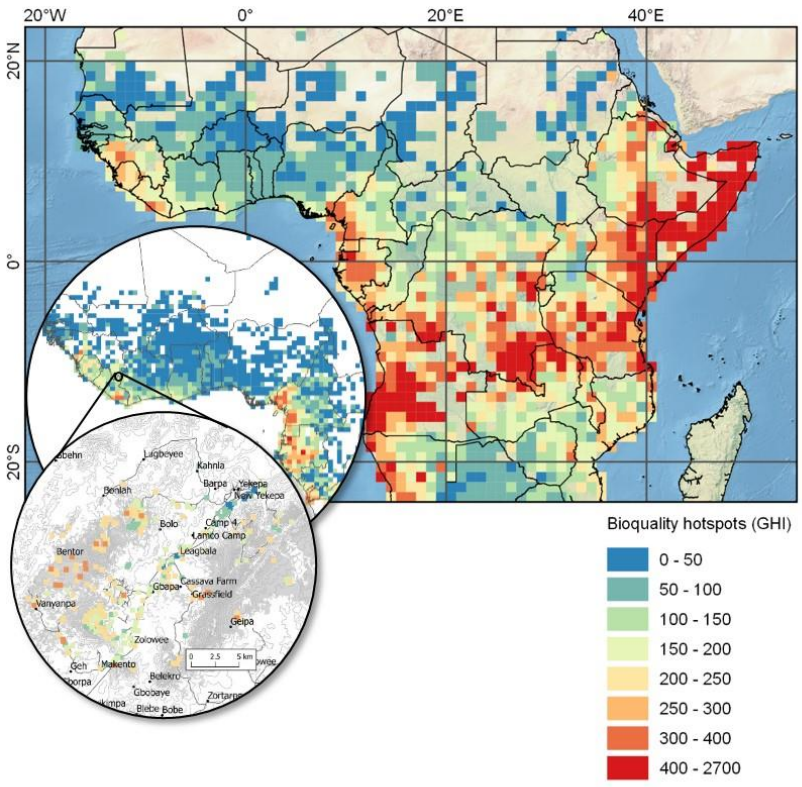
Endemism hotspots and floristics of Upper Guinea in the context of tropical Africa

Cicely A. M. Marshall

Department of Plant Sciences, University of Oxford

Merton College



Thesis submitted for the Degree of Doctor of Philosophy

Trinity Term 2017

I thank my supervisor, Stephen Harris, for agreeing to take me on as a DPhil student, and for supporting me to complete this thesis. I thank William Hawthorne for inspiring me so thoroughly, and for teaching me most of what I know of African botany. My family, Beth and Blake, thank you for your love and encouragement. I thank my friends in the office for making the department a happy place to work.

I thank Robert Scotland for reading and questioning my work over the years, Lindsay Turnbull and Andy Hector for bringing out my quantitative side, Mike Swaine for discussions and data, Jan Wieringa, for prompt and apt advice and encouragement, Jonathan Dabo and Markfred Mensah for their fieldwork expertise, Cyrille Chatelain for sharing the African Plants Database, and Denis Filer and Andrew Liddell for website and database help and developments.

I thank Samartex, FORIG and the Ghana Herbarium for facilitating my field work, and in particular acknowledge the support of Patrick Ekpe and Musah Abu-Juam, John Howell and Jamison Suter of ArcelorMittal Liberia and BHP Billiton for supporting the field work in Liberian Nimba and promoting its use in science. This thesis benefits from the work of hundreds of botanists and RBS team members who have collected specimens and made their data available.

This thesis was funded by the Clarendon Fund, with matched support from Merton College. I am very grateful for the opportunity afforded to me. Additional funding for the new field work in SW Ghana was given by the Oxford University Expeditions Council, Wilson Fund, Department of Plant Sciences and Merton College.

Contents

Appendices

Supplementary Material

Supplementary material for each thesis chapter can be found in the accompanying CD; supplementary appendices are described following the Chapter to which they relate.

Cicely A. M. Marshall, Department of Plant Sciences, University of Oxford
Submitted for the degree of Doctor of Philosophy, Trinity Term 2017.

