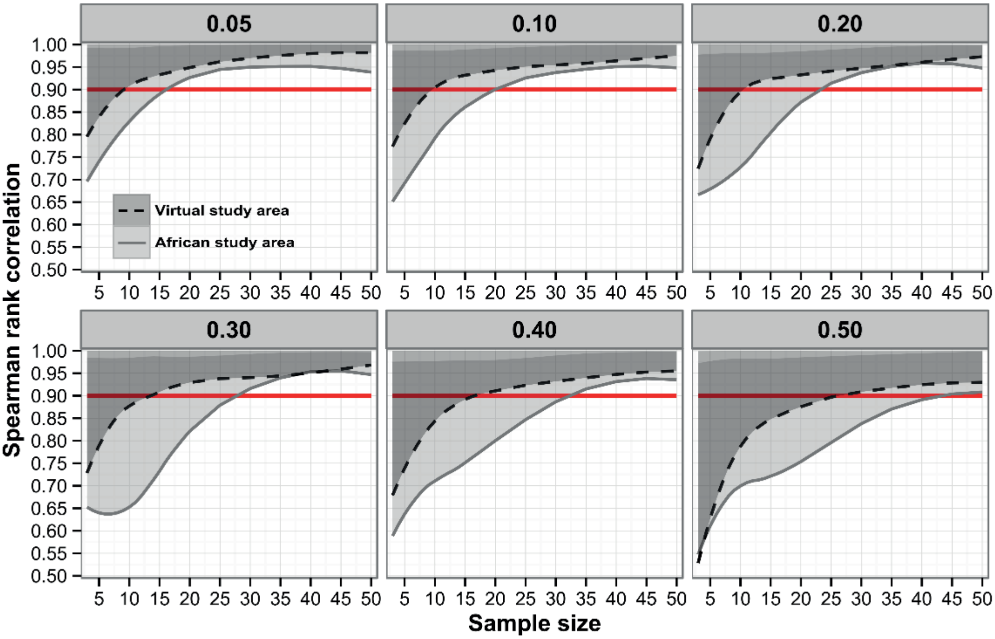


Figure 4. Model performance as a function of species' prevalence class and sample size in a virtual and in an African study area based on Spearman rank correlation values *between defined and predicted habitat suitability with separate panels for each prevalence class. Ribbons show the upper one-sided 95% range of the observed values, excluding the 5% worst performing models; darker lines show the lower range limit; dashed lines for the virtual study area; solid lines for the African study area. Red, horizontal lines show the Spearman rank correlation value of 0.9.*

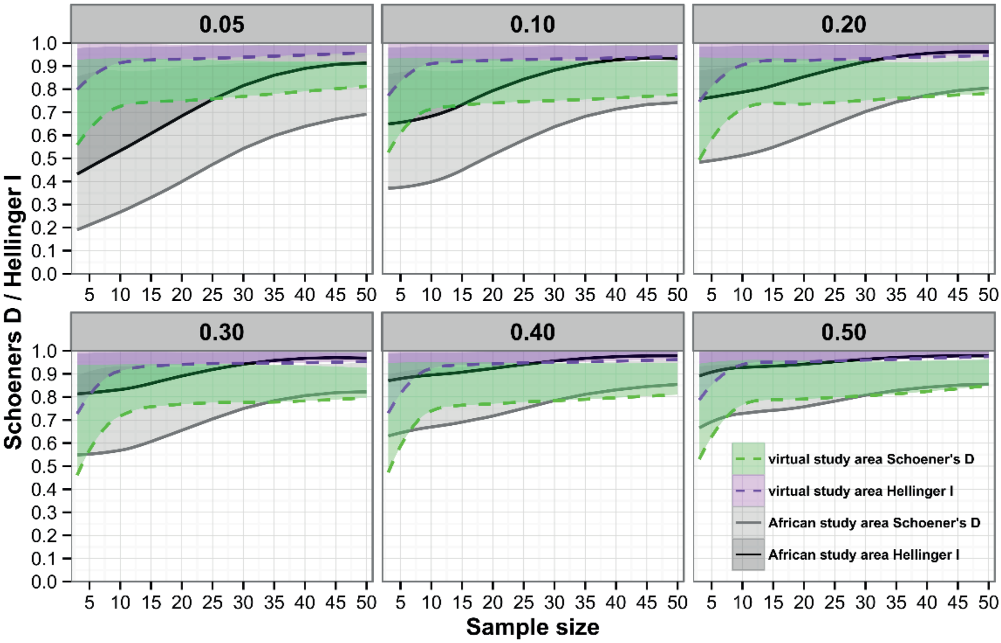


Figure 5. Model performance as a function of species' prevalence class and sample size in a virtual and in an African study area based on Schoener's* D *and Hellinger distance* I *values *with separate panels for each prevalence class. Ribbons show the upper one-sided 95% range of the observed values, excluding the 5% worst performing models; darker lines show the lower range limit; dashed lines for the virtual study area; solid lines for the African study area.*

The required minimum number of records or lower limits of sample size for each prevalence class were identified for both the virtual and the African study area based on our three pre-defined decision rules: the lower limit of the upper 95% range values of 1) real AUC > 0.9; 2) real AUC ranks > 95; 3) Spearman rank correlation > 0.9 (Table 1). Note that although we do show results for prevalence class 0.05 (Table 1, Figs 2-5), we discuss minimum sample sizes for prevalence classes 0.1-0.5 only, as prevalences below 0.1 should be avoided (see Discussion). For the most ideal situation, that of our virtual study area with balanced, orthogonal gradients, we observed that for criterion 1 (real AUC > 0.9, Fig. 2), the minimum sample size ranged from 3 for narrow-ranged species (prevalence class 0.1) to 13 for widespread species (prevalence class 0.5). However, when using criterion 2 (real AUC rank > 95, Fig. 3), the required minimum numbers of records were substantially lower for species of larger prevalence classes: 3 for species in prevalence class 0.1 to 8 for species in prevalence class 0.5. In contrast, based on criterion 3 (Spearman > 0.9, Fig. 4), the required minimum numbers of records were considerably higher and ranged from 10 for species of prevalence class 0.1 to 30 for species in prevalence class 0.5. For the African study area, the minimum sample size based on criterion 1 (real AUC > 0.9) ranged from 14 for species with a prevalence of 0.1 to 25 for species with a prevalence of 0.5. When using criterion 2 (real AUC rank > 95), minimum sample sizes ranged from 6 for species in prevalence class 0.1 to 7 for species in prevalence class 0.5 (the observed sample size of 8 for prevalence class 0.3 might be caused by a stochastic effect). Here again, based on criterion 3 (Spearman > 0.9), minimum sample sizes were considerably higher and ranged from 20 for species with prevalence 0.1 to 45 for species in prevalence class 0.5.

Table 1. The minimum number of records required for building accurate Species Distribution Models *for a virtual study area (a) and an African study area (b) based on critical minimal values of model performance using the following indicators of model performance: real AUC, real AUC rank, MaxEnt AUC rank, and Spearman rank correlation. Minimum sample sizes are based on the lower limit of the upper one-sided 95% range of the model performance values.*

a: virtual study area				
Species prevalence	**Real AUC > 0.90**	**Real AUC rank > 95**	**MaxEnt AUC rank > 95**	**Spearman *rho* > 0.90**
0.05	3	3	3	10
0.10	3	3	3	10
0.20	4	5	5	11
0.30	5	6	7	14
0.40	8	7	9	17
0.50	13	8	12	30
b: African study area				
Species prevalence	**Real AUC > 0.90**	**Real AUC rank > 95**	**MaxEnt AUC rank > 95**	**Spearman *rho* > 0.90**
0.05	11	6	7	17
0.10	14	6	7	20
0.20	25	6	9	25
0.30	25	8	12	30
0.40	25	7	12	35
0.50	25	7	15	45

Discussion

Novel method to identify the required minimum number of presence records

The novel method presented here enables users to rigorously identify the required minimum number of records to generate accurate SDMs for species of different species' prevalence classes under ideal conditions. It does so by using simulated species for which presence and absence is defined and quantifies the effect of sample size and species' prevalence on model performance. As such, this study supplies handles for anyone using SDMs to assess whether their data allows generating accurate SDMs. Our results (Table 1) corroborate two main aspects addressed by others before. First, model performance strongly depends on sample size and a small increase in the smallest sample sizes results in a large increase of model performance (Loiselle *et al.*, 2008, Papeş & Gaubert, 2007, Pearson *et al.*, 2007, Stockwell & Peterson, 2002, Wisz *et al.*, 2008). Second, model performance decreases with increasing species' prevalence when using the same number of records to train the model (Hernandez *et al.*, 2006, Lobo & Tognelli, 2011, Mateo *et al.*, 2010, Stockwell & Peterson, 2002, Tessarolo *et al.*, 2014).

Minimum sample sizes required for model calibration

The minimum number of records required to generate accurate SDMs differs between our virtual and real African study areas (Table 1) and can be different for other study areas. Our virtual study area represents an ideal situation with balanced, orthogonal gradients, and the required minimum sample size for each species' prevalence class should thus be regarded as a theoretical absolute lower limit. In our African study area the lower limits are considerably higher (Figs. 2-4; Table 1), as a result of the non-uniform frequency distribution of environments in this study area. For both study areas, the minimum required sample size is higher for widespread than for narrow-ranged species, as in general the ecological niche of the latter is comparatively better covered by the samples. Therefore, studies applying a generalized a priori defined minimum number of records (Algar *et al.*, 2009, Raes *et al.*, 2009, Raes *et al.*, 2013, Schmidt *et al.*, 2005) will lead to erroneous exclusion of models for narrow-ranged species when setting the limit too high and/or erroneous acceptance of those for widespread species for which too few records are used. Where scientists are usually concerned about data-deficiency for rare, narrow-ranged species, data quantity for widespread species appears to be crucial too. Therefore, estimating the prevalence of a modelled species is essential to determine the required minimum number of records. Although typically unknown, the species' prevalence can be estimated e.g. by calculating the Extent Of Occurrence (EOO) or Area Of Occupancy (AOO) (IUCN, 2001) or by calculating the predicted presence fraction based on an exploratory SDM to which a threshold is applied (Syfert *et al.*, 2014).

The differences in required minimum number of records based on the indicators of model performance applied here are the result of the different nature of these indicators. At the top end of the scale of model performance are thresholds such as the real AUC value of 0.9 and a Spearman rank correlation value of 0.9, which both classify a model as 'good'. Note that a Spearman correlation test compares absolute ranks, whereas the AUC only compares the relative ranks between presence and absence. Testing an SDM against null models (rank AUC) informs us if the SDM is 'significantly better than random expectation', which is not the same as 'good'. Obviously, the desired model accuracy strongly depends on the application (Guisan & Zimmermann, 2000, Jiménez-Valverde *et al.*, 2011, Liu *et al.*, 2011, Peterson, 2006). One may choose to accept a reasonably accurate model as an indication for where a species occurs, but rely exclusively on highly accurate models when more precise predictions are needed.

Factors that increase the minimum required sample size

We stress that the minimum numbers of records listed in Table 1 increases when working with real species and real spatial data related to factors addressed here. First, real species possibly correlate to a larger and more complex set of environmental predictors, whose signal is not fully represented by the two orthogonal PCA-based variables that were used as a proxy for environmental conditions here. To test the effect of the number of included variables on the required minimum number of records, we repeated our simulations using the first three and first four PCA axes, which together explained resp. 55% and 65% of the variance instead of the 43% explained by the first two PCA axes only. These analyses gave similar threshold values for the minimum required number of records, indicating that the thresholds will not substantially change by including more environmental variables (Supplementary material Appendix 2). Related to this and deserving future research is the question on the effect of including a variable important for the occurrence of a species that is not part of the model variables, e.g. a variable not included in global climate and soil data sets. In addition, multicollinearity of environmental predictors will have a negative effect on model accuracy (Dormann *et al.*, 2013), which was not an issue in our study using orthogonal variables. Third, we defined all our simulated species to be in equilibrium with climate (Araújo *et al.*, 2005), and therefore to demonstrate niche stability and complete niche-filling. The effects of time scale and biotic interactions on niche stability, as well as the effects of dispersal limitations on niche-filling have serious impact on the correctness of an SDM considered to reflect a species' distribution (Nogués-Bravo, 2009, Saupe *et al.*, 2012). Fourth, natural history specimens are commonly subject to a collecting bias (Reddy & Davalos, 2003, ter Steege *et al.*, 2011), often representing a geographical bias (Kadmon *et al.*, 2004, Loiselle *et al.*, 2008). When a collecting bias translates into an environmental bias – an unbalanced or partial coverage of the ecological niche – models show falsely inflated AUC values, but are actually performing worse compared to models based on unbiased data sets (Bean *et al.*, 2012, Merckx *et al.*, 2011, Raes, 2012). Consequently, the minimum required number of records increases (Feeley & Silman, 2011). Finally, other factors with a negative impact on model performance based on real data include misidentifications, incorrect georeferencing (Graham *et al.*, 2008, Moudry & Simova, 2012), and uncertainty in environmental variables and accuracy of climate data (Hijmans *et al.*, 2005). Although these aspects warrant future research, we feel safe to ignore them here, when specifically addressing the questions of our current study working with simulated species under optimal orthogonal bivariate environmental conditions.

AUC values based on presence-only data

Our results show that MaxEnt AUC values based on sampled presence vs. background data, differ from real AUC values based on true presence vs. absence data (Fig. 2). This supports previous critics on the use of AUC as indicator of model performance without further analysis (Jiménez-Valverde, 2012, Lobo *et al.*, 2008). Applying a generalized AUC threshold value – commonly set at 0.7 – results in the erroneous acceptance of SDMs based on small sample sizes as these have inflated MaxEnt AUC values and dismissal of good SDMs for widespread species that theoretically can never reach an AUC of 0.7. The aspect of the unknown maximum value of the MaxEnt AUC due to its dependence on the unknown real species' prevalence disqualifies the MaxEnt AUC as a reliable indicator of model accuracy if treated without further evaluation such as i.e. a null model test (Raes and ter Steege 2007). The exceptionally high values of both real AUC and MaxEnt AUC for models based on small sample sizes of narrow-ranged species (Fig. 2) should be treated with caution. Species with a prevalence of 0.05 show a strong imbalance between true presences and absences: 5% presences vs. 95% absences. The chance that a random presence has a higher probability of occurrence than a random absence for such species is high, resulting in high AUC values. These statistical artefacts inflate AUC values for SDMs of species with a prevalence below 0.1 (Lobo & Tognelli, 2011, McPherson & Jetz, 2007, McPherson *et al.*, 2004), which is confirmed by our results. It is therefore recommended to choose the study area proportionally to the presence area of the assessed species, so that a species' prevalence between 0.1 and 0.9 is achieved (McPherson *et al.*, 2004). Note that the species' dispersal capacity should be leading in defining the width of the border around the presence localities (Barve et al. 2011). Unfortunately, the true distribution and prevalence of a species is usually not known– after all, that is what we are trying to assess. In the current study, we assessed the reliability of an alternative: to evaluate SDMs using the MaxEnt AUC in a comparative way by testing it against a null model. For the virtual study area, our results show a strong congruence between the behavior of the real AUC and the MaxEnt AUC rank for species of all prevalence classes as well as required minimum sample sizes based on them (Figs. 2 & 3; Table 1). In contrast, for the real African study area, the behavior of these indicators of model performance differs and the minimum sample sizes required to obtain an accurate SDM based on the lower limit of the upper 95% range values of real AUC > 0.9 are on average twice as high as those based on the lower range limit of MaxEnt AUC rank values > 95. This difference is the result of the different nature of these indicators of model performance as addressed above ('good' vs. 'significantly better than random expectation'). We conclude that the use of a null model test (Raes & ter Steege, 2007) is an appropriate method to evaluate SDMs using the MaxEnt AUC value in a comparative way, provided that the nature of this indicator is respected: to identify SDMs that perform 'significantly better than random expectation'.

Aspects of modelling narrow-ranged species

The difficulties with accurate modelling of species with a narrow ecological niche are illustrated in our results by the very low Schoener's *D* and Hellinger distance *I* values for narrow-ranged species in the African study area, indicating that SDMs for narrow-ranged species perform worse than those of widespread species when using the same sample size (Fig. 5). In contrast, AUC values for narrow-ranged species are high, although a large spread in values is visible for smaller sample sizes (Fig. 2). This contrast between low *D* and *I* values and high AUC values can be explained by the large number of absences of which many are correctly predicted as absences (high specificity), but small numbers of presences of which only few are predicted correctly as presences (low sensitivity). The above-mentioned statistical artefact for species with a prevalence below 0.1 should be noted too. Other explanations for the high AUC values and AUC rank values of these SDMs could be spatial autocorrelation (Merckx *et al.*, 2011) and collecting bias (Phillips *et al.*, 2009) in the training samples, although the latter only applies to most situations when working with datasets of real species. Null models are based on randomly sampled records from the entire study area, with less or no spatial autocorrelation and no sampling bias. Consequently, a random null-model test results in a too optimistic acceptance of SDMs based on few samples. To counter the effect of collecting bias in real datasets, we recommend evaluating SDMs by using bias-corrected null-models (Raes & ter Steege, 2007) based on target-group background sampling, which has been shown to be effective (Phillips *et al.*, 2009).

Conclusions

We conclude that applying a lower limit to the sample size used in SDMs is essential for generating accurate SDMs. The lower limit strongly depends on the species' prevalence and the specific features of the targeted study area. The required minimum numbers of records for species of different prevalence classes based on analyses in our virtual study area apply only to an ideal, balanced, orthogonal world. These numbers strongly increase for an irregular real study area like our African study area. MaxEnt AUC values cannot be used for model evaluation as such, but testing these against random or bias-corrected null models provides a reliable alternative method. Generating and evaluating SDMs for narrow-ranged species, those with a prevalence below 0.1, is difficult and should be avoided by selecting a study area proportionally to the species' presence area and with respect to the species dispersal capacity.

The novel method presented here is applicable to any taxonomic clade or other group, study area and past, current or future climate scenario. The R-scripts with detailed stepwise methodology and a brief manual on how to apply these scripts to given data are provided in the Supplementary materials. We advocate the use of our method as a routine procedure prior (or in retrospect) to any SDM study. This will aid in verifying if required levels of data quantity and quality are met and will improve the reliability of SDMs as well as the results of all future studies involving SDMs.

Acknowledgements

We thank the subject editor and two anonymous reviewers for their valuable comments on the manuscript. NR was supported by NWO-ALW grant 819.01.014. All authors declare to have no conflict of interest.

Supplementary material

All supplementary materials (appendices 1-9) can be found in the published version of this article online: Appendix ECOG-01509 at www.ecography.org/readers/appendix

Chapter 3

Prevalence estimators put to the test

André S.J. van Proosdij[1,2], Jan J. Wieringa[1,2], Niels Raes[2], Marc S.M. Sosef[1,3]

1 Biosystematics Group, Wageningen University, Droevendaalsesteeg 1, 6708 PB Wageningen, the Netherlands

2 Naturalis Biodiversity Center, Darwinweg 2, 2333 CR Leiden, the Netherlands

3 Botanic Garden Meise, Nieuwelaan 38, 1860 Meise, Belgium

Abstract

1. Macroecological, biogeographical and conservation research depend on basic information about the geographic distribution and prevalence of a species: the fraction of the study area where the species is present. Typically, distributional data of most species is highly fragmented, especially for those occurring in the tropics. Therefore, accurate methods to estimate a species' prevalence are highly wanted.

2. Using simulated species, for which by definition the distribution and prevalence are known, we assess the accuracy and consistency of 10 existing and one novel method to estimate species' prevalence, which are based on different principles. Some methods use spatial data related to known occurrences, while others use ecological data linked to occurrences to quantify a species' niche width as a proxy for prevalence. Our novel method estimates prevalence as the fraction of raster cells within the minimum convex hull of species' samples, when all cells from the study area are plotted in normalized, 2-dimensional, environmental parameter space. We assess prevalence estimator' consistency for different sample sizes of simulated species ranging from narrow-ranged to widespread at spatial resolutions ranging from 2.5 arc-minute to 0.25 degree.

3. Our results show that prevalence estimators based on ecological data in general outperform those based on spatial data only. Our novel method is the most consistent estimator at all spatial resolutions for all but the smallest sample sizes, for which it ranks second best. Consistency of all prevalence estimators primarily depends on the used sample size. None of the estimators is accurate and can therefore be used as a direct estimator. However, by running simulations and using consistent estimators, accurate and consistent estimations of species' prevalence can be obtained indirectly for which we provide a step-wise procedure.

4. We recommend reconsidering the current use of the Extent Of Occurrence (EOO) and Area Of Occupancy (AOO) as estimators of the geographic range and actual occupied area respectively in macroecology, biogeography and conservation. Alternatively, we recommend to estimate the EOO from the predicted distribution based on a thresholded Species Distribution Model and to estimate the AOO by using the novel method presented here.

Key words

Geographic range, niche width, simulated species, Extent Of Occurrence, Area Of Occupancy, convex hull, IUCN Red List, rare, conservation

Introduction

Knowledge on the spatial distribution of species and their prevalence – the fraction of the study area where the species is present (McPherson *et al.*, 2004) – is crucial for research in many disciplines, including macroecology and biogeography (Gaston, 2003, Lomolino *et al.*, 2010). Setting priorities in conservation, especially in the light of climate and land use change heavily depends on knowledge of species distributions (Brooks *et al.*, 2006, Margules & Pressey, 2000). Yet, for many species, information on their distribution is scarce and highly fragmented, the so-called 'Wallacean shortfall' (Bini *et al.*, 2006, Lomolino, 2004, Whittaker *et al.*, 2005). Species Distribution Models (SDMs), which predict the distribution of a species based on a limited number of observations, are widely used to overcome this lack of information (Franklin, 2009, Guisan & Zimmermann, 2000). In order to generate SDMs with a sufficient level of accuracy, the model needs to be trained on a minimum number of records. This number is not fixed, but increases with increasing species' prevalence (Pearson *et al.*, 2007, van Proosdij *et al.*, 2016b). Hence, having knowledge of the species' prevalence is essential to develop accurate SDMs. As exactly this knowledge is typically missing for most species, accurate and consistent estimators of species' prevalence are of great value. Quantitative measures of prevalence are known to be sensitive to the applied spatial resolution, resulting in an increased prevalence for lower spatial resolutions (Azaele *et al.*, 2012, Hartley & Kunin, 2003, Hurlbert & Jetz, 2007). Ideally, estimators of species' prevalence should be accurate over a wide range of spatial resolutions.

Many different estimators of species' prevalence have been developed. Which estimate of prevalence best matches a species' true prevalence can be tested quantitatively using simulated species, for which niche dimensions and prevalence are pre-defined. Simulated species definitions like these have been recommended and used as a method to systematically assess specific aspects of SDMs (Miller, 2014, van Proosdij *et al.*, 2016b, Zurell *et al.*, 2010). In the present study, we use simulated species to rigorously assess the accuracy and consistency of different prevalence estimators over four spatial resolutions ranging from 2.5 arc-minute (~5 km) to 0.25 degree (~30 km, hereafter 15 arc-minute).

Review of prevalence estimators

We review and subsequently test 10 widely used prevalence estimators and present one novel method, 'Fraction MCP PCA', to estimate species' prevalence (Table 1). These 11 methods build on different principles and are tested using simulated species. Prevalence estimators may use either spatial or ecological information derived from known presence localities of a species. Some estimators incorporate the distribution of samples over environmental parameters (‘niches’). Few incorporate the density or availability of niche space in the study area; the number of raster cells in the study area that represents specific niches. Two classic and frequently used estimators using spatial information are the Extent Of Occurrence (EOO) and the Area Of Occupancy (AOO). EOO is the area limited by the geographically outermost known occurrences and as such quantifies the geographical range of the known occurrences, whereas AOO quantifies the area that is actually occupied by that species based on known occurrences and a buffer area around each presence locality (Gaston, 1991, Gaston & Fuller, 2009, Lomolino *et al.*, 2010). These two estimators are implemented in the Geospatial Conservation Assessment Tool (GeoCAT; (Bachman *et al.*, 2011) and are the standard metric for IUCN Red List assessments (IUCN, 2014).

1. EOO: The most straightforward and most commonly used definition of EOO is the area of the minimum convex polygon (MCP) or convex hull that includes all sampled localities without omitting obvious unsuitable areas (Gaston, 1991).
2. AOO 2KM: The AOO is calculated by adding a buffer to the known presence localities and then summing the total area (Bachman *et al.*, 2011, Edelsbrunner *et al.*, 1983, IUCN, 2001). Although the spatial resolution strongly influences the value of the AOO (Hartley & Kunin, 2003, Willis *et al.*, 2003), commonly a buffer of 2 km is applied, as recommended by the IUCN.
3. AOO 10PCT: Alternatively, one can add a buffer of 10% of the minimum distance between the most distant pair of occurrences applying an equal area projection (Bachman *et al.*, 2011, Willis *et al.*, 2003). However, this approach is not accepted by the current IUCN Red List Criteria (IUCN, 2014).

Analysing the predicted distribution based on an SDM offers opportunities for a retrospective approach to estimate the species’ prevalence. In a case study on Red Listing of Neotropical plants, the EOO was inferred using the predicted distributions (Syfert *et al.*, 2014).

4. EOO predicted: The EOO of a predicted distribution is computed by applying a convex hull to the predicted presence distribution (Syfert *et al.*, 2014). Presence and absence predictions are derived from an SDM to which a threshold is applied.
5. AOO predicted: Using presence/absence maps based on predicted distribution, the AOO of a predicted distribution is computed as the fraction of raster cells where the species is predicted to be present.

Alternatively, prevalence can be estimated using the species’ niche width (or niche breadth). The niche is a concept defined in many different ways and subject of much debate, especially in the field of species distribution modelling and/or ecological niche modelling (Araújo & Peterson, 2012, Peterson & Soberón, 2012, Warren, 2012). In general, species with a wider ecological niche have a larger prevalence (Kadmon *et al.*, 2003, Pulliam, 2000, Slatyer *et al.*, 2013). Several methods to measure species' niche dimensions have been described and criticized (Chejanovski & Wiens, 2014, Colwell & Futuyma, 1971, Doledec *et al.*, 2000, Feinsinger *et al.*, 1981). We used the following methods as prevalence estimators based on niche width. The environmental variables are defined as the first two axes of a Principal Component Analysis of climatic, altitude and soil variables (Explained in detail under 'Methods'). The following prevalence estimators operate in normalized 2-dimensional environmental parameter space defined by the first two PCA axes (hereafter 'PCA space').

6. Fraction PCA1: Feeley & Silman (2011) used a simple method: climatic niche width defined as the fraction of an ecological gradient covered by the samples. We applied their method to the first PCA axis.
7. Maximum Euclidean distance: A method using multiple gradients defining the study area is to measure niche width as the maximum Euclidean distance between samples plotted in PCA space (Merckx *et al.*, 2011).
8. 2SD PCA1: The method used by Thuiller *et al.* (2004) incorporates variation in ecological values of the sampled localities by taking the fraction of an ecological gradient in the study area covered by *mean* ± 1 *standard deviation* (SD), which we here apply to the first PCA axis.
9. Inverse kernel height: Broennimann *et al.* (2012) fitted a standard normal density kernel on the samples plotted in normalized, 2-dimensional, environmental parameter space and used that as a proxy for niche width. This method incorporates multiple gradients as well as information on the variation in ecological values of the sampled localities. The inverse of the height of the kernel serves as a proxy for niche width, as a narrower species' niche results in a narrower and higher kernel. We apply this estimator in PCA space.
10. Area MCP PCA: Beck *et al.* (2013) used a minimum convex polygon (MCP) on the samples plotted in PCA space and took the area of the MCP as a measure of niche width.

Novel method

11. Fraction MCP PCA: In addition to the methods above, we here present a novel method. We define niche width as the fraction of raster cells from the study area which are located inside the MCP encompassing all samples when these are plotted in PCA space. This method builds upon the method of Beck *et al.* (2013) ('Area MCP PCA'), but corrects for the unequal density or availability of

Table 1. Overview of assessed estimators of species' prevalence with description of the methods and applied R-functions, aspects of data type and variance in the samples as well as in the study area.

Prevalence estimator	Description of the methods and applied R-functions	Type of data	Variance of samples included?	Distribution of niches in study area included?	Reference
1: EOO	Extent Of Occurrence, the area of the minimum convex polygon encompassing all records using an equal area projection: functions *chull* of the *grDevices* library (R Core Team, 2016) and *areaPolygon* of the *geosphere* library (Hijmans, 2014).	Spatial	No	No	(IUCN, 2014)
2: AOO 2KM	Area Of Occupancy, the total area based on a 2 km buffer around each record using an equal area projection and counting areas of overlapping buffers only once: *gBuffer* and *gArea* functions of the *rgeos* library (Bivand & Rundel, 2014).	Spatial	No	No	(IUCN, 2014)
3: AOO 10PCT	Area Of Occupancy, the total area based on a customized buffer around each record. Buffers equal 10% of the minimum distance between the two most distant records using an equal area projection. Overlapping buffer areas are counting only once: functions *distm* of the *geosphere* library (Hijmans, 2014), *gBuffer* and *gArea* of the *rgeos* library (Bivand & Rundel, 2014).	Spatial	No	No	(IUCN, 2014)
4: EOO predicted	Extent Of Occurrence or the area of the minimum convex polygon bounding all predicted presences based on an SDM to which the 'maximum training sensitivity plus specificity' threshold rule was applied and using an equal area projection: functions *chull* of the *grDevices* library (R Core Team, 2016) and *areaPolygon* of the *geosphere* library (Hijmans, 2014).	Ecological & Spatial	Yes	Yes	(Syfert *et al.*, 2014)

5: AOO predicted	Fraction of the raster cells predicted as presences based on an SDM to which the 'maximum training sensitivity plus specificity' threshold rule was applied.	Ecological & Spatial	Yes	Yes	
6: Fraction PCA1	Fraction of the 1st PCA axis' range covered by the sampled records.	Ecological	No	No	(Feeley & Silman, 2011)
7: Maximum Euclidean distance	Maximum Euclidean distance between records in normalized parameter space: function *rdist* of the *fields* library (Nychka *et al.*, 2013).	Ecological	No	No	(Merckx *et al.*, 2011)
8: 2SD PCA1	Fraction of the 1st PCA axis' range covered by *mean* ± 1 *SD* of the records.	Ecological	Yes	No	(Thuiller *et al.*, 2004)
9: Inverse kernel height	Inverse of the height of a standard normal density kernel fitted on the records plotted in normalized parameter space: function *kernelUD* of the *adehabitat* libray (Calenge, 2006).	Ecological	Yes	Yes	(Broennimann *et al.*, 2012)
10: Area MCP PCA	Area of the minimum convex polygon encompassing all records plotted in normalized parameter space: functions *chull* of the *grDevices* libray (R Core Team, 2016) and *areapl* of the *splancs* library (Rowlingson & Diggle, 2013).	Ecological	No	No	(Beck *et al.*, 2013)
11: Fraction MCP PCA	The fraction of raster cells located inside the minimum convex polygon encompassing all records plotted in normalized parameter space: functions *chull* of the *grDevices* libray (R Core Team, 2016) and *pnt.in.poly* of the *SDMTools* library (VanDerWal *et al.*, 2014).	Ecological	No	Yes	

niches in the study area; in PCA space, two equally large MCPs can represent different numbers of raster cells, hence have different values of Fraction MCP PCA.

Our main research question is to assess which of the above prevalence estimators provides the best estimate of species' true prevalence for spatial resolutions ranging between 2.5 and 15 arc-minute. We identify the most consistent estimators by comparing the estimated and predefined prevalence values of the simulated species. We provide a step-wise procedure to correct for inaccuracy in the prevalence estimations. Finally, we address the importance of prevalence estimators for macroecology, biogeography and conservation.

Material and methods

All analyses were performed in R (R Core Team, 2016). Scripts used for this study are available as Supplementary Material Appendices 15-17 which provide details on the applied R-functions. To enhance readability, 'simulated species' are referred to as 'species', unless stated otherwise.

Study area and simulated species

Although universally applicable, we selected tropical Africa as a study area ranging from 15°N to 19°S and from 17.5°W to 43°E. Tropical Africa is illustrative for many tropical regions as quantity and quality of botanical specimens data strongly differ between areas due to (regional) undersampling or limited digital accessibility of data (Graham *et al.*, 2004, Küper *et al.*, 2006).

Environmental variables included WORLDCLIM bioclimatic variables (Hijmans *et al.*, 2005), soil variables (Harmonized World Soil Database; (FAO/IIASA/ISRIC/ISSCAS/JRC, 2012) and SRTM 90 m resolution elevation data (srtm.csi.cgiar.org) (Supplementary Material Appendix 1 & 15). Variables were prepared at four spatial resolutions: 2.5, 5, 10, and 15 arc-minutes. For each spatial resolution, we extracted two orthogonal gradients from the 39 variables by means of a Principal Component Analysis, explaining 40% of the total environmental variation (41% for 10 and 15 arc-minute resolution). These two PCA-based variables shape the normalized, 2-dimensional, environmental parameter space (PCA space) and were used to define the simulated species. In general, a species' response to environmental variables can be described as a multivariate normal function (Boucher-Lalonde *et al.*, 2012), which has been applied in several other studies (Broennimann *et al.*, 2012, Duan *et al.*, 2015). Here, we defined habitat suitability of the species for each raster cell as a bivariate normal response to the two orthogonal PCA-based predictors using the *dmvnorm* function of the R-library *mvtnorm* (Gentz *et al.*, 2014). This is similar to the 'Artificial bell-shaped response method' of Varela *et al.* (2014) and Duan *et al.* (2015) although they applied it directly to environmental variables. Our species were defined to be present in raster cells whose environmental bivariate variables are within the central circle of the bivariate normal density that has probability 68%. Here, using standardized, fully orthogonal PCA-based variables, this is represented by a circle cutting the PCA axes at the optima ± 1 *standard deviation* (*SD*). We defined the following ten species' prevalence classes (fraction of presence cells): 0.05, 0.10, 0.15, 0.20, 0.25, 0.30, 0.35, 0.40, 0.45, and 0.50. The defined prevalences were realized by varying the *SD* of the species to both PCA gradients simultaneously. Species' optima were unique and were selected by taking the values of the two PCA-based predictors on a randomly selected locality in the study area (Supplementary Material Appendix 17). We assume that each species inhabits all environmentally suitable areas.

Sample sizes and replications

We used 12 sample sizes, 5-10, 15, 20, 25, 50, 75, and 100 records, which were based on sampling without replacement from the pre-defined presence localities. Sampling probability equalled the defined habitat suitability score of each raster cell. For each of the ten prevalence classes, we replicated species definitions 100 times, resulting in 1000 simulated species. Sampling of the 12 sample sizes was done for each simulated species, yielding a total of 12,000 individual samples. This procedure was replicated for each of the four spatial resolutions.

Species Distribution Models

We selected MaxEnt to generate the SDMs (Phillips *et al.*, 2006) that were used to assess the performance of those prevalence estimators that are based on predicted presence localities (Table 1, #4-5). MaxEnt is widely used and has shown to outperform other methods, especially when dealing with presence-only data (Aguirre-Gutiérrez *et al.*, 2013, Elith *et al.*, 2006). We set MaxEnt to use linear and quadratic features only for all sample sizes following recommendations by Merow *et al.* (2013). All other settings were kept as default. As environmental variables the SDMs used the 2 PCA-based variables. To convert the continuous MaxEnt predictions into discrete predicted presence/absence maps, we applied the 'maximum training sensitivity plus specificity' rule as threshold value (Liu *et al.*, 2013).

Testing estimator performance

Estimator performance depends on both consistency and accuracy. Consistency of the 11 prevalence estimators summarized in Table 1 was assessed by the fit of a linear regression between the estimated and the pre-defined prevalence values using the function *lm* of the R-library *stats* (R Core Team, 2016) with separate regressions for each sample size and each spatial resolution. To compensate for heteroscedasticity, estimated prevalence values of most estimators were arcsine transformed using the square-rooted values (asin sqrt) (Fowler *et al.*, 1998)(Supplementary Material Appendix 16). Accuracy of the estimators and a step-wise procedure to compensate for inaccuracy is discussed.

Results

Analyses of the 12,000 individual samples per spatial resolution resulted in estimator values for each of the 11 assessed methods. Accuracy of the estimators at 5 arc-minute resolution is visualised in Figure 1 by showing the regression lines for selected sample sizes with facets for each estimator. All estimators show a significant, positive, linear relation with pre-defined prevalence of the simulated species, except AOO 2KM for which no significant relation was found for most sample sizes (Table 2). Intercept and slope of the regression lines depend on the choice of estimator, sample size and pre-defined prevalence (regression statistics for all sample sizes and estimators in Supplementary Material Appendix 4). For some estimators, sample size influences the intercept and slope for fitted regressions. EOO, AOO 10PCT, Fraction MCP PCA and to a lesser extent Fraction PCA1, Maximum Euclidean distance and Area MCP PCA show a larger intercept and larger slope for increasing sample size, thus predicting a larger prevalence. In contrast, EOO predicted, AOO predicted, 2SD PCA1 and Inverse kernel height are largely insensitive to sample size.

Consistency of the assessed estimators at 5 arc-minute spatial resolution is presented in Figure 2 with separate panels for six selected sample sizes (results for all spatial resolutions in Supplementary Material Appendix 2). R^2 values, intercept and slope for the regression based on the largest sample size (n = 100) are summarized in Table 2 for 5 arc-minute spatial resolution (results for other resolutions in Supplementary Material Appendix 3). Consistency of the estimators increases with increasing sample sizes for all estimators except AOO 2KM.

The most consistent estimator is marked in bold and the second best estimator is underlined for each sample size in Table 2. The rank based on consistency for specific sample sizes differs, but rank differences are small. The novel method presented here, Fraction MCP PCA, is the most consistent for all but the smallest sample sizes (n = 5 and n = 6) for which it closely ranks second after #9 Inverse kernel height. Consistency for Fraction MCP PCA at 5 arc-minute spatial resolution is visualized in Figure 3 by showing the 95% range of the observed values and regression lines in separate facets for six selected sample sizes (other estimators at 5 arc-minute spatial resolution in Supplementary Material Appendices 5-14). For all assessed estimators, the range in estimated prevalence values is larger for larger prevalence classes, although the effect on the fitted regressions is small and most of this heteroscedasticity is compensated for by the applied arcsine transformation on the square-rooted values.