Abstract

This thesis documents patterns in plant species distribution across tropical Africa. Geographic patterns in the distribution of globally rare plants within Upper Guinea are emphasised, and correlates with these patterns are investigated. This thesis argues that globally rare species can and should be emphasised in conservation strategy.

Approximately 3.7 million global occurrence records of 28,803 tropical mainland African vascular plant species are compiled into a database framework. The database is used to propose an updated biogeographic framework for tropical Africa, which is sympathetic to previous chorological frameworks but maximises regional endemism within quantitatively defined boundaries. A definition of Upper Guinea as the forests of West Africa between Sierra Leone and Ghana is recovered.

The concentration of globally rare species in the tropical African flora (bioquality as measured by the GHI) is calculated and mapped at one degree square and half degree square resolution, revealing high bioquality of the horn of Africa. Bioquality is calculated by categorising the global area of occupancy of all tropical African taxa into a Star rating, with the result that a bioquality score can now be calculated at a local scale anywhere in tropical Africa to inform conservation strategy.

At the local scale, variation in bioquality is modelled in two areas of high global endemism within Upper Guinea: the Nimba Mountains and SW Ghana. Disturbance is the only significant variable retained in both models, and shows a strong negative relationship with bioquality. Bioquality scores in forest reserves of SW Ghana are shown to have been stable over 20 years, although our perception of the global rarity and identity of species within the area has altered substantially. This finding supports the GHI as a metric of conservation priority in the face of partial information.

At half degree square resolution, 55% of tropical African cells are estimated to have less than 2.5% of their likely species richness documented. A regression tree model is used to interpolate bioquality scores for cells lacking species distribution data, making use of a range of modern climatic, paleo climatic, geographic and biogeographic variables, found to be predictive of bioquality scores. Areas showing the most stable climates over the past 21,000 years are shown to have the highest modern-day bioquality across tropical Africa.

Areas with relatively stable climates during periods of past climate change have been hypothesised to show high modern endemism as the result of increased speciation in isolated refugia, and/or a lower rate of extinction within the refugia. Two dated phylogenies are estimated for African magnoliids and Rubiaceae species, which show that most of Upper Guinea's globally rare species are young, vicariant species, although at least one species is likely to be a paleo-endemic, supporting both hypotheses.

Areas with historically stable climates showing present day climatic or topological isolation explain areas of high global endemism in tropical Africa, although the introduction of a high disturbance regime can remove this pattern. A conservation strategy which promotes the protection of globally rare species in tropical Africa is feasible, given what we now know of the African flora, and is wise, at least given the success of the Ghanaian programme over 20 years.

Chapter 1

General Introduction

The uneven distribution of plant biodiversity across Earth has absorbed biologists for centuries, and mapping and interpreting patterns of plant species distribution remains a central research area in tropical botany. Knowledge of the patterns and processes that have generated and maintain our tropical plant diversity are key to conservation theory and practice, and contribute to our general understanding of macroecological and evolutionary patterns. Understanding the evolutionary processes that have generated the richness of the tropical flora is predicated on accurate information about species identity and distribution. Studies in biogeography, taxonomy and ecology, combined with molecular studies of phylogeny, have much to contribute to our understanding of where and why plant species live where they do.

The flora of tropical Africa is better recorded than that of the tropical Americas or Asia. 90% of the tropical rain forest of Africa is found in the Guineo-Congolian region, which is home to c. 8000 angiosperm species (White 1983). The Upper Guinean forests are a subcentre of endemism within this large forest block, extending along the West African coast from Sierra Leone to Ghana. They are home to c. 2800 species, of which c. 25% are endemic (Jongkind 2004).

My aim is to illuminate the floristic patterns in Upper Guinea *sensu lato*, and in particular to account for patterns in the occurrence of globally rare species within it. My thesis is organised around three themes: Upper Guinea, its definition and botanic affinities; patterns of endemism within tropical Africa, and within hotspots of Upper Guinea; and the drivers of high endemism, at broad and local geographic scales, and over long and recent time scales. I hope that the research will result in improved conservation outcomes, benefitting African plants, wildlife, and the people who depend on them.