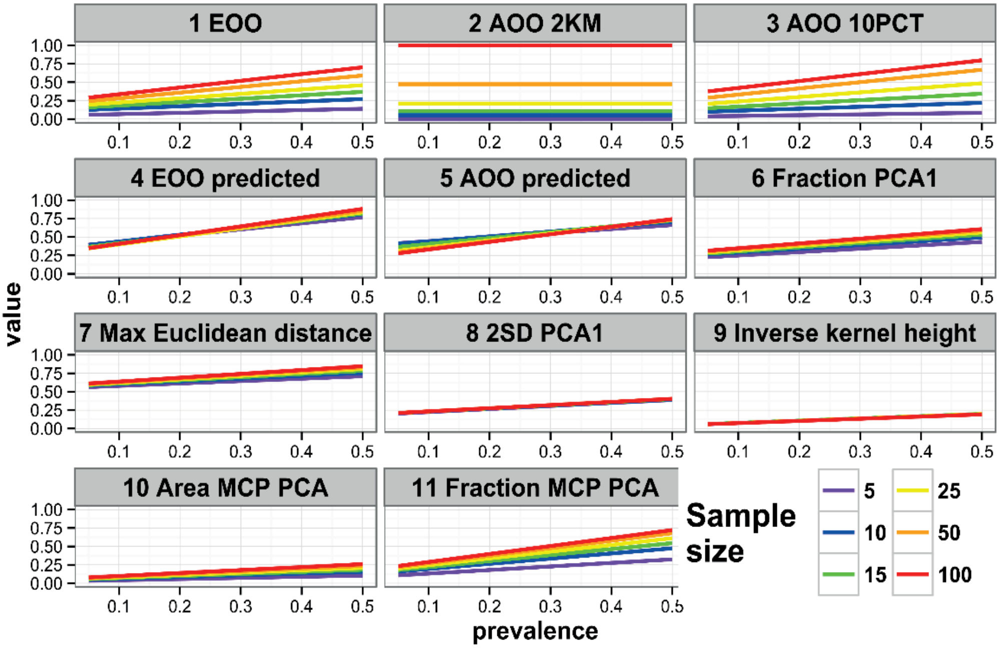


Figure 1. Accuracy of 11 prevalence estimators *with estimated prevalence values (y-axis) as a function of pre-defined prevalence values (x-axis) at 5 arc-minute spatial resolution with facets for each assessed estimator and regression lines for selected sample sizes (n = 5, 10, 15, 25, 50, 100). To facilitate visual comparison EOO, AOO 2KM, AOO 10PCT and EOO predicted were rescaled to (0-1).*

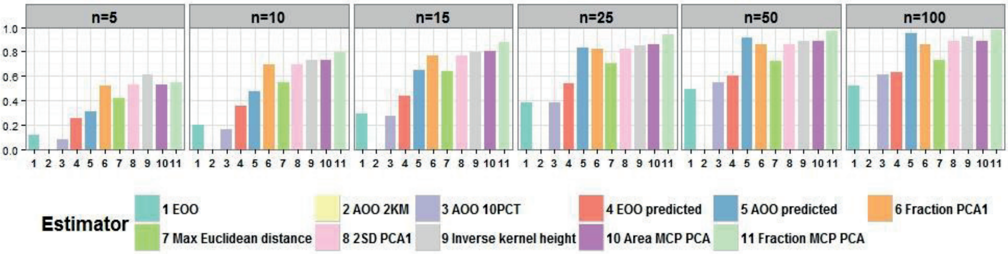


***Figure 2. Consistency of 11 prevalence estimators** at 5 arc-minute spatial resolution with consistency defined as the R^2 values of the linear regression fitted on the estimated and predefined prevalence values with facets for six different sample sizes (5, 10, 15, 25, 50, and 100).*

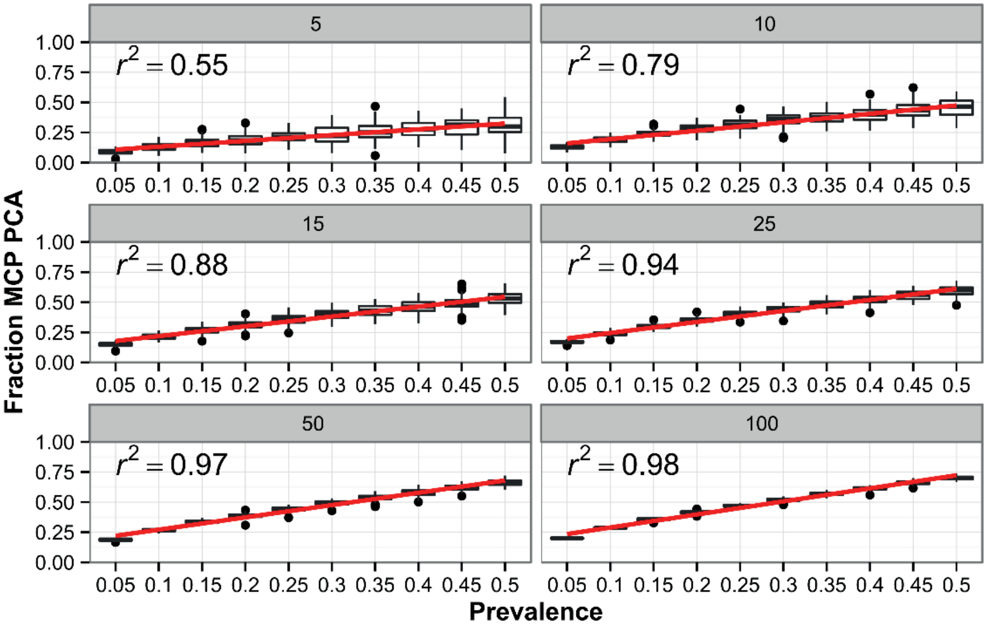


***Figure 3. Consistency of the prevalence estimator 'Fraction MCP PCA'** (y-axis) as a function of defined prevalence values (x-axis) at 5 arc-minute spatial resolution with facets for six selected sample sizes (n = 5, 10, 15, 25, 50, 100). Fraction MCP PCA is defined as fraction of raster cells located inside the minimum convex polygon encompassing all records plotted in normalized parameter space. In each facet, boxplots show the 95% range of the observed values, regression lines are given in red, and R^2 values of the regressions are provided.*

Table 2. Consistency of assessed estimators of species' prevalence at 5 arc-minute spatial resolution*. For each estimator, intercept, slope and confidence interval for the prediction are given for the fitted linear regressions based on sample size n = 100. For each estimator, R^2 values of the linear regressions are given for six selected sample sizes (n = 5, 10, 15, 25, 50, 100) with significance levels p < 0.01(*) and p < 0.001(**). For each sample size, the most consistent estimator is marked in bold, the second best estimator is underlined.*

Prevalence estimator	Type of data	Data trans-formation	Intercept n = 100	Slope n = 100	CI n = 100	R^2 n = 5	R^2 n = 10	R^2 n = 15	R^2 n = 25	R^2 n = 50	R^2 n=100
1: EOO	Spatial	None	4.0E+12	1.5E+13	4.0E+12	0.12**	0.20**	0.29**	0.38**	0.49**	0.52**
2: AOO 2KM	Spatial	None	1.2E+09	0	0	0	0	0	0	0	0
3: AOO 10PCT	Spatial	None	7.9E+12	2.2E+13	5.1E+12	0.08**	0.16**	0.27**	0.38**	0.55**	0.61**
4: EOO predicted	Ecological & Spatial	None	7.1E+12	2.2E+13	4.7E+12	0.25**	0.35**	0.44**	0.54**	0.60**	0.63**
5: AOO predicted	Ecological & Spatial	arcsin sqrt	0.23	1.03	0.07	0.31**	0.47**	0.65**	0.83**	0.91**	0.95**
6: Fraction PCA1	Ecological	arcsin sqrt	0.28	0.65	0.07	0.52**	0.69**	0.77**	0.82**	0.86**	0.86**
7: Maximum Euclidean distance	Ecological	arcsin sqrt	0.59	0.52	0.09	0.42**	0.55**	0.64**	0.70**	0.72**	0.73**
8: 2SD PCA1	Ecological	arcsin sqrt	0.19	0.43	0.04	0.53**	0.69**	0.77**	0.82**	0.86**	0.89**
9: Inverse kernel height	Ecological	arcsin sqrt	0.04	0.29	0.03	__0.61**__	0.73**	0.79**	0.85**	0.89**	0.92**
10: Area MCP PCA	Ecological	arcsin sqrt	0.06	0.40	0.04	0.53**	0.73**	0.80**	0.86**	0.89**	0.89**
11: Fraction MCP PCA	Ecological	arcsin sqrt	0.18	1.08	0.04	0.55**	__0.79**__	__0.88**__	__0.94**__	__0.97**__	__0.98**__

Discussion

Evaluating estimators of prevalence

Our results show that prevalence estimators based on ecological information in general outperform estimators based on spatial information only (Figure 2 & Table 2). This can be explained by their insensitivity for patchiness of the species' distribution, as well as their ability to include information on the unequal density or availability of niches in the study area (Broennimann *et al.*, 2012), two aspects commonly mentioned as reasons of concern (Beck *et al.*, 2013, Gaston & Fuller, 2009). The novel estimator introduced here, Fraction MCP PCA, is the most consistent at every applied spatial resolution and for all but the smallest sample sizes, with R^2 values up to 0.98 (Table 2). This estimator is least sensitive for the two above-mentioned aspects.

A few aspects should be considered. To test the effect of non-normal species' responses to environmental variables, we repeated the analyses using a uniform sampling probability instead of a probability equal to the defined habitat suitability. A uniform sampling probability in our analyses mimics the situation where the species shows a thresholded response to environmental variables and, consequently, its abundance and detectability is uniform for all presence localities. The results were similar, confirming our conclusions.

The use of a bivariate density kernel (Table 1, #9: 'Inverse kernel height') contains an aspect of circularity as the simulated species is defined using a bivariate normal response and the kernel is fitted with a bivariate normal response too. In general, however, real species show similar responses (Boucher-Lalonde *et al.*, 2012) and the method was previously used in other studies based on simulated species (Broennimann *et al.*, 2012, Varela *et al.*, 2014). For the overall best performing estimator, Fraction MCP PCA, this potential pitfall is not an issue, as the method is insensitive to the type of the species' response function to environmental predictors.

The weak performance of EOO and AOO in our analysis supports earlier critics on their use for conservation purposes, especially for species with non-convex shaped spatial distributions (Burgman & Fox, 2003, Gaston & Fuller, 2009, Syfert *et al.*, 2014). The alpha-convex hull and alpha shape have been suggested as alternatives (Burgman & Fox, 2003). Both use a Delauney triangulation where *alpha* defines the radius of the half-planes that limit the boundaries of the hull and hence the level of detail in the shape (Edelsbrunner *et al.*, 1983). However, as the value for *alpha* strongly influences the outcome and conventions on its preferred value are lacking, we regard the use of alpha shape and alpha-convex hull as not useful to estimate species' prevalence.

How to estimate the prevalence of real species?

None of the assessed methods can directly estimate species' prevalence accurately. Some methods, however, are highly consistent indirect prevalence estimators with R^2 values as high as 0.98 (Table 2). By using the fitted regression equations based on simulations (Table 2 for n = 100, Supplementary Material Appendix 4 for other sample sizes) and the estimated prevalence values based on real species sample data, an accurate estimation of the true prevalence can be obtained. The following step-wise procedure can be used to obtain an accurate estimation of the species' prevalence using the R scripts provided in Supplementary Material Appendix 15-17: 1) extract PCA-based environmental variables, 2) run simulations with virtual species for different classes of species' prevalence and sample sizes, 3) fit the linear regression to the estimated and pre-defined prevalence values, 4) apply the prevalence estimator to real species sample data, and 5) use the fitted regression equation and estimated prevalence value to compute an accurate estimation of the species' real prevalence. When estimating the prevalence of real species, we recommend applying the Fraction MCP PCA estimator. However, when the relative contribution of each environmental variable to the species' response is unclear, the AOO predicted might be applied to the original environmental variables, as MaxEnt will weigh environmental variables relative to their contribution to the model (Phillips *et al.*, 2006). Two aspects should be considered: natural habitats loss and incomplete range filling, which both decrease a species' real and estimated prevalence.

Effect of spatial resolution

Species' prevalence is known to be sensitive to the spatial resolution (Azaele *et al.*, 2012, Hartley & Kunin, 2003, Hurlbert & Jetz, 2007). However, we found only minor differences between estimator accuracy, consistency, regression intercept, and slope at four different spatial resolutions. This can be explained by the definition of our presence localities. We defined an entire raster cell to be either a presence or an absence locality. In reality, a species is represented by individuals who are commonly not evenly distributed, resulting in species-dependent scale-area curves (Hartley & Kunin, 2003). This is beyond the scope of this study, however, in which we assessed how accurate and consistent a suite of methods can estimate the prevalence of simulated species for which the prevalence is known.

Bias and errors in sample data

Commonly, specimens data sets of natural history collections show a collecting bias caused by the tendency to collect close to villages, roads, rivers, and in national parks (Hortal *et al.*, 2007, Kadmon *et al.*, 2004, Reddy & Davalos, 2003). In addition, uneven effort in data mobilization causes large differences in availability of specimens data between countries (Beck *et al.*, 2014). This bias influences estimations of prevalence based on spatial data including the EOO and AOO, nicely illustrated by Wieringa & Mackinder (2012). Collecting bias can result in ecological bias when parts of the species' niche are underrepresented by the sampled records (Loiselle *et al.*, 2008), although this is not necessarily the case (Kadmon *et al.*, 2004). Incomplete niche coverage can lead to an underestimation of the species' prevalence, even when estimators use ecological data (Raes, 2012). Errors in georeferencing and identification are typically present in real species data, but absent in samples in our analyses based on simulated species. When generating presence/absence maps from SDMs, commonly, the 'ten percentile training presence' threshold is applied. This forces 10% of the actual presence localities outside the predicted presence area, allowing that 10% of the records may be wrongly georeferenced or identified without serious consequences for the model.

Implications for biogeography, macroecology and conservation

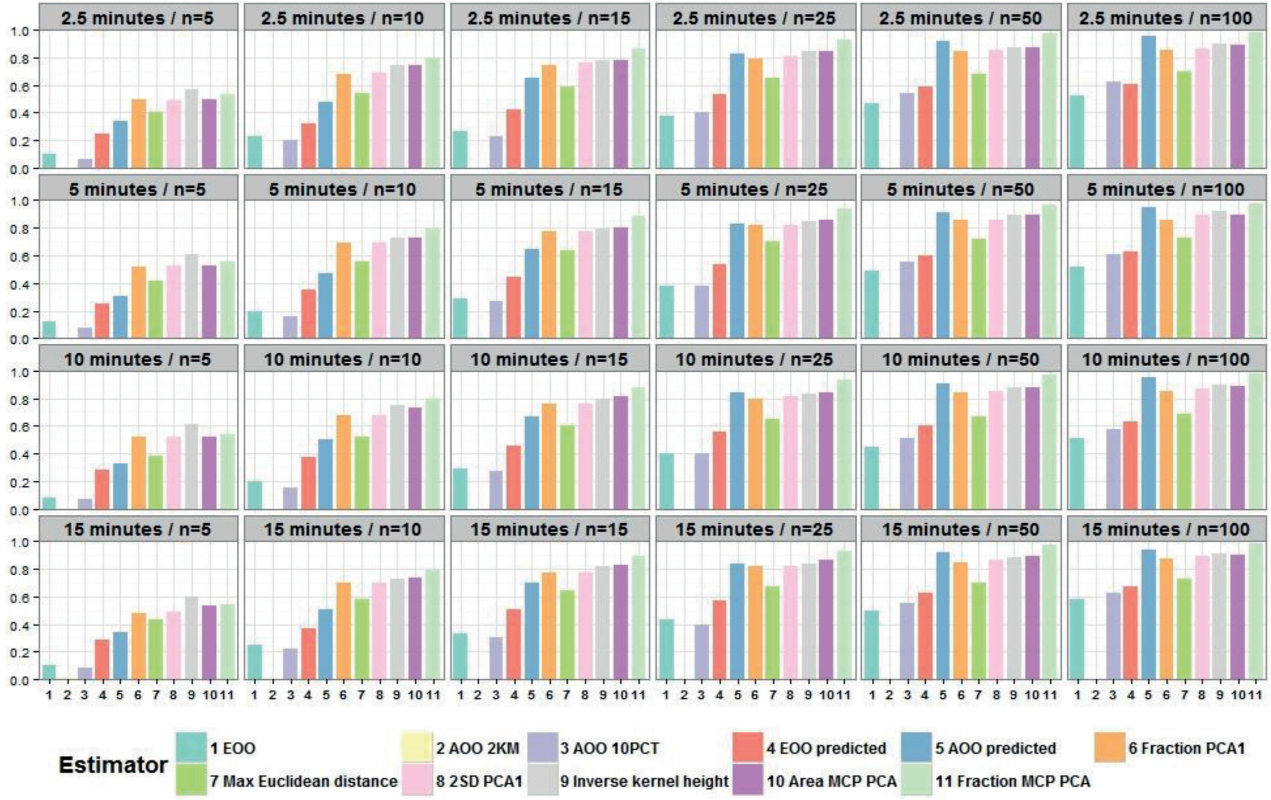
Macroecology, biogeography and conservation build on reliable information on species' prevalence. Here, we show that the use of commonly applied methods such as EOO and AOO may result in incorrect prevalence data, possibly leading to false conclusions. Particularly when conservation priorities are set and resources are allocated based on such incorrect prevalence data, this may jeopardize these exact priorities. In Red List assessments of species, EOO and AOO based on known occurrences are currently used as standard evaluation criteria (IUCN, 2001). However, our results indicate that estimators based on ecological data clearly outperform them, and that AOO 2KM has no relation to species' prevalence at all. On the other hand, the IUCN Red List Guidelines do allow alternatives and explicitly state that ''Both AOO and EOO may be estimated based on known, inferred or projected sites of present occurrences of a taxon'', where "projected" refers to "spatially predicted sites on the basis of habitat maps or models" (IUCN, 2001, IUCN, 2014). It therefore seems logical to adapt the IUCN Red List Guidelines and prescribe better performing estimators of the EOO and AOO.

Conclusions and recommendations

We assessed the performance of prevalence estimators using simulated species in a tropical African study area at four different spatial resolutions. Our results show that estimators based on ecological information outperform those based on spatial information only. Particularly, the novel estimator presented here – the fraction of raster cells in the study area located inside the minimum convex polygon when plotted in normalized parameter space ('Fraction MCP PCA') – has shown to outperform all other assessed estimators at each spatial resolution for all but the smallest sample sizes. A more accurate estimation of a species' prevalence, as achieved by this method in the here described step-wise procedure, will aid in a better understanding of the rarity of species. The default use of range and prevalence estimators based on spatial data only – the EOO and AOO – in the field of biogeography, macroecology and species conservation, including IUCN Red List assessments, should be reconsidered. We recommend to estimate the species' range or EOO by the area of the minimum convex polygon that includes all presence localities based on its SDM to which the 'ten percentile training presence' threshold is applied (Table 1, #4: 'EOO predicted'). To obtain a more accurate estimation of the species' actual prevalence or AOO, we recommend using the Fraction MCP PCA.

Supplementary material

All supplementary materials (appendices 1-17) are available upon request.



***Supplementary Material Appendix 2. Consistency of 11 assessed prevalence estimators as a function of sample size and spatial resolution**. Consistency is defined as the R^2 values of the linear regression fitted on the estimated and predefined prevalence values. Results for six different sample sizes (6 columns: 5, 10, 15, 25, 50, and 100) and four different spatial resolutions (4 rows: 2.5, 5, 10, and 15 arc-minute) are given. Note that 2 AOO 2KM is not missing, but has only values of 0 or 0.01.*

Supplementary Material Appendix 3. Consistency of assessed estimators of species' prevalence at 2.5, 5, 10, and 15 arc-minute spatial resolution. *For each estimator, intercept, slope and confidence interval for the prediction are given for the fitted linear regressions based on sample size n = 100. For each estimator, R^2 values of the linear regressions are given for six selected sample sizes (n = 5, 10, 15, 25, 50, 100) with significance levels* $p < 0.01$*(*) and* $p < 0.001$*(**). For each sample size, the most accurate estimator is marked in bold, the second best estimator is underlined.*

15 arc-minute											
Prevalence estimator	**Type of data**	**Data trans-formation**	**Intercept n = 100**	**Slope n = 100**	**CI n = 100**	**R^2 n = 5**	**R^2 n = 10**	**R^2 n = 15**	**R^2 n = 25**	**R^2 n = 50**	**R^2 n=100**
1: EOO	Spatial	None	3.8E+12	1.5E+13	3.7E+12	0.10**	0.25**	0.33**	0.43**	0.50**	0.58**
2: AOO 2KM	Spatial	None	1.2E+09	0	0	0.01	0	0	0	0	0
3: AOO 10PCT	Spatial	None	7.7E+12	2.3E+13	4.9E+12	0.08**	0.22**	0.30**	0.40**	0.55**	0.63**
4: EOO predicted	Ecological & spatial	None	5.9E+12	2.2E+13	4.4E+12	0.29**	0.37**	0.51**	0.57**	0.63**	0.67**
5: AOO predicted	Ecological & spatial	arcsin sqrt	0.23	1.03	0.07	0.34**	0.51**	0.70**	0.84**	0.92**	0.94**
6: Fraction PCA1	Ecological	arcsin sqrt	0.31	0.71	0.08	0.48**	0.70**	0.77**	0.82**	0.85**	0.87**
7: Maximum Euclidean distance	Ecological	arcsin sqrt	0.50	0.62	0.11	0.43**	0.58**	0.64**	0.67**	0.70**	0.73**
8: 2SD PCA1	Ecological	arcsin sqrt	0.21	0.45	0.05	0.49**	0.70**	0.77**	0.82**	0.86**	0.89**
9: Inverse kernel height	Ecological	arcsin sqrt	0.05	0.34	0.03	**0.60****	0.73**	0.82**	0.84**	0.88**	0.91**
10: Area MCP PCA	Ecological	arcsin sqrt	0.07	0.47	0.04	0.53**	0.74**	0.83**	0.86**	0.89**	0.90**
11: Fraction MCP PCA	Ecological	arcsin sqrt	0.18	1.09	0.04	0.54**	**0.79****	**0.89****	**0.93****	**0.97****	**0.98****

10 arc-minute											
Estimator of prevalence	**Type of data**	**Data trans-formation**	**Intercept n = 100**	**Slope n = 100**	**CI n= 100**	**R^2 n = 5**	**R^2 n = 10**	**R^2 n = 15**	**R^2 n = 25**	**R^2 n = 50**	**R^2 n = 100**
1: EOO	Spatial	None	4.0E+12	1.5E+13	4.1E+12	0.08**	0.20**	0.29**	0.40**	0.45**	0.51**
2: AOO 2KM	Spatial	None	1.2E+09	0	0	0	0	0	0	0.01*	0
3: AOO 10PCT	Spatial	None	7.8E+12	2.2E+13	5.3E+12	0.07**	0.15**	0.27**	0.40**	0.51**	0.58**
4: EOO predicted	Ecological & spatial	None	6.4E+12	2.2E+13	4.7E+12	0.28**	0.37**	0.46**	0.56**	0.60**	0.63**
5: AOO predicted	Ecological & spatial	arcsin sqrt	0.23	1.03	0.07	0.33**	0.50**	0.67**	0.84**	0.91**	0.95**
6: Fraction PCA1	Ecological	arcsin sqrt	0.30	0.68	0.08	0.52**	0.68**	0.76**	0.80**	0.84**	0.85**
7: Maximum Euclidean distance	Ecological	arcsin sqrt	0.53	0.56	0.10	0.38**	0.52**	0.60**	0.65**	0.67**	0.69**
8: 2SD PCA1	Ecological	arcsin sqrt	0.21	0.44	0.05	0.52**	0.68**	0.76**	0.81**	0.85**	0.87**
9: Inverse kernel height	Ecological	arcsin sqrt	0.05	0.32	0.03	**0.61****	0.75**	0.80**	0.83**	0.88**	0.90**
10: Area MCP PCA	Ecological	arcsin sqrt	0.07	0.45	0.05	0.52**	0.73**	0.81**	0.84**	0.88**	0.89**
11: Fraction MCP PCA	Ecological	arcsin sqrt	0.18	1.09	0.04	0.54**	**0.80****	**0.88****	**0.93****	**0.97****	**0.98****

5 arc-minute											
Estimator of prevalence	Type of data	Data trans-formation	Intercept n = 100	Slope n = 100	CI n= 100	R^2 n = 5	R^2 n = 10	R^2 n = 15	R^2 n = 25	R^2 n = 50	R^2 n = 100
1: EOO	Spatial	None	4.0E+12	1.5E+13	4.0E+12	0.12**	0.20**	0.29**	0.38**	0.49**	0.52**
2: AOO 2KM	Spatial	None	1.2E+09	0	0	0	0	0	0	0	0
3: AOO 10PCT	Spatial	None	7.9E+12	2.2E+13	5.1E+12	0.08**	0.16**	0.27**	0.38**	0.55**	0.61**
4: EOO predicted	Ecological & spatial	None	7.1E+12	2.2E+13	4.7E+12	0.25**	0.35**	0.44**	0.54**	0.60**	0.63**
5: AOO predicted	Ecological & spatial	arcsin sqrt	0.23	1.03	0.07	0.31**	0.47**	0.65**	0.83**	0.91**	0.95**
6: Fraction PCA1	Ecological	arcsin sqrt	0.28	0.65	0.07	0.52**	0.69**	0.77**	0.82**	0.86**	0.86**
7: Maximum Euclidean distance	Ecological	arcsin sqrt	0.59	0.52	0.09	0.42**	0.55**	0.64**	0.70**	0.72**	0.73**
8: 2SD PCA1	Ecological	arcsin sqrt	0.19	0.43	0.04	0.53**	0.69**	0.77**	0.82**	0.86**	0.89**
9: Inverse kernel height	Ecological	arcsin sqrt	0.04	0.29	0.03	0.61**	0.73**	0.79**	0.85**	0.89**	0.92**
10: Area MCP PCA	Ecological	arcsin sqrt	0.06	0.40	0.04	0.53**	0.73**	0.80**	0.86**	0.89**	0.89**
11: Fraction MCP PCA	Ecological	arcsin sqrt	0.18	1.08	0.04	0.55**	0.79**	0.88**	0.94**	0.97**	0.98**

2.5 arc-minute											
Estimator of prevalence	**Type of data**	**Data trans-formation**	**Intercept n = 100**	**Slope n = 100**	**CI n= 100**	**R^2 n = 5**	**R^2 n = 10**	**R^2 n = 15**	**R^2 n = 25**	**R^2 n = 50**	**R^2 n = 100**
1: EOO	Spatial	None	4.1E+12	1.5E+13	4.0E+12	0.10**	0.23**	0.27**	0.38**	0.47**	0.52**
2: AOO 2KM	Spatial	None	1.2E+09	5.36	39.2	0	0	0	0	0	0
3: AOO 10PCT	Spatial	None	7.8E+12	2.2E+13	4.9E+12	0.06**	0.20**	0.23**	0.40**	0.54**	0.62**
4: EOO predicted	Ecological & spatial	None	7.6E+12	2.1E+13	4.7E+12	0.25**	0.32**	0.42**	0.53**	0.59**	0.61**
5: AOO predicted	Ecological & spatial	arcsin sqrt	0.23	1.03	0.07	0.34**	0.48**	0.65**	0.83**	0.92**	0.95**
6: Fraction PCA1	Ecological	arcsin sqrt	0.28	0.61	0.07	0.50**	0.68**	0.74**	0.79**	0.84**	0.85**
7: Maximum Euclidean distance	Ecological	arcsin sqrt	0.62	0.48	0.09	0.40**	0.54**	0.59**	0.65**	0.68**	0.70**
8: 2SD PCA1	Ecological	arcsin sqrt	0.19	0.40	0.04	0.49**	0.69**	0.76**	0.81**	0.85**	0.86**
9: Inverse kernel height	Ecological	arcsin sqrt	0.04	0.27	0.03	__0.57**__	0.74**	0.78**	0.84**	0.87**	0.90**
10: Area MCP PCA	Ecological	arcsin sqrt	0.06	0.37	0.04	0.50**	0.74**	0.78**	0.84**	0.87**	0.89**
11: Fraction MCP PCA	Ecological	arcsin sqrt	0.18	1.08	0.04	0.53**	__0.80**__	__0.86**__	__0.93**__	__0.97**__	__0.98**__

Chapter 4

Unequal contribution of widespread and narrow-ranged species to botanical diversity patterns

André S.J. van Proosdij[1,2], Niels Raes[2], Jan J. Wieringa[1,2],
Marc S.M. Sosef[1,3]

Published in PLoS One, 2016, 11 (12): e0169200

DOI: 10.1371/journal.pone.0169200

1 Biosystematics Group, Wageningen University, Wageningen, the Netherlands

2 Naturalis Biodiversity Center, Leiden, the Netherlands

3 Botanic Garden Meise, Meise, Belgium

Abstract

In conservation studies, solely widespread species are often used as indicators of diversity patterns, but narrow-ranged species can show different patterns. Here, we assess how well subsets of narrow-ranged, widespread or randomly selected plant species represent patterns of species richness and weighted endemism in Gabon, tropical Africa. Specifically, we assess the effect of using different definitions of widespread and narrow-ranged and of the information content of the subsets. Finally, we test if narrow-ranged species are overrepresented in species-rich areas. Based on distribution models of Gabonese plant species, we defined sequential subsets from narrow-ranged-to-widespread, widespread-to-narrow-ranged, and 100 randomly arranged species sequences using the range sizes of species in tropical Africa and within Gabon. Along these sequences, correlations between subsets and the total species richness and total weighted endemism patterns were computed. Random species subsets best represent the total species richness pattern, whereas subsets of narrow-ranged species best represent the total weighted endemism pattern. For species ordered according to their range sizes in tropical Africa, subsets of narrow-ranged species represented the total species richness pattern better than widespread species subsets did. However, the opposite was true when range sizes were truncated by the Gabonese national country borders. Correcting for the information content of the subset results in a skew of the sequential correlations, its direction depending on the range-size frequency distribution. Finally, we find a strong, positive, non-linear relation between weighted endemism and total species richness. Observed differences in the contribution of narrow-ranged, widespread and randomly selected species to species richness and weighted endemism patterns can be explained by the range-size frequency distribution and the use of different definitions of widespread or narrow-ranged. We call for a reconsideration of the use of widespread species as an indicator of diversity patterns, and advocate using the full ranges of species when assessing diversity patterns.

Keywords

Diversity patterns, Species Distribution Models, centres of endemism, biodiversity hotspots, Gabon, widespread, narrow-ranged, weighted endemism, species richness

Introduction

The current biodiversity crisis and limited availability of resources forces governments and NGOs to define conservation priorities (Margules & Pressey, 2000). Commonly, highly biodiverse regions (harbouring many species), centres of endemism (harbouring many narrow-ranged species), and crisis ecoregions (regions under threat of habitat conversion and climate change) are identified as priority areas for conservation (Brooks *et al.*, 2006, Pitman & Jorgensen, 2002, Sala *et al.*, 2000). Unfortunately, for many parts of the world, especially the tropics, little is known about the spatial distribution of most individual species or of the spatial distribution of diversity; a phenomenon known as the Wallacean shortfall (Lomolino, 2004). Most species are narrow-ranged, resulting in a right-skewed range-size frequency distribution (Magurran & Henderson, 2003, ter Steege *et al.*, 2013). Several studies have shown that species richness patterns based on narrow-ranged species differ from those based on widespread species and that most patterning in species richness is caused by a comparatively small subset of widespread species (Jetz & Rahbek, 2002, Kreft *et al.*, 2006, Lennon *et al.*, 2004, Mazaris *et al.*, 2008). Generally, the distribution of narrow-ranged species appears less correlated with climatic variables, but more strongly correlated with topographic and historical factors (Jetz & Rahbek, 2002, Kreft *et al.*, 2006). Therefore, using a subset of relatively common, widespread species as an indicator of species richness may well yield inappropriate conservation priorities for rare, narrow-ranged species.

Consideration of endemism has also been suggested as a replacement for assessment of total species richness in the context of identifying conservation priorities (Loyola *et al.*, 2007, Pitman & Jorgensen, 2002). Levels of endemism have been calculated in various ways including measures that weigh each species according to its rarity (Crisp *et al.*, 2001, Wieringa & Poorter, 2004). Several studies have shown a positive, non-linear relationship between the number of narrow-ranged species and the total number of species in an area, resulting in species-rich areas having a higher proportion of narrow-ranged species than average (Jetz *et al.*, 2004, Raes *et al.*, 2009). However, studies on vertebrates have shown that centres of endemism are not necessarily congruent with centres of species richness (Ceballos & Ehrlich, 2006, Grenyer *et al.*, 2006, Lamoreux *et al.*, 2006, Orme *et al.*, 2005, Villalobos *et al.*, 2013).

The contribution of each species to the pattern of species richness depends on the individual prevalence of species (Lennon *et al.*, 2011, Lennon *et al.*, 2004, Mazaris *et al.*, 2013), with prevalence defined as the fraction of the study area where the species occurs (McPherson *et al.*, 2004). A species present in 50% of the study area has the highest contribution to the richness pattern, whereas species present in 10% or 90% have an equally lower contribution of information to the pattern. This effect is known as the information content of a set of species and is defined as Σ(p*(1-p)) with p being the fraction of presence cells of each species (Lennon *et al.*, 2004). The difference between species richness patterns based on subsets of widespread and narrow-ranged species is only partly explained by differences in information content of these subsets (Kreft *et al.*, 2006, Lennon *et al.*, 2004, Mazaris *et al.*, 2008, Vázquez & Gaston, 2004). Often, when assessing richness patterns, the range sizes or prevalences are calculated for areas defined by political boundaries, thus not encompassing the full ranges of species. This logically leads to patterns only applicable at a local scale, though these may be important for political reasons. However, those interested in global diversity patterns need to take into account the full ranges of species (2006), which is what we aim for in our present study of Gabonese plant species.

For most species, preserved collections are not adequate reflections of species distribution patterns. By contrast, species Distribution Models (SDM) offer a solution as these predict the spatial distribution of species by linking a limited number of observations to environmental data with high spatial resolution (Franklin, 2009). Typically, the constantly growing body of digitized presence-only specimen data from natural history collections are used as observations (Graham *et al.*, 2004). Diversity patterns can be inferred by stacking SDMs that are converted into binary presence/absence maps (Calabrese *et al.*, 2014, Raes *et al.*, 2009). This method offers unique opportunities to assess congruence between diversity patterns based on different subsets of species.

Here, using SDMs of plant species from Gabon, central Africa, we infer patterns of species richness and weighted endemism for Gabon. More specifically, we address the following questions: 1) Do diversity patterns based on subsets of narrow-ranged or widespread plant species differ from those based on random subsets? 2) Are these differences still apparent when corrected for the information content of each subset? 3) Are these differences sensitive to the extent of the study area in which the range sizes are defined, here Gabon versus tropical Africa as a whole? 4) Are narrow-ranged species overrepresented in species-rich areas?

Materials and Methods

Study area

We selected Gabon to serve as a case study. Gabon is a highly biodiverse country in the Lower Guinean phytogeographical region (Kier *et al.*, 2005, White, 1979) with around 80% of its 267,667 km^2 covered by lowland rain forest and the remaining 20% mainly by savannahs and urban areas

(S1 Fig). It hosts an estimated number of 7000-7500 vascular plants species (Sosef *et al.*, 2006), of which 5323 have been recorded so far. Of these, 13% are endemic or near endemic to Gabon and many more are native only in the Lower Guinean biogeographic region (Sosef *et al.*, 2006), showing the importance of the contribution of narrow-ranged species to diversity patterns. In contrast to most other species-rich, tropical African countries, the botanical diversity of Gabon is well-documented with > 95% of the known herbarium collections digitally available through the Naturalis Biodiversity Center database. This renders Gabon an excellent study area to address the research questions formulated above. We defined our African study area from 15ºN to 19ºS and from 17.5ºW to 43ºE, encompassing the known range of the majority of Gabonese plant species and covering 180,399 raster cells at 5 arc-minute spatial resolution (excluding oceans and large water bodies).

Species distribution data

To avoid the exclusion of species known to occur in neighbouring countries and possibly also to be found in Gabon, but not yet collected there, we selected all plant species recorded at least once from Gabon including a buffer area of five degrees (approx. 600 km). Species known to only occur in cultivation in Gabon were excluded. Subspecific taxa were combined in the germane species. From the species list so compiled, we used all available herbarium specimen data from Gabon and other tropical African countries to avoid modelling truncated niches of species (Raes, 2012) and to make use of all available data for model training (van Proosdij *et al.*, 2016b). Records comprising doubtful identifications as well as duplicate records from the same raster cell were excluded. Only records with latitude/longitude data accurate to at least five arc-minute spatial resolution were used.

Environmental data and two model training areas

We used WorldClim temperature data (Hijmans *et al.*, 2005), CHIRPS precipitation data (Deblauwe *et al.*, 2016, Funk *et al.*, 2014), and quantitative soil data from the Harmonized World Soil Database (FAO/IIASA/ISRIC/ISSCAS/JRC, 2012). Environmental data layers were cropped to the extent of the study area (hereafter ‘African training area’) and, where necessary, aggregated to five arc-minute spatial resolution. As a measure of topographic heterogeneity we used the standard deviation of altitude based on the 90 m SRTM altitude data (<srtm.csi.cgiar.org>) within each five arc-minute raster cell. Out of the 39 original variables we selected those correlated with Spearman’s |*rho*| < 0.7 (Dormann *et al.*, 2013), to avoid overfitting of models due to multi-collinearity, resulting in 15 selected variables (S2 Table & S6 File).

We adjusted the extent of the training area of species with a prevalence < 0.1 or > 0.9 to avoid statistical artefacts in modelling these species (McPherson *et al.*, 2004). The prevalence of species was estimated by using the fraction of raster cells where the species was predicted as present in tropical Africa based on a thresholded SDM (Syfert *et al.*, 2014). For species with a predicted prevalence < 0.1 in the African training area, we used the smaller training area of Gabon including a buffer area of five degrees (hereafter 'Gabonese training area') resulting in 18,144 5-arc minute raster cells and using the same selected environmental variables. No species had a prevalence > 0.9.

Model building

SDMs were generated using MaxEnt (Phillips *et al.*, 2006), which has shown to outperform other methods when using presence-only data like ours, even when applied to small data sets (Elith *et al.*, 2006). We modified the MaxEnt default settings by allowing only linear and quadratic features for all sample sizes, and excluding hinge, product and threshold features to prevent over-parameterization of the models (Merow *et al.*, 2013). To compensate for a potential collecting bias in our specimen data, possibly resulting in an ecological bias (Loiselle *et al.*, 2008, Reddy & Davalos, 2003), we applied the same bias to the background data used to train the models by means of target background sampling (Phillips *et al.*, 2009). Consequently, pseudo-absences were selected from raster cells with at least one herbarium record. The logistic MaxEnt output for each species was converted into a binary presence/absence map by applying the 'ten percentile training presence' threshold. This threshold forces 10% of the training records to fall outside the predicted suitable area, which is thought to allow for 10% of the records to contain identification, georeferencing or other errors without serious consequences for the model (Liu *et al.*, 2013, Merow *et al.*, 2013). A Multivariate Environmental Similarity Surface analysis (Elith *et al.*, 2010) showed considerable areas with negative MESS values for models trained on the Gabonese training area (S3 Fig), which is why SDMs trained on the Gabonese training area were projected on the larger tropical African area without extrapolation to environmental conditions not present in the smaller training area.

Model evaluation

Models were evaluated using two criteria. First, each model was tested against a bias-corrected null model following Raes & ter Steege (2007) and accepted if its AUC value ranked > 95 when grouped with the 99 null model AUC values. This implies that the model performed significantly better than random expectation ($p < 0.05$). Second, from the significant

SDMs, a model was accepted when the number of unique training records equalled or exceeded the minimum number of records required to generate models significantly better than random expectation. This minimum number of records increases with increasing prevalence of the species (van Proosdij *et al.*, 2016b). Following the procedure of van Proosdij *et al.* (van Proosdij *et al.*, 2016b), we identified the following required minimum numbers of records for species of different prevalence classes for the models trained on the African training area and between brackets the minimum numbers for the Gabonese training area: 7 (5) for prevalence < 0.1, 7 (8) for prevalence 0.1-0.2, 9 (10) for prevalence 0.2-0.3, 12 (11) for prevalence 0.3-0.4, 12 (14) for prevalence 0.4-0.5, and 15 (17) for prevalence > 0.5.

Patterns of species richness and weighted endemism

Three types of diversity patterns were computed by stacking the selected thresholded SDMs. Firstly, total species richness was computed by summing the number of species predicted to be present in each raster cell. Secondly, weighted endemism was computed following Crisp *et al.* (2001) and Wieringa & Poorter (2004) by summing up the rarity values of the species present in a unit or raster cell, with rarity value defined as the inverse of the number of presence cells. Finally, residuals of weighted endemism were defined as the weighted endemism relative to the species richness of the raster cell (Raes *et al.*, 2009), also termed corrected weighted endemism (Linder, 2001) (hereafter called 'residual weighted endemism'). We computed the residual weighted endemism values by first fitting a curve to the values of weighted endemism plotted against total species richness. Akaike Information Criterion was used to select the best polynomial curve. Then, relative residuals were computed by taking for each cell the difference between the actual weighted endemism value and the fitted value, relative to the fitted value. The resulting three diversity patterns were cropped to the national borders of Gabon.

Species sequences and correlation with species richness and weighted endemism

Species with accepted SDMs were ranked according to their predicted prevalence in tropical Africa. We generated one narrow-ranged to widespread sequence, one widespread to narrow-ranged sequence, and 100 random sequences (Evans *et al.*, 2005, Lennon *et al.*, 2004, Mazaris *et al.*, 2013). This procedure was repeated by ranking the species according to their prevalence within Gabon. For subsets of *n* species, with increasing values of *n* along the sequences, species richness maps ('n_richness') and weighted endemism maps ('n_weighted_endemism') were generated. Along the sequences we computed the Pearson correlation of n_richness with the total species richness pattern and of n_weighted_endemism with the total weighted endemism pattern, all cropped to the national borders of Gabon. Resulting Pearson's *r* values of the subsets along the sequences were plotted against the number of species as well as against the information content of the subsets. The information content of a subset was computed by summing the information contents of the species in the subset. All analysis were performed in *R*, using functions provided in the *R* script available as Supporting Information (S7 File).

Results

In total, our dataset contains 5323 species from Gabon and an additional 3361 from the five degrees buffer zone. A total of 317,582 herbarium specimen records related to these 8684 species were aggregated in our dataset and used for model-building. 3572 species did not have sufficient records to model a reliable SDM; for another 2628, their SDMs did not pass the null model test, while 395 of the remaining SDMs predicted the species to be absent from Gabon (although present in the buffer zone). In total, SDMs of 2089 species were used for further analyses including one liverwort species, 22 moss species, eight clubmoss species, 63 fern species, one gymnosperm species and 1994 angiosperm species (S4 Table, SDMs available from the Dryad Digital Repository: http://dx.doi.org/10.5061/dryad.v4f53). When trained on tropical Africa, SDMs of 1306 species resulted in a predicted prevalence < 0.1, and hence their SDMs were rerun using the smaller Gabonese training area. Of these new SDMs, 624 also had a predicted prevalence of < 0.1 in the Gabonese training area, which we regard as acceptable given the scope of this study. The range size frequency distribution based on the predicted prevalences for both tropical Africa and Gabon is strongly right-skewed towards narrow-ranged species (Fig 1). It is to be noted that for range sizes based on tropical Africa, the apparent peak at a prevalence of 0.10-0.15 is actually caused by the exclusion of many species with a prevalence < 0.10. These excluded species are recorded from the five degrees buffer zone but are predicted to be absent from Gabon or have too few records inside the Gabonese training area to generate a significant SDM (S5 Fig).

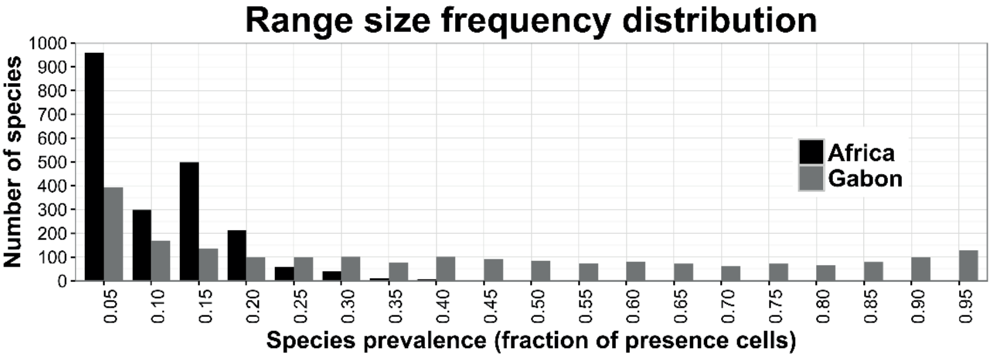


Fig 1. Range size frequency distribution of Gabonese plant species. *The range size frequency distribution is shown for the Gabonese plant species with accepted SDMs. Range size or prevalence is defined as the fraction of raster cells where the species is predicted present in tropical Africa (black) and Gabon (grey) respectively.*

The highest species richness in Gabon is predicted for north-western Gabon (foothills of the Crystal Mountains and the vicinity of Libreville), as well as hills in central and western Gabon (Doudou Mountains and western parts of Chaillu Massif) (Fig 2A). Areas with high values of weighted endemism are largely congruent with centres of species richness with maximum values in the Crystal Mountains and the vicinity of Libreville (Fig 2B). Species richness and weighted endemism show a strong positive, non-linear relation, best represented by a fourth-order polynomial function (Fig 2C, $Y = 2.583\text{e-}04*X - 6.645\text{e-}07*X^2 + 6.379\text{e-}10*X^3 - 8.581\text{e-}14*X^4$, adjusted $R^2 = 0.92$, $p < 0.001$). Figure 2D shows high positive values of residual weighted endemism in the two aforementioned centres of endemism and in the coastal region south of one degree south latitude, meaning that in those areas more narrow-ranged species are present than would be expected from the species richness.

Along each of the sequences based on the prevalence of species in tropical Africa, correlation values of n_richness patterns to total richness pattern increase, but they do so more rapidly for the narrow-ranged-to-widespread sequence (Fig 3A, Kolmogorov-Smirnov test: D = 0.37, $p < 0.001$). A correlation of $r = 0.7$ is achieved with the 5% most narrow-ranged species, versus the 35% most widespread species. However, these subsets are both outperformed by random subsets. When corrected for the information content of the subsets, the narrow-ranged-to-widespread sequence performs as well as the random sequences, while the performance of the widespread-to-narrow-ranged sequence decreases (Fig 3B).

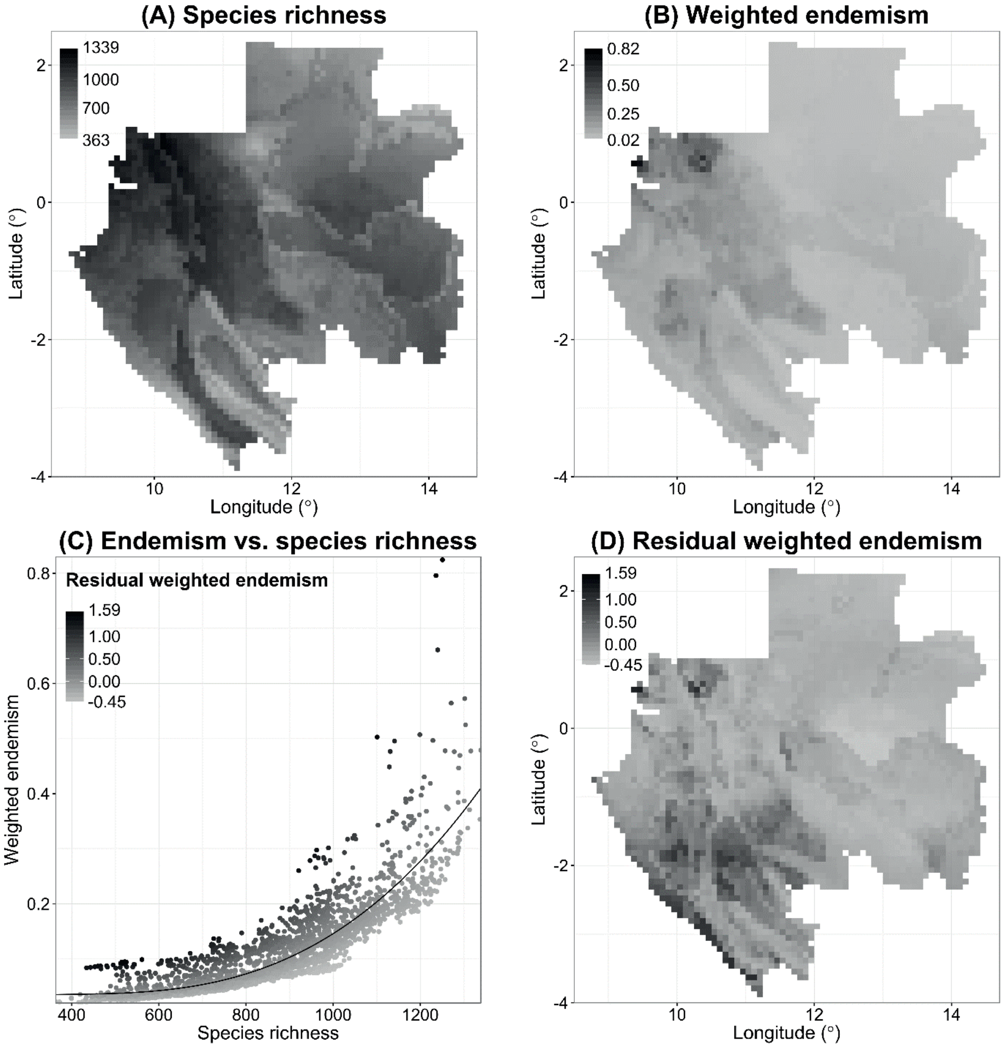


Fig 2. Botanical diversity patterns for Gabon based on 2089 species. *The following diversity patterns are shown based on thresholded SDMs of 2089 Gabonese plant species: (A) total species richness; (B) weighted endemism; (C) weighted endemism (y-axis) plotted against total species richness (x-axis) with shades of grey indicating values of residual weighted endemism and the black curve representing a fourth-order polynomial function; (D) residual weighted endemism.*

By contrast, along each of the sequences based on the prevalence within Gabon, correlation values increase more rapidly for the widespread-to-narrow-ranged sequence (Fig 3E, Kolmogorov-Smirnov test: D = 0.18, p < 0.001). Here, a correlation of 0.7 is achieved with the 20% most widespread species versus the 35% most narrow-ranged ones. Here too, both are outperformed by random subsets. Correcting for the information content of the subsets results in narrow-ranged species slightly outperforming widespread ones (Fig 3F).

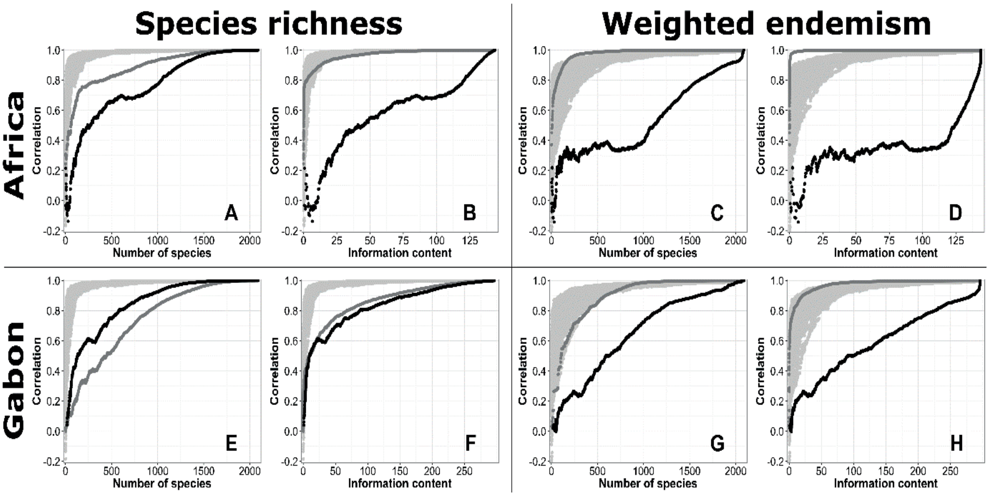


Fig 3. Correlations between subset and total diversity patterns. *Correlations are presented between species richness patterns based on subsets of n Gabonese plant species (n_richness) and total Gabonese species richness (A,B,E,F), as well as between weighted endemism patterns based on subsets (n_weighted_endemism) and total weighted endemism (C,D,G,H). Subsets were composed along the narrow-ranged to widespread sequence (dark grey lines), widespread to narrow-ranged sequence (black lines), and 100 random sequences (light grey lines). Defining the species sequences was done on the prevalences of species in either tropical Africa (A-D), or Gabon (E-H). Correlations are plotted against the number of species (A,C,E,G) or the information content of the subset (B,D,F,H).*

Using sequences based on the prevalence in tropical Africa, we found, as was to be expected, patterns of n_weighted_endemism based on narrow-ranged species to be more strongly correlated with the total weighted endemism pattern than were patterns based on widespread species (Fig 3C, Kolmogorov-Smirnov test: D = 0.90, p < 0.001), even outperforming random subsets. When corrected for the information content, narrow-ranged species remain more strongly correlated with weighted endemism, whereas the correlation of widespread species to weighted endemism decreases (Fig 3D). For the sequences based on prevalence within Gabon, patterns of n_weighted_endemism based on narrow-ranged species are more strongly correlated with the total weighted endemism pattern than are patterns based on widespread species (Fig 3G, Kolmogorov-Smirnov test: D = 0.54, $p < 0.001$), but are outperformed by random subsets when these contain less than 25% of the species. Correcting for the information content of the subsets results in subsets of narrow-ranged species showing the strongest correlation with the total weighted endemism pattern (Fig 3H).

Discussion

Diversity patterns in Gabon

The inferred pattern of plant species richness with centres of diversity in the vicinity of Libreville as well as the mountains of central and western Gabon confirms previous findings based on legumes (de la Estrella *et al.*, 2012) or endemic species (Walters *et al.*, 2016). These centres of species richness and of weighted endemism coincide with the hypothesised Last Glacial Maxima forest refugia in the Crystal Mountains, western parts of the Chaillu Massif and the Doudou Mountains (Hardy *et al.*, 2013, Maley, 1996). The high levels of residual weighted endemism in the coastal region south of one degree south latitude illustrates the uniqueness of this relatively species-poor area that is floristically not related to other parts of central Africa and contains a comparatively high number of endemic species (Harris *et al.*, 2012, Wieringa & Sosef, 2011).

Widespread versus narrow-ranged

Our results confirm that richness patterns based on narrow-ranged species differ from those based on widespread species (Jetz *et al.*, 2004, Lennon *et al.*, 2011, Villalobos *et al.*, 2013). However, the correlation of each of these patterns with the total species richness pattern depends on the extent of the study area used to define the prevalence of species. When prevalence was defined for tropical Africa, we found patterns of narrow-ranged species in Gabon to be more strongly correlated with the pattern of total species richness. This contradicts the results of previous studies which found patterns of widespread species being more strongly correlated with total species richness patterns (Jetz & Rahbek, 2002, Kreft *et al.*, 2006, Lennon *et al.*, 2004, Mazaris *et al.*, 2013, Perez-Quesada & Brazeiro, 2013, Vázquez & Gaston, 2004). In addition to the unique suite of species and habitats in each study area, four other matters need further consideration so as to put our results into perspective.

Firstly, the range size frequency distribution of the species influences the sequential correlations and depends on the study area and species group.

Our data set is strongly right-skewed and thus similar to the dataset of Uruguayan plants used by Perez-Quesada & Brazeiro (2013), whose results are in line with ours. By contrast, Kreft *et al.* (2006) found patterns of widespread species that were more strongly correlated with the total species richness pattern using a Neotropical palm data set with an approximately normally distributed range size frequency. The work of Lennon *et al.* (2004) then, on birds from Scotland, the united Kingdom as a whole, and South Africa, presents results similar to those of Kreft *et al*. for Scottish and British birds, but contrasting results for South African birds. The sequential correlation of their South African bird data set plotted against the information content of the subsets is higher for narrow-ranged species than for widespread species. From their three data sets, the South African birds data set is the most strongly right-skewed (Lennon *et al.*, 2004). Based on these and our results from different study areas and different species groups, we conclude that strongly right-skewed range size frequency distributions result in stronger correlations between narrow-ranged species subsets and the total species richness pattern.

The second matter is the range size or prevalence criterion that is applied to define the species sequences, a matter to which little attention has been paid up to now. Most studies order species based on their prevalence in the study area alone, which can be much smaller than the full range size of the species (Evans *et al.*, 2005, Lennon *et al.*, 2011, Lennon *et al.*, 2004, Mazaris *et al.*, 2013, Perez-Quesada & Brazeiro, 2013, Vázquez & Gaston, 2004), with few positive exceptions (Kreft *et al.*, 2006). For example, widespread African species are sometimes rare in Gabon and Gabonese endemics sometimes have a large prevalence within the country. We assessed both of these by ordering species based on their prevalence in both tropical Africa and in Gabon and found contradicting results. We conclude that for a correct comparison of aspects of narrow-ranged and widespread species, species should be ordered according to their entire range size.

The third matter to consider is that 6595 species (76%) were excluded from our analysis as their models did not meet the criteria of model accuracy, or the species were recorded only from the five degrees buffer zone but not predicted to be present in Gabon. Little can be said with confidence on the overall distribution of these excluded species, but since 3572 were excluded because of insufficient records, we expect the majority to be narrow-ranged. In general, we expect that if these apparently rare species could be included in the analysis, this would result in an even larger difference between diversity patterns based on narrow-ranged species versus those based on widespread species.

Thirdly, our results are based on the use of SDMs, which usually do not take into account biotic interactions, historical constraints, and dispersal limitations (Araújo & Peterson, 2012). Therefore, the actual prevalence of species limited in their distribution by such factors, may well be (much) smaller than predicted here, resulting in an even more skewed range size frequency distribution. Ignoring dispersal limitations might also affect the calculated species composition of ecologically isolated areas.

Random subsets

We found species richness patterns based on random subsets of species to be more strongly correlated with the total species richness pattern than were patterns based on either narrow-ranged or widespread species alone. Some studies report a stronger correlation with the total species richness pattern for widespread species subsets over random subsets (Kreft *et al.*, 2006, Mazaris *et al.*, 2013), but others show contradictory results (Lennon *et al.*, 2011). With respect to the correlation of subsets with weighted endemism, we found, as expected, random subsets being outperformed by those of narrow-ranged species when species are ordered according to their full range size. However, when ordered on prevalence within Gabon, again, random species subsets better represent the total weighted endemism pattern. Comparing the sequential correlation curves of our study with those reported by others cited above, we see strong similarities between the curves of random species subsets and large differences between the sequential correlation curves of widespread and of narrow-ranged species subsets. These differences can be explained by the matters addressed above: the range size frequency distribution of the assessed species and the applied criterion to define species sequences from narrow-ranged to widespread and vice versa.

Information content

Correcting for the information content of the subsets influences the sequential correlation curves. The magnitude of this correction depends on the information content of the species included in the subset with species present in 50% of the study area (prevalence = 0.5) contributing the largest amount of information (Lennon *et al.*, 2004). Here, the prevalence in Africa of all but a few species is < 0.5 and hence the group of species with the largest information content consists of those species with the largest prevalence (prev. 0.3-0.5). Correcting for the information content in our study resulted in a skew to the right for the narrow-ranged-to-widespread sequence and a skew to the left for the widespread-to-narrow-ranged sequence (Figs 3B and D). The skew is less strong for the curves based on Gabonese prevalences, which contain many species with a prevalence value > 0.5. The sequential correlation curves of random subsets did not change when corrected for information content. The differences in skew found by us and by others (Kreft *et al.*, 2006, Lennon *et al.*, 2004) can be explained by the specific range size frequency distributions.

Overrepresentation of narrow-ranged species in species-rich areas

In Gabon, narrow-ranged plant species are overrepresented in species-rich areas resulting in a strong, positive, non-linear relation. Therefore, estimating total plant species richness in Gabon based on the number of widespread plant species in an area will result in an underestimate of species richness in Gabon's centres of diversity. Our results thus confirm similar findings for African birds (Jetz *et al.*, 2004), North American vertebrates and invertebrates (Rickets, 2001), and vascular plants from the United Kingdom (Lennon *et al.*, 2011) and Borneo (Raes *et al.*, 2009). By contrast, other studies have shown that centres of species richness and centres of endemism are not congruent (Ceballos & Ehrlich, 2006, Lamoreux *et al.*, 2006, Orme *et al.*, 2005) or only partially so (Villalobos *et al.*, 2013). These seemingly contradictory results underline the difficulties of identifying universal estimators for patterns of species richness and weighted endemism, but can be explained by some factors that are often ignored, including differences in the suite of species and habitat types present in the study areas, differences in applied spatial resolution and differences in the extent of the study areas (Kreft *et al.*, 2006, Rahbek, 2005, Rahbek & Graves, 2001). In addition, concordance of the species richness pattern and endemism pattern is low when only the few most species-rich and most rare–species-rich-cells are compared, but is high when correlation is computed over all cells (Rickets, 2001). Others have found a small overlap between the most species-rich cells with those containing the most rare species (Ceballos & Ehrlich, 2006, Orme *et al.*, 2005), as well as a weak, or no, correlation between patterns of total species richness and endemism when this is computed over all cells (Lamoreux *et al.*, 2006, Orme *et al.*, 2005). Furthermore, congruence is higher when endemism is defined as weighted endemism including all species (Rickets, 2001), as we report here.

Implications for conservation

Setting priorities in conservation is topical, especially for the Tropics, that harbour by far the most species, but face the highest extinction risks (Vamosi & Vamosi, 2008). If one aims to identify the most species-rich areas using small subsets of species, random subsets of species best represent these areas given that the range size or prevalence of the targeted species is defined over their entire range. However, if one aims to identify areas containing the highest endemicity values, and applying the same range size criterion, subsets of narrow-ranged species are to be preferred. Both criteria may ignore areas with high values of residual weighted endemism, thus harbouring only few species but a disproportionally high number of species not present elsewhere, as we have demonstrated here for the coastal zone of Gabon and has also been shown for other areas, including e.g. Borneo (Raes *et al.*, 2009). These areas deserve priority for conservation too, as they contain disproportionally many species not present elsewhere.

Conclusions

For Gabon we have shown that patterns of plant species richness based on subsets of narrow-ranged species differ substantially from those based on subsets of widespread species. If species are ordered according to their full range size, subsets of narrow-ranged species represent the total species richness pattern better, but both are outperformed by random subsets. However, if ordered on range sizes truncated by the country borders of Gabon, subsets of narrow-ranged species are outperformed by subsets of widespread species. This difference in the ordering of species from narrow-ranged to widespread, in concert with the unique range size frequency distribution, suite of species and habitats present in a study area, influences the correlation of subsets of species with patterns of total species richness and weighted endemism. Correcting for the unequal information content of subsets of narrow-ranged and widespread species influences the sequential correlation with diversity patterns, and the exact effect of this correction depends on the range size frequency distribution of the species.

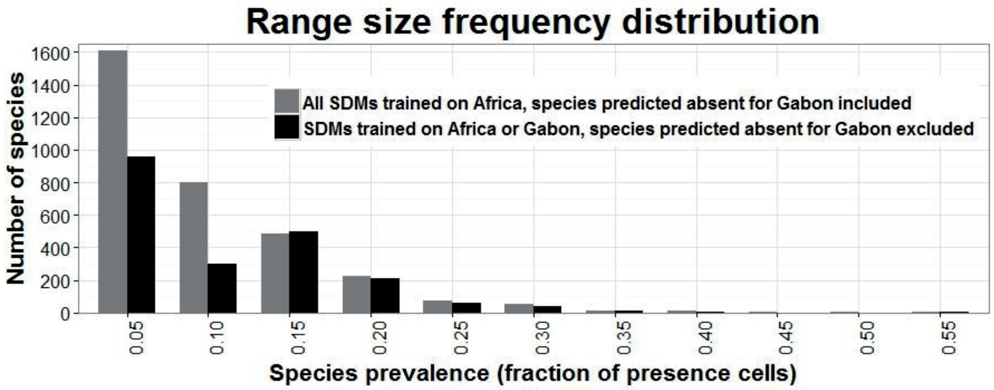
In Gabon, narrow-ranged plant species are overrepresented in species-rich areas. Omitting narrow-ranged species from diversity assessments will result in an underestimate of species richness in species-rich areas. In addition, some centres of residual weighted endemism contain few species in total but a disproportionally high number of narrow-ranged species and hence can be overlooked too when narrow-ranged species are omitted from diversity assessments. We call for a reconsideration of the use of richness patterns based on a selection of widespread species as a measure of total species richness, as this is not universally applicable to all taxonomic groups or study areas. Secondly, we argue for an analysis of the range size frequency distribution of the species and always to use the full ranges of species when assessing diversity patterns and correlations with possible explanatory environmental variables.

Acknowledgements

The "Gabon database" is compiled with specimen data from BM, BR, BRLU, LBV, MO, P & WAG. We thank collectors and curators of the herbaria B, BNRH, BR, BRLU, E, FR, G, K, MA, MO, P, PRE, UPS for their generous exchange of herbarium specimen data with NHN-Naturalis during duplicate exchange of African collections. MA and MO sent additional African specimen data specifically for this project, which is gratefully acknowledged. David Mabberley is thanked for proofreading the manuscript and improving the English language. The subject editor and an anonymous reviewer are thanked for their valuable comments.

Supplementary material

All supplementary materials (appendices 1-7) can be found in the published version of this article online: DOI: 10.1371/journal.pone.0169200



S5 Fig. Comparison of range size frequency distributions. *The range size frequency distribution of the species with accepted SDMs is shown, with range size or prevalence of the species defined as the fraction of raster cells where the species is predicted to be present in tropical Africa. In black the original RSFD values based on Species Distribution Models trained on either tropical Africa or Gabon and including only species with accepted SDMs that are predicted to be present in Gabon (same as in main text Figure 1). In grey the RSFD values based on SDMs which are all trained on tropical Africa and including all species with accepted SDMs, thus including those species recorded from the five degree buffer zone but predicted to be absent for Gabon itself.*

Chapter 5

Climate change is predicted to cause major turnover in plant species composition in Gabon

André S.J. van Proosdij[1,2], Jan J. Wieringa[1,2], Marc S.M. Sosef[1,3], Niels Raes[2]

1 Biosystematics Group, Wageningen University, Droevendaalsesteeg 1, 6708 PB Wageningen, the Netherlands

2 Naturalis Biodiversity Center, Postbus 9517, 2300 RA, Leiden, the Netherlands

3 Botanic Garden Meise, Nieuwelaan 38, 1860 Meise, Belgium

Abstract

To set conservation priorities, knowing the effect of predicted climate change on patterns of species richness and the expected level of extinction is important. Here we analyse these effects for Gabonese plant species and quantify the explanatory power of individual climate anomalies on species gain, loss and turnover, and quantify the additional effect of dispersal limitations. For 2,137 species of Gabonese plants we generated species distribution models (SDMs) using georeferenced herbarium records and high spatial resolution soil and climate data. SDMs are projected to 2085 under two representative concentration pathways (RCPs) assuming either full or no dispersal. Patterns of future species richness, gain, loss and turnover are generated and correlations with climate anomalies are computed. In Gabon, predicted loss of plant species varies between just over 5% for 2085 under RCP 4.5 assuming full dispersal and almost 10% by 2085 under RCP 8.5 assuming no dispersal. As for many rare, narrow-ranged species no significant SDM could be generated, species losses are likely to be even higher than computed here. Species loss is best explained by increased precipitation in the dry season, whereas species gain and turnover are correlated with a shift from extreme to average values of annual temperature range. Whereas other regions with tropical rainforests are facing warmer and drier conditions, Gabon is already experiencing warmer and wetter conditions. Our models predict that this situation is driving a nation-wide loss of species. At the same time, average species-richness per area is predicted to increase, but only if species are able to migrate timely into future suitable habitats. Therefore, dispersal limitations are posing an additional threat to the survival of plant species in Gabon. We advocate for the identification and protection of potential refugia with particular attention for microrefugia inside as well as outside the current network of protected areas.

Keywords

Diversity patterns, species richness, climate change, species distribution models, dispersal, microrefugia, Africa, rainforest, extinction, precipitation

Introduction

Worldwide, biodiversity is under increasing pressure of climate change, acting on all levels from population genetic to biome scale (Scheffers *et al.*, 2016). Climate change alters the ranges, growth and abundance of individual species (Lenoir & Svenning, 2015, Parmesan & Yohe, 2003), leading to changes in patterns of species richness and local or global extinctions, although locally species richness may increase (Brown *et al.*, 2015, Fitzpatrick *et al.*, 2008). Setting priorities for conservation requires knowledge on which species are likely to be affected by climate change and where changes in species richness are predicted to occur. Global climate change scenarios are defined by the Intergovernmental Panel on Climate Change (IPCC) by Representative Concentration Pathways (RCP 2.6, 4.5, 6.0, and 8.5), which represent scenarios ranging from a ‘mitigating climate change’ scenario to a ‘business as usual’ scenario (IPCC, 2013). The RCP 4.5 and RCP 8.5 scenarios project atmospheric CO_2 concentrations rising from the pre-industrial level of 280 ppm to 650 and 1,370 ppm by 2100 respectively, (Moss *et al.*, 2010), resulting in an increase in global mean annual temperature of 2.4 °C and 4.9 °C above pre-industrial levels (Rogelj *et al.*, 2012).

Such changes are likely to affect the patterns of species richness as well as total species richness in central African lowland rainforests. We here focus on Gabon, which is part of the Lower Guinean phytogeographic region (White, 1979), which is botanically one of the most species-rich regions in tropical Africa, especially when its relative small surface area is considered (Sayer *et al.*, 1992, Sosef *et al.*, 2017, Sosef *et al.*, 2006). Approximately 80% of the country’s land surface is covered by primary or slightly disturbed tropical lowland rainforest. During past climate change events, Gabon has experienced major changes in vegetation coverage and species composition, with large regional differences. During Pleistocene glacial maxima, the rainforest contracted to forest refugia located in montane regions of Gabon (Fig. 1a) followed by expansions during interglacials. On top of these major events, shorter climatic changes invoked rainforest expansions like during the African Humid Period (11,000 – 8,000 years BP), or contractions during drier conditions like that between approx. 4,000 and 2,000 years BP (Maley, 1996, Willis *et al.*, 2013). In contrast to many other tropical lowland rainforests, including large parts of the Amazon (Malhi *et al.*, 2008, Zelazowski *et al.*, 2011, Zhang *et al.*, 2015) and the Congo basin (Zhou *et al.*, 2014) that are predicted to face drier future conditions, climatic conditions for Gabon are predicted to become wetter while periods of aridity will become shorter (Platts *et al.*, 2015). However, what this implies in terms of the effects on botanical diversity in Gabon is still unknown.

Several studies have estimated the level of species extinction as result of climate change, e.g. 0-15% by 2070 for New Caledonian tree species (Pouteau & Birnbaum, 2016), 5-25% by 2080 for *Banksia* species in Southwest Australia (Fitzpatrick *et al.*, 2008), 15-37% by 2050 for endemic animal and plant species in a variety of biotopes worldwide (Thomas *et al.*, 2004), and 39-43% of endemic plant and vertebrate species in biodiversity hotspots worldwide under the worst-case climate change scenarios (Malcolm *et al.*, 2006). The effects of climate change on species' range sizes is not uniform and differs between ecosystems and taxonomic groups. The degree of range shifts can largely be explained by the level of habitat specialisation of the individual species and the proximity and range size of future analogue biotopes (Bellard *et al.*, 2012, Pouteau & Birnbaum, 2016). Species that have specific habitat preferences, a restricted range or limited dispersal capacity are particularly prone to the effects of climate change as these are not able to adapt to novel climates, or migrate timely to suitable areas with analogue future climates (Malcolm *et al.*, 2006).

Few studies identified the climate anomalies that drive predicted species loss, gain and turnover to understand the effect of climate change on biodiversity. Feeley *et al.* (2012) showed that the ranges of Amazonian plant species will decrease if they are not able to tolerate or adapt to the expected drier and warmer conditions. Zhang *et al.* (2014) found predicted species losses in Yunnan, China to be associated with increased temperature variability and decreased dry season precipitation. To our knowledge, no studies addressed the quantitative importance of individual future climate anomalies to predicted changes in biodiversity.

The effect of regional differences in climate anomalies on the distribution of species and patterns of species richness are usually ignored, but these can have great impact (Lenoir & Svenning, 2015, VanDerWal *et al.*, 2013). Recently, remotely sensed temperature and precipitation data at high spatial resolution have become available (Platts *et al.*, 2015). These remotely sensed data have shown to outperform model-based interpolations of weather station data in macroecological studies as they result in better fit and transferability of species distribution models (Deblauwe *et al.*, 2016). Remotely sensed current climatic data and regionally downscaled predictions for future climate scenarios calibrated on these remotely sensed current data better capture regional differences in climatic conditions and therefore enable assessments of the regional effects of climate change on the distribution of individual species and patterns of species richness.

The spatial distribution of the vast majority of species is still unknown. Species Distribution Models (SDMs) can be used to overcome this lack of data as they predict a species' distribution by correlating known occurrences to environmental variables at high spatial resolution (Elith &

Leathwick, 2009). Commonly, bioclimatic variables are used that represent annual, seasonal and extreme or limiting environmental factors based on temperature and precipitation values. SDMs are commonly applied to assess the effects of different climate change scenarios on the distribution of species (Guisan & Thuiller, 2005). The capacity of species to respond to changing climatic conditions by shifting their distributions strongly depends on their dispersal capacities. Recently developed *R* packages enable the integration of dispersal constraints in SDMs (Engler *et al.*, 2012, Nobis & Normand, 2014). Unfortunately, the dispersal capacity of many species is unknown, thus limiting the use of SDMs to predict future distributions (Corlett & Westcott, 2013). Consequently, most studies model future distributions assuming either unlimited or no dispersal (Bateman *et al.*, 2013), but see Fitzpatrick *et al.* (2008) and Hsu *et al.* (2012) for two examples where estimates of dispersal capacities were applied.

Here, we use SDMs of Gabonese plant species to assess the regional fingerprint of predicted climate change on the total number of species and patterns of species richness in Gabon for 2085 under two representative concentration pathways and assuming no and full dispersal. We quantify the correlations of individual future climate anomalies with patterns of species gain, loss and turnover. Specifically, we address the following research questions: 1) What is the effect of predicted climate change on the total number of plant species in Gabon? 2) What is the effect of predicted climate change on patterns of species richness in Gabon in terms of species gain, loss and turnover? 3) What is the contribution of individual bioclimatic anomalies in explaining patterns of future species gain, loss and turnover? and 4) What is the potential effect of dispersal limitations on patterns of future species richness in Gabon?

Methods

Study area and species data

Our study area ranged from 15ºN to 19ºS and from 17.5ºW to 43ºE. This covers 180,294 raster cells at a 5 arc-minute spatial resolution (excluding large water bodies) and encompasses the documented records of most Gabonese plant species. Gabon hosts an estimated number of 6,100-7,500 vascular plant species (Sosef *et al.*, 2017, Sosef *et al.*, 2006), of which 5,323 have been documented to date. The collections database of Naturalis Biodiversity Center contains data of > 90% of all herbarium specimens ever made in Gabon. We selected all species recorded for Gabon plus a 5 degrees buffer zone (approx. 600 km). We added the buffer to avoid omitting species known from areas close to Gabon and possibly also occurring in Gabon, although not yet collected inside Gabon, or species that at present do not occur in Gabon but may arrive in the future as result of climate change. Data was taken at species level and species only known from cultivation in Gabon were excluded. From this list of species we used all available herbarium specimen data present at Naturalis Biodiversity Center and 13 additional institutes (herbarium acronyms listed in Acknowledgements) with confirmed identification at species level and latitude/longitude data accurate to at least 5 arc-minute spatial resolution, resulting in 321,275 presence records. The inclusion of data from outside Gabon concerning species present in Gabon increases the number of records to train the models (van Proosdij *et al.*, 2016b) and reduces the risk of modelling truncated species' niches (Raes, 2012). Duplicate records from the same raster cell were removed. For records from locations with no climatic data, but within ten kilometres of cells with data, we used the average of the data of these data-containing, neighbouring cells. This was the case for records projected in water bodies or oceans due to small georeferencing errors or incompleteness of the layers in areas with water bodies.

Environmental data and two training areas

Current bioclimatic variables were obtained from the Africlim 3.0 database (**https://webfiles.york.ac.uk/KITE/AfriClim/)** and are based on CHIRPS precipitation and WorldClim temperature baseline data (Platts *et al.*, 2015). Potential evapotranspiration ratio was computed following Loiselle *et al*. (2008) using Holdridge *et al.* (1971) using the CHIRPS-based Africlim precipitation data. From the same source, we obtained future ensemble climate projections for 2085 for RCP 4.5 and 8.5. The Africlim climate projections are based on five dynamically down-scaled, bias-corrected regional climate models for Africa nested within ten global climate models. Quantitative soil data were retrieved from the Harmonised World Soil Database (FAO/IIASA/ISRIC/ISSCAS/JRC, 2012). As a proxy for topographic heterogeneity we used the standard deviation of altitude based on the 90 m SRTM altitude data (<http://srtm.csi.cgiar.org>) within each 5 arc-minute raster cell. All environmental data layers were cropped to the extent of the study area and, where necessary, aggregated to 5 arc-minute spatial resolution. From the 41 initial variables, we selected 16 uncorrelated ones using Spearman's |*rho*|< 0.7 to avoid overfitting of models due to multi-collinearity (Dormann *et al.*, 2013) (Table S1, SM Script 1).

All models were initially trained on the above defined full study area (hereafter 'African training area'). Models for species with a species' range size or prevalence < 0.1 or > 0.9 can show statistical artefacts, with prevalence defined as the fraction of raster cells where the species is present (McPherson *et al.*, 2004). Species' prevalence was estimated based on a thresholded SDM (Syfert *et al.*, 2014). Therefore, for species with a predicted prevalence < 0.1 in tropical Africa, we used a smaller training area defined as Gabon plus a 5 degrees buffer zone following the country borders (hereafter 'Gabonese training area') consisting of 18,112 raster cells containing the same selected environmental variables as the African training area. None of the species had a predicted prevalence > 0.9.

Species Distribution Models

For model building, we selected MaxEnt (Phillips *et al.*, 2006), as this algorithm has shown to outperform other methods using presence-only data (Elith *et al.*, 2006), including ensemble models (Aguirre-Gutiérrez *et al.*, 2013) and when tested with small sample sizes (Wisz *et al.*, 2008). We used target-group background sampling by limiting the selection of pseudo-absences to raster cells where herbarium records were made (Phillips *et al.*, 2009), hence applying the same geographical and environmental bias to the background data that is present in the specimen data. MaxEnt's default settings were applied, except that we only allowed linear and quadratic features for all sample sizes and excluded other features to prevent over-parameterization of the models (Merow *et al.*, 2013). Binary presence/absence maps for each species were computed by applying the 'ten percentile training presence' threshold to the logistic MaxEnt output, which excludes 10% of the records with the lowest habitat suitability values (Liu *et al.*, 2013, Merow *et al.*, 2013). Models trained on either tropical Africa or Gabon were projected to the year 2085 for RCP 4.5 and RCP 8.5 allowing extrapolation to novel future bioclimatic conditions. Species' response functions to environmental variables were clamped at the minimum and maximum variable values in the training area as a Multivariate Environmental Similarity Surface (MESS) analysis showed anomalies of up to 19% beyond the current range of values in environmental variables (Fig. S1).