My particular interest is the Upper Guinean flora, especially its endemic elements, and most of my botanic experience has been in Upper Guinea. As the definition of Upper Guinea is somewhat controversial, the definition of Guineo-Congolia also becomes controversial, and a much broader study area must be considered when Upper Guinea is the subject of analysis. Upper Guinea is allied floristically to the Guineo-Congolian subregions of Lower Guinea and Congo, and Guineo-Congolia is allied to its adjacent floras in Eastern Africa, the Zambezian region, and the Sudanian region (White 1979, 1983, 1993) (Figure 1.1). Such a scope comprises much of tropical Africa. The tropical African flora intergrades with the North African flora and the South African flora, and the Madagascan flora is allied to, in particular within Africa, the coastal forests of East Africa; these floras in turn have affinities outside Africa. It is not easy, therefore, to set *a priori* limits on a study area. Truly though, the North African, South African and Madagascan floras have limited botanic affinity to Upper Guinea, and the high percentage endemism and species richness of Madagascar, southern Africa and the Mediterranean region of North Africa mean that those floras must be tackled outside of this thesis. Thus, the widest geographic scope tackled in this thesis is mainland tropical Africa, defined literally between 24°N and 24°S.

Species distribution data are the mainstay of phytogeography. Aspects of species ranges have been used to define regions, hotspots, species themselves, and conservation priorities; they have inspired evolutionary and ecological theories and have been called upon to reconstruct historical climate changes. The last major effort at compiling species distribution data for tropical Africa and Upper Guinea was c. 10 years ago, and resulted in around 90,000 records of 6000 species for tropical Africa (BIOTA-BISAP: Linder et al. 2005, Küper et al. 2006), and for Upper Guinea around 48,000 records (Poorter et al. 2004, Wieringa & Poorter 2004). The availability of botanic data has increased dramatically in 10 years, with notable advances and contributions coming from the digitisation and online dissemination of Lebrun & Stork (1991-2012) through the African Plants Database (APD), the NHN collection, the completion of the Flora of Tropical East Africa (FTEA, Royal Botanic Gardens Kew

1948-2012), and the digitisation of African type specimens through the Global Plants Initiative. In Chapter 2 I describe the assembly of a 40,000-taxon distribution database for tropical African vascular plants (TRAFRICA), and report databased taxon richness at tropical African, country, one degree square and half degree square resolutions. The database is compared to the RAINBIO dataset (Dauby et al. 2016, Sosef et al. 2017), a concurrent effort to compile and analyse a large database for tropical African plants. TRAFRICA was compiled as the framework of the thesis, and is almost certainly the largest database of tropical African plants to have been analysed, both in terms of species coverage and distribution data coverage. Such an effort is of significant utility to those working on the tropical African flora, in both academic and practical pursuits.

Until 2012, Upper Guinea was accepted as an outpost of the Guineo-Congolian forest region (Eig 1931, Lebrun 1961, Engler 1964, Good 1974, White 1979, Takhtajan 1986). A quantitative reappraisal of species distributions cast doubt on this interpretation, by subsuming Upper Guinea within the Sudanian region (Linder et al. 2012). I investigate whether this result was justified in Chapter 3, by re-evaluating Upper Guinea as a phytochorion using the TRAFRICA database. This serves to define my study region, and is intended to resolve some long-standing issues surrounding the northern and eastern termination of Upper Guinea (Mangenot 1955, Aubréville 1962, Léonard 1965, Clayton & Hepper 1974, Wickens 1976, Hall & Swaine 1981, Takhtajan 1986, Linder et al. 2005, Fayolle et al. 2014, Droissart et al. 2017).

White's (1983) biogeographic regionalisation for Africa is widely held to be the most authoritative biogeographic framework for African plants, although it is was drawn more than 30 years ago, without the benefit of multivariate methods and large distribution databases. I explore the extent to which his framework can be recovered within the TRAFRICA dataset, and present a modified framework which is more predictive of species distributions within this dataset (Chapter 3). This biogeographic analysis may stand as a useful framework for others studying tropical African biodiversity.