Models were only accepted if two criteria were met. First, the number of unique training records should equal or exceed the minimum sample size required to generate models that are significantly better than random expectation. This minimum sample size increases with increasing species' prevalence (van Proosdij *et al.*, 2016b), and was identified for the following prevalence classes in the African training area (minimum sample sizes for the Gabonese training area between brackets): 7(5) for prevalence < 0.1, 7(8) for prevalence 0.1-0.2, 9(10) for prevalence 0.2-0.3, 12(11) for prevalence 0.3-0.4, 12(14) for prevalence 0.4-0.5, and 15(17) for prevalence > 0.5. Second, the model AUC value should rank > 95 when grouped with the 99 null models AUC values derived from a bias-corrected null model (Raes & ter Steege, 2007), meaning that the model performs significantly better than chance alone ($p < 0.05$).

Dispersal capacity

For the vast majority of species insufficient information is available to estimate the species' dispersal capacity, let alone estimates of the rare but important long distance dispersal events. We are aware of the criticism on the use of correlative species distribution models for predicting future species ranges, as these ignore dispersal limitations (Dormann *et al.*, 2012, Zurell *et al.*, 2016). Therefore, we modelled future distribution assuming either full or no dispersal, representing the most opportunistic and most conservative estimates of the species' future ranges. Future presence/absence maps assuming no dispersal were computed by limiting the predicted presence localities of individual species to raster cells where that species was predicted to be present under both the current and the future climate scenario. All resulting distribution maps were cropped to the country borders of Gabon.

Changes in species richness patterns

Current and future patterns of species richness were computed by stacking and summing the predicted presence/absence maps of significant SDMs. Second, for each 5 arc-minute raster cell for both RCP scenarios, species gain, loss and turn-over were computed. Species gain is defined as the number of species that are predicted to be present in the future but currently absent, species loss as the number of species currently present but absent in the future, and percentage of species turnover as 100 × (loss + gain) / (current richness + gain) (Thuiller *et al.*, 2005). This procedure was repeated for both RCP scenarios assuming no dispersal. Logically, under the no dispersal assumption, species gain is zero and the percentage of species turnover can be simplified as 100 × loss / current richness.

Explanatory power of climate anomalies

Climate anomalies for the two RCP scenarios for the year 2085 were computed by subtracting the future bioclimatic values from the current ones for each raster cell (Fig. S2). We assessed the correlation of bioclimatic anomalies with patterns of species gain, loss and turnover assuming either full or no dispersal. Quadratic terms of climate anomalies were included to account for non-linear responses to predicted climate change. Out of the 21 anomaly surfaces and their corresponding quadratic terms, we selected the following 11 uncorrelated variables based on a Spearman's |*rho*| < 0.7 under RCP 8.5 (data not shown): mean annual temperature (Bio01), mean diurnal temperature range (Bio02), isothermality (Bio03 = mean diurnal temperature range / annual temperature range), temperature seasonality (Bio04, standard deviation over monthly values), minimum temperature of the coldest month (Bio06), precipitation seasonality (Bio15, standard deviation over monthly values), precipitation of the driest quarter (Bio17), length of the longest dry season (LLDS), quadratic values of annual temperature range (Bio07^2), quadratic values of the number of dry months (DM^2), and quadratic values of the annual moisture index (MI^2). Correlations between anomalies and their quadratic terms for RCP 4.5 slightly differed from those for RCP 8.5, but for reasons of consistency we used the same selection of variables.

By their nature, biogeographic patterns are spatially auto-correlated, meaning that values of variables are not independent, as values at locations nearby are more similar than those at distant locations ('first law of geography') (Tobler, 1970). The presence of residual spatial autocorrelation (RSA) between predictions and data violates the assumption of statistical independence of observations and inflates type I errors (Dormann *et al.*, 2007). Although the results of spatial regression analyses are not seriously affected by the presence of short-distance spatial autocorrelation (Hawkins *et al.*, 2007), to avoid such, we accounted for spatial dependency in our correlation analyses by including the nine terms of the third order polynomial trend-surface regression equation of latitude and longitude (Borcard *et al.*, 1992).

We first modelled future species gain, loss and turnover as a function of climate anomalies (listed in Table 2) using a forward-backward stepwise multiple regression by applying the *lm* function of the *R*-library *stats* (R Core Team, 2016) and the *stepAIC* function of the *R*-library *MASS* (Venables & Ripley, 2002). Final regression models were computed by omitting non-significant variables. The residuals were tested for spatial autocorrelation (Rangel *et al.*, 2010) by building spatial autoregressive (SAR) lag models that compute the Moran's I value as a function of the lag distance. For this, we applied the *lagsarlm* function of the *R*-library *spdep* (Bivand & Piras, 2015) to the regression models computed above and explanatory variables using lag distances from 0 to 300 km. Secondly, we quantified the relative explanatory power of each individual climate anomaly to species gain, loss and turnover by applying the *calc.relimp* function of the *R*-library *relaimpo* (Gromping, 2006) to the computed regression models. All analyses were run in *R* (R Core Team, 2016), using the script provided as SM Script 2.

Results

We modelled the distribution of the 8,684 plant species, of which 5,323 have at least one occurrence in Gabon while the remaining 3,361 non-Gabonese species were recorded in the five degrees buffer zone. A total of 321,275 herbarium specimen records were used for model building. For

5,095 species, the number of records to train the model equals or exceeds the required minimum number of records. Of these species, the SDMs of 2,377 species, trained on 101,949 records, performed significantly better than null models. Out of the 2,377 species, 2,080 species are predicted present in Gabon today. The remaining 297 species are predicted present in the buffer zone, of which 57 are predicted to shift their ranges to within Gabon under future climate scenarios. These 2,137 Gabonese species (2,080 + 57) include nine Lycopodiophyta (club and spike mosses), 67 Pteridophyta (ferns), two Pinophyta (gymnosperms), and 2,059 Magnoliophyta (angiosperms) (listed in Table S2, maps with predicted current and future distributions available via [DRYAD insert link]). The SDMs of 1,271 of these 2,137 species were trained on the Gabonese training area as their initial models trained on tropical Africa resulted in a predicted species prevalence < 0.1. Of the final models trained on the Gabonese training area, 645 had a predicted prevalence of < 0.1, which we regard as acceptable given the scope of this study.

Changes in species richness

Assuming full dispersal, the total number of species predicted present in Gabon by 2085 decreased for both RCP scenarios. The loss of species is higher than the gain (arrival of species already present in the 5 degrees buffer zone but currently not recorded from Gabon itself; Table 1). In contrast, the predicted mean number of species in a raster cell increases from 796 to 866 by 2085 under RCP 8.5, whereas the maximum number decreased from 1,311 to 1,240. Under RCP 4.5, predicted mean richness is 869 and predicted maximum richness is 1,222 species. However, in the absence of dispersal, under RCP 8.5 the mean and maximum species numbers per raster cell decrease to 619 and 1,075 respectively (689 and 1,117 for RCP 4.5).

Table 1: The number of species predicted present in Gabon *for 2085 under representative concentration pathways 4.5 and 8.5, assuming either full or no dispersal. Percentages of species gained or lost are given between brackets.*

Climate scenario	Full dispersal	No dispersal
Present	2,080	2,080
2085 RCP 4.5	2,006 (+2.1% / -5.7%)	1,948 (-6.3%)
2085 RCP 8.5	1,955 (+2.7% / -8.8%)	1,878 (-9.7%)

Changes in species richness patterns

The current pattern of predicted plant species richness in Gabon is presented in Fig. 1b. In general, Gabonese mountains, as well as the area around Libreville are more species-rich compared to other lowland areas with the Crystal, Doudou and Pelé Mountains and the north-western part of the Chaillu Massif being particularly species-rich (see Fig. 1a). Areas in the north-east, as well as several coastal and inland savannahs in the South-West contain the lowest numbers of species.

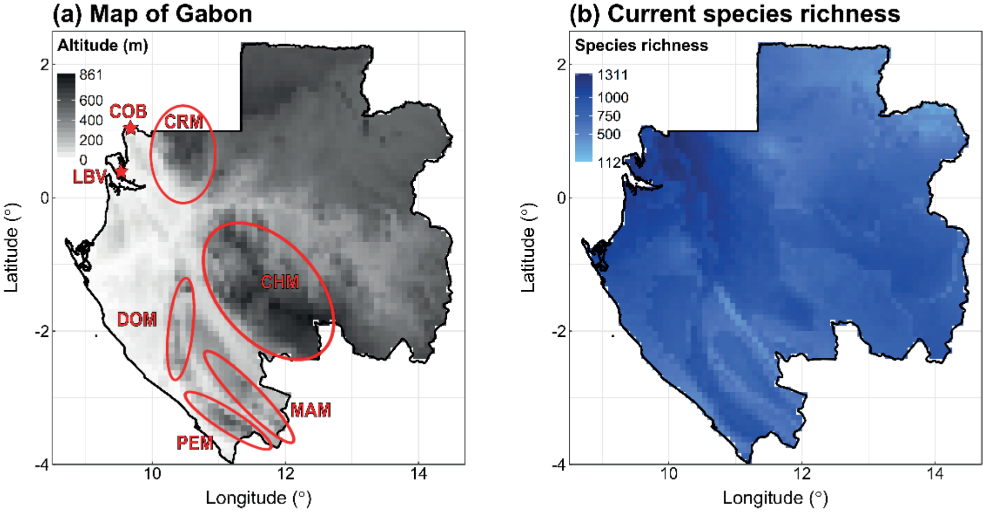
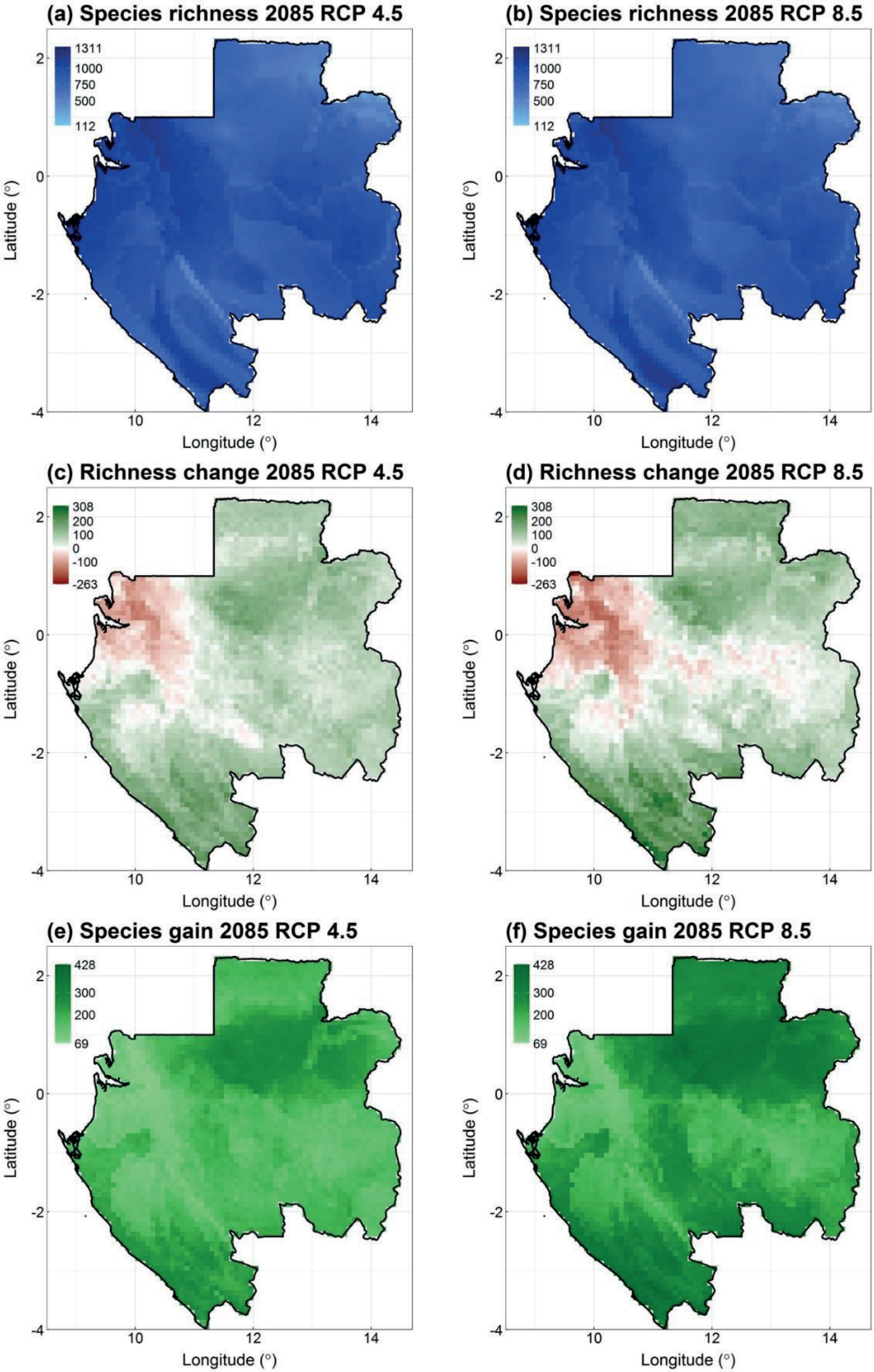


Figure 1: Map of Gabon showing altitude (a) and current plant species richness (b). *Altitude (Worldclim data) is shown in meters (Hijmans et al., 2005) and the following places are marked by red stars and mountains ranges indicated by red polygons: Libreville, the capital of Gabon (LBV), Coco Beach (COB), Crystal Mountains (CRM), Chaillu Massif (CHM), Doudou Mountains (DOM), Pelé Mountains (PEM) and Mabanda Mountains (MAM). Plant species richness (b) is based on stacked SDMs of 2,080 species.*

By 2085 under RCP 8.5 and assuming full dispersal, species richness in Gabon is predicted to have increased in the Central-North, in the Doudou, Pelé and Mabanda mountains, as well as in the coastal and inland savannahs bordering the Doudou and Pelé Mountains (Fig. 2b,d,f). In contrast, species richness is predicted to have decreased in the Crystal Mountains and in the area around Libreville, resulting from a much larger loss of species which is only partially compensated by the gain of other species (Fig. 2d,f,h). Species loss is also high in the north-western part of



(a) Species richness 2085 RCP 4.5

(b) Species richness 2085 RCP 8.5

(c) Richness change 2085 RCP 4.5

(d) Richness change 2085 RCP 8.5

(e) Species gain 2085 RCP 4.5

(f) Species gain 2085 RCP 8.5

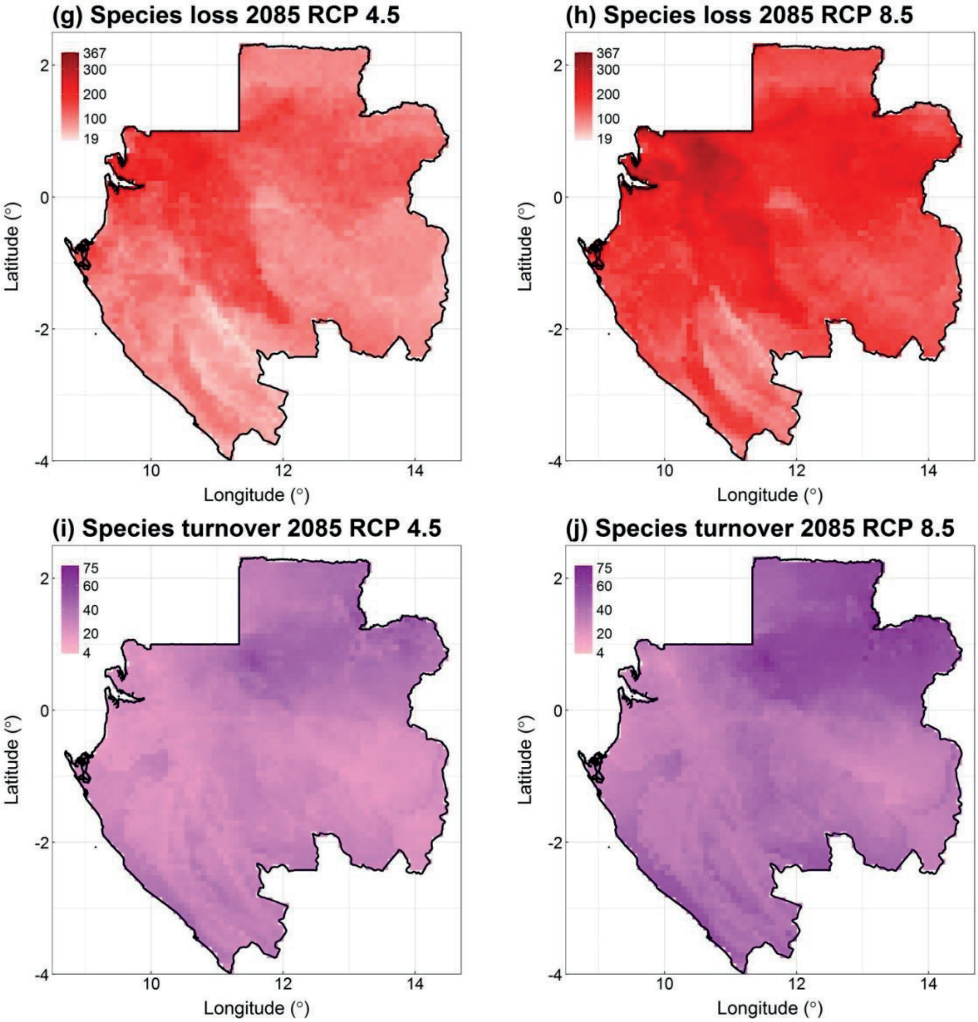

(g) Species loss 2085 RCP 4.5
(h) Species loss 2085 RCP 8.5
367
300
200
100
19
(i) Species turnover 2085 RCP 4.5
(j) Species turnover 2085 RCP 8.5
75
60
40
20
4
Latitude (°)
Longitude (°)
2
0
-2
-4
10
12
14

Figure 2 (this and previous page): Patterns of species richness (a&b), change in species richness (c&d), species gain (e&f), loss (g&h) and turnover (in %) (i&j) in Gabon for 2085 *under the representative concentration pathways 4.5 (a,c,e,g,i) and 8.5 (b,d,f,h,j) assuming full dispersal.*

the Chaillu Massif, in the Doudou and Pelé mountains and around Coco Beach in the far North-West. In concert, these gains and losses of species result in levels of species turnover as high as 75% in the north of Gabon, in the Chaillu Massif and in the coastal and inland savannahs in the South-West (Fig. 2j).

By 2085 under RCP 4.5 and assuming full dispersal, patterns of species richness, change of species richness, as well as species gain, loss and turnover are similar, although the values of species gain, loss and turnover are lower than for the RCP 8.5 scenario (Fig. 2c,e,g,i). For the RCP 4.5 scenario, values of species richness are intermediate between those of RCP 8.5 and the current situation (Fig. 2a).

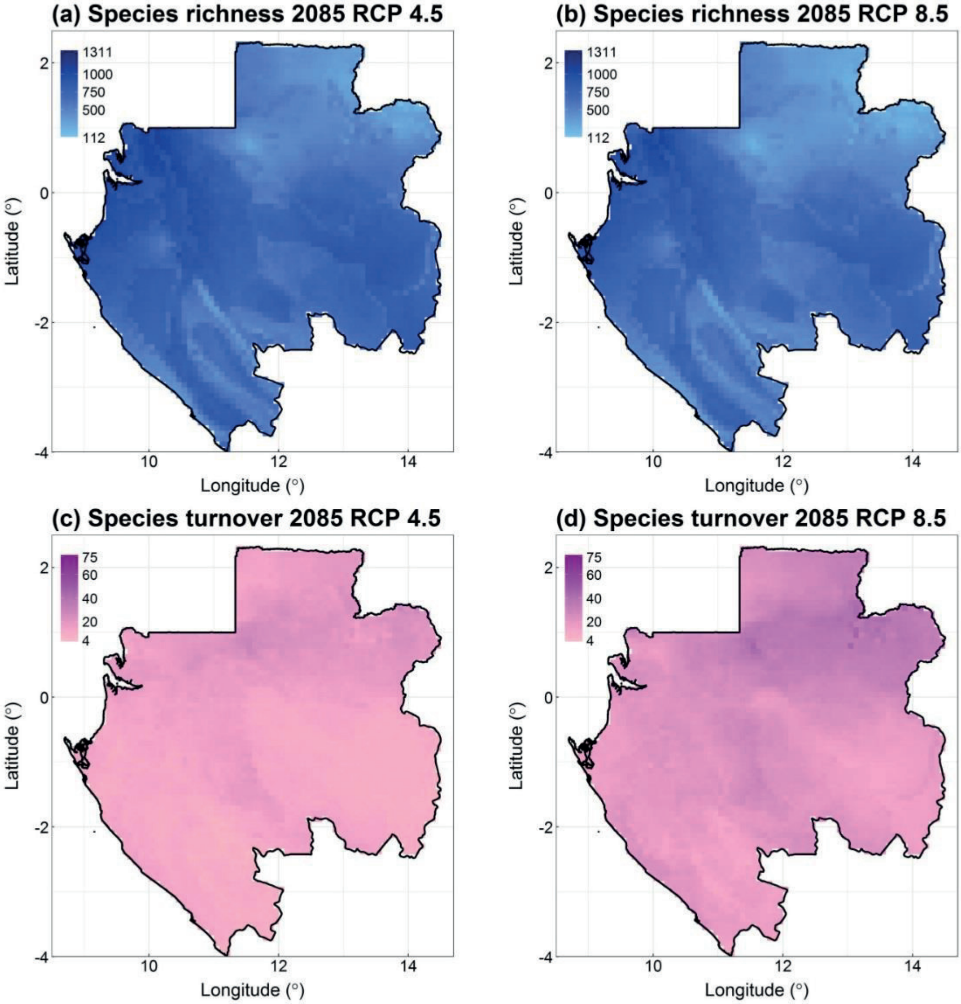


Figure 3: Patterns of species richness (a,b) and turnover (in %) (c&d) in Gabon for 2085 under the representative concentration pathways 4.5 (a,c) and 8.5 (b,d) assuming no dispersal.

Effect of dispersal limitations

The effect of dispersal limitations on species richness patterns is visualised by the patterns of species richness and turnover assuming no dispersal for 2085 under RCP 4.5 and 8.5 (Fig. 3). Obviously, future species richness is generally lower in the situation of no dispersal compared to the situation of full dispersal, but species-poor areas show greater relative decrease of richness than species-rich areas. By 2085 under RCP 8.5, species richness of the most species-poor areas is predicted to have decreased with 46% and richness of the most species-rich areas with 18%. The absence of gained species in the case of no dispersal results in overall lower turnover values, particularly in the south-western coastal areas. Some areas are more affected by dispersal limitations than others. Whereas species richness in the Central-North, as well as the coastal and inland savannahs bordering the Doudou and Pelé Mountains is predicted to increase in the case of full dispersal by 2085 under RCP 8.5, these areas face a strong decrease in species richness in the absence of dispersal. Here too, for RCP 4.5, patterns of species richness and turnover are similar to the RCP 8.5 scenario (Fig. 3a,c). As expected, in the absence of dispersal, the future total number of species predicted present in Gabon is lower compared to the situation of full dispersal (Table 1).

Correlation with environmental variables

Spatial autocorrelation is present for the first 100-140 km in the residuals of the final stepwise multiple regression models of species gain, loss and turnover for the year 2085 under RCP 8.5 (Fig. S3).

Some climate anomalies show a strong correlation with predicted patterns of species gain, loss and turnover for the year 2085 under RCP 8.5 (Table 2). Species gain is related to increased quadratic values of annual temperature range (Table 2 – Bio07^2, $R_{adj.}^2 = 0.19$). Species loss is best explained by increased precipitation in the driest quarter (Table 2 – Bio17, $R_{adj.}^2 = 0.32$). Species turnover under the assumption of full dispersal is correlated with increased quadratic values of annual temperature range (Table 2 – Bio07^2, $R_{adj.}^2 = 0.17$), and with increased mean diurnal temperature range (Table 2 – Bio02, $R_{adj.}^2 = 0.09$). In the absence of dispersal, species turnover is most strongly correlated with increased mean diurnal temperature range (Table 2 – Bio02, $R_{adj.}^2 = 0.14$) and increased mean annual temperature (Table 2 – Bio01, $R_{adj.}^2 = 0.11$).

Correlations with anomalies under the RCP 4.5 scenario are different and generally weaker than under the RCP 8.5 scenario (Table S3). Species gain is weakly correlated with decreased precipitation in the driest quarter (Table S3 – Bio17, $R_{adj.}^2 = 0.07$) and with increased quadratic values of annual temperature range (Table S3 – Bio07^2, $R_{adj.}^2 = 0.06$). Species

Table 2: Coefficients and explanatory power (adjusted R^2) of individual uncorrelated climate anomalies and their quadratic terms by 2085 under RCP 8.5 *to patterns of species gain (a), loss (b) and turnover assuming full dispersal (c) and turnover assuming no dispersal (d). Anomalies are ordered according to their explanatory power in the final stepwise multiple regression model and coefficients of this final model are given. Variables not contributing significantly, and therefore excluded from the final model, are omitted from this table. The intercept and $R_{adj.}^2$ of the final model are included.*

(a) Gain			**(b) Loss**			**(c) Turnover full dispersal**			**(d) Turnover no dispersal**		
Full model	Intercept	$R_{adj.}^2$	Full model	Intercept	$R_{adj.}^2$	Full model	Intercept	$R_{adj.}^2$	Full model	Intercept	$R_{adj.}^2$
	235.140	0.372		-96.168	0.583		-15.693	0.540		-49.859	0.592
	Coefficient	$R_{adj.}^2$		Coefficient	$R_{adj.}^2$		Coefficient	$R_{adj.}^2$		Coefficient	$R_{adj.}^2$
Bio07^2	2.266	0.194	Bio17	1.826	0.321	Bio07^2	0.277	0.174	Bio02	0.957	0.138
Bio17	-0.490	0.046	Bio02	9.835	0.066	Bio02	0.769	0.090	Bio01	2.335	0.105
MI^2	-0.153	0.036	Bio03	-1.052	0.051	Bio01	3.031	0.073	Bio07^2	0.111	0.076
Bio02	4.430	0.028	LLDS	14.022	0.034	Bio06	-1.574	0.045	Bio04	0.884	0.060
Bio03	1.576	0.018	MI^2	-0.200	0.028	MI^2	-0.026	0.044	Bio17	0.145	0.059
Bio06	-4.109	0.018	Bio06	6.358	0.027	Bio04	1.019	0.038	DM^2	1.820	0.054
Bio01	4.893	0.014	Bio01	-1.420	0.027	Bio17	0.023	0.026	Bio06	-0.626	0.049
Bio15	3.001	0.009	DM^2	5.155	0.027	DM^2	2.414	0.023	MI^2	-0.026	0.040
LLDS	-16.086	0.006	Bio15	0.479	0.005	Bio03	0.156	0.021	Bio15	0.170	0.008
DM^2	7.602	0.003				Bio15	0.327	0.007	LLDS	1.157	0.006

loss is most strongly correlated with increased precipitation in the driest quarter (Table S3 – Bio17, $R_{adj.}^2 = 0.27$) and weakly with decreased isothermality (Table S3 – Bio03, $R_{adj.}^2 = 0.08$). Species turnover under full dispersal is weakly correlated with increased quadratic values of annual temperature range (Table S3 – Bio07^2, $R_{adj.}^2 = 0.08$) and with decreased isothermality (Table S3 – Bio03, $R_{adj.}^2 = 0.07$). Species turnover assuming no dispersal is most strongly correlated with decreased isothermality (Table S3 – Bio03, $R_{adj.}^2 = 0.13$) and with increased mean diurnal temperature range (Table S3 – Bio02, $R_{adj.}^2 = 0.09$).

Discussion

Species gain, loss and turnover driven by bioclimatic anomalies

Using the most complete collection record data set of Gabonese vascular plants species together with the latest regional climate models predicting future bioclimatic conditions, we show that climate change is predicted to have major consequences for the pattern of species richness (Fig. 2,3), as well as for the total number of plant species present occurring in Gabon (Table 1). Using climate change anomalies we quantify the contribution of individual climate variables in explaining the predicted changes in species richness (Table 2).

Future species gain under RCP 8.5 is most strongly correlated with increased quadratic values of the annual temperature range (Bio07^2), which captures two different processes that were not captured by anomalies in the annual temperature range itself. Both in areas where the future values of Bio07 decrease (South-West) and increase (North-East, Fig. S2 – Bio07), captured by the quadratic term of Bio07 are predicted to gain species under future conditions. According to our predictions under RCP 8.5, species colonize areas with an annual temperature range that is currently either too small, which is the case in the lowland rainforests in north-eastern Gabon, or that is currently too large, which is the case in the savannahs in south-western Gabon. Species expanding their ranges in these areas predominantly inhabit semi-wet to wet forests with small annual temperature ranges. These findings underline the difficulty of modelling general regional responses to climate change as species each have their own ecological niche and respond differently to changing climatic conditions.

Species loss appears to be mainly driven by increased precipitation in the driest quarter (Bio17). Although the actual maximum increase in precipitation for the driest quarter of 212 mm (Fig. S2 – Bio17) is small compared to the over 3,000 mm of annual precipitation (Fig. S2 – Bio12), this is still a 3-4-fold increase in the dry season. Consequently, the intensity of the dry season is strongly reduced resulting in a loss of species that are adapted to those conditions.

Species turnover under the assumption of full dispersal is dominated by species gain, which explains the correlation with increased values of the quadratic term of annual temperature range (Bio07^2), followed by increased mean diurnal temperature range (Bio02). In the absence of dispersal, species turnover is based on species loss and appears to be driven by increased diurnal temperature range (Table 2 – Bio02) and increased mean annual temperature (Bio01). The predicted high levels of turnover, of up to 75% for RCP 8.5 assuming full dispersal, are reason for concern as these result in future plant communities characterized by novel species compositions for which the effect on novel biotic interactions are unknown. Although the highest levels of turnover are predicted for relatively species-poor areas, such effects could be detrimental to the flora of these regions. This is especially the case for the coastal savannahs in south-western Gabon that contain disproportionally high numbers of endemic species (Harris *et al.*, 2012, van Proosdij *et al.*, 2016a, Wieringa & Sosef, 2011). For the local people this implies that they will lose a major part of their traditionally used natural products, hence part of their cultural heritage.

Only few other studies address correlations of climate anomalies with patterns of species gain, loss and turnover. In Western Australia, species loss is correlated with decreasing precipitation (Fitzpatrick *et al.*, 2008), and in Yunnan, China it is correlated with increasing temperature variability and decreasing precipitation during the dry season (Zhang *et al.*, 2014). For Gabon, we conclude that, species loss is mainly related to precipitation anomalies and to a lesser extent to future changes in temperature.

Effects of climate change on Gabonese plant species

In Gabon, the predicted increase in precipitation of up to 500 mm (Fig. S2 – Bio12) is in contrast with other major tropical rainforest regions. The Amazon is predicted to experience warmer and drier conditions, particularly characterized by longer and more intense periods of water deficit (Zelazowski *et al.*, 2011), although some models predict regional wetter conditions (Feeley *et al.*, 2012). West African rainforests are thought to have been adapted to regular water deficits during their long-term exposure to droughts (Asefi-Najafabady & Saatchi, 2013) and have experienced a continuous drying trend starting in the 1970s. The same applies to northern Congolese rainforests, for which a widespread decline in greenness over the past decade was documented (Zhou *et al.*, 2014).

West Central Africa, including Gabon, experienced regular water deficits throughout their history too, but in contrast, it has faced relatively wetter conditions in the past 15 years (Asefi-Najafabady & Saatchi, 2013) and the bioclimatic anomalies we used here indicate a continuation of this warmer and wetter climate (Platts *et al.*, 2015).

The effect of global climate change on the predicted number of species present in Gabon by 2085 is clearly negative: species numbers decrease in both climate scenarios ranging from over 5% under RCP 4.5 and full dispersal to almost 10% under RCP 8.5 assuming no dispersal. In contrast, our results show a predicted increase in average species richness per raster cell, but only under the condition that species will be able to reach the newly suitable habitats in time.

Effect of dispersal limitations

The substantial difference in patterns of species richness in Gabon between full vs. no dispersal illustrates the additional threat dispersal limitations pose to the survival of species under climate change. In the absence of dispersal, the higher number of species predicted to go extinct in Gabon and lacking arrival of foreign species in concert, lead to a clear decrease in species richness. These results are in contrast to the findings of Zhang *et al.* (2014), who found almost identical results for models with or without dispersal. This can, at least partly, be explained by Zhang *et al.*'s selection of species based on the extent of the study area. Zhang *et al.* (2014) included only species recorded from the Yunnan province and excluded species known from neighbouring provinces and countries, resulting in an underestimation of future species richness in the presence of dispersal. In contrast, in our study we included all Gabonese species as well as those recorded from the 5 degrees buffer zone. Although many species will have limited dispersal capacity and therefore might not be able to reach Gabon, ignoring possible future immigrants from neighbouring areas leads to erroneous conclusions. Fitzpatrick *et al.* (2008) found little effect of dispersal limitations in the future distribution of 100 endemic Australian *Banksia* species. However, these species mostly face a range collapse, thus minimizing the effect of dispersal limitations on future distributions. Here, we assessed two extreme situations, one where species have full dispersal capacity and hence fill their entire potential range in contrast to one with no dispersal where the future spatial distribution is limited to those areas that are suitable both now and in the future. The truth will lie somewhere in between, where some species are poor dispersers and others will easily reach any potential area in Gabon or even Africa. The simple lack of data on the dispersal capacity of each species hampers more refined analyses. For those species with limited dispersal capacity that are predicted to go extinct, facilitated migration might offer opportunities for survival.

Caveats and considerations

Climate change drives modifications to the distribution of each individual species and can only partly explain changes in patterns of species richness. The lack of a collective response of all species in a specific community is here best illustrated by the low correlation of species gain with climate anomalies. As species respond individually to climate change and climate change creates non-analogue climate-soil combinations, in concert this results in non-analogue future plant communities. Therefore, to assess the effect of global climate change on diversity patterns, we advocate to assess the response of individual species like we did here.

Some aspects not included in our analyses are expected to further increase the numbers of species lost from Gabon and further reduce the range sizes of many species. First of all, the SDMs of 75% of the species were not included because they did not pass the null model test or were trained on too few records. As most of these species are rare and narrow-ranged, and hence more vulnerable to change, we expect that the actual percentage of species disappearing in Gabon due to climate change will be (much) higher than the 10% computed in our analyses. The richness pattern of narrow-ranged species differs from that of widespread species, with the former concentrated in centres of plant species richness (van Proosdij *et al.*, 2016a). Here, we show that these species-rich areas are predicted to face the largest species losses. Second, we ignored biotic interactions such as competition and pollination. In reality, limited dispersal capacity and interspecific differences in dispersal capacity coupled with different competitive strengths and the interaction with pollinators will result in higher extinction rates than the ones we modelled. Notably, species that have a narrow niche are slow dispersers or weak competitors (Urban *et al.*, 2012). Consequently, our predicted numbers of species present in Gabon under different future climate scenarios are arguably too optimistic and species loss will be (much) higher in many areas. In addition, if loss of habitat is taken into account, species loss will be even stronger, particularly on a local scale (Newbold *et al.*, 2015). And finally, reduced geneflow and fragmentation of populations may cause these to drop below a viable minimum size resulting in species entering an extinction vortex (Bellard *et al.*, 2012).

On a positive side, microrefugia (small favourable pockets caused by habitat heterogeneity, Leal 2001) are often not visible at the spatial scale of the analyses but have prevented species from extinction in the past and might do so too in the future (Gillingham *et al.*, 2012, Hylander *et al.*, 2015). For Gabon, the presence of microrefugia outside the main rainforest refugia has been postulated for Gabonese trees (Leal, 2001, Wieringa, 1999) and for *Begonia* species (Sosef, 1996). Such microrefugia were often gallery forests or otherwise wetter areas; the future microrefugia are of a different nature and need to be intrinsically drier areas. Possibly inselbergs, commonly surrounded by shallow soils, may play a role here. We advocate for additional research on the location and stability of microrefugia under climate change, particularly in relation to the current network of protected areas.

Acknowledgements

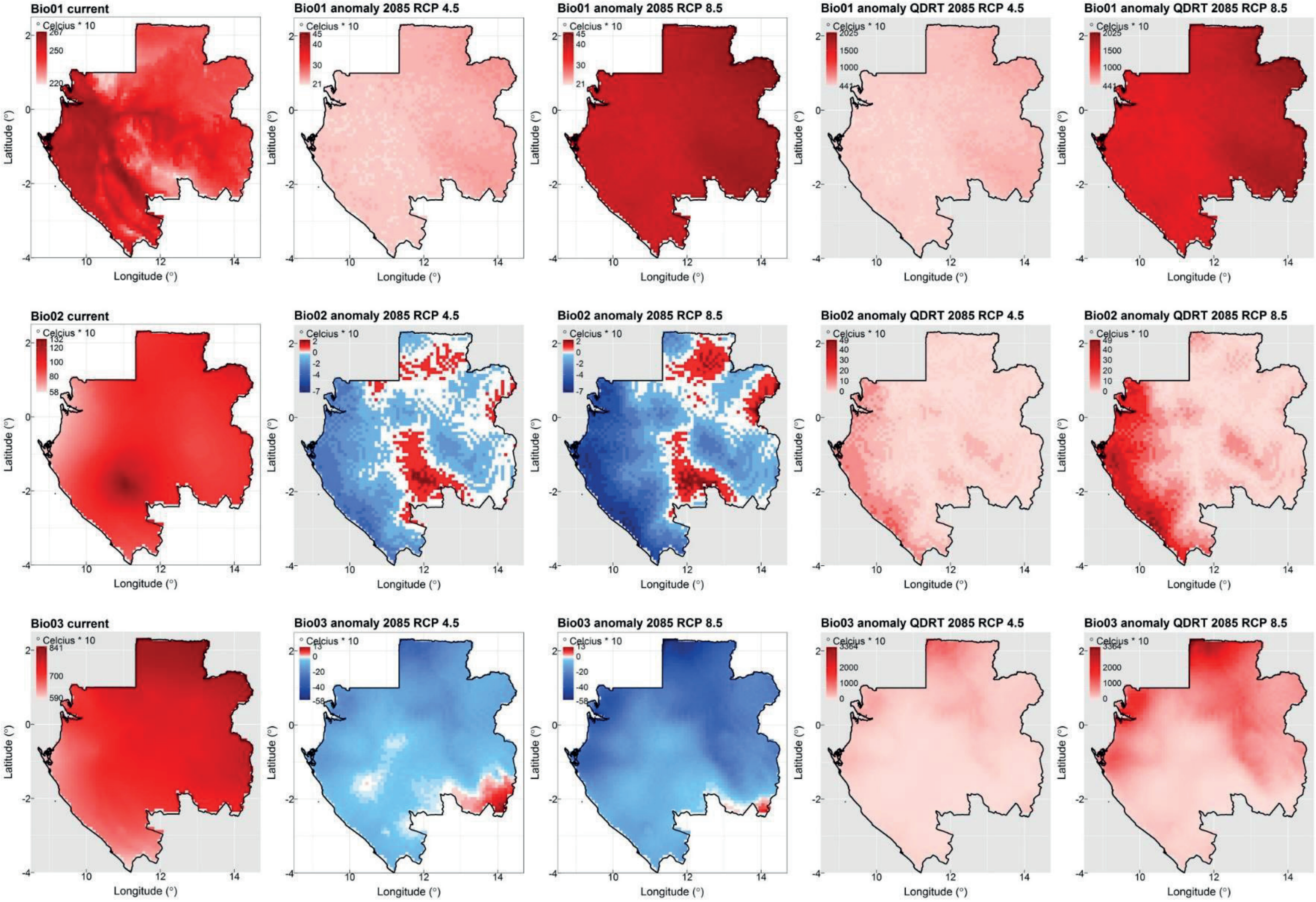
The “Gabon database” is compiled in a joined project of the herbaria BM, BR, BRLU, LBV, MO, P & WAG. We thank collectors and curators of the herbaria B, BNRH, BR, BRLU, E, FR, G, K, MA, MO, P, PRE, UPS for their generous exchange of herbarium specimen data with NHN-Naturalis during duplicate exchange of African collections, which eased our job of getting our specimens in the database. MA and MO sent additional African specimen data specifically for this project, which is gratefully acknowledged. The National Herbarium of Gabon, part of IPHAMETRA, and its staff has been instrumental during many field missions of WAG, MO, and BR in the past 35 years and this collaboration has resulted in the majority of the Gabonese specimens used in this study.

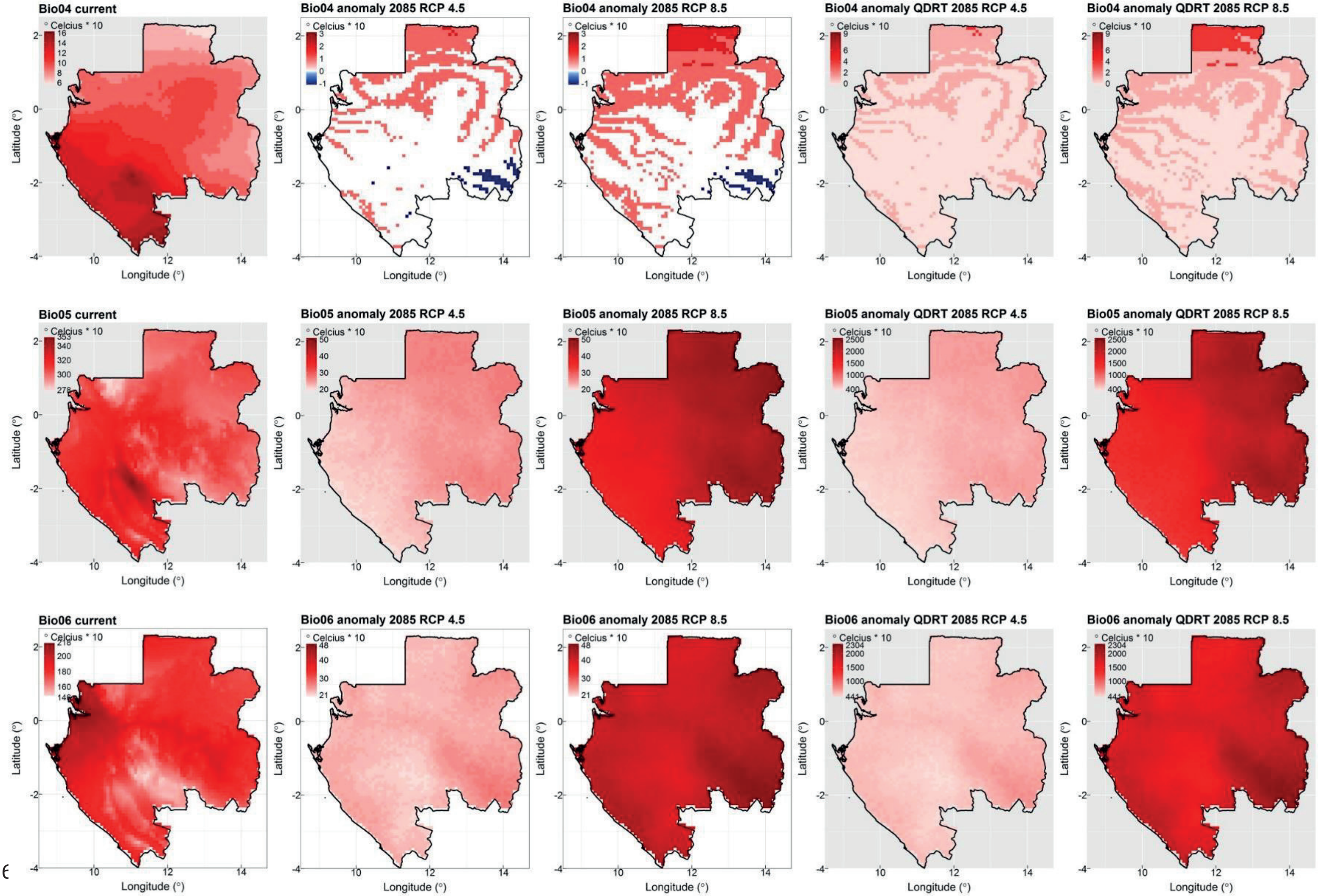
Supplementary material

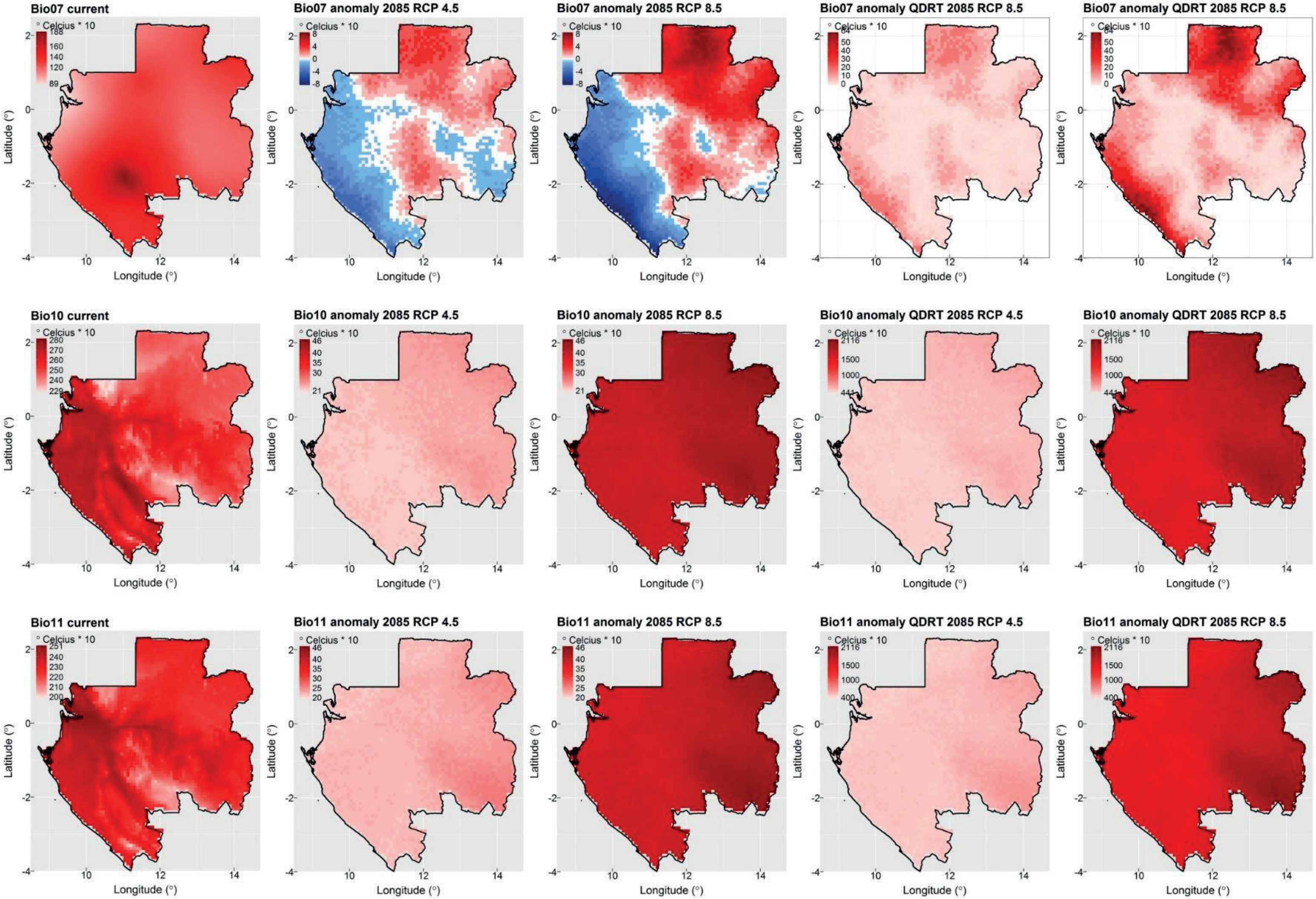
All supplementary materials (appendices 1-8) are available upon request.

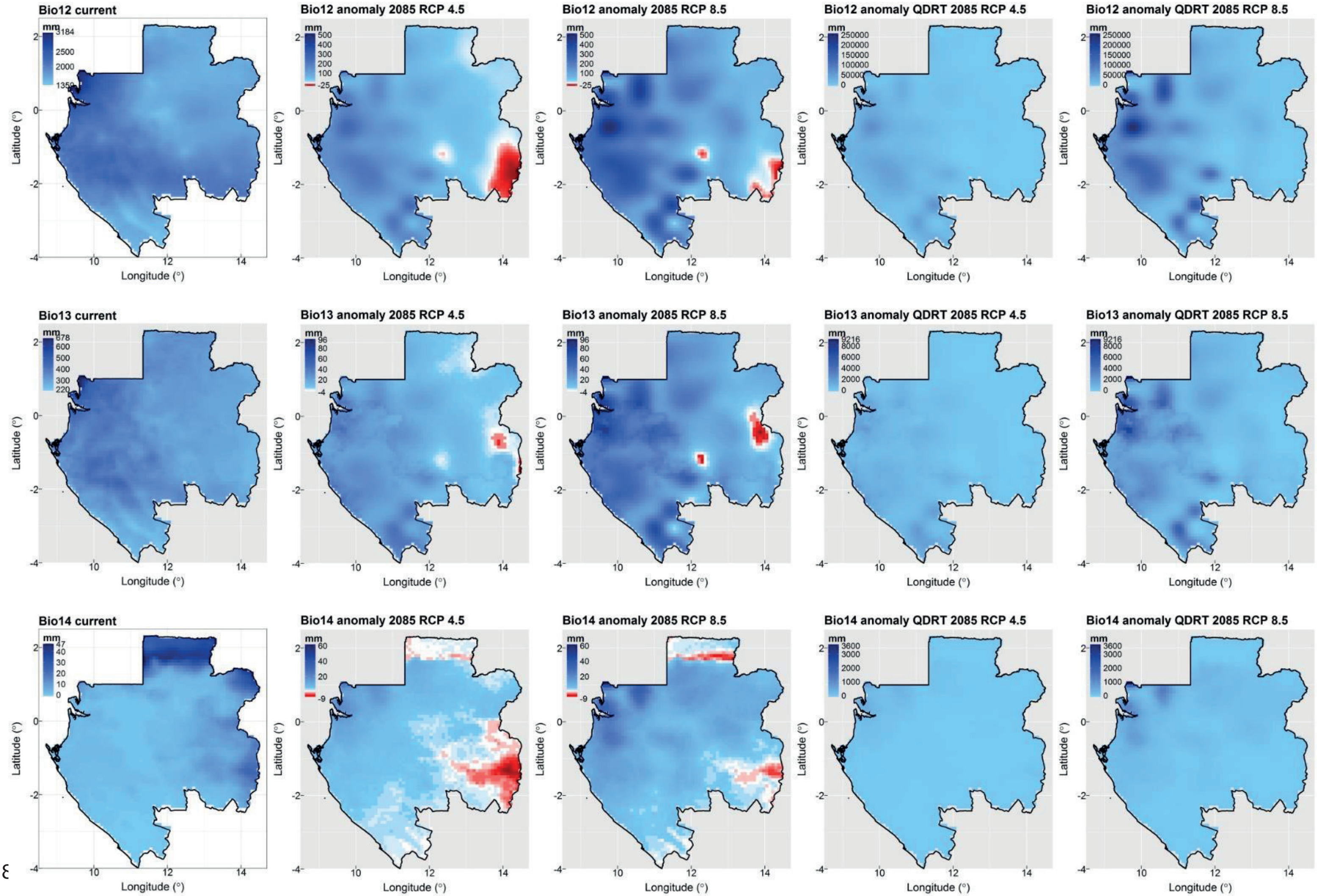
***Figure S2 (next pages): Climate anomalies.** Presented are current climate variables (left panels), climate anomalies for 2085 under representative concentration pathways RCP 4.5 and 8.5 (middle panels), as well as quadratic terms of climate anomalies (right panels) in Gabon. For the following climatic variables, current data and future ensemble projections were obtained from the Africlim 3.0 database **https://webfiles.york.ac.uk/KITE/AfriClim/,** based on CHIRPS precipitation and WorldClim temperature baseline data (Platts* et al.*, 2015). The potential evapotranspiration (PET) ratio was computed following Holdrigde* et al. *(1971). Redder colours represent warmer conditions, drier conditions or larger diurnal or annual temperature or precipitation ranges. Bluer colours represent colder conditions, wetter conditions or smaller ranges. Variables printed on a white background are used for analysis, variables printed on a grey background are excluded due to collinearity with included variables.*

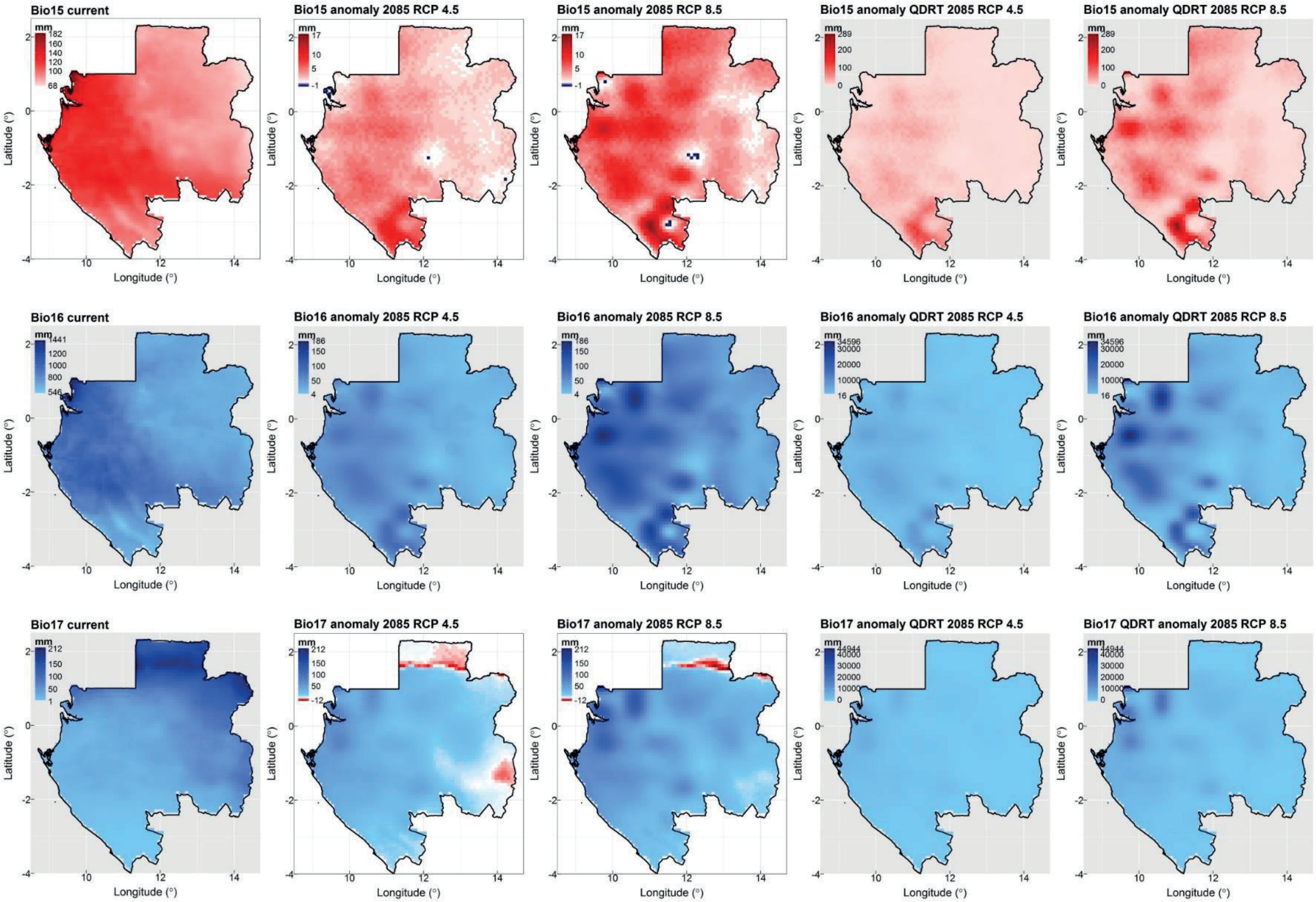
- Bio01: Mean annual temperature (° Celcius * 10)
- Bio02: Mean diurnal temperature range (° Celcius * 10)
- Bio03: Isothermality (* 10 = Bio02 / Bio07 * 100)
- Bio04: Temperature seasonality (° Celcius * 10 = standard deviation over monthly values)
- Bio05: Maximum temperature warmest month (° Celcius * 10)
- Bio06: Minimum temperature coolest month (° Celcius * 10)
- Bio07: Annual temperature range (° Celcius * 10)
- Bio10: Mean temp warmest quarter (° Celcius * 10)
- Bio11: Mean temp coolest quarter (° Celcius * 10)
- Bio12: Mean annual precipitation (mm)
- Bio13: Precipitation wettest month (mm)
- Bio14: Precipitation driest month (mm)
- Bio15: Precipitation seasonality (mm = standard deviation over monthly values)
- Bio16: Precipitation wettest quarter (mm)
- Bio17: Precipitation driest quarter (mm)
- PET: Potential evapotranspiration (mm)
- DM: Number of dry months (months)
- LLDS: Length of longest dry season (months)
- MI: Annual moisture index (index * 100 = Bio12 / PET)
- MIMQ: Moisture index moist quarter (index * 100)
- MIAQ: Moisture index arid quarter (index * 100)

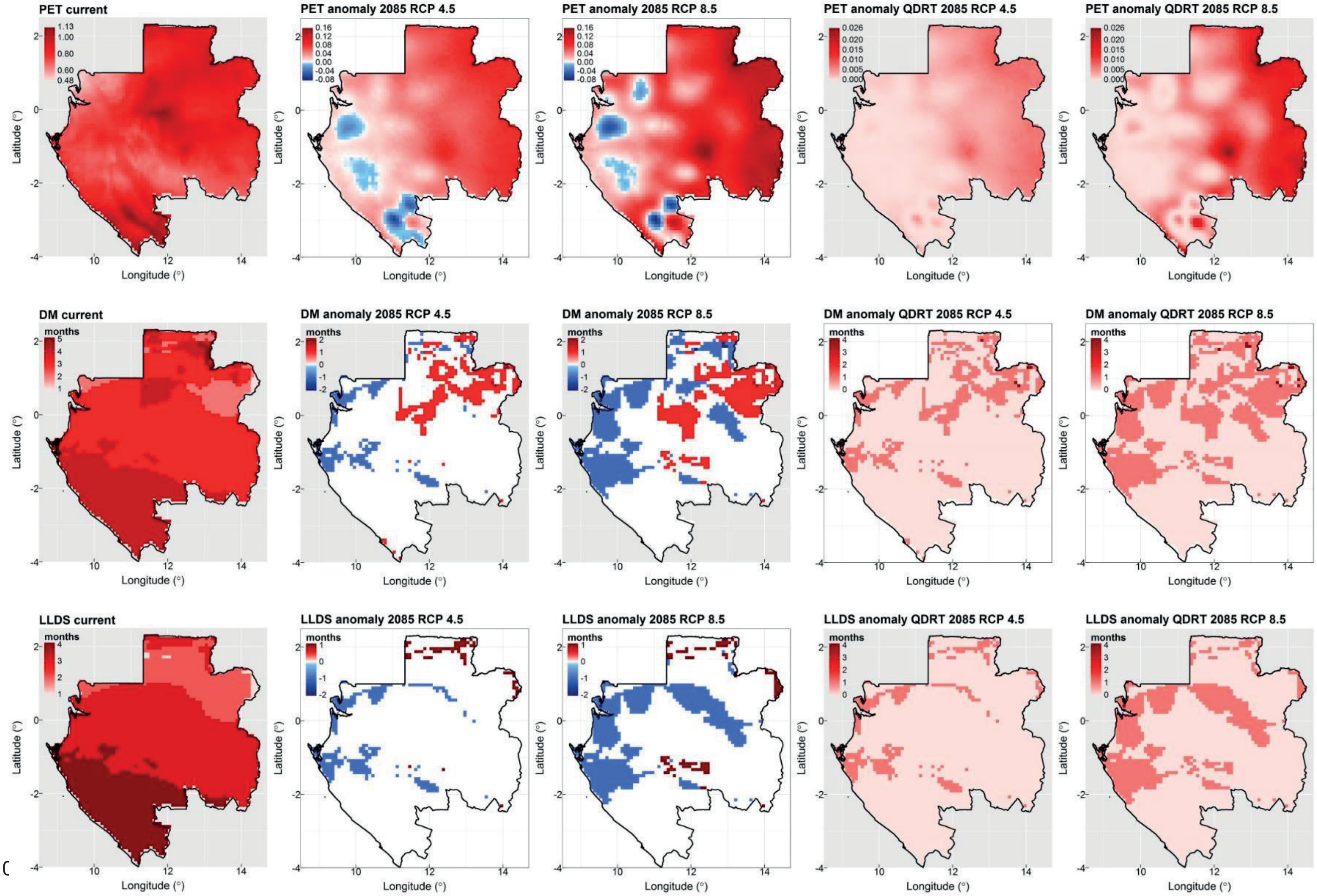

Bio01 current
° Celcius * 10
267
250
220
Bio01 anomaly 2085 RCP 4.5
° Celcius * 10
45
40
30
21
Bio01 anomaly 2085 RCP 8.5
° Celcius * 10
45
40
30
21
Bio01 anomaly QDRT 2085 RCP 4.5
° Celcius * 10
2025
1500
1000
441
Bio01 anomaly QDRT 2085 RCP 8.5
° Celcius * 10
2025
1500
1000
441
Bio02 current
° Celcius * 10
132
120
100
80
58
Bio02 anomaly 2085 RCP 4.5
° Celcius * 10
2
0
-2
-4
-7
Bio02 anomaly 2085 RCP 8.5
° Celcius * 10
2
0
-2
-4
-7
Bio02 anomaly QDRT 2085 RCP 4.5
° Celcius * 10
49
40
30
20
10
0
Bio02 anomaly QDRT 2085 RCP 8.5
° Celcius * 10
49
40
30
20
10
0
Bio03 current
° Celcius * 10
841
700
590
Bio03 anomaly 2085 RCP 4.5
° Celcius * 10
13
0
-20
-40
-58
Bio03 anomaly 2085 RCP 8.5
° Celcius * 10
13
0
-20
-40
-58
Bio03 anomaly QDRT 2085 RCP 4.5
° Celcius * 10
3364
2000
1000
0
Bio03 anomaly QDRT 2085 RCP 8.5
° Celcius * 10
3364
2000
1000
0
Latitude (°)
2
0
-2
-4
Longitude (°)
10
12
14


Bio04 current
° Celcius * 10
16
14
12
10
8
6
Bio04 anomaly 2085 RCP 4.5
° Celcius * 10
3
2
1
0
-1
Bio04 anomaly 2085 RCP 8.5
° Celcius * 10
3
2
1
0
-1
Bio04 anomaly QDRT 2085 RCP 4.5
° Celcius * 10
9
6
4
2
0
Bio04 anomaly QDRT 2085 RCP 8.5
° Celcius * 10
9
6
4
2
0
Bio05 current
° Celcius * 10
353
340
320
300
278
Bio05 anomaly 2085 RCP 4.5
° Celcius * 10
50
40
30
20
Bio05 anomaly 2085 RCP 8.5
° Celcius * 10
50
40
30
20
Bio05 anomaly QDRT 2085 RCP 4.5
° Celcius * 10
2500
2000
1500
1000
400
Bio05 anomaly QDRT 2085 RCP 8.5
° Celcius * 10
2500
2000
1500
1000
400
Bio06 current
° Celcius * 10
218
200
180
160
146
Bio06 anomaly 2085 RCP 4.5
° Celcius * 10
48
40
30
21
Bio06 anomaly 2085 RCP 8.5
° Celcius * 10
48
40
30
21
Bio06 anomaly QDRT 2085 RCP 4.5
° Celcius * 10
2304
2000
1500
1000
441
Bio06 anomaly QDRT 2085 RCP 8.5
° Celcius * 10
2304
2000
1500
1000
441
Latitude (°)
2
0
-2
-4
10
12
14
Longitude (°)


Bio07 current
° Celcius * 10
188
160
140
120
89
Bio07 anomaly 2085 RCP 4.5
° Celcius * 10
8
4
0
-4
-8
Bio07 anomaly 2085 RCP 8.5
° Celcius * 10
8
4
0
-4
-8
Bio07 anomaly QDRT 2085 RCP 8.5
° Celcius * 10
64
50
40
30
20
10
0
Bio07 anomaly QDRT 2085 RCP 8.5
° Celcius * 10
64
50
40
30
20
10
0
Bio10 current
° Celcius * 10
280
270
260
250
240
229
Bio10 anomaly 2085 RCP 4.5
° Celcius * 10
46
40
35
30
21
Bio10 anomaly 2085 RCP 8.5
° Celcius * 10
46
40
35
30
21
Bio10 anomaly QDRT 2085 RCP 4.5
° Celcius * 10
2116
1500
1000
444
Bio10 anomaly QDRT 2085 RCP 8.5
° Celcius * 10
2116
1500
1000
444
Bio11 current
° Celcius * 10
251
240
230
220
210
200
Bio11 anomaly 2085 RCP 4.5
° Celcius * 10
46
40
35
30
25
20
Bio11 anomaly 2085 RCP 8.5
° Celcius * 10
46
40
35
30
25
20
Bio11 anomaly QDRT 2085 RCP 4.5
° Celcius * 10
2116
1500
1000
400
Bio11 anomaly QDRT 2085 RCP 8.5
° Celcius * 10
2116
1500
1000
400
Latitude (°)
2
0
-2
-4
Longitude (°)
10
12
14


Bio12 current
mm
3184
2500
2000
1359
Bio12 anomaly 2085 RCP 4.5
mm
500
400
300
200
100
-25
Bio12 anomaly 2085 RCP 8.5
mm
500
400
300
200
100
-25
Bio12 anomaly QDRT 2085 RCP 4.5
mm
250000
200000
150000
100000
50000
0
Bio12 anomaly QDRT 2085 RCP 8.5
mm
250000
200000
150000
100000
50000
0
Bio13 current
mm
678
600
500
400
300
220
Bio13 anomaly 2085 RCP 4.5
mm
96
80
60
40
20
-4
Bio13 anomaly 2085 RCP 8.5
mm
96
80
60
40
20
-4
Bio13 anomaly QDRT 2085 RCP 4.5
mm
9216
8000
6000
4000
2000
0
Bio13 anomaly QDRT 2085 RCP 8.5
mm
9216
8000
6000
4000
2000
0
Bio14 current
mm
47
40
30
20
10
0
Bio14 anomaly 2085 RCP 4.5
mm
60
40
20
-9
Bio14 anomaly 2085 RCP 8.5
mm
60
40
20
-9
Bio14 anomaly QDRT 2085 RCP 4.5
mm
3600
3000
2000
1000
0
Bio14 anomaly QDRT 2085 RCP 8.5
mm
3600
3000
2000
1000
0
Latitude (°)
2
0
-2
-4
Longitude (°)
10
12
14


Bio15 current
mm
182
160
140
120
100
68
Bio15 anomaly 2085 RCP 4.5
mm
17
10
5
-1
Bio15 anomaly 2085 RCP 8.5
mm
17
10
5
-1
Bio15 anomaly QDRT 2085 RCP 4.5
mm
289
200
100
0
Bio15 anomaly QDRT 2085 RCP 8.5
mm
289
200
100
0
Bio16 current
mm
1441
1200
1000
800
546
Bio16 anomaly 2085 RCP 4.5
mm
186
150
100
50
4
Bio16 anomaly 2085 RCP 8.5
mm
186
150
100
50
4
Bio16 anomaly QDRT 2085 RCP 4.5
mm
34596
30000
20000
10000
16
Bio16 anomaly QDRT 2085 RCP 8.5
mm
34596
30000
20000
10000
16
Bio17 current
mm
212
150
100
50
1
Bio17 anomaly 2085 RCP 4.5
mm
212
150
100
50
-12
Bio17 anomaly 2085 RCP 8.5
mm
212
150
100
50
-12
Bio17 anomaly QDRT 2085 RCP 4.5
mm
44944
40000
30000
20000
10000
0
Bio17 QDRT anomaly 2085 RCP 8.5
mm
44944
40000
30000
20000
10000
0
Latitude (°)
2
0
-2
-4
Longitude (°)
10
12
14


PET current
1.13
1.00
0.80
0.60
0.48
PET anomaly 2085 RCP 4.5
0.16
0.12
0.08
0.04
0.00
-0.04
-0.08
PET anomaly 2085 RCP 8.5
PET anomaly QDRT 2085 RCP 4.5
0.026
0.020
0.015
0.010
0.005
0.000
PET anomaly QDRT 2085 RCP 8.5
DM current
months
5
4
3
2
1
DM anomaly 2085 RCP 4.5
DM anomaly 2085 RCP 8.5
DM anomaly QDRT 2085 RCP 4.5
DM anomaly QDRT 2085 RCP 8.5
LLDS current
LLDS anomaly 2085 RCP 4.5
LLDS anomaly 2085 RCP 8.5
LLDS anomaly QDRT 2085 RCP 4.5
LLDS anomaly QDRT 2085 RCP 8.5
Latitude (°)
Longitude (°)
10
12
14
2
0
-2
-4

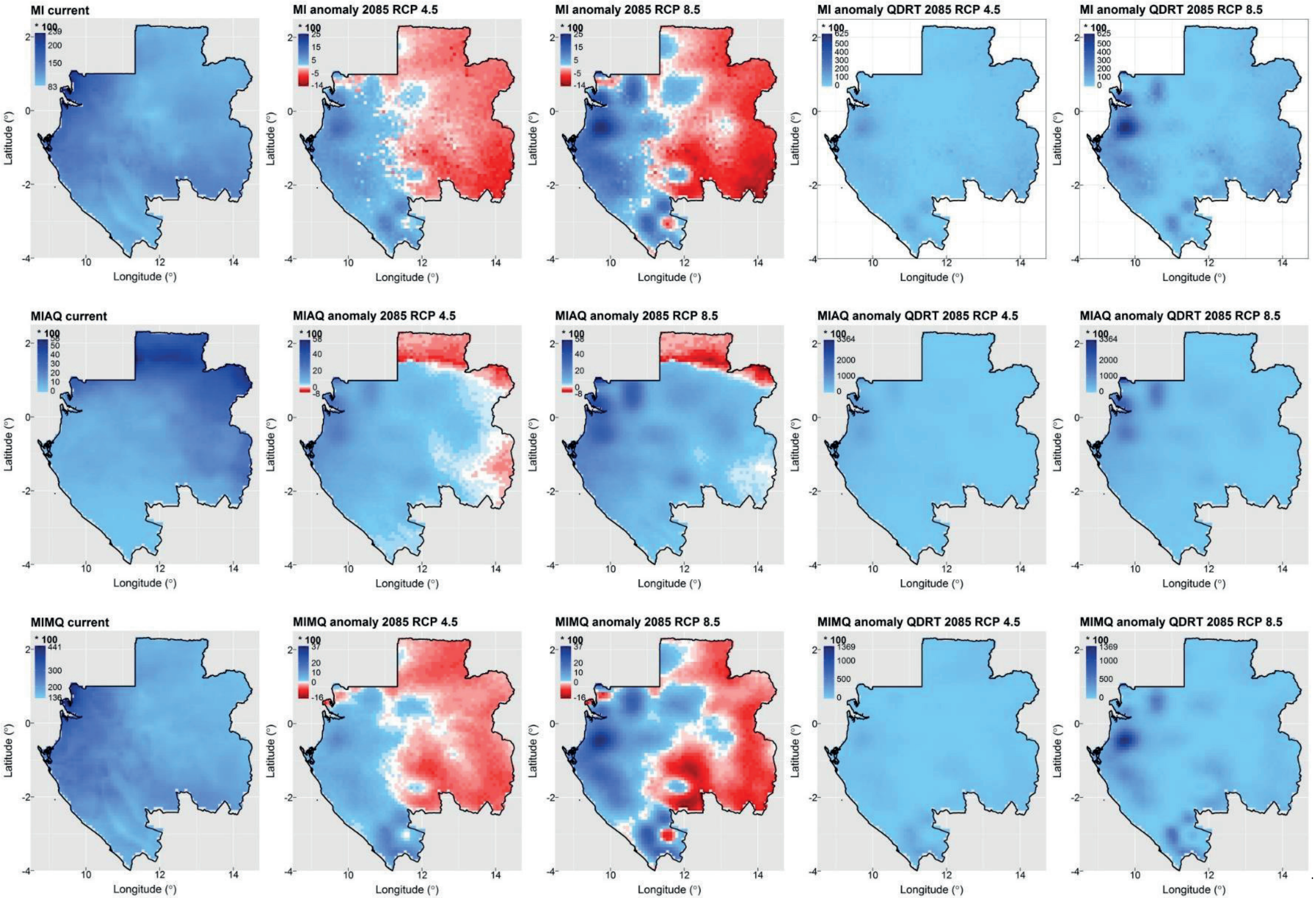

MI current
* 100
239
200
150
83
MI anomaly 2085 RCP 4.5
* 100
25
15
5
-5
-14
MI anomaly 2085 RCP 8.5
* 100
25
15
5
-5
-14
MI anomaly QDRT 2085 RCP 4.5
* 100
625
500
400
300
200
100
0
MI anomaly QDRT 2085 RCP 8.5
* 100
625
500
400
300
200
100
0
MIAQ current
* 100
58
50
40
30
20
10
0
MIAQ anomaly 2085 RCP 4.5
* 100
58
40
20
0
-8
MIAQ anomaly 2085 RCP 8.5
* 100
58
40
20
0
-8
MIAQ anomaly QDRT 2085 RCP 4.5
* 100
3364
2000
1000
0
MIAQ anomaly QDRT 2085 RCP 8.5
* 100
3364
2000
1000
0
MIMQ current
* 100
441
300
200
136
MIMQ anomaly 2085 RCP 4.5
* 100
37
20
10
0
-16
MIMQ anomaly 2085 RCP 8.5
* 100
37
20
10
0
-16
MIMQ anomaly QDRT 2085 RCP 4.5
* 100
1369
1000
500
0
MIMQ anomaly QDRT 2085 RCP 8.5
* 100
1369
1000
500
0
Latitude (°)
2
0
-2
-4
Longitude (°)
10
12
14

Figure S3: Spatial autocorrelation. *Presented is spatial autocorrelation as a function of lag distance for species gain (a), loss (b) and turnover (c) assuming full dispersal and of species turnover assuming no dispersal (d) under the representative concentration pathway 8.5 (red lines and squares), as well as for the residuals of the stepwise multiple regression model with climate anomalies and uncorrelated quadratic terms of climate anomalies (blue lines and triangles).*

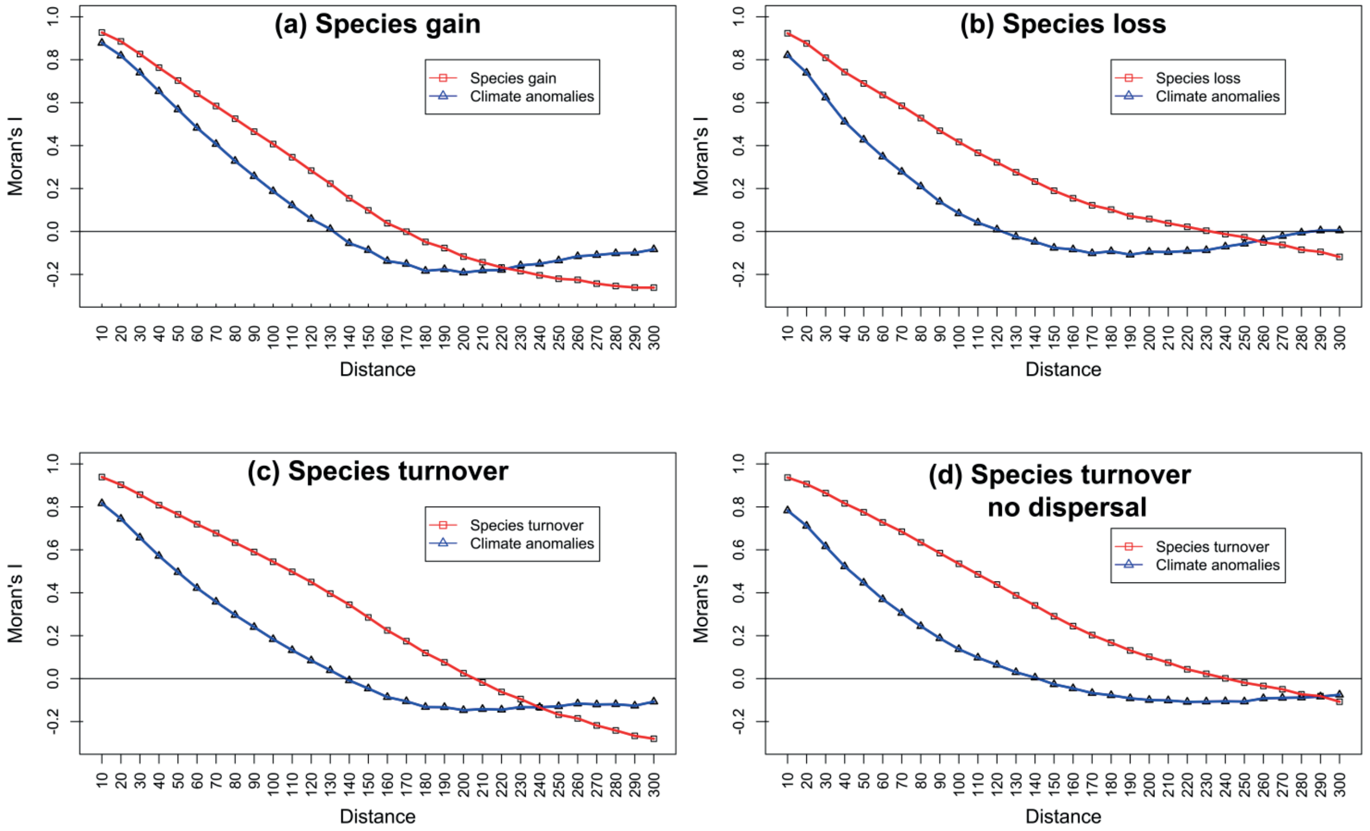


Table S3: Correlation with climate anomalies. *Presented are coefficients and explanatory power (adjusted R^2) of individual uncorrelated climate anomalies and quadratic terms by 2085 under the representative concentration pathway 4.5 to patterns of species gain (a), loss (b) and turnover (c) assuming full dispersal and of species turnover assuming no dispersal (d). Anomalies are ordered on their explanatory power in the final stepwise multiple regression model and coefficients of this final model are given. Variables not contributing significantly and therefore excluded from the final model are omitted from this table. The intercept and adjusted R^2 of the final model are included. Uncorrelated anomalies include mean annual temperature (Bio01), Isothermality (Bio03 = mean diurnal temperature range / annual temperature range), temperature seasonality (Bio04, standard deviation over monthly values), minimum temperature of the coldest month (Bio06), precipitation of the driest month (Bio14), number of dry months (DM), length of the longest dry season (LLDS), potential evapotranspiration (PET), quadratic values of the mean diurnal temperature range (Bio02), quadratic values of annual temperature range (Bio07), quadratic values of the number of dry months (DM^2), quadratic values of the length of the longest dry season (LLDS^2), and quadratic values of the annual moisture index (MI^2).*

(a) Gain			**(b) Loss**			**(c) Turnover full dispersal**			**(d) Turnover no dispersal**		
Full model	Intercept	$R_{adj.}^2$	Full model	Intercept	$R_{adj.}^2$	Full model	Intercept	$R_{adj.}^2$	Full model	Intercept	$R_{adj.}^2$
	68.765	0.241		-244.521	0.504		-18.154	0.428		-39.222	0.527
	Coefficient	$R_{adj.}^2$		Coefficient	$R_{adj.}^2$		Coefficient	$R_{adj.}^2$		Coefficient	$R_{adj.}^2$
Bio17	-0.864	0.068	Bio17	1.951	0.269	Bio07^2	0.424	0.075	Bio03	-0.239	0.130
Bio07^2	2.613	0.057	Bio03	-2.221	0.075	Bio03	-0.200	0.066	Bio02	1.506	0.086
Bio15	7.048	0.035	Bio02	12.653	0.039	Bio01	2.689	0.050	Bio17	0.146	0.053
MI^2	-0.317	0.030	MI^2	-0.485	0.030	MI^2	-0.062	0.042	Bio04	0.880	0.051
Bio01	4.446	0.015	Bio01	3.210	0.029	DM^2	3.498	0.042	Bio01	1.839	0.045
DM^2	13.715	0.013	Bio06	9.928	0.026	Bio04	1.276	0.039	MI^2	-0.052	0.044
Bio03	-0.553	0.012	Bio15	2.434	0.014	Bio02	1.067	0.036	DM^2	1.470	0.042
LLDS	-7.874	0.004	Bio07^2	-1.003	0.012	Bio17	-0.026	0.035	Bio06	0.288	0.026
Bio04	-3.335	0.004	Bio04	2.064	0.011	Bio06	-0.683	0.023	Bio07^2	0.092	0.024
Bio02	1.345	0.004				LLDS	1.819	0.012	Bio15	0.298	0.022
						Bio15	0.768	0.010	LLDS	1.309	0.004

Chapter 6

General discussion

The past three centuries of exploration of life on Earth produced a vast amount of data on the distribution of species. These data are captured in millions of natural history specimens curated in herbaria and musea. In the past two decades, data on such collections have become digitally available at an ever increasing speed (Blagoderov *et al.*, 2012, Graham *et al.*, 2004). This development coupled with novel methods and an exponential increase in computational power offered opportunities never experienced before to generate spatial distribution models for species, to infer patterns of biodiversity using these models, and to assess the driving forces of these patterns (Franklin, 2009, Lomolino *et al.*, 2010). A better understanding of these biodiversity patterns and their causes is crucial in setting priorities for conservation and to efficiently spend the limited resources available in the current era of global change (Boitani *et al.*, 2011, Meyer *et al.*, 2016, Newbold, 2010).