Globally rare species should be central to conservation strategy. Endemic species may be locally rare as well as globally rare, and both these aspects of their distribution make them vulnerable to extinction (Gaston 1998). Prioritising the conservation of globally rare species is an efficient way to maintain species richness globally (Brooks et al. 2006). Unlike species richness or diversity, the endemism of an area is not increased by disturbance or alien invasion; these properties make areas with high endemism a rightful priority for conservation and research efforts (Mittermeier et al. 2011). The number of species in an area, whether rare, threatened or simply present (richness), is now generally recognised as a poor metric for identifying conservation priorities, because richness alone reveals little more than the availability of data, the size and shape of the area under consideration, and the biome type (Brummitt & Lughadha 2003). Several endemism measures differ mostly in the degree to which they take account of, or correct for, species richness (Kier & Barthlott 2001).

In Chapter 4, the global ranges of c. 40,000 tropical African vascular plant taxa are calculated and summarised into one of four groups, called 'Stars'. Black Star species are the world's rarest plants, found in 2.7 degree squares on average; Gold, Blue and Green Star species have larger global ranges (Hawthorne 1996). The Stars are used to calculate and map *bioquality* across mainland tropical Africa: the weighted average of species present in an area, where the weight is inverse to the global range of the species present. Bioquality as measured by the GHI (Hawthorne 1996) is similar to range size rarity (Rabinowitz 1981, after Kier & Barthlott 2001), except that (i) species' ranges are measured globally at one degree square resolution rather than within the study area at any resolution, (ii) species ranges are categorised into 4 groups, called Stars, rather than being treated continuously, (iii) weightings for the Star categories are standardised globally, and (iv) the mean of rarity scores is calculated, rather than the sum, as is sometimes the case in range size rarity metrics. As species distribution data is far from complete, a minimum reliable metric for bioquality is defined in this Chapter. Bioquality, and minimum bioquality, are mapped at one degree square resolution

across mainland tropical Africa, and at a finer resolution across northern Nimba County, Liberia, to reveal patterns of endemism across the continent at these different geographic scales. Identifying the patterns of global endemism across tropical Africa is useful for prioritising conservation efforts, and help us to understand how plant biodiversity is distributed. The Star definitions themselves allow local bioquality scores to be calculated anywhere in mainland tropical Africa, allowing the global conservation value of a local area to be measured, useful for example in the assessment of impact on vegetation from major developments.

The recognition of Upper Guinea as a subcentre of endemism (Chapter 3, White 1979) implies a high degree of endemism, and the whole of Upper Guinea has been described as a hotspot (Myers et al. 2000; incorporating forest loss in the definition). Endemism in Upper Guinea is amongst the highest in Africa, although endemism is highest in the Cape (Kier & Barthlott 2001, Linder 2001, Küper et al. 2004, Kier et al. 2009). Although several authors have mapped endemism values within Upper Guinea (and further afield), most studies were conducted at a coarse resolution (2.5-degree squares or country level: Clayton & Hepper 1974, Brenan 1978, Linder et al. 2001) and revealed little more than the endemism of the region. For plant distribution data in Africa, a resolution of one degree square represents a reasonable compromise between accuracy and detail (Kreft & Jetz 2010).

Two studies (Küper et al. 2004, Wieringa & Poorter 2004) identified hotspots of endemism in Upper Guinea at a 1-degree square scale and a 0.5-degree square scale respectively. Range-size rarity in the BIOTA-BISAP dataset (Küper et al. 2004) peaked in the Nimba Mountains, Monrovia, SE Liberia/SW Ivory coast, and Abidjan, with slightly lower range size rarity in SW Ghana. This technique was sensitive to the different collection intensities for each cell, which probably accounts for the high range size rarity associated with Monrovia and Abidjan. Wieringa & Poorter (2004) mapped the range size rarity of species across Upper Guinea at half-degree square resolution, controlling for collection intensity, and identified high values in the Nimba Mountains, SE Liberia, SW Ivory Coast, and SW Ghana. Marshall et al. (2016) (Chapter 4) find high *bioquality* within Upper Guinea in Nimba,

SE Liberia, SW Ivory Coast and SW Ghana, as well as generally elevated scores for western Liberia and Sierra Leone, confirming the patterns shown by Wieringa & Poorter (2004).