In the next paragraphs, I place the results presented in this thesis in a broader context. First, I reflect on the contributions I made to the pipeline of methods to generate species distribution models (SDMs) and infer diversity patterns from them. In addition, I highlight possible lines of future research for which the novel methods presented in this thesis offer opportunities. Secondly, I reflect on the factors that shape current and future patterns of botanical diversity in Gabon, Central Africa, as well as the factors that influence our knowledge of these patterns and causes. I discuss how the results presented in this thesis contribute to a better understanding of Central African botanical diversity patterns and will aid in setting priorities in conservation. Thirdly, I comment on the future of African rainforests and plants in general and those of Gabon in particular with respect to the findings of this thesis and discuss relevant conservation aspects. I conclude by commenting on the applicability of SDMs in a wider scientific and societal context.

A pipeline for generating SDMs

In chapter 2 and 3 of this thesis, I contribute to the further improvement of the pipeline for generating accurate SDMs and inferring diversity patterns based on SDMs (Fig. 2, Chapter 1). I address two methodological matters addressed below using simulated species, following the virtual ecologist approach (Zurell *et al.*, 2010). The use of simulated species offers opportunities to quantify the effects of these and other aspects in a fully controlled environment, both individually as well as in concert (Miller, 2014, Saupe *et al.*, 2012). The novel method I used to simulate species in either a virtual environment or in a selected study area, here Central Africa, offers opportunities to quantitatively assess matters of the pipeline mentioned in the Introduction and addressed in more detail below.

I – Sample size

In chapter 2, I quantify the effect of sample size on model accuracy. Sample size has been shown to have a substantial effect on model accuracy (Hernandez *et al.*, 2006, Wisz *et al.*, 2008), although so far, this has only been assessed by comparing performance of models based on subsets of species occurrence records of decreasing sizes. Still, most researchers working with SDMs simply ignore the effect of sample size on model accuracy, although some apply a minimum sample size for all species to be modelled, often arbitrarily chosen without further explanation or valorisation. Using simulated species, I show that the minimum sample size required to generate accurate SDMs differs for species of different prevalence or range size classes. As could be expected, more widespread species require more occurrence records. Thus, ignoring the reported differences in data requirements for widespread and narrow-ranged species by applying a uniform minimum sample size will often lead to erroneous acceptance of SDMs for widespread species and erroneous rejection of SDMs of narrow-ranged species. As the required minimum number of records also depends on the specific study area, I present a novel method using simulated species that enables the identification of this minimum sample size for species of every possible prevalence class and for every possible study area (chapter 2). One could argue that testing whether an SDM deviates significantly from random chance by applying a null model test (Raes & ter Steege, 2007), qualifies the application of a range size dependent minimum sample size as redundant. However, in chapter 4 and 5 I show that these two criteria are not mutually exclusive, underlining the importance of applying range size dependent minimum sample sizes. Other factors addressed in more detail below are likely to influence the minimum sample size too.

II – Range size

Range size or prevalence is an important feature of species. An accurate estimation of this parameter is crucial for conservation, biogeographical and macroecological research (Gaston, 2003). In addition, as I show in chapter 2, the prevalence of species also determines the minimum number of records required to generate accurate SDMs. In chapter 3, I present a novel method to estimate the range size or prevalence of a species. Using simulated species, I quantitatively evaluate the accuracy of this new method as well as that of ten existing methods. I show that traditional, spatial methods to estimate range sizes are clearly outperformed by estimators operating in parameter space. The novel method presented in chapter 3 produces the most accurate estimations. These results challenge the current IUCN recommendations for methods to estimate the Extent of Occurrence (EOO) and Area of Occupancy (AOO) of species for the purpose of Red List assessments (Bachman *et al.*, 2011, IUCN, 2014). I advocate to reconsider the use of these methods and instead apply the novel method presented in chapter 3 to estimate the AOO and to estimate the EOO from thresholded SDMs. One aspect limiting the accuracy of each of the assessed methods is the spatial occurrence or clustering of individuals within a raster cell. At microscale, the occurrence of species inside an area obviously differs, with some species being clustered in small pockets only, whereas others are scattered over the entire raster cell. This scale-dependent, time-dependent, and species-specific difference might hamper the effective use of indicators of species range size and warrants additional assessments.

III – Spatial resolution

Spatial resolution strongly influences the accuracy of SDMs (Guisan & Thuiller, 2005) and its impact appears to be driven by contradicting factors. The use of larger cell sizes results in more accurate models as in smaller cells non-climatic factors such as biotic interactions and stochastic events play an important role (Kadmon *et al.*, 2003). In addition, not all species for which the area is suitable can actually be continuously present in each particular cell due to the carrying capacity of a community or area-dependence of species richness (Guisan & Rahbek, 2011). On the other hand, loss of information on e.g. habitat heterogeneity when using larger cells reduces model accuracy at coarser spatial resolution (Pearson *et al.*, 2004, Rengstorf *et al.*, 2012). A recent study assessed the effect of spatial scale on the area predicted suitable for 52 Californian plant species under climate change and reported only modest agreement between the results under different spatial resolutions (Franklin *et al.*, 2013). A possible future line of research is a quantitative assessment of the impact of spatial resolution using multiple species for which the true distribution is known. The use of simulated species will enable such a quantification of the effect of spatial resolution on model accuracy in the following way. First, species are simulated at the highest possible spatial resolution. Then, the thus defined occurrences are aggregated in a series of lower spatial resolutions. SDMs at each respective spatial resolution are then generated from records sampled from that spatial resolution. By comparing the modelled distribution with the defined distribution, the effect of spatial resolution on model accuracy is quantified.

IV – Collecting bias

Bias in species occurrence data is an often mentioned, but rarely quantified matter. It is well-known that most species occurrence data from herbaria and other natural history museum collections are biased towards specific taxa, periods in history, countries, and easy accessible areas such as near roads, cities and rivers, as well as in national parks

(Beck *et al.*, 2014, Reddy & Davalos, 2003). These biases may well lead to such records incompletely and inaccurately representing the niche of the species, which negatively affects the accuracy of SDMs (Feeley *et al.*, 2016, Hortal *et al.*, 2008). Collecting bias can be reduced through spatial filtering, or preferably by ecological filtering of the species occurrence data (Kramer-Schadt *et al.*, 2013, Varela *et al.*, 2014). When sample size does not allow filtering, the same bias can be applied to the background data (Syfert *et al.*, 2013). Some studies assessed the effect of bias and methods to compensate for this bias, but these studies are based on few individual species only (see e.g. Fourcade et *al.*, 2014). Unfortunately, an exact quantification of the bias as well as of the bias reduction by these methods based on an assessment of a large number of species is still lacking. Future research could address this matter, here again, by using simulated species. By using simulated species and bias files based on historical sampling localities, the effects of the taxonomic, temporal and spatial bias on model accuracy can be quantified for every possible study area and taxonomic group and similarly, the effect of data filtering, target-group background sampling, or other methods to correct for bias can be quantified. In addition, the effect of each type of collecting bias on the required minimum sample size as well as on the accuracy of estimated of EOO and AOO can be quantified.

V – Spatial errors

In addition to bias, the effect of spatial errors in species occurrence data on model accuracy warrants a quantitative assessment. Species records are known to contain spatial errors due to i.a. inaccuracy of the measurement in the field, rounding of coordinates, as well as errors made by the retrospective interpretation (georeferencing) or operating of often imprecise label information (Graham *et al.*, 2004, Newbold, 2010, Wieczorek *et al.*, 2004). Spatial errors in species records have been found to substantially affect estimations of species traits as well as species distributions (Feeley & Silman, 2010b), particularly when models are transferred in time and space (Gould *et al.*, 2014). However, modelling algorithms vary in their sensitivity to such errors (Graham *et al.*, 2008). High-performing models can be generated from data that contain minor spatial errors (Mitchell *et al.*, 2017), particularly if occurrence data originate from areas with high spatial autocorrelation in the environmental variables (Graham *et al.*, 2008, Naimi *et al.*, 2011). Where these studies used data of real species that were degraded by applying additional spatial error, Velasquez Tibata *et al.*, (2015) and Naimi *et al.*, (2011) used simulated species in a simplified virtual landscape. Nonetheless, a quantification of the effect of spatial error on model accuracy in a real study area for large groups of species for which the true distribution in known, is still lacking. The novel method to simulate species presented in chapter 2 offers opportunities for such quantitative assessments. First, spatial errors of different sizes are applied to records sampled from simulated species. Then, differences in model accuracy due to these spatial errors are quantified by comparing the modelled spatial distributions based on either the original or the manipulated records with the defined distribution. Finally, the additional effect of spatial errors on the minimum number of records required for generating accurate SDMs as well as on the accuracy of estimations of EOO and AOO can be quantified.

VI – Dispersal limitations

Species distribution models are built on the assumption that species are in equilibrium with their environment, thus filling their entire niche and entire range. However, this is often not the case due to biotic interactions (competition, pollination, food availability, etc.) or dispersal (migratory) limitations including historical constraints. This concept is conceptualized by the BAM diagram (Biotic, Abiotic and Migratory factors) of Soberón et Peterson (2005). The effect of dispersal limitations on range filling are e.g. illustrated by data on European tree species facing a post glacial dispersal lag (Svenning & Skov, 2004). Model performance has been shown to be strongly affected by dispersal limitations (Saupe *et al.*, 2012), but can be improved significantly by incorporating distance constraints in the model (Allouche *et al.*, 2008). Alternatively, the study area can be defined so that it represents the area that is actually accessible for the modelled species (M in the BAM diagram) (Anderson & Raza, 2010, Barve *et al.*, 2011). Related to the above mentioned equilibrium assumption, as well as the matter of collecting bias, is the assumption that species occurrence data represent the entire niche of the species. This assumption is often violated, when species occurrence data are absent from specific areas due to e.g. incomplete range fill by the species, inaccessibility of areas for researchers, or unavailability of data due to incomplete digitization of specimen data from herbaria and natural history museum collections. Models trained on such data sets not representing the species' entire niche can be biased (Raes, 2012). The use of simulated species enables a quantitative assessment of the effect of incomplete niche sampling and dispersal limitations on model accuracy. This can be done e.g. by defining simulated species under conditions that prevailed during the last glacial maximum (LGM, 18,000 year BP) as well as under current climatic conditions and applying dispersal limitations of different magnitudes from localities defined suitable in both current and LGM climatic conditions. The effect of dispersal limitations on niche fill and range fill, as well as the effect on model accuracy can then be quantified. In addition, the minimum number of records required for generating accurate SDMs (chapter 2) can be identified for species facing different levels of dispersal limitations. Recently developed *R* packages such as *KISSMig* (Nobis & Normand, 2014) or *MIGCLIM* (Engler & Guisan, 2009, Engler *et al.*, 2012) apply iterative migration functions that mimic range filling through time. However, the lack of data on the dispersal kernel of most species including the rare, but important events of long distance dispersal strongly limits the applicability of these methods.

VII – Stochastic effects

Whether a species is actually present in a raster cell for which a high habitat suitability is predicted, depends i.a. on dispersal limitations and biotic interactions, but also on temporal and stochastic effects. Temporal effects are best illustrated by the spatial distribution of widespread species with a wide ecological niche but low abundance (Rabinowitz, 1981), for which the actual presence, particularly for long-living species, should be considered over a longer temporal window. Similarly, the gradual response of many species to the suitability of habitats is ignored when a threshold is applied to SDMs to convert gradual habitat suitability values into discrete presence/absence values (Meynard & Kaplan, 2012). The influence of such stochastic effects on the minimum number of records required to generate accurate SDMs as well as on the accuracy of EOO and AOO estimations is unknown, but assumed to be substantial (Syfert *et al.*, 2014). A quantification of these effects can be obtained by repeating the analyses of chapters 2 and 3 with the inclusion of different levels of stochasticity in the defined presences and absences of the simulated species.

VIII – Inferring patterns of species richness

In this thesis I inferred patterns of species richness by superimposing (stacking) SDMs to which a threshold is applied and then summing the number of predicted presences in each raster cell (chapter 4 & 5). Notwithstanding its wide application to different taxonomic groups and study areas at various spatial resolution, this method (S-SDMs) has been criticized for consistently overestimating species richness (Calabrese *et al.*, 2014, D'Amen *et al.*, 2015a, Mateo *et al.*, 2012), although D’Amen *et al.* (2015b) did not report overpredictions. It should be noted though that overprediction is relative to the number of species for which significant SDMs could be generated, as often for many species insufficient records are available or models do not perform significantly better than random chance (chapter 4 and 5). Such overprediction is suggested to be caused by dispersal limitations, biotic interactions and constraints related to the carrying capacity and dynamics of the community (Guisan & Rahbek, 2011), by a methodological bias resulting from the selection of threshold (Calabrese *et al.*, 2014), or by overprediction of individual species’ distributions related to low quality of SDMs (D'Amen *et al.*, 2015b). More realistic values of species richness are said to be computed by stacking SDMs and summing the unthresholded habitat suitability values (P-SDMs), instead of binary presence/absence scores (Aranda & Lobo, 2011, Dubuis *et al.*, 2011), although at the cost of not knowing the identity of the species present in individual raster cells. However, in studies using well-sampled sites, P-SDMs are found to overpredict species richness at species-poor sites and underpredict richness at species-rich sites (Calabrese *et al.*, 2014, Dubuis *et al.*, 2011). Alternatively, species richness patterns can be inferred by using macroecological models (MEM) that model richness directly as a function of environmental factors (Gotelli *et al.*, 2009). Similar to P-SDMs, MEM tend to overestimate species richness in species-poor sites and underestimate it for species-rich sites (Calabrese *et al.*, 2014). Finally, recently developed methods integrate S-SDMs with MEM. Species richness values based on S-SDMs are constrained by richness values generated through MEM, reflecting the effect of biotic filtering (Calabrese *et al.*, 2014, D'Amen *et al.*, 2015a, Gavish *et al.*, 2017, Guisan & Rahbek, 2011). However, the number of studies verifying such predictions with observed data from well-sampled areas is limited and this certainly warrants more assessments (D'Amen *et al.*, 2017).

Other new methodological extensions of the SDM concept include dynamic SDMs that incorporate temporal variation in the distribution of species (Merow *et al.*, 2011), joint SDMs that generate distribution models for multiple species simultaneously to more accurately estimate the occupancy or density of rare species (Dorazio & Royle, 2005, Thorson *et al.*, 2015), geostatistical models that include aspects of spatial autocorrelation (Conn *et al.*, 2015), and joint dynamic SDMs that account for all these three aspects (Thorson *et al.*, 2016). Such joint mechanistic models can help to address some of the matters described above, e.g. species detectability, collecting bias as well as changes in species abundance and distribution due to climate change. However, the aspects assessed in this thesis address the presence or absence of species in a specific raster cell and not their respective abundance. When addressing biogeographical questions like the ones I address in this thesis, the identity of the species in each raster cell is crucial and hence S-SDMs are required. Furthermore, estimations of species richness based on S-SDMs are strongly correlated to observed species richness (Calabrese *et al.*, 2014), justifying the use of S-SDMs to infer patterns of species richness, although with the consideration that absolute richness values may or may not be overestimated.

What drives plant species richness in Central Africa?

Gabonese patterns of plant species richness and endemism

Using significant SDMs of over 2,000 plant species, I inferred patterns of botanical species richness and of weighted endemism for Gabon. These are largely congruent with previously published theories on botanical species richness patterns in Gabon, as summarized in the Introduction of this thesis. To the previously hypothesized centres of species richness located in the Crystal Mountains, western parts of the Chaillu Massif and the Doudou Mountains (Maley, 1996), as well as the wider vicinity of Libreville, I add the Pelé and Mabanda Mountains that extend from the Doudou Mountains southward towards Congo-Brazzaville (chapter 4 & 5).

Centres of weighted endemism as presented in chapter 4 are largely congruent with these centres of species richness and in addition confirm previously reported high levels of endemism for the Monda forest northwest of Libreville (Walters *et al.*, 2016) as well as the wider vicinity of Libreville (Lachenaud *et al.*, 2013). Centres of residual weighted endemism show higher levels of weighted endemism than expected based on the total number of species present there. These include the centres of species richness mentioned above. Next, the high values of residual weighted endemism for the Loango National Park in the coastal region south of one degree south latitude confirm the presence of an exceptional high number of species in this otherwise rather species-poor area that are not found elsewhere (Harris *et al.*, 2012, Wieringa & Sosef, 2011).

The centres of species richness and of weighted endemism presented in this thesis coincide well with previously hypothesized LGM forest refugia in the Crystal Mountains, western parts of the Chaillu Massif and Doudou Mountains (Maley, 1996, Pietsch & Gautam, 2013, Sosef, 1996). In order to further improve the understanding of Quaternary rainforest dynamics in Africa, studies on the location of hypothesized LGM forest refugia are recommended, preferably by inferring the LGM species richness pattern using SDMs. Unfortunately, until today, this is hampered by the lower accuracy of LGM climate data models in contrast to the higher accurate current and future climate data models such as the AFRICLIM data (Platts *et al.*, 2015).

Congruence of S-SDMs with patterns of genetic diversity

The results of recent genetic studies offer knowledge on the localities of LGM forest refugia from a genetic perspective. As most plant species evolved before the Quaternary, variation within-species rather than among-species patterns can provide detailed information about Late

Quaternary rainforest dynamics. A large number of studies showed that spatial patterns of genetic diversity are correlated with patterns of species richness, the so-called species-gene diversity correlation (SGDC, (Vellend, 2003), reviewed by Vellend *et al.*, 2014). Based on the SGDC, genetic diversity has been suggested as an indicator of species richness (Papadopoulou *et al.*, 2011). Genetic diversity within species has been used to identify centres of genetic diversity (Ley *et al.*, 2014), colonization routes (Tian *et al.*, 2015), and genetic relatedness of populations and areas (Demenou *et al.*, 2016). The theory of rainforest contraction into forest refugia during the LGM followed by expansion is supported by a variety of phylogeographic studies on rainforest tree species (reviewed by Hardy *et al.* 2013). The existence of such isolated forests in Gabon during the LGM is supported by genetic data of various tree species, e.g. *Greenwayodendron suaveolens* (Dauby *et al.*, 2010), *Santiria trimera* (Koffi *et al.*, 2011), and *Erythrophleum* species (Duminil *et al.*, 2010). Similar results were found for gorillas in Central Africa (Anthony *et al.*, 2007, Clifford *et al.*, 2004). Genetic discontinuities between areas with hypothesized LGM forest refugia have been reported for a selection of non-related Central African plant species (Dauby *et al.*, 2014, Faye *et al.*, 2016, Hardy *et al.*, 2013). The patterns of genetic diversity in these studies are largely congruent with the pattern of species richness presented in chapter 4 & 5. Bringing together palynological data, SDMs and phylogeographic studies has been recommended to identify past climate refugia (Gavin *et al.*, 2014). In line with this recommendation, as a possible line of future research, I suggest to assess the congruence of genetic and floristic resemblance between hypothesized LGM forest refugia, e.g. by following the Relative Floristic Resemblance method of Wieringa & Sosef (2011).

Contribution of narrow-ranged species to diversity patterns

In this thesis, I show that narrow-ranged species contribute differently to patterns of species richness and weighted endemism than widespread species and that the former are overrepresented in species-rich areas. This confirms previous findings on plants for other regions (Raes *et al.*, 2009), as well as for e.g. birds in Sub-Saharan Africa (Jetz *et al.*, 2004). However, in chapter 4, I show that it is important to clearly define what is meant by narrow-ranged and widespread. It might seem obvious to use the full range of species to define their range size or prevalence. However, unfortunately, too often, researchers use study areas limited by political, rather than natural boundaries and hence ignore large parts of species' ranges.

The results presented here are based on SDMs of a subset of all Gabonese plant species, as for 75% of the species no significant SDM could be generated. It remains unclear how species for which no significant SDM could be generated (due to lack of data) contribute to the inferred patterns of species richness and weighted endemism. As most of these species have few records, the majority is expected to be narrow-ranged, although some will be widespread but represented by too few records. To overcome this problem, species of which the collecting localities are clustered within few raster cells at low spatial resolution could be modelled at a higher spatial resolution if this results in a higher number of records available for training the model. Alternatively, giving priority to digitize herbarium specimen records of these species or collect additional records in the field further increases the available number of records. When significant SDMs can be generated for all species, I expect the species richness gradient in Gabon to be steeper than the gradient presented in chapters 4 and 5.

SDMs in chapter 4 and 5 for current climate conditions are generated without considering dispersal limitations, thus assuming full range fill. In reality, many species will not have been able to fill their entire potential current range. As the current centres of species richness and of weighted endemism are congruent with hypothesized LGM forest refugia, I expect the true ranges of species with limited dispersal capacity to be restricted towards these LGM refugia. This leads to levels of species richness outside the LGM forest refugia being lower than predicted in this thesis. As the levels of species richness inside these refugia is less affected by dispersal limitations, the predicted gradient in species richness and in weighted endemism may well be steeper than presented here. For future climate scenarios, the effect of dispersal limitations has been addressed in chapter 5.

African tropical rainforest diversity: the odd man out

The question what drives plant species richness in African rainforests takes a central position in macroecological and conservation research. Although the latitudinal diversity gradient (Rosenzweig, 1995), driven primarily by higher energy levels in the Tropics (Brown, 2014), explains general higher levels of species richness in the Tropics, it does not explain the relative lower richness of African rainforests compared to Neotropical and Asian ones. Africa's position as "odd man out" is still not fully understood (Parmentier *et al.*, 2007). Couvreur (2015) reviewed this topic and highlighted some important but often neglected aspects, including the need to correct for area as well as to include all species instead of only trees with DBH higher than 10 cm. Recently, the role of megafauna on species richness in African rainforests has been addressed, but no clear picture has yet emerged (Terborgh *et al.*, 2016a, Terborgh *et al.*, 2016b).

Some other aspects may play a role too in explaining the lower species richness of African rainforests. The African rainforest region is much smaller in size and much more fragmented than the larger and continuous Amazonian rain forest. These differences in size and shape may well contribute to the lower species richness of African rainforests.

When comparing species richness of Upper Guinean, Lower Guinean and Congolian rainforests, a few aspects may account for the higher level of species richness of Lower Guinean rainforest. First, Lower Guinean rainforests share many endemic species with either Congolian or Upper Guinean rainforest and hence the geographical central position of the Lower Guinean rainforest region may in part explain its higher species richness. Second, the Upper Guinean rainforest region is smaller than either the Lower Guinean or the Congolian rainforest region. The smaller area of Upper Guinean rainforest region may host fewer species. Thirdly, the elongated shape of the Upper Guinean rainforest region may contribute to its lower species richness compared to the Lower Guinean and Congolian rainforests regions that each have a more compact shape. Finally, the East-West orientation of the Upper Guinean coastline hampered southwards migration of species during arid glacial periods and thus may well have led to increased levels of extinction, possibly explaining the lower species richness of Upper Guinean rainforests. Phylogenetic diversity assessments of each of these rainforest regions may address potential differences in extinction levels between the African rainforests regions.

To this debate on the relative species-poorness of African rainforests and differences of species richness between individual African rainforest regions, I contribute by addressing the potential role of niche size of individual species and of the total available niche space in each rainforest region. A potentially important difference in climatic conditions between the tropical rainforests of the three continents is that Africa is largely lacking the mega-wet conditions that are widely present in the Neotropics and South-East Asia (Malhi & Wright, 2004). On the one hand, for areas of medium precipitation and temperature values (1,500-2,500 mm and 22-25 °C), Parmentier *et al.* (2007) observed levels of alpha diversity in African rainforests that are similar or even higher than those for Amazonian rainforests. However, on the other hand, differences in the total available niche space in each rainforest region may explain the lower species richness of African rainforest in several ways. To assess this, available niche space can be quantified by applying niche density kernels following the method of Broennimann *et al.* (2012) or the novel method presented in chapter 3 to all raster cells of each rainforest region plotted in multidimensional parameter space. First, African rainforests may be species-poor due to smaller current available niche space. Secondly, they may be species-poor due to smaller available niche space during the LGM, which can be assessed by comparing available niche space in that period across the different tropical regions. Thirdly, they may be species-poor due to a shift of available niche space in Asia and the Neotropics towards warmer and wetter conditions, thus with higher energy levels. An alternative approach is to assess if the niche size of individual species of the respective rainforest regions may explain the differences in species richness between the three rainforests. If species in Africa tend to have wider niches, this potentially allows for fewer species to co-exist.

The future of the Gabonese rainforest and its plant species

Safeguarding Gabonese rainforests

In Africa, the increase of human populations and/or the level of their prosperity will come at the cost of the loss of tropical rainforests. This will be the case for Gabonese rainforests too, although not as severely as for West African rain forests (Poorter *et al.*, 2004). In the latter, the rainforests in some unprotected areas have been dramatically reduced, best illustrated e.g. by the 79% loss of forest in the Guiglo-Taï region of Ivory Coast (Chatelain *et al.*, 1996). The future of Gabonese rainforests is influenced by factors that at least partially have a contrasting effect. First, recent data on climate change show an increase in precipitation (Asefi-Najafabady & Saatchi, 2013) and recent models predict a continuation of this increase (Platts *et al.*, 2015), favouring forest growth in current savannahs. However, in chapter 5, I show that this increased precipitation, particularly in the dry season, results in species loss and high levels of species turnover. Loss of habitat due to logging, mining, urbanization and agriculture including large-scale production of palm-oil will contribute to a decreasing extent of rainforest. In contrast, the reduced number of large herbivores or even extinction of megafauna lowers the grazing pressure on trees and hence favours the recruitment of new trees, although the impact on species composition, community dynamics and ecosystem functions remains largely unknown (Malhi *et al.*, 2016). All in one, the future of Gabonese rain forests remains unknown, but clearly, some of the main threats have been identified and important steps towards long-term conservation have been taken. The current network of national parks and other protected areas in Gabon hosts a wide variety of biotopes and many of its species. Nevertheless, the status of national park does not fully protect species against poaching, harvesting or human-induced fires. In addition, an assessment of the effectiveness of the current network of protected areas in the light of systematic conservation planning (Margules & Pressey, 2000) deserves high priority. Such assessments have been done for several areas and taxa including e.g. amphibians (Chen *et al.*, 2017) and vascular plants in China (Zhang *et al.*, 2012), terrestrial vertebrates in Europe (Maiorano *et al.*, 2015), and animals and plants in Guyana (McPherson, 2014), and medicinal plants in Egypt (Kaky & Gilbert, 2016), to name but a few. The SDMs of Gabonese plant species presented in this thesis enable such an assessment and the identification of priority areas for conservation with recommendations for potential additional protected areas.

Safeguarding Gabonese plant species

In the light of global climate change, the future of individual Gabonese plant species largely depends on their ability to either adapt to changes in their current biotope or to migrate in time to suitable future habitats. In chapter 5, I show that dispersal limitations pose a severe additional threat to the future survival of plant species in Gabon. This is expected to be even more the case for the species for which no significant SDM could be generated, representing 75% of the total number of species, and the majority of which is expected to be narrow-ranged.

A crucial step in the conservation of Gabonese plant species is an IUCN Red List assessment (IUCN, 2001, IUCN, 2014). For most species the level of threat is unknown, underlining the "urgent need to produce fast, objective and consistent Red List assessment" (Willis *et al.*, 2003). However, traditional Red List assessments are time-consuming and therefore the majority of plant species worldwide has not been assessed yet. An efficient and reliable method to execute a preliminary Red List assessment for large numbers of species would enable a quick identification of those species most probably being classified in the highest risk categories. The use of SDMs and the estimators of species range size or prevalence discussed in chapter 3 offer opportunities for such rapid, preliminary Red List assessments. First, an SDM is generated and the EOO is computed using the SDM to which a threshold is applied. The next step is to compute the AOO, defined as the predicted fraction of raster cells where the species is predicted present. Third, the number of populations is counted using both the predicted presences and the known occurrences. The predicted decline (or increase) in AOO and number of populations due to climate change and change of land use is computed by joining SDM projections to future climate scenarios and maps of foreseen future land use. These basic data enable a rapid, preliminary IUCN Red List classification for species that have sufficient records to generate significant SDMs. This may aid in listing many species with few data in an appropriate Red List category rather than as 'data deficient' as most of these species will be narrow-ranged, understudied and facing a higher risk of extinction that other species (Roberts *et al.*, 2016).

Novel applications of SDMs

Nowadays, SDMs have been widely used to infer the spatial distribution of indigenous and invasive species, assess the impact of climate change on these distributions, identify geographic areas where a target species may be found (gap analysis), as well as to infer patterns of species richness and weighted endemism and quantify the effectiveness of protected areas in safeguarding the future of species. However, other potential applications of SDMs are still in their infancy. Promising new applications of SDMs include modelling the distribution of phylogenetic lineages within species (D'Amen *et al.*, 2013), modelling the distribution of functional groups such as vegetation types (Elias *et al.*, 2016, Zhang *et al.*, 2013), or modelling the spatial distribution of social values for ecosystem services (Sherrouse *et al.*, 2014). The overall principle remains the same: known occurrences of the target, whether lineage, species, function, group or else is linked to high resolution spatial data on climate, soil, elevation, distance, biotic competitors, social, economic or other variables.

In agriculture, SDM received little attention so far, but may prove to be of great future value. In many tropical regions, including Central Africa, people still largely depend on non-timber forest products (NTFP) for food, medicine, and religious purposes. Such NTFP are mainly harvested from the wild, posing a threat to the survival of many NTFP species. A recent study on West-African medicinal plants used SDMs to map the potential distribution of 12 commercially important medicinal plants (van Andel *et al.*, 2015). As such, SDMs help to document the distribution of NTFPs and set priorities in their conservation for future generations. Similarly, SDMs have been used successfully to assess the distribution of crop wild relatives (CWR) and identify geographic locations for further collecting of genetic resources (Cobben *et al.*, 2014, Khoury *et al.*, 2015). Following the line of thought that crops grow bests at the smallest costs and smallest effort when the agricultural conditions are closest to their natural niche optimum, SDMs offer opportunities to identify areas suitable for the future production of crops or of NTFP if the latter are to be planted in semi-natural environments such as permacultures in forests. Such was done for coffee in Nicaragua (Laderach *et al.*, 2017), but studies on other crops and regions are likely to follow soon.

Concluding remarks

In this thesis, I contribute to a better documentation and understanding of biodiversity patterns in Gabon, Central Africa. The central question “What determines plant species richness in Gabon?” could only partly be answered. Obviously, a question stimulating researchers, governmental and non- governmental professionals, and nature lovers for centuries to explore Central Africa's tropical rainforest, cannot be fully answered within the scope of a single PhD project. Still, I have contributed to a better understanding by assessing for the first time the patterns of plant species richness and weighted endemism in Gabon using all plant species for which significant SDMs could be generated. This is the first step and provides a supportive framework for future biogeographical, macroecological and conservation research. In addition, as I showed in the first chapters of this thesis, methodological matters have a substantial impact of usually unknown size on SDMs. By using simulated species, I enabled a quantitative assessment of some of these matters. Thereby, I contributed to the production of better, more accurate SDMs and hence estimations of EOO and AOO. In return, these improved methodologies contribute in answering that enigmatic question: "What determines plant species richness in Gabon?".

References

Aguirre-Gutiérrez J, Carvalheiro LG, Polce C, Van Loon EE, Raes N, Reemer M, Biesmeijer JC (2013) Fit-for-Purpose: Species Distribution Model Performance Depends on Evaluation Criteria - Dutch Hoverflies as a Case Study. *Plos One,* 8, e63708.

Aguirre-Gutiérrez J, Serna-Chavez HM, Villalobos-Arambula AR, Pérez De La Rosa JA, Raes N (2014) Similar but not equivalent: ecological niche comparison across closely–related Mexican white pines. *Diversity and Distributions*, n/a-n/a.

Aiello-Lammens ME, Boria RA, Radosavljevic A, Vilela B, Anderson RP (2015) spThin: an R package for spatial thinning of species occurrence records for use in ecological niche models. *Ecography,* 38, 541-545.

Algar AC, Kharouba HM, Young ER, Kerr JT (2009) Predicting the future of species diversity: macroecological theory, climate change, and direct tests of alternative forecasting methods. *Ecography,* 32, 22-33.

Allouche O, Steinitz O, Rotem D, Rosenfeld A, Kadmon R (2008) Incorporating distance constraints into species distribution models. *Journal of Applied Ecology,* 45, 599-609.

Allouche O, Tsoar A, Kadmon R (2006) Assessing the accuracy of species distribution models: prevalence, kappa and the true skill statistic (TSS). *Journal of Applied Ecology,* 43, 1223-1232.

Amaral AG, Munhoz CBR, Walter BMT, Aguirre-Gutiérrez J, Raes N, De Cáceres M (2017) Richness pattern and phytogeography of the Cerrado herb-shrub flora and implications for conservation. *Journal of Vegetation Science*, n/a-n/a.

Anderson RP (2012) Harnessing the world's biodiversity data: promise and peril in ecological niche modeling of species distributions. *Ann N Y Acad Sci,* 1260, 66-80.

Anderson RP (2017) When and how should biotic interactions be considered in models of species niches and distributions? *Journal of Biogeography,* 44, 8-17.

Anderson RP, Raza A (2010) The effect of the extent of the study region on GIS models of species geographic distributions and estimates of niche evolution: preliminary tests with montane rodents (genus *Nephelomys*) in Venezuela. *Journal of Biogeography,* 37, 1378-1393.

Anhuf D, Ledru MP, Behling H *et al.* (2006) Paleo-environmental change in Amazonian and African rainforest during the LGM. *Palaeogeography Palaeoclimatology Palaeoecology,* 239, 510-527.

Anthony NM, Johnson-Bawe M, Jeffery K *et al.* (2007) The role of Pleistocene refugia and rivers in shaping gorilla genetic diversity in central Africa. *Proc Natl Acad Sci U S A,* 104, 20432-20436.

Aranda SC, Lobo JM (2011) How well does presence-only-based species distribution modelling predict assemblage diversity? A case study of the Tenerife flora. *Ecography,* 34, 31-38.

Araújo MB, Guisan A (2006) Five (or so) challenges for species distribution modelling. *Journal of Biogeography,* 33, 1677-1688.

Araújo MB, Pearson RG, Thuiller W, Erhard M (2005) Validation of species-climate impact models under climate change. *Global Change Biology,* 11, 1504-1513.

Araújo MB, Peterson AT (2012) Uses and misuses of bioclimatic envelope modeling. *Ecology,* 93, 1527-1539.

Asefi-Najafabady S, Saatchi S (2013) Response of African humid tropical forests to recent rainfall anomalies. *Philos Trans R Soc Lond B Biol Sci,* 368, 20120306.

Austin MP, Belbin L, Meyers JA, Doherty MD, Luoto M (2006) Evaluation of statistical models used for predicting plant species distributions: Role of artificial data and theory. *Ecological Modelling,* 199, 197-216.

Azaele S, Cornell SJ, Kunin WE (2012) Downscaling species occupancy from coarse spatial scales. *Ecological Applications,* 22, 1004-1014.

Bachman S, Moat J, Hill A, De La Torre J, Scott B (2011) Supporting Red List threat assessments with GeoCAT: geospatial conservation assessment tool. *Zookeys,* 150, 117-126.

Barthlott W, Hostert A, Kier G *et al.* (2007) Geographic patterns of vascular plant diversity at continental to global scales. *Erdkunde,* 61, 305-315.