Thus, within Upper Guinea are a number of hotspots of endemism, including the Nimba Mountains, SW Ghana, SE Liberia, and western Liberia. The designation of these general areas as 'hotspots' does not imply that globally rare species are found throughout them. In Chapter 5, the 'anatomy' of two of Upper Guinea's hotspots is explored, to investigate whether globally rare species are restricted to particular vegetation types within the hotspot. This is achieved through linear modelling of the GHIs of 114 and 454 vegetation samples within two of Upper Guinea's hotspots (south west Ghana, Nimba Mountains). 27 new samples were enumerated in SW Ghana, during the course of this thesis, for the purpose. The advantage of modelling in this context is that we can draw conclusions on the drivers of high endemism.

There is the possibility that the hotspots we see are an artefact of our current (incomplete) knowledge of species' true ranges (Nelson et al. 1990, Brown & Lomolino 1998, Lomolino 2004, Whittaker et al. 2005). The stability of species' Star ratings, and GHI scores, over 20 years, is explored for SW Ghana in Chapter 5. Assessing the stability of hotspots in the face of dynamic data is necessary for any interpretation of hotspot assemblage or maintenance, and has implications for the relevance of conservation priority setting over time.

Hotspots of endemism in Upper Guinea have been interpreted as a consequence of species adaptation to unusual modern-day ecogeographic conditions (Linder 2001, Mutke et al. 2001, Wieringa & Poorter 2004, Barthlott et al. 2007). A range of modern climatic, altitude, latitude, longitude, disturbance and local landscape position variables were collected to explore the correlation of each with bioquality: here we are asking, can the modern environment predict hotspots in Upper Guinea? By examining two of Upper Guinea's hotspots, we can be more confident of the generalisability of our conclusions. By working at the sample level with a continuous metric of

global endemism we can be precise about the relationship between endemism and its putative correlates.

In Chapter 5, hotspot scores were modelled without reference to past climate changes, because paleo-climate models cannot be rendered reliably at a resolution finer than 0.5 degree square resolution (Lima-Ribeiro et al. 2015, Varela et al. 2015). In Chapter 6, the role of climate stability in predicting bioquality is explored at the continental scale, since the last glacial maximum (21 kya), through the mid-Holocene (6 kya) interglacial phase to the present (1950-1999 average). Climatic stability, or change, has been related conceptually to areas of modern day high endemism (Harrison & Noss 2017) through refuge theory – the hypotheses being that species would have gone extinct outside of areas with stable climates (Stebbins 1974, Lovett & Friis 1996, Fjeldså & Lovett 1997), or that climatic change might have driven speciation through the necessity of adaptation (Haffer 1969, Hamilton 1976, Prance 1982, Robbrecht 1996).

An unsupervised machine-learning method is used to investigate the relative importance of climatic changes since the last glacial maximum in predicting high global endemism at half degree square resolution across tropical Africa, alongside 45 other putative environmental, geophysical, biogeographic, and modern climatic explanatory variables. The model is used to predict bioquality patterns across tropical Africa at half degree square resolution, as less than half of cells have >10 species records at this resolution. These predictions help to fill the gaps in our map until species distribution data are representative for tropical Africa as a whole. They also serve as testable hypotheses of our model of global endemism within tropical Africa – if our model is good, the GHI predictions should stand the test of time, and we have accurately captured the drivers of global endemism in tropical Africa. In the meantime, we have tested our model by removing portions of the training data. The model is used additionally to extrapolate to northern and southern Africa: these results are subject to considerably more uncertainty, but could be tested by other researchers with data from those regions.

There is as yet no consensus on the processes that have generated the diversity of Upper Guinean or African angiosperms, either inside hotspots or for the flora in general, over evolutionary timescales. Particular hypotheses have been presented and tested piecemeal, typically for one or a few genera under study, and taken together they represent a tiny fraction of African species or landmass. The relative ease with which dated phylogenies for taxa can be generated makes the task of dating species origins with respect to previous climate