Barve N, Barve V, Jiménez-Valverde A *et al.* (2011) The crucial role of the accessible area in ecological niche modeling and species distribution modeling. *Ecological Modelling,* 222, 1810-1819.

Bateman BL, Murphy HT, Reside AE, Mokany K, Vanderwal J (2013) Appropriateness of full-, partial- and no-dispersal scenarios in climate change impact modelling. *Diversity and Distributions,* 19, 1224-1234.

Bean WT, Stafford R, Brashares JS (2012) The effects of small sample size and sample bias on threshold selection and accuracy assessment of species distribution models. *Ecography,* 35, 250-258.

Beck J, Ballesteros-Mejia L, Nagel P, Kitching IJ (2013) Online solutions and the 'Wallacean shortfall': what does GBIF contribute to our knowledge of species' ranges? *Diversity and Distributions,* 19, 1043-1050.

Beck J, Boller M, Erhardt A, Schwanghart W (2014) Spatial bias in the GBIF database and its effect on modeling species' geographic distributions. *Ecological Informatics,* 19, 10-15.

Bellard C, Bertelsmeier C, Leadley P, Thuiller W, Courchamp F (2012) Impacts of climate change on the future of biodiversity. *Ecology Letters,* 15, 365-377.

Bini LM, Diniz JaF, Rangel T, Bastos RP, Pinto MP (2006) Challenging Wallacean and Linnean shortfalls: knowledge gradients and conservation planning in a biodiversity hotspot. *Diversity and Distributions,* 12, 475-482.

Bivand R, Piras G (2015) Comparing Implementations of Estimation Methods for Spatial Econometrics. *Journal of Statistical Software,* 63, 1-36.

Bivand R, Rundel C (2014) rgeos: Interface to Geometry Engine - Open Source (GEOS). R package version 0.3-4. pp Page.

Blagoderov V, Kitching IJ, Livermore L, Simonsen TJ, Smith VS (2012) No specimen left behind: industrial scale digitization of natural history collections. *Zookeys*, 133-146.

Blois JL, Zarnetske PL, Fitzpatrick MC, Finnegan S (2013) Climate change and the past, present, and future of biotic interactions. *Science,* 341, 499-504.

Boitani L, Maiorano L, Baisero D, Falcucci A, Visconti P, Rondinini C (2011) What spatial data do we need to develop global mammal conservation strategies? *Philos Trans R Soc Lond B Biol Sci,* 366, 2623-2632.

Bookhagen B, Strecker MR (2008) Orographic barriers, high-resolution TRMM rainfall, and relief variations along the eastern Andes. *Geophysical Research Letters,* 35.

Borcard D, Legendre P, Drapeau P (1992) Partialling out the Spatial Component of Ecological Variation. *Ecology,* 73, 1045-1055.

Boucher-Lalonde V, Morin A, Currie DJ (2012) How are tree species distributed in climatic space? A simple and general pattern. *Global Ecology and Biogeography,* 21, 1157-1166.

Boulangeat I, Gravel D, Thuiller W (2012) Accounting for dispersal and biotic interactions to disentangle the drivers of species distributions and their abundances. *Ecol Lett,* 15, 584-593.

Bradshaw CJA, Sodhi NS, Brook BW (2009) Tropical turmoil: a biodiversity tragedy in progress. *Frontiers in Ecology and the Environment,* 7, 79-87.

Broennimann O, Fitzpatrick MC, Pearman PB *et al.* (2012) Measuring ecological niche overlap from occurrence and spatial environmental data. *Global Ecology and Biogeography,* 21, 481-497.

Brook BW, Sodhi NS, Bradshaw CJ (2008) Synergies among extinction drivers under global change. *Trends Ecol Evol,* 23, 453-460.

Brooks TM, Mittermeier RA, Da Fonseca GA *et al.* (2006) Global biodiversity conservation priorities. *Science,* 313, 58-61.

Brown JH (2014) Why are there so many species in the tropics? *J Biogeogr,* 41, 8-22.

Brown JH, Lomolino MV (1998) *Biogeography,* Sunderland, Massachusetts, Sinauer Associates.

Brown KA, Parks KE, Bethell CA, Johnson SE, Mulligan M (2015) Predicting Plant Diversity Patterns in Madagascar: Understanding the Effects of Climate and Land Cover Change in a Biodiversity Hotspot. *Plos One,* 10, e0122721.

Burgman MA, Fox JC (2003) Bias in species range estimates from minimum convex polygons: implications for conservation and options for improved planning. *Animal Conservation,* 6, 19-28.

Calabrese JM, Certain G, Kraan C, Dormann CF (2014) Stacking species distribution models and adjusting bias by linking them to macroecological models. *Global Ecology and Biogeography,* 23, 99-112.

Calenge C (2006) The package adehabitat for the R software: a tool for the analysis of space and habitat use by animals. *Ecological Modelling,* 197, 516-519.

Ceballos G, Ehrlich PR (2006) Global mammal distributions, biodiversity hotspots, and conservation. *Proceedings of the National Academy of Sciences of the United States of America,* 103, 19374-19379.

Chatelain C, Gautier L, Spichiger R (1996) A recent history of forest fragmentation in southwestern Ivory Coast. *Biodiversity and Conservation,* 5, 37-53.

Chejanovski ZA, Wiens JJ (2014) Climatic niche breadth and species richness in temperate treefrogs. *Journal of Biogeography*, n/a-n/a.

Chen GK, Kery M, Plattner M, Ma KP, Gardner B (2013) Imperfect detection is the rule rather than the exception in plant distribution studies. *Journal of Ecology,* 101, 183-191.

Chen Y, Zhang J, Jiang J, Nielsen SE, He F, Robertson M (2017) Assessing the effectiveness of China's protected areas to conserve current and future amphibian diversity. *Diversity and Distributions,* 23, 146-157.

Clifford SL, Anthony NM, Bawe-Johnson M *et al.* (2004) Mitochondrial DNA phylogeography of western lowland gorillas (Gorilla gorilla gorilla). *Mol Ecol,* 13, 1551-1565, 1567.

Cobben MMP, Van Treuren R, Castañeda-Álvarez NP, Khoury CK, Kik C, Van Hintum TJL (2014) Robustness and accuracy of Maxent niche modelling for *Lactuca* species distributions in light of collecting expeditions. *Plant Genetic Resources,* 13, 153-161.

Cohen J (1960) A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement,* 20, 37-46.

Collen B, Ram M, Zamin T, Mcrae L (2008) The tropical biodiversity data gap: addressing disparity in global monitoring. *Tropical Conservation Science,* 1, 75-88.

Colwell RK, Futuyma DJ (1971) On the Measurement of Niche Breadth and Overlap. *Ecology,* 52, 567-576.

Colyn M, Gautierhion A, Verheyen W (1991) A Reappraisal of Paleoenvironmental History in Central Africa - Evidence for a Major Fluvial Refuge in the Zaire Basin. *Journal of Biogeography,* 18, 403-407.

Conn PB, Johnson DS, Hoef JMV, Hooten MB, London JM, Boveng PL (2015) Using spatiotemporal statistical models to estimate animal abundance and infer ecological dynamics from survey counts. *Ecological Monographs,* 85, 235-252.

Corlett RT, Westcott DA (2013) Will plant movements keep up with climate change? *Trends Ecol Evol,* 28, 482-488.

Costello MJ, May RM, Stork NE (2013) Can we name Earth's species before they go extinct? *Science,* 339, 413-416.

Couvreur TLP (2015) Odd man out: why are there fewer plant species in African rain forests? *Plant Systematics and Evolution,* 301, 1299-1313.

Crisp MD, Laffan S, Linder HP, Monro A (2001) Endemism in the Australian flora. *Journal of Biogeography,* 28, 183-198.

D'amen M, Dubuis A, Fernandes RF, Pottier J, Pellissier L, Guisan A (2015a) Using species richness and functional traits predictions to constrain assemblage predictions from stacked species distribution models. *Journal of Biogeography,* 42, 1255-1266.

D'amen M, Pradervand JN, Guisan A (2015b) Predicting richness and composition in mountain insect communities at high resolution: a new test of the SESAM framework. *Global Ecology and Biogeography,* 24, 1443-1453.

D'amen M, Rahbek C, Zimmermann NE, Guisan A (2017) Spatial predictions at the community level: from current approaches to future frameworks. *Biol Rev Camb Philos Soc,* 92, 169-187.

D'amen M, Zimmermann NE, Pearman PB (2013) Conservation of phylogeographic lineages under climate change. *Global Ecology and Biogeography,* 22, 93-104.

Darwin CR (1859) *On the origin of species by means of natural selection, or the preservation of favoured races in the struggle for life.,* London, John Murray.

Dauby G, Duminil J, Heuertz M, Hardy OJ (2010) Chloroplast DNA Polymorphism and Phylogeography of a Central African Tree Species Widespread in Mature Rainforests: *Greenwayodendron suaveolens* (Annonaceae). *Tropical Plant Biology,* 3, 4-13.

Dauby G, Duminil J, Heuertz M, Koffi GK, Stevart T, Hardy OJ (2014) Congruent phylogeographical patterns of eight tree species in Atlantic Central Africa provide insights into the past dynamics of forest cover. *Mol Ecol,* 23, 2299-2312.

Dauby G, Zaiss R, Blach-Overgaard A *et al.* (2016) RAINBIO: a mega-database of tropical African vascular plants distributions. *PhytoKeys,* 74, 1-18.

De La Estrella M, Mateo RG, Wieringa JJ, Mackinder B, Munoz J (2012) Legume diversity patterns in West Central Africa: influence of species biology on distribution models. *Plos One,* 7, e41526.

Deblauwe V, Droissart V, Bose R *et al.* (2016) Remotely sensed temperature and precipitation data improve species distribution modelling in the tropics. *Global Ecology and Biogeography,* 25, 443-454.

Demenou BB, Piñeiro R, Hardy OJ (2016) Origin and history of the Dahomey Gap separating West and Central African rain forests: insights from the phylogeography of the legume tree Distemonanthus benthamianus. *Journal of Biogeography,* 43, 1020-1031.

Deutsch CA, Tewksbury JJ, Huey RB, Sheldon KS, Ghalambor CK, Haak DC, Martin PR (2008) Impacts of climate warming on terrestrial ectotherms across latitude. *Proc Natl Acad Sci U S A,* 105, 6668-6672.

Doledec S, Chessel D, Gimaret-Carpentier C (2000) Niche separation in community analysis: A new method. *Ecology,* 81, 2914-2927.

Dorazio RM (2014) Accounting for imperfect detection and survey bias in statistical analysis of presence-only data. *Global Ecology and Biogeography,* 23, 1472-1484.

Dorazio RM, Royle JA (2005) Estimating size and composition of biological communities by modeling the occurrence of species. *Journal of the American Statistical Association,* 100, 389-398.

Dormann CF, Elith J, Bacher S *et al.* (2013) Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. *Ecography,* 36, 27-46.

Dormann CF, Mcpherson JM, Araújo MB *et al.* (2007) Methods to account for spatial autocorrelation in the analysis of species distributional data: a review. *Ecography,* 30, 609-628.

Dormann CF, Schymanski SJ, Cabral J *et al.* (2012) Correlation and process in species distribution models: bridging a dichotomy. *Journal of Biogeography,* 39, 2119-2131.

Duan R-Y, Kong X-Q, Huang M-Y, Wu G-L, Wang Z-G (2015) SDMvspecies: a software for creating virtual species for species distribution modelling. *Ecography,* 38, 108-110.

Duan Z, Bastiaanssen WGM (2013) First results from Version 7 TRMM 3B43 precipitation product in combination with a new downscaling–calibration procedure. *Remote Sensing of Environment,* 131, 1-13.

Dubuis A, Pottier J, Rion V, Pellissier L, Theurillat JP, Guisan A (2011) Predicting spatial patterns of plant species richness: a comparison of direct macroecological and species stacking modelling approaches. *Diversity and Distributions,* 17, 1122-1131.

Duminil J, Heuertz M, Doucet JL *et al.* (2010) CpDNA-based species identification and phylogeography: application to African tropical tree species. *Mol Ecol,* 19, 5469-5483.

Edelsbrunner H, Kirkpatrick DG, Seidel R (1983) On the shape of a set of points in the plane. *Ieee Transactions on Information Theory,* 29, 551-559.

Elias RB, Gil A, Silva L, Fernandez-Palacios JM, Azevedo EB, Reis F (2016) Natural zonal vegetation of the Azores Islands: characterization and potential distribution. *Phytocoenologia,* 46, 107-123.

Elith J, Graham CH, Anderson RP *et al.* (2006) Novel methods improve prediction of species' distributions from occurrence data. *Ecography,* 29, 129-151.

Elith J, Kearney M, Phillips S (2010) The art of modelling range-shifting species. *Methods in Ecology and Evolution,* 1, 330-342.

Elith J, Leathwick JR (2009) Species Distribution Models: Ecological Explanation and Prediction Across Space and Time. In: *Annual Review of Ecology Evolution and Systematics.* pp Page.

Engler R, Guisan A (2009) MIGCLIM: Predicting plant distribution and dispersal in a changing climate. *Diversity and Distributions,* 15, 590-601.

Engler R, Hordijk W, Guisan A (2012) The MIGCLIM R package - seamless integration of dispersal constraints into projections of species distribution models. *Ecography,* 35, 872-878.

Evans KL, Greenwood JJD, Gaston KJ (2005) Relative contribution of abundant and rare species to species-energy relationships. *Biology Letters,* 1, 87-90.

Fao/Iiasa/Isric/Isscas/Jrc (2012) Harmonized World Soil Database (version 1.2). pp Page, Rome, Italy & Laxenburg, Austria, FAO & IIASA.

Faye A, Deblauwe V, Mariac C, Richard D, Sonke B, Vigouroux Y, Couvreur TL (2016) Phylogeography of the genus *Podococcus* (Palmae/Arecaceae) in Central African rain forests: Climate stability predicts unique genetic diversity. *Mol Phylogenet Evol,* 105, 126-138.

Feeley KJ (2012) Distributional migrations, expansions, and contractions of tropical plant species as revealed in dated herbarium records. *Global Change Biology,* 18, 1335-1341.

Feeley KJ (2015) Are we filling the data void? An assessment of the amount and extent of plant collection records and census data available for tropical South America. *Plos One,* 10, e0125629.

Feeley KJ, Malhi Y, Zelazowski P, Silman MR (2012) The relative importance of deforestation, precipitation change, and temperature sensitivity in determining the future distributions and diversity of Amazonian plant species. *Global Change Biology,* 18, 2636-2647.

Feeley KJ, Silman MR (2009) Extinction risks of Amazonian plant species. *Proceedings of the National Academy of Sciences of the United States of America,* 106, 12382-12387.

Feeley KJ, Silman MR (2010a) Land-use and climate change effects on population size and extinction risk of Andean plants. *Global Change Biology,* 16, 3215-3222.

Feeley KJ, Silman MR (2010b) Modelling the responses of Andean and Amazonian plant species to climate change: the effects of georeferencing errors and the importance of data filtering. *Journal of Biogeography,* 37, 733-740.

Feeley KJ, Silman MR (2011) Keep collecting: accurate species distribution modelling requires more collections than previously thought. *Diversity and Distributions,* 17, 1132-1140.

Feeley KJ, Stroud JT, Perez TM, Kühn I (2016) Most 'global' reviews of species' responses to climate change are not truly global. *Diversity and Distributions*, n/a-n/a.

Feinsinger P, Spears EE, Poole RW (1981) A simple measure of niche breadth. *Ecology,* 62, 27-32.

Fielding AH, Bell JF (1997) A review of methods for the assessment of prediction errors in conservation presence/absence models. *Environmental Conservation,* 24, 38-49.

Fitzpatrick MC, Gove AD, Sanders NJ, Dunn RR (2008) Climate change, plant migration, and range collapse in a global biodiversity hotspot: the *Banksia* (Proteaceae) of Western Australia. *Global Change Biology,* 14, 1337-1352.

Flora of North America Editorial Committee E (1993+) Flora of North America North of Mexico. (ed Flora of North America Editorial Committee E) pp Page, New York and Oxford, Flora of North America Association.

Fourcade Y, Engler JO, Rodder D, Secondi J (2014) Mapping species distributions with MAXENT using a geographically biased sample of presence data: a performance assessment of methods for correcting sampling bias. *Plos One,* 9, e97122.

Fowler J, Cohen L, Jarvis P (1998) *Practical statistics for field biology.,* Chichester, West Sussex, United Kingdom, John Wiley & Sons.

Franklin J (2009) *Mapping Species Distributions - Spatial Inference and Prediction*, Cambridge University Press.

Franklin J (2010) Moving beyond static species distribution models in support of conservation biogeography. *Diversity and Distributions,* 16, 321-330.

Franklin J, Davis FW, Ikegami M, Syphard AD, Flint LE, Flint AL, Hannah L (2013) Modeling plant species distributions under future climates: how fine scale do climate projections need to be? *Glob Chang Biol,* 19, 473-483.

Funk CC, Peterson PJ, Landsfeld MF *et al.* (2014) A quasi-global precipitation time series for drought monitoring. pp Page.

Gaston KJ (1991) How large is a species' geographic range? *Oikos,* 61, 434-438.

Gaston KJ (1994) *Rarity*, Springer Netherlands.

Gaston KJ (1996a) *Biodiversity*, Wiley.

Gaston KJ (1996b) Species-range-size distributions: patterns, mechanisms and implications. *Trends Ecol Evol,* 11, 197-201.

Gaston KJ (2000) Global patterns in biodiversity. *Nature,* 405, 220-227.

Gaston KJ (2003) *The structure and dynamics of geographic ranges,* Oxford, UK, Oxford University Press.

Gaston KJ, Fuller RA (2009) The sizes of species' geographic ranges. *Journal of Applied Ecology,* 46, 1-9.

Gavin DG, Fitzpatrick MC, Gugger PF *et al.* (2014) Climate refugia: joint inference from fossil records, species distribution models and phylogeography. *New Phytol,* 204, 37-54.

Gavish Y, Marsh CJ, Kuemmerlen M, Stoll S, Haase P, Kunin WE, Freckleton R (2017) Accounting for biotic interactions through alpha-diversity constraints in stacked species distribution models. *Methods in Ecology and Evolution*, n/a-n/a.

Gentry AH (1992) Tropical Forest Biodiversity - Distributional Patterns and Their Conservational Significance. *Oikos,* 63, 19-28.

Gentz A, Bretz F, Miwa T, Mi X, Leisch F, Scheipl F, Hothorn T (2014) mvtnorm: Multivariate Normal and t Distributions. R package version 0.9-9997. pp Page.

Giannini TC, Chapman DS, Saraiva AM, Alves-Dos-Santos I, Biesmeijer JC (2013) Improving species distribution models using biotic interactions: a case study of parasites, pollinators and plants. *Ecography,* 36, 649-656.

Gillingham PK, Huntley B, Kunin WE, Thomas CD (2012) The effect of spatial resolution on projected responses to climate warming. *Diversity and Distributions,* 18, 990-1000.

Gilman SE, Urban MC, Tewksbury J, Gilchrist GW, Holt RD (2010) A framework for community interactions under climate change. *Trends Ecol Evol,* 25, 325-331.

Godsoe W, Murray R, Plank MJ (2015) Information on biotic interactions improves transferability of distribution models. *Am Nat,* 185, 281-290.

Gotelli NJ, Anderson MJ, Arita HT *et al.* (2009) Patterns and causes of species richness: a general simulation model for macroecology. *Ecol Lett,* 12, 873-886.

Gould SF, Beeton NJ, Harris RM, Hutchinson MF, Lechner AM, Porfirio LL, Mackey BG (2014) A tool for simulating and communicating uncertainty when modelling species distributions under future climates. *Ecology and Evolution,* 4, 4798-4811.

Graham CH, Elith J, Hijmans RJ, Guisan A, Peterson AT, Loiselle BA, Nceas Predect Species Working G (2008) The influence of spatial errors in species occurrence data used in distribution models. *Journal of Applied Ecology,* 45, 239-247.

Graham CH, Ferrier S, Huettman F, Moritz C, Peterson AT (2004) New developments in museum-based informatics and applications in biodiversity analysis. *Trends Ecol Evol,* 19, 497-503.

Grenyer R, Orme CDL, Jackson SF *et al.* (2006) Global distribution and conservation of rare and threatened vertebrates. *Nature,* 444, 93-96.

Gromping U (2006) Relative importance for linear regression in R: The package relaimpo. *Journal of Statistical Software,* 17, 27.

Guillera-Arroita G (2017) Modelling of species distributions, range dynamics and communities under imperfect detection: advances, challenges and opportunities. *Ecography,* 40, 281-295.

Guisan A, Rahbek C (2011) SESAM - a new framework integrating macroecological and species distribution models for predicting spatio-temporal patterns of species assemblages. *Journal of Biogeography,* 38, 1433-1444.

Guisan A, Thuiller W (2005) Predicting species distribution: offering more than simple habitat models. *Ecology Letters,* 8, 993-1009.

Guisan A, Tingley R, Baumgartner JB *et al.* (2013) Predicting species distributions for conservation decisions. *Ecol Lett,* 16, 1424-1435.

Guisan A, Zimmermann NE (2000) Predictive habitat distribution models in ecology. *Ecological Modelling,* 135, 147-186.

Hardy OJ, Born C, Budde K *et al.* (2013) Comparative phylogeography of African rain forest trees: A review of genetic signatures of vegetation history in the Guineo-Congolian region. *Comptes Rendus Geoscience,* 345, 284-296.

Harris DJ, Armstrong KE, Walters GM *et al.* (2012) Phytogeographical analysis and checklist of the vascular plants of Loango National Park, Gabon. *Plant Ecology and Evolution,* 145, 242-257.

Hartley S, Kunin WE (2003) Scale Dependency of Rarity, Extinction Risk, and Conservation Priority. *Conservation Biology,* 17, 1559-1570.

Hassan R, Scholes RJ, Ash N (2005) *Ecosystems and human well-being: current state and trends* Washington, DC, Island Press.

Hawkins BA, Diniz JaF, Bini LM, De Marco P, Blackburn TM (2007) Red herrings revisited: spatial autocorrelation and parameter estimation in geographical ecology. *Ecography,* 30, 375-384.

Heibl C, Calenge C (2013) phyloclim: Integrating Phylogenetics and Climatic Niche Modeling. R package version 0.9-4. pp Page.

Hernandez PA, Graham CH, Master LL, Albert DL (2006) The effect of sample size and species characteristics on performance of different species distribution modeling methods. *Ecography,* 29, 773-785.

Hijmans RJ (2014) geosphere: Spherical Trigonometry. R package version 1.3-11. pp Page.

Hijmans RJ, Cameron SE, Parra JL, Jones PG, Jarvis A (2005) Very high resolution interpolated climate surfaces for global land areas. *International Journal of Climatology,* 25, 1965-1978.

Hijmans RJ, Phillips S, Leathwick JR, Elith J (2013) dismo: Species distribution modeling. R package version 0.8-17. pp Page.

Hirzel AH, Helfer V, Metral F (2001) Assessing habitat-suitability models with a virtual species. *Ecological Modelling,* 145, 111-121.

Holdridge LR, Grenke WH, Hatheway WH, Al. (1971) *Forest environments in tropical life zones: a pilot study*, Pergamon.

Hooper DU, Adair EC, Cardinale BJ *et al.* (2012) A global synthesis reveals biodiversity loss as a major driver of ecosystem change. *Nature,* 486, 105-108.

Hortal J, Jimenez-Valverde A, Gomez JF, Lobo JM, Baselga A (2008) Historical bias in biodiversity inventories affects the observed environmental niche of the species. *Oikos,* 117, 847-858.

Hortal J, Lobo JM, Jiménez-Valverde A (2007) Limitations of biodiversity databases: case study on seed-plant diversity in Tenerife, Canary Islands. *Conserv Biol,* 21, 853-863.

Hsu RCC, Tamis WLM, Raes N, De Snoo GR, Wolf JHD, Oostermeijer G, Lin S-H (2012) Simulating climate change impacts on forests and associated vascular epiphytes in a subtropical island of East Asia. *Diversity and Distributions,* 18, 334-347.

Huang JH, Huang JH, Liu CR, Zhang JL, Lu XH, Ma KP (2016) Diversity hotspots and conservation gaps for the Chinese endemic seed flora. *Biological Conservation,* 198, 104-112.

Hubbell SP, He F, Condit R, Borda-De-Agua L, Kellner J, Ter Steege H (2008) How many tree species are there in the Amazon and how many of them will go extinct? *Proc Natl Acad Sci U S A,* 105 Suppl 1, 11498-11504.

Hurlbert AH, Jetz W (2007) Species richness, hotspots, and the scale dependence of range maps in ecology and conservation. *Proceedings of the National Academy of Sciences of the United States of America,* 104, 13384-13389.

Hylander K, Ehrlen J, Luoto M, Meineri E (2015) Microrefugia: Not for everyone. *AMBIO,* 44 Suppl 1, S60-68.

Ipcc (2013) Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change. (eds Stocker TF, Qin D, Plattner G-K, Tignor M, Allen SK, Boschung J, Nauels A, Xia Y, Bex V, Midgley PM) pp Page, Cambridge, United Kingdom and New York, NY, USA.

IUCN (2001) *IUCN Red List Categories and Criteria, version 3.1.,* Cambridge, UK, IUCN Species Survival Commision.

IUCN (2014) *Guidelines for Using the IUCN Red List Categories and Criteria. Version 11.*, IUCN Standards and Petitions Subcommittee.

Jetz W, Rahbek C (2002) Geographic range size and determinants of avian species richness. *Science,* 297, 1548-1551.

Jetz W, Rahbek C, Colwell RK (2004) The coincidence of rarity and richness and the potential signature of history in centres of endemism. *Ecology Letters,* 7, 1180-1191.

Jiménez-Valverde A (2012) Insights into the area under the receiver operating characteristic curve (AUC) as a discrimination measure in species distribution modelling. *Global Ecology and Biogeography,* 21, 498-507.

Jiménez-Valverde A, Lobo JM, Hortal J (2009) The effect of prevalence and its interaction with sample size on the reliability of species distribution models. *Community Ecology,* 10, 196-205.

Jiménez-Valverde A, Peterson AT, Soberón J, Overton JM, Aragon P, Lobo JM (2011) Use of niche models in invasive species risk assessments. *Biological Invasions,* 13, 2785-2797.

Joppa LN, Roberts DL, Pimm SL (2011) How many species of flowering plants are there? *Proc Biol Sci,* 278, 554-559.

Kadmon R, Farber O, Danin A (2003) A systematic analysis of factors affecting the performance of climatic envelope models. *Ecological Applications,* 13, 853-867.

Kadmon R, Farber O, Danin A (2004) Effect of roadside bias on the accuracy of predictive maps produced by bioclimatic models. *Ecological Applications,* 14, 401-413.

Kaky E, Gilbert F (2016) Using species distribution models to assess the importance of Egypt's protected areas for the conservation of medicinal plants. *Journal of Arid Environments,* 135, 140-146.

Kennedy D, Norman C (2005) What don't we know? *Science,* 309, 75.

Khoury CK, Castaneda-Alvarez NP, Achicanoy HA *et al.* (2015) Crop wild relatives of pigeonpea [*Cajanus cajan* (L.) Millsp.]: Distributions, ex situ conservation status, and potential genetic resources for abiotic stress tolerance. *Biological Conservation,* 184, 259-270.

Kier G, Mutke J, Dinerstein E, Ricketts TH, Kuper W, Kreft H, Barthlott W (2005) Global patterns of plant diversity and floristic knowledge. *Journal of Biogeography,* 32, 1107-1116.

Kissling WD, Dormann CF, Groeneveld J *et al.* (2012) Towards novel approaches to modelling biotic interactions in multispecies assemblages at large spatial extents. *Journal of Biogeography,* 39, 2163-2178.

Klopper RR, Gautier L, Chatelain C, Smith GF, Spichiger R (2007) Floristics of the angiosperm flora of sub-Saharan Africa: an analysis of the African Plant Checklist and Database. *Taxon,* 56, 201-208.

Koffi KG, Hardy OJ, Doumenge C, Cruaud C, Heuertz M (2011) Diversity gradients and phylogeographic patterns in *Santiria trimera* (Burseraceae), a widespread African tree typical of mature rainforests. *Am J Bot,* 98, 254-264.

Kramer-Schadt S, Niedballa J, Pilgrim JD *et al.* (2013) The importance of correcting for sampling bias in MaxEnt species distribution models. *Diversity and Distributions,* 19, 1366-1379.

Kreft H, Sommer JH, Barthlott W (2006) The significance of geographic range size for spatial diversity patterns in Neotropical palms. *Ecography,* 29, 21-30.

Küper W, Sommer JH, Lovett JC, Barthlott W (2006) Deficiency in African plant distribution data - missing pieces of the puzzle. *Botanical Journal of the Linnean Society,* 150, 355-368.

Küper W, Sommer JH, Lovett JC *et al.* (2004) Africa's hotspots of biodiversity redefined. *Annals of the Missouri Botanical Garden,* 91, 525-535.

Lachenaud O, Stévart TOB, Ikabanga D, Ndjabounda ECN, Walters G (2013) The littoral forests of the Libreville area (Gabon) and their importance for conservation: description of a new endemic *Psychotria* species (Rubiaceae). *Plant Ecology and Evolution,* 146, 68-74.

Laderach P, Ramirez-Villegas J, Navarro-Racines C, Zelaya C, Martinez-Valle A, Jarvis A (2017) Climate change adaptation of coffee production in space and time. *Climatic Change,* 141, 47-62.

Lamoreux JF, Morrison JC, Ricketts TH, Olson DM, Dinerstein E, Mcknight MW, Shugart HH (2006) Global tests of biodiversity concordance and the importance of endemism. *Nature,* 440, 212-214.

Leal ME (2001) Microrefugia, Small Scale Ice Age Forest Remnants. *Systematics and Geography of Plants,* 71, 1073-1077.

Leal ME (2004) The African rain forest during the Last Glacial Maximum, an archipelago of forests in a sea of grass. Unpublished PhD Wageningen University, Wageningen.

Lennon JJ, Beale CM, Reid CL, Kent M, Pakeman RJ (2011) Are richness patterns of common and rare species equally well explained by environmental variables? *Ecography,* 34, 529-539.

Lennon JJ, Koleff P, Greenwood JJD, Gaston KJ (2004) Contribution of rarity and commonness to patterns of species richness. *Ecology Letters,* 7, 81-87.

Lenoir J, Svenning JC (2015) Climate-related range shifts - a global multidimensional synthesis and new research directions. *Ecography,* 38, 15-28.

Leroy B, Meynard CN, Bellard C, Courchamp F (2016) virtualspecies, an R package to generate virtual species distributions. *Ecography,* 39, 599-607.

Ley AC, Dauby G, Kohler J, Wypior C, Roser M, Hardy OJ (2014) Comparative phylogeography of eight herbs and lianas (Marantaceae) in central African rainforests. *Front Genet,* 5, 403.

Linder HP (2001) Plant diversity and endemism in sub-Saharan tropical Africa. *Journal of Biogeography,* 28, 169-182.

Linder HP (2014) The evolution of African plant diversity. *Frontiers in Ecology and Evolution,* 2.

Liu CR, White M, Newell G (2011) Measuring and comparing the accuracy of species distribution models with presence-absence data. *Ecography,* 34, 232-243.

Liu CR, White M, Newell G (2013) Selecting thresholds for the prediction of species occurrence with presence-only data. *Journal of Biogeography,* 40, 778-789.

Lobo JM, Jiménez-Valverde A, Real R (2008) AUC: a misleading measure of the performance of predictive distribution models. *Global Ecology and Biogeography,* 17, 145-151.

Lobo JM, Tognelli MF (2011) Exploring the effects of quantity and location of pseudo-absences and sampling biases on the performance of distribution models with limited point occurrence data. *Journal for Nature Conservation,* 19, 1-7.

Loiselle BA, Jorgensen PM, Consiglio T, Jimenez I, Blake JG, Lohmann LG, Montiel OM (2008) Predicting species distributions from herbarium collections: does climate bias in collection sampling influence model outcomes? *Journal of Biogeography,* 35, 105-116.

Lomba A, Pellissier L, Randin C, Vicente J, Moreira F, Honrado J, Guisan A (2010) Overcoming the rare species modelling paradox: A novel hierarchical framework applied to an Iberian endemic plant. *Biological Conservation,* 143, 2647-2657.

Lomolino MV (2004) Conservation biogeography. In: *Conservation biogeography. Frontiers of Biogeography: new directions in the geography of nature.* (eds Lomolino MV, Heaney LR) pp Page. Sunderland, MA., Sinauer Associates.

Lomolino MV, Riddle BR, Whittaker RJ, Brown JH (2010) *Biogeography,* Sunderland, Sinauer Associates, Inc.

Longino JT, Coddington J, Colwell RK (2002) The ant fauna of a tropical rain forest: Estimating species richness three different ways. *Ecology,* 83, 689-702.

Loyola RD, Kubota U, Lewinsohn TM (2007) Endemic vertebrates are the most effective surrogates for identifying conservation priorities among Brazilian ecoregions. *Diversity and Distributions,* 13, 389-396.

Magurran AE, Henderson PA (2003) Explaining the excess of rare species in natural species abundance distributions. *Nature,* 422, 714-716.

Maiorano L, Amori G, Montemaggiori A, Rondinini C, Santini L, Saura S, Boitani L (2015) On how much biodiversity is covered in Europe by national protected areas and by the Natura 2000 network: insights from terrestrial vertebrates. *Conserv Biol,* 29, 986-995.

Malcolm JR, Liu C, Neilson RP, Hansen L, Hannah L (2006) Global warming and extinctions of endemic species from biodiversity hotspots. *Conserv Biol,* 20, 538-548.

Maley J (1996) The African rain forest - main characteristics of changes in vegetation and climate from the Upper Cretaceous to the Quaternary. *Proc. Roy. Soc. Edinb.,* 104, 31-73.

Maley J (2002) A catastrophic destruction of African forests about 2,500 years ago still exerts a major influence on present vegetation formations. *Ids Bulletin-Institute of Development Studies,* 33, 13-+.

Malhi Y, Doughty CE, Galetti M, Smith FA, Svenning JC, Terborgh JW (2016) Megafauna and ecosystem function from the Pleistocene to the Anthropocene. *Proc Natl Acad Sci U S A,* 113, 838-846.

Malhi Y, Roberts JT, Betts RA, Killeen TJ, Li W, Nobre CA (2008) Climate change, deforestation, and the fate of the Amazon. *Science,* 319, 169-172.

Malhi Y, Wright J (2004) Spatial patterns and recent trends in the climate of tropical rainforest regions. *Philos Trans R Soc Lond B Biol Sci,* 359, 311-329.

Manel S, Williams HC, Ormerod SJ (2001) Evaluating presence-absence models in ecology: the need to account for prevalence. *Journal of Applied Ecology,* 38, 921-931.

Margules CR, Pressey RL (2000) Systematic conservation planning. *Nature,* 405, 243-253.

Marshall CA, Wieringa JJ, Hawthorne WD (2016) Bioquality Hotspots in the Tropical African Flora. *Curr Biol*.

Mateo RG, Felicísimo AM, Muñoz J (2010) Effects of the number of presences on reliability and stability of MARS species distribution models: the importance of regional niche variation and ecological heterogeneity. *Journal of Vegetation Science,* 21, 908-922.

Mateo RG, Felicisimo AM, Pottier J, Guisan A, Munoz J (2012) Do Stacked Species Distribution Models Reflect Altitudinal Diversity Patterns? *Plos One,* 7, 9.

Mazaris A, Tzanopoulos J, Kallimanis A, Matsinos Y, Sgardelis S, Pantis J (2008) The contribution of common and rare species to plant species richness patterns: the effect of habitat type and size of sampling unit. *Biodiversity and Conservation,* 17, 3567-3577.

Mazaris AD, Tsianou MA, Sigkounas A, Dimopoulos P, Pantis JD, Sgardelis SP, Kallimanis AS (2013) Accounting for the capacity of common and rare species to contribute to diversity spatial patterns: Is it a sampling issue or a biological effect? *Ecological Indicators,* 32, 9-13.

Mccarthy KP, Fletcher RJ, Jr., Rota CT, Hutto RL (2012) Predicting species distributions from samples collected along roadsides. *Conserv Biol,* 26, 68-77.

Mcclean CJ, Lovett JC, Kuper W *et al.* (2005) African plant diversity and climate change. *Annals of the Missouri Botanical Garden,* 92, 139-152.

Mcpherson JM, Jetz W (2007) Effects of species' ecology on the accuracy of distribution models. *Ecography,* 30, 135-151.

Mcpherson JM, Jetz W, Rogers DJ (2004) The effects of species' range sizes on the accuracy of distribution models: ecological phenomenon or statistical artefact? *Journal of Applied Ecology,* 41, 811-823.

Mcpherson TY (2014) Landscape scale species distribution modeling across the Guiana Shield to inform conservation decision making in Guyana. *Biodiversity and Conservation,* 23, 1931-1948.

Merckx B, Steyaert M, Vanreusel A, Vincx M, Vanaverbeke J (2011) Null models reveal preferential sampling, spatial autocorrelation and overfitting in habitat suitability modelling. *Ecological Modelling,* 222, 588-597.

Merow C, Lafleur N, Silander JA, Jr., Wilson AM, Rubega M (2011) Developing dynamic mechanistic species distribution models: predicting bird-mediated spread of invasive plants across northeastern North America. *Am Nat,* 178, 30-43.

Merow C, Smith MJ, Silander JA (2013) A practical guide to MaxEnt for modeling species' distributions: what it does, and why inputs and settings matter. *Ecography,* 36, 1058-1069.

Metz CE (1978) Basic principles of ROC analysis. *Seminars in Nuclear Medicine,* VIII, 283-298.

Meyer C, Kreft H, Guralnick R, Jetz W (2015) Global priorities for an effective information basis of biodiversity distributions. *Nat Commun,* 6, 8221.

Meyer C, Weigelt P, Kreft H (2016) Multidimensional biases, gaps and uncertainties in global plant occurrence information. *Ecology Letters*, n/a-n/a.

Meynard CN, Kaplan DM (2012) The effect of a gradual response to the environment on species distribution modeling performance. *Ecography,* 35, 499-509.

Meynard CN, Kaplan DM (2013) Using virtual species to study species distributions and model performance. *Journal of Biogeography,* 40, 1-8.

Meynard CN, Quinn JF (2007) Predicting species distributions: a critical comparison of the most common statistical models using artificial species. *Journal of Biogeography,* 34, 1455-1469.

Miller JA (2014) Virtual species distribution models: Using simulated data to evaluate aspects of model performance. *Progress in Physical Geography,* 38, 117-128.

Mitchell PJ, Monk J, Laurenson L (2017) Sensitivity of fine-scale species distribution models to locational uncertainty in occurrence data across multiple sample sizes. *Methods in Ecology and Evolution,* 8, 12-21.

Moilanen A, Arponen A (2011) Setting conservation targets under budgetary constraints. *Biological Conservation,* 144, 650-653.

Mora C, Tittensor DP, Adl S, Simpson AG, Worm B (2011) How many species are there on Earth and in the ocean? *PLoS Biol,* 9, e1001127.

Moss RH, Edmonds JA, Hibbard KA *et al.* (2010) The next generation of scenarios for climate change research and assessment. *Nature,* 463, 747-756.

Moudry V, Simova P (2012) Influence of positional accuracy, sample size and scale on modelling species distributions: a review. *International Journal of Geographical Information Science,* 26, 2083-2095.

Myers N, Mittermeier RA, Mittermeier CG, Da Fonseca GA, Kent J (2000) Biodiversity hotspots for conservation priorities. *Nature,* 403, 853-858.

Naimi B, Skidmore AK, Groen TA, Hamm NaS (2011) Spatial autocorrelation in predictors reduces the impact of positional uncertainty in occurrence data on species distribution modelling. *Journal of Biogeography,* 38, 1497-1509.

Newbold T (2010) Applications and limitations of museum data for conservation and ecology, with particular attention to species distribution models. *Progress in Physical Geography,* 34, 3-22.

Newbold T, Hudson LN, Hill SL *et al.* (2015) Global effects of land use on local terrestrial biodiversity. *Nature,* 520, 45-50.

Ngomanda A, Chepstow-Lusty A, Makaya M *et al.* (2009) Western equatorial African forest-savanna mosaics: a legacy of late Holocene climatic change? *Climate of the Past,* 5, 647-659.

Nobis MP, Normand S (2014) KISSMig - a simple model for R to account for limited migration in analyses of species distributions. *Ecography,* 37, 1282-1287.

Nogués-Bravo D (2009) Predicting the past distribution of species climatic niches. *Global Ecology and Biogeography,* 18, 521-531.

Nychka D, Furrer R, Sain S (2013) fields: Tools for spatial data. R package version 6.9.1. pp Page.

Orme CDL, Davies RG, Burgess M *et al.* (2005) Global hotspots of species richness are not congruent with endemism or threat. *Nature,* 436, 1016-1019.

Oslisly R, White L, Bentaleb I, Favier C, Fontugne M, Gillet JF, Sebag D (2013) Climatic and cultural changes in the west Congo Basin forests over the past 5000 years. *Philos Trans R Soc Lond B Biol Sci,* 368, 20120304.

Papadopoulou A, Anastasiou I, Spagopoulou F, Stalimerou M, Terzopoulou S, Legakis A, Vogler AP (2011) Testing the species-genetic diversity correlation in the Aegean archipelago: toward a haplotype-based macroecology? *Am Nat,* 178, 241-255.

Papeş M, Gaubert P (2007) Modelling ecological niches from low numbers of occurrences: assessment of the conservation status of poorly known viverrids (Mammalia, Carnivora) across two continents. *Diversity and Distributions,* 13, 890-902.

Parmentier I, Malhi Y, Senterre B *et al.* (2007) The odd man out? Might climate explain the lower tree alpha-diversity of African rain forests relative to Amazonian rain forests? *Journal of Ecology,* 95, 1058-1071.

Parmesan C (2006) Ecological and evolutionary responses to recent climate change. *Annual Review of Ecology Evolution and Systematics,* 37, 637-669.

Parmesan C, Yohe G (2003) A globally coherent fingerprint of climate change impacts across natural systems. *Nature,* 421, 37-42.

Paton AJ (2013) From Working List to Online Flora of All Known Plants-Looking Forward with Hindsight. *Annals of the Missouri Botanical Garden,* 99, 206-213.

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

Pearson RG, Dawson TP, Liu C (2004) Modelling species distributions in Britain: a hierarchical integration of climate and land-cover data. *Ecography,* 27, 285-298.

Pearson RG, Raxworthy CJ, Nakamura M, Peterson AT (2007) Predicting species distributions from small numbers of occurrence records: a test case using cryptic geckos in Madagascar. *Journal of Biogeography,* 34, 102-117.

Pennisi E (2005) What determines species diversity. *Science,* 309, 90-90.

Perez-Quesada A, Brazeiro A (2013) Contribution of rarity and commonness to patterns of species richness in biogeographic transitions regions: Woody plants of Uruguay. *Austral Ecology,* 38, 639-645.

Peterson AT (2006) Uses and requirements of ecological niche models and related distributional models. *Biodiversity Informatics,* 3, 59-72.

Peterson AT, Soberón J (2012) Species Distribution Modeling and Ecological Niche Modeling: Getting the Concepts Right. *Natureza & Conservacao,* 10, 102-107.

Phillips SJ, Anderson RP, Schapire RE (2006) Maximum entropy modeling of species geographic distributions. *Ecological Modelling,* 190, 231-259.

Phillips SJ, Dudik M, Elith J, Graham CH, Lehmann A, Leathwick J, Ferrier S (2009) Sample selection bias and presence-only distribution models: implications for background and pseudo-absence data. *Ecological Applications,* 19, 181-197.

Pietsch SA, Gautam S (2013) Ancient origin of a rainforest in Gabon as revealed by carbon isotope data of vegetation and soil. *Holocene,* 23, 1778-1785.

Pimm SL, Jenkins CN, Abell R *et al.* (2014) The biodiversity of species and their rates of extinction, distribution, and protection. *Science,* 344, 1246752.

Pitman NC, Jorgensen PM (2002) Estimating the size of the world's threatened flora. *Science,* 298, 989.

Plana V (2004) Mechanisms and tempo of evolution in the African Guineo-Congolian rainforest. *Philos Trans R Soc Lond B Biol Sci,* 359, 1585-1594.

Platts PJ, Omeny PA, Marchant R (2015) AFRICLIM: high-resolution climate projections for ecological applications in Africa. *African Journal of Ecology,* 53, 103-108.

Poorter L, Bongers F, Kouame FN, Hawthorne WD (2004) *Biodiversity of West African forests: an ecological atlas of woody plant species,* Wallingford & Cambridge, CABI Publisher.

Pouteau R, Birnbaum P (2016) Island biodiversity hotspots are getting hotter: vulnerability of tree species to climate change in New Caledonia. *Biological Conservation,* 201, 111-119.

Pulliam HR (2000) On the relationship between niche and distribution. *Ecology Letters,* 3, 349-361.

R Core Team (2016) R: A language and environment for statistical computing. pp Page, R Foundation for Statistical Computing, Vienna, Austria.

Rabinowitz D (1981) Seven forms of rarity. In: *The Biological Aspects of Rare Plants Conservation.* (ed Synge H) pp Page., John Wiley & Sons Ltd.

Raes N (2012) Partial versus Full Species Distribution Models. *Natureza & Conservacao,* 10, 127-138.

Raes N, Roos MC, Slik JWF, Van Loon EE, Ter Steege H (2009) Botanical richness and endemicity patterns of Borneo derived from species distribution models. *Ecography,* 32, 180-192.

Raes N, Saw LG, Van Welzen PC, Yahara T (2013) Legume diversity as indicator for botanical diversity on Sundaland, South East Asia. *South African Journal of Botany,* 89, 265-272.

Raes N, Ter Steege H (2007) A null-model for significance testing of presence-only species distribution models. *Ecography,* 30, 727-736.

Rahbek C (2005) The role of spatial scale and the perception of large-scale species-richness patterns. *Ecology Letters,* 8, 224-239.

Rahbek C, Graves GR (2001) Multiscale assessment of patterns of avian species richness. *Proceedings of the National Academy of Sciences of the United States of America,* 98, 4534-4539.

Randin CF, Dirnbock T, Dullinger S, Zimmermann NE, Zappa M, Guisan A (2006) Are niche-based species distribution models transferable in space? *Journal of Biogeography,* 33, 1689-1703.

Rangel TF, Diniz JaF, Bini LM (2010) SAM: a comprehensive application for Spatial Analysis in Macroecology. *Ecography,* 33, 46-50.

Reddy S, Davalos LM (2003) Geographical sampling bias and its implications for conservation priorities in Africa. *Journal of Biogeography,* 30, 1719-1727.

Rengstorf AM, Grehan A, Yesson C, Brown C (2012) Towards High-Resolution Habitat Suitability Modeling of Vulnerable Marine Ecosystems in the Deep-Sea: Resolving Terrain Attribute Dependencies. *Marine Geodesy,* 35, 343-361.

Richards PW (1973) Africa, the "Odd man out". In: *Tropical Forest Ecosystems in Africa and South America: A Comparative Review.* (eds Meggers BJ, Ayensu ES, Duckworth WD) pp Page. Washington DC, Smithsonian Institution Press.

Rickets TH (2001) Aligning conservation goals: are patterns of species richness and endemism concordant at regional scales? *Animal Biodiversity and Conservation,* 24, 91-99.

Ricketts TH, Dinerstein E, Boucher T *et al.* (2005) Pinpointing and preventing imminent extinctions. *Proc Natl Acad Sci U S A,* 102, 18497-18501.

Robbrecht E (1996) Geography of African Rubiaceae with reference to glacial rain forest refuges. In: *The Biodiversity of African Plants: Proceedings XIVth AETFAT Congress 22–27 August 1994, Wageningen, The Netherlands.* (eds Van Der Maesen LJG, Van Der Burgt XM, Van Medenbach De Rooy JM) pp Page. Dordrecht, Springer Netherlands.

Roberts DL, Taylor L, Joppa LN, Beggs J (2016) Threatened or Data Deficient: assessing the conservation status of poorly known species. *Diversity and Distributions,* 22, 558-565.

Rödder D, Engler JO (2011) Quantitative metrics of overlaps in Grinnellian niches: advances and possible drawbacks. *Global Ecology and Biogeography,* 20, 915-927.

Rogelj J, Meinshausen M, Knutti R (2012) Global warming under old an new scenarios using IPCC climate sensitivity range estimates. *Nature Climate Change,* 2, 248-253.

Roos MC (1993) State of affairs regarding Flora Malesiana: progress in revision work and publication schedule *Flora Malesiana Bulletin,* 11, 133-142.

Rosenzweig ML (1995) *Species Diversity in Space and Time*, Cambridge University Press.

Rowlingson B, Diggle P (2013) splancs: Spatial and Space-Time Point Pattern Analysis. R package version 2.01-34. pp Page.

Sala OE, Chapin FS, 3rd, Armesto JJ *et al.* (2000) Global biodiversity scenarios for the year 2100. *Science,* 287, 1770-1774.

Sánchez-Fernández D, Lobo JM, Hernández-Manrique OL (2011) Species distribution models that do not incorporate global data misrepresent potential distributions: a case study using Iberian diving beetles. *Diversity and Distributions,* 17, 163-171.

Saupe EE, Barve V, Myers CE *et al.* (2012) Variation in niche and distribution model performance: The need for a priori assessment of key causal factors. *Ecological Modelling,* 237, 11-22.

Sayer JA, Harcourt CS, Collins NM (1992) *The Conservation Atlas of Tropical Forests: Africa*, Springer.

Scheffers BR, De Meester L, Bridge TC *et al.* (2016) The broad footprint of climate change from genes to biomes to people. *Science,* 354.

Schipper J, Chanson JS, Chiozza F *et al.* (2008) The status of the world's land and marine mammals: diversity, threat, and knowledge. *Science,* 322, 225-230.

Schmidt-Lebuhn AN, Knerr NJ, Gonzalez-Orozco CE (2012) Distorted perception of the spatial distribution of plant diversity through uneven collecting efforts: the example of Asteraceae in Australia. *Journal of Biogeography,* 39, 2072-2080.

Schmidt M, Kreft H, Thiombiano A, Zizka G (2005) Herbarium collections and field data-based plant diversity maps for Burkina Faso. *Diversity and Distributions,* 11, 509-516.

Schoener TW (1970) Nonsynchronous spatial overlap of lizards in patchy habitats. *Ecology,* 51, 408-418.

Schulze ED, Mooney HA (1994) *Biodiversity and Ecosystem Function,* Berlin / Heidelberg, Springer-Verlag.

Schweiger O, Heikkinen RK, Harpke A *et al.* (2012) Increasing range mismatching of interacting species under global change is related to their ecological characteristics. *Global Ecology and Biogeography,* 21, 88-99.

Sherrouse BC, Semmens DJ, Clement JM (2014) An application of Social Values for Ecosystem Services (SolVES) to three national forests in Colorado and Wyoming. *Ecological Indicators,* 36, 68-79.

Slatyer RA, Hirst M, Sexton JP (2013) Niche breadth predicts geographical range size: a general ecological pattern. *Ecology Letters,* 16, 1104-1114.

Slik JW, Arroyo-Rodriguez V, Aiba S *et al.* (2015) An estimate of the number of tropical tree species. *Proc Natl Acad Sci U S A,* 112, 7472-7477.

Smith AB (2013) On evaluating species distribution models with random background sites in place of absences when test presences disproportionately sample suitable habitat. *Diversity and Distributions,* 19, 867-872.

Soberón J, Peterson AT (2005) Interpretation of models of fundamental ecological niches and species' distributional areas. *Biodiversity Informatics,* 2, 1-10.

Sosef MSM, Dauby G, Blach-Overgaard A *et al.* (2017) Exploring the floristic diversity of tropical Africa. *BMC Biol,* 15, 15.

Sosef MSM (1994) Refuge begonias. Taxonomy, phylogeny and historical biogeography of *Begonia* sect. *Loasibegonia* and sect. *Scutobegonia* in relation to glacial rain forest refuges in Africa. Unpublished PhD Thesis, Wageningen Agricultural University.

Sosef MSM (1996) Begonias and African rain forest refuges: general aspects and recent progress. In: *The biodiversity of African plants. Proceedings XIVth AETFAT Congress.* (eds Van Der Maesen LJG, Van Der Burgt XM, Van Medenbach De Rooy JM) pp Page, Wageningen, the Netherlands, Kluwer Academic Publishers, Dordrecht.

Sosef MSM, Wieringa JJ, Jongkind CCH *et al.* (2006) Check-list des plantes vasculaires du Gabon / Checklist of Gabonese vascular plants. *Scripta Botanica Belgica,* 35, 1-438.

Stockwell DRB, Peterson AT (2002) Effects of sample size on accuracy of species distribution models. *Ecological Modelling,* 148, 1-13.

Stropp J, Ladle RJ, Malhado ACM *et al.* (2016) Mapping ignorance: 300 years of collecting flowering plants in Africa. *Global Ecology and Biogeography,* 25, 1085-1096.

Svenning JC, Skov F (2004) Limited filling of the potential range in European tree species. *Ecology Letters,* 7, 565-573.

Swets JA (1988) Measuring the accuracy of diagnostic systems. *Science,* 240, 1285-1293.

Syfert MM, Joppa L, Smith MJ, Coomes DA, Bachman SP, Brummitt NA (2014) Using species distribution models to inform IUCN Red List assessments. *Biological Conservation,* 177, 174-184.

Syfert MM, Smith MJ, Coomes DA (2013) The effects of sampling bias and model complexity on the predictive performance of MaxEnt species distribution models. *Plos One,* 8, e55158.

Ter Steege H, Haripersaud PP, Banki OS, Schieving F (2011) A model of botanical collectors' behavior in the field: never the same species twice. *Am J Bot,* 98, 31-37.

Ter Steege H, Pitman NC, Sabatier D *et al.* (2013) Hyperdominance in the Amazonian tree flora. *Science,* 342, 1243092.

Terborgh J, Davenport LC, Niangadouma R, Dimoto E, Mouandza JC, Scholtz O, Jaen MR (2016a) Megafaunal influences on tree recruitment in African equatorial forests. *Ecography,* 39, 180-186.

Terborgh J, Davenport LC, Niangadouma R, Dimoto E, Mouandza JC, Schultz O, Jaen MR (2016b) The African rainforest: odd man out or megafaunal landscape? African and Amazonian forests compared. *Ecography,* 39, 187-193.

Tessarolo G, Rangel TF, Araújo MB, Hortal J (2014) Uncertainty associated with survey design in Species Distribution Models. *Diversity and Distributions*, n/a-n/a.

Thomas CD, Cameron A, Green RE *et al.* (2004) Extinction risk from climate change. *Nature,* 427, 145-148.

Thomas WW (1999) Conservation and monographic research on the flora of Tropical America. *Biodiversity and Conservation,* 8, 1007-1015.

Thorson JT, Ianelli JN, Larsen EA, Ries L, Scheuerell MD, Szuwalski C, Zipkin EF (2016) Joint dynamic species distribution models: a tool for community ordination and spatio-temporal monitoring. *Global Ecology and Biogeography,* 25, 1144-1158.

Thorson JT, Scheuerell MD, Shelton AO, See KE, Skaug HJ, Kristensen K (2015) Spatial factor analysis: a new tool for estimating joint species distributions and correlations in species range. *Methods in Ecology and Evolution,* 6, 627-637.

Thuiller W, Lavorel S, Araujo MB, Sykes MT, Prentice IC (2005) Climate change threats to plant diversity in Europe. *Proc Natl Acad Sci U S A,* 102, 8245-8250.

Thuiller W, Lavorel S, Midgley G, Lavergne S, Rebelo T (2004) Relating plant traits and species distributions along bioclimatic gradients for 88 Leucadendron taxa. *Ecology,* 85, 1688-1699.

Thuiller W, Pollock LJ, Gueguen M, Münkemüller T (2015) From species distributions to meta-communities. *Ecology Letters*, n/a-n/a.

Tian S, Lei S-Q, Hu W *et al.* (2015) Repeated range expansions and inter-/postglacial recolonization routes of *Sargentodoxa cuneata* (Oliv.) Rehd. et Wils. (Lardizabalaceae) in subtropical China revealed by chloroplast phylogeography. *Molecular Phylogenetics and Evolution,* 85, 238-246.

Tobler WR (1970) A Computer Movie Simulating Urban Growth in the Detroit Region. *Economic Geography,* 46, 234.

Tutin TG (1964-1980) *Flora Europaea*, Cambridge University Press.

Urban MC, Tewksbury JJ, Sheldon KS (2012) On a collision course: competition and dispersal differences create no-analogue communities and cause extinctions during climate change. *Proc Biol Sci,* 279, 2072-2080.

Vamosi JC, Vamosi SM (2008) Extinction risk escalates in the tropics. *Plos One,* 3, e3886.

Van Andel TR, Croft S, Van Loon EE, Quiroz D, Towns AM, Raes N (2015) Prioritizing West African medicinal plants for conservation and sustainable extraction studies based on market surveys and species distribution models. *Biological Conservation,* 181, 173-181.

Van Proosdij ASJ, Raes N, Wieringa JJ, Sosef MSM (2016a) Unequal Contribution of Widespread and Narrow-Ranged Species to Botanical Diversity Patterns. *Plos One,* 11, e0169200.

Van Proosdij ASJ, Sosef MSM, Wieringa JJ, Raes N (2016b) Minimum required number of specimen records to develop accurate species distribution models. *Ecography,* 39, 542-552.

Vanderwal J, Falconi L, Januchowski S, Shoo L, Storlie C (2014) SDMTools: Species Distribution Modelling Tools: Tools for processing data associated with species distribution modelling exercises. R package version 1.1-20. pp Page.

Vanderwal J, Murphy HT, Kutt AS, Perkins GC, Bateman BL, Perry JJ, Reside AE (2013) Focus on poleward shifts in species' distribution underestimates the fingerprint of climate change. *Nature Climate Change,* 3, 239-243.

Vanderwal J, Shoo LP, Graham C, William SE (2009) Selecting pseudo-absence data for presence-only distribution modeling: How far should you stray from what you know? *Ecological Modelling,* 220, 589-594.

Varela S, Anderson RP, García-Valdès R, Fernàndez-González F (2014) Environmental filters reduce the effects of sampling bias and improve predictions of ecological niche models. *Ecography,* 37, 1084-1091.

Vázquez LB, Gaston KJ (2004) Rarity, commonness, and patterns of species richness: the mammals of Mexico. *Global Ecology and Biogeography,* 13, 535-542.

Velásquez-Tibatá J, Graham CH, Munch SB (2015) Using measurement error models to account for georeferencing error in species distribution models. *Ecography*, n/a-n/a.

Vellend M (2003) Island biogeography of genes and species. *Am Nat,* 162, 358-365.

Vellend M, Lajoie G, Bourret A, Murria C, Kembel SW, Garant D (2014) Drawing ecological inferences from coincident patterns of population- and community-level biodiversity. *Mol Ecol,* 23, 2890-2901.

Venables WN, Ripley BD (2002) *Modern Applied Statistics with S.*, Springer, New York.

Villalobos F, Dobrovolski R, Provete DB, Gouveia SF (2013) Is Rich and Rare the Common Share? Describing Biodiversity Patterns to Inform Conservation Practices for South American Anurans. *Plos One,* 8, e56073.

Wallace AR (1876) *The geographical distribution of animals; with a study of the relations of living and extinct faunas as elucidating the past changes of the Earth's surface.,* London, Macmillan & Co.

Walters G, Ndjabounda EN, Ikabanga D *et al.* (2016) Peri-urban conservation in the Mondah forest of Libreville, Gabon: Red List assessments of endemic plant species, and avoiding protected area downsizing. *Oryx,* 50, 419-430.

Warren DL (2012) In defense of ‘niche modeling’. *Trends in Ecology & Evolution,* 27, 497-500.

Warren DL, Glor RE, Turelli M (2008) Environmental niche equivalency versus conservatism: quantitative approaches to niche evolution. *Evolution,* 62, 2868-2883.

White F (1979) The Guineo-Congolian region and its relationship to other phytochoria. *Bull. Jard. Bot. Nat. Belg.,* 49, 11-55.

Whittaker RJ, Araújo MB, Paul J, Ladle RJ, Watson JEM, Willis KJ (2005) Conservation Biogeography: assessment and prospect. *Diversity and Distributions,* 11, 3-23.

Wieczorek J, Guo QG, Hijmans RJ (2004) The point-radius method for georeferencing locality descriptions and calculating associated uncertainty. *International Journal of Geographical Information Science,* 18, 745-767.

Wieringa JJ (1999) *Monopetalanthus* exit. A systematic study of *Aphanocalyx*, *Bikinia*, *Icuria*, *Michelsonia* and *Tetraberlinia* (Leguminosae,Caesalpinioideae). Agricultural University, Wageningen.

Wieringa JJ, Mackinder BA (2012) Novitates Gabonensis 79: *Hymenostegia elegans* and *H. robusta* spp. nov (Leguminosae-Caesalpinioideae) from Gabon. *Nordic Journal of Botany,* 30, 144-152.

Wieringa JJ, Mackinder BA, Van Proosdij ASJ (2013) *Gabonius* gen. nov. (Leguminosae, Caesalpinioideae, Detarieae), a distant cousin of *Hymenostegia* endemic to Gabon. *Phytotaxa,* 142, 15-24.

Wieringa JJ, Poorter L (2004) Biodiversity hotspots in West Africa; patterns and causes. In: *Biodiversity of West African forests: an ecological atlas of woody plant species.* (eds Poorter L, Bongers F, Kouamé FN, Hawthorne WD) pp Page. Wallingford, CABI Publishing.

Wieringa JJ, Sosef MSM (2011) The applicability of Relative Floristic Resemblance to evaluate the conservation value of protected areas. *Plant Ecology and Evolution,* 144, 242-248.

Willig MR, Kaufman DM, Stevens RD (2003) Latitudinal gradients of biodiversity: Pattern, process, scale, and synthesis. *Annual Review of Ecology Evolution and Systematics,* 34, 273-309.

Willis F, Moat J, Paton A (2003) Defining a role for herbarium data in Red List assessments: a case study of *Plectranthus* from eastern and southern tropical Africa. *Biodiversity and Conservation,* 12, 1537-1552.

Willis KJ, Bennett KD, Burrough SL, Macias-Fauria M, Tovar C (2013) Determining the response of African biota to climate change: using the past to model the future. *Philosophical Transactions of the Royal Society B-Biological Sciences,* 368.

Wisz MS, Hijmans RJ, Li J, Peterson AT, Graham CH, Guisan A, Group NPSDW (2008) Effects of sample size on the performance of species distribution models. *Diversity and Distributions,* 14, 763-773.

Wisz MS, Pottier J, Kissling WD *et al.* (2013) The role of biotic interactions in shaping distributions and realised assemblages of species: implications for species distribution modelling. *Biol Rev Camb Philos Soc,* 88, 15-30.

Zelazowski P, Malhi Y, Huntingford C, Sitch S, Fisher JB (2011) Changes in the potential distribution of humid tropical forests on a warmer planet. *Philos Trans A Math Phys Eng Sci,* 369, 137-160.

Zhang K, Castanho ADD, Galbraith DR *et al.* (2015) The fate of Amazonian ecosystems over the coming century arising from changes in climate, atmospheric CO2, and land use. *Global Change Biology,* 21, 2569-2587.

Zhang MG, Slik JW, Ma KP (2016) Using species distribution modeling to delineate the botanical richness patterns and phytogeographical regions of China. *Sci Rep,* 6, 22400.

Zhang MG, Zhou ZK, Chen WY, Cannon CH, Raes N, Slik JWF (2014) Major declines of woody plant species ranges under climate change in Yunnan, China. *Diversity and Distributions,* 20, 405-415.

Zhang MG, Zhou ZK, Chen WY, Slik JWF, Cannon CH, Raes N (2012) Using species distribution modeling to improve conservation and land use planning of Yunnan, China. *Biological Conservation,* 153, 257-264.

Zhang ZD, Zang RG, Convertino M (2013) Predicting the distribution of potential natural vegetation based on species functional groups in fragmented and species-rich forests. *Plant Ecology and Evolution,* 146, 261-271.

Zhou L, Tian Y, Myneni RB *et al.* (2014) Widespread decline of Congo rainforest greenness in the past decade. *Nature,* 509, 86-90.

Zimmermann NE, Edwards TC, Graham CH, Pearman PB, Svenning JC (2010) New trends in species distribution modelling. *Ecography,* 33, 985-989.

Zurell D, Berger U, Cabral JS *et al.* (2010) The virtual ecologist approach: simulating data and observers. *Oikos,* 119, 622-635.

Zurell D, Thuiller W, Pagel J *et al.* (2016) Benchmarking novel approaches for modelling species range dynamics. *Global Change Biology*, n/a-n/a.

Summary

Planet Earth hosts an incredible biological diversity. Estimated numbers of species occurring on Earth range from 5 to 11 million eukaryotic species including 400,000-450,000 species of plants. Much of this biodiversity remains poorly known and many species have not yet been named or even been discovered. This is not surprising, as the majority of species is known to be rare and ecosystems are generally dominated by a limited number of common species.

Tropical rainforests are the most species-rich terrestrial ecosystems on Earth. The general higher level of species richness is often explained by higher levels of energy near the Equator (latitudinal diversity gradient). However, when comparing tropical rainforest biomes, African rainforests host fewer plant species than either South American or Asian ones. The Central African country of Gabon is situated in the Lower Guinean phytochorical region. It is largely covered by what is considered to be the most species-rich lowland rainforest in Africa while the government supports an active conservation program. As such, Gabon is a perfect study area to address that enigmatic question that has triggered many researchers before: “What determines botanical species richness?”.

In the past 2.5 million years, tropical rainforests have experienced 21 cycles of global glaciations. They responded to this by contracting during drier and cooler glacials into larger montane and smaller riverine forest refugia and expanding again during warmer and wetter interglacials. The current rapid global climate change coupled with change of land use poses new threats to the survival of many rainforest species. The limited availability of resources for conservation forces governments and NGOs to set priorities. Unfortunately, for many plant species, lack of data on their distribution hampers well-informed decision making in conservation.

Species distribution models (SDMs) offer opportunities to bridge at least partly this knowledge gap. SDMs are correlative models that infer the

spatial distribution of species using only a limited set of known species occurrence records coupled with high resolution environmental data. SDMs are widely applied to study the past, present and future distribution of species, assess the risk of invasive species, infer patterns of species richness and identify hotspots, as well as to assess the impact of climate change. The currently available methods form a pipeline, with which data are selected and cleaned, models selected, parameterized, evaluated and projected to other areas and climatic scenarios, and biodiversity patterns are computed from these SDMs. In this thesis, SDMs of all Gabonese plant species were generated and patterns of species richness and of weighted endemism were computed (chapter 4 & 5).

Although this pipeline enables the rapid generation of SDMs and inferring of biodiversity patterns, its effective use is limited by several matters of which three are specifically addressed in this thesis. Not knowing the true distribution limits the opportunities to assess the accuracy of models and assess the impact of assumptions and limitations of SDMs. The use of simulated species has been advocated as a method to systematically assess the impact of specific matters of SDMs (virtual ecologist). Following this approach, in chapter 2, I present a novel method to simulate large numbers of species that each have their own unique niche.

One matter of SDMs that is usually ignored but has been shown to be of great impact on model accuracy is the number of species occurrence records used to train a model. In chapter 2, I quantify the effect of sample size on model accuracy for species of different range size classes. The results show that the minimum number of records required to generate accurate SDMs is not uniform for species of every range size class and that larger sample sizes are required for more widespread species. By applying a uniform minimum number of records, SDMs of narrow-ranged species are incorrectly rejected and SDMs of widespread species are incorrectly accepted. Instead, I recommend to identify and apply the unique minimum numbers of required records for each individual species. The method presented here to identify the minimum number of records for species of particular range size classes is applicable to any species group and study area.

The range size or prevalence is an important plant feature that is used in IUCN Red List classifications. It is commonly computed as the Extent Of Occurrence (EOO) and Area Of Occupancy (AOO). Currently, these metrics are computed using methods based on the spatial distribution of

the known species occurrences. In chapter 3, using simulated species again, I show that methods based on the distribution of species occurrences in environmental parameter space clearly outperform those based on spatial data. In this chapter, I present a novel method that estimates the range size of a species as the fraction of raster cells within the minimum convex hull of the species occurrences, when all cells from the study area are plotted in environmental parameter space. This novel method outperforms all ten other assessed methods. Therefore, the current use of EOO and AOO based on spatial data alone for the purpose of IUCN Red List classification should be reconsidered. I recommend to use the novel method presented here to estimate the AOO and to estimate the EOO from the predicted distribution based on a thresholded SDM.

In chapter 4, I apply the currently best possible methods to generate accurate SDMs and estimate the range size of species to the large dataset of Gabonese plant species records. All significant SDMs are used here to assess the unique contribution of narrow-ranged, widespread, and randomly selected species to patterns of species richness and weighted endemism. When range sizes of species are defined based on their full range in tropical Africa, random subsets of species best represent the pattern of species richness, followed by narrow-ranged species. Narrow-ranged species best represent the weighted endemism pattern. Moreover, the results show that the applied criterion of widespread and narrow-ranged is crucial. Too often, range sizes of species are computed on their distribution within a study area defined by political borders. I recommend to use the full range size of species instead. Secondly, the use of widespread species, of which often more data are available, as an indicator of diversity patterns should be reconsidered.

The effect of global climate change on the distribution patterns of Gabonese plant species is assed in chapter 5 using SDMs projected to the year 2085 for two climate change scenarios assuming either full or no dispersal. In Gabon, predicted loss of plant species ranges from 5% assuming full dispersal to 10% assuming no dispersal. However, these numbers are likely to be substantially higher, as for many rare, narrow-ranged species no significant SDMs could be generated. Predicted species turnover is as high as 75% and species-rich areas are predicted to loose many species. The explanatory power of individual future climate anomalies to predicted future species richness patterns is quantified. Species loss is best explained by increased precipitation in the dry season.

Species gain and species turnover are correlated with a shift from extreme to average values of annual temperature range.

In the final chapter, the results are placed in a wider scientific context. First, the results on the methodological aspects of SDMs and their implications of the SDM pipeline are discussed. The method presented in this thesis to simulate large numbers of species offers opportunities to systematically investigate other matters of the pipeline, some of which are discussed here. Secondly, the factors that shape the current and predicted future patterns of plant species richness in Gabon are discussed including the location of centres of species richness and of weighted endemism in relation to the hypothesized location of glacial forest refugia. Factors that may contribute to the lower species richness of African rainforests compared with South American and Asian forests are discussed. I conclude by reflecting on the conservation of the Gabonese rainforest and its plant species as well as on the opportunities SDMs offer for this in the wider socio-economic context of a changing world with growing demand for food and other ecosystem services.

Acknowledgements

Doing this PhD study and writing this thesis pushed me out of my comfort zone, which was exactly what I set out for. Getting familiar with and succeeding in a scientific discipline entirely new to me, has been a wonderful, albeit not easy journey.

Four aspects have been crucial for the work presented in this thesis and therefore I start by acknowledging the people related to these aspects. First of all, all people in Central Africa, Gabon in particular, are thanked for their hospitality and cooperation. I feel privileged having visited the stunning Gabonese rainforest three times during my MSc and PhD study, each time being welcomed and helped by the Gabonese. Secondly, this study used herbarium specimen records that have been collected over the past three centuries. All plant collectors and herbarium curators are thanked for their incredible effort to document the distribution of plant species in tropical Africa and sharing these data. The third crucial aspect is the digital availability of herbarium specimen data. Everyone who contributed to the digitalization of the millions of specimens worldwide are thanked for disclosing this mind-boggling volume of data. Fourthly, my research depended on the availability of seriously strong computers and clever *R* functions. All the computer geeks out there are thanked for developing these great tools and sharing so much *R* programming code online.

More in detail, I thank the following people. In Gabon, Henri-Paul Bourobou-Bourobou, Flore Koumba Pambo, Kath Jeffrey, Lee White, Thomas Nzabi, Pulcherie Bissiengou, and all staff of the Herbier National du Gabon, as well as Ard and Marijke Louis are thanked. All colleagues from Naturalis in Leiden and Wageningen University are thanked for their friendliness and help in finding publications, solving computer issues and for digitizing and curating the herbarium specimens. The Nationaal Herbarium Nederland, now part of Naturalis, is thanked for funding this PhD study and the Alberta Mennega Foundation is thanked for financially supporting both my field work and congress visit.

All colleagues from the Biosystematics group in Wageningen are thanked. Without forgetting others, I here thank Wilma, Setareh, Sara, Kitty, Lotte, Ximena, Ed, Lars, Freek, Roel, Tao, Robin, Pieter, Eric, as well as Erik, Theo, Jos, Frans, Lubbert, Paul and Hiltje. The students I supervised, Erik-Jan, Sander and Wessel, are thanked for their help, numerous questions and clever ideas that improved the quality of this work. The colleagues of the *R* user group are thanked: Phuong, Geovana, Jenny, Madelon, Karen, Mart and Paul. A special thank you for my two paranymphs, Floris and Edwin.

A special word of gratitude to Marc, Jan and Niels, my promotor and co-promotors. Marc: thank you for your confidence, patience and advice when I needed it most. Niels: thank you for your endless comments on methodological steps and *R* scripts. Although I did not always like getting so many of them at first, I did appreciate it in the end. Jan: doing fieldwork together was great and your expertise of the tropical African plant collection is unequalled.

My friends were always there to remind of life outside university and helped me to stay healthy, clear my mind and simply have fun. Friends from tennis clubs Smashing Pink Amsterdam, NVLTB Wageningen and de Hoge Wick Oosterhout: see you all soon on & off court. Hein Jan and Alex: thanks for being such wonderful GLTA tournament co-directors. To the members of the Humboldt Society, David, Pieter, Ton, Erik, Jan & Loic: thanks guys for sharing our love for plants, good wine and botanists. Jesse: thanks for sending odd plant pics for ID and discussing our mutual research projects. Frank: thanks for teaching me pinball, we should practice more often though. Dirk-Jan: I lost count of the number of beautiful places we've seen, let's discover more.

My parents were the first to notice my passion for nature in general and plants in particular. They took me out and taught me about plants and animals and always encouraged me to follow my heart, for which I'm deeply grateful. Also thank you to my wider family, Gerard & Mirjam, Tebbe & Carla, Bas & Anne, Koen & Maaike. And finally, the last and most important thank-you is for you Allan. We've said it many times before: "never a dull moment". Well, this PhD added a whole new dimension to our lives. Thank you for understanding, supporting, pushing me when I needed it and celebrating the successes. I'm looking forward to our next adventure, I'm sure there are many more to come.

Education statement

PE&RC Training and Education Statement

With the training and education activities listed below the PhD candidate has complied with the requirements set by the C.T. de Wit Graduate School for Production Ecology and Resource Conservation (PE&RC) which comprises of a minimum total of 32 ECTS (= 22 weeks of activities)

Review of literature (6 ECTS)

- Review of literature, presented in own chair group and for annual PhD day of research school 'Biodiversity' (2010)

Writing of project proposal (4.5 ECTS)

- NWO project proposal 'What determines plant species diversity in Central Africa?' (2010)

Post-graduate courses (4.1 ECTS)

- Geo-Ecological Data Analysis; Amsterdam Graduate School of Sciences (2010)
- Geostatistics; PE&RC (2010)
- Introduction to R for statistical analysis; PE&RC (2011)

Invited review of (unpublished) journal manuscript (3 ECTS)

- Annals of Botany: phylogeny and biogeography (2010)
- Plant Ecology and Evolution: predicting distribution of functional groups (2013)
- Journal of Biogeography: species richness patterns based on species distribution models (2015)

Deficiency, refresh, brush-up courses (9 ECTS)

- Advanced Statistics; WUR (2010)
- R Courses; self-study (2010-2014)

Competence strengthening / skills courses (1.5 ECTS)

- Organising the Caribbean Botany symposium and book presentation; Naturalis, Leiden (2012)
- The essentials of scientific writing and presenting; Wageningen in'to Languages (2014)

PE&RC Annual meetings, seminars and the PE&RC weekend (1.3 ECTS)

- PE&RC Introduction weekend (2011)
- PE&RC Day (2012)
- PE&RC Introduction weekend; oral presentation (2014)

Discussion groups / local seminars / other scientific meetings (5 ECTS)

- Spatial Methods, PE&RC discussion group (2010-2013)
- R User Group; PE&RC discussion group (2010-2014)
- Wageningen Evolution and Ecology Seminars series (2010-2014)
- Journal club; Biosystematics Group (2010-2016)

International symposia, workshops and conferences (6.7 ECTS)

- Botanical Diversity Symposium; Meise, Belgium (2010)
- Caribbean Botany symposium and book presentation; Naturalis, Leiden (2012)
- AETFAT (Association for the Taxonomic Study of the Flora of Tropical Africa); Stellenbosch, South Africa (2014)
- Netherlands Annual Ecology Meeting; Lunteren, the Netherlands (2015)

Lecturing / supervision of practicals / tutorials (25,8 ECTS)

- Diversity of the Netherlands (2011-2012)
- Webs of terrestrial diversity (2011-2013)
- Advanced biosystematics (2011-2013)

Supervision of MSc students (6 ECTS)

- Testing the testers: chaff and wheat in ecological niche modelling
- Incorporation of dispersal barriers and long distance dispersal in species distribution models of Gabonese taxa

The research described in this thesis was financially supported by the Nationaal Herbarium Nederland and the Alberta Mennega Foundation.

Financial support from Wageningen University for printing this thesis is gratefully acknowledged.

Cover photographs by Andreas S.J. van Proosdij

Printed by Digiforce

Propositions

1. In multi-species distribution modelling studies, applying a uniform lower limit to the required sample size results in an incorrect evaluation of a number of the models obtained.
   (this thesis)

2. Using the Extent Of Occurrence and Area Of Occupancy to assign an IUCN Red List category to species hampers the conservation of these species, in particular of rare, but widespread species.
   (this thesis)

3. The increasing life expectancy of humans enables scientists to concatenate scientific careers in multiple disciplines leading to new opportunities for top-quality inter- and multi-disciplinary research.

4. Despite the increasing availability of open source species occurrence data, that improve the basis for data-driven decision making on conservation priorities, such decisions will still be greatly influenced by public emotion.

5. The ever increasing over-protection of today's children from accidents during playing harms the exploring nature of tomorrow's scientists.

6. The appreciation of products, values and people is subject to the time, place and context of their visibility.

Propositions belonging to the thesis, entitled

What determines plant species diversity in Central Africa?

Andreas Simon Johan van Proosdij
Wageningen, 6 October 2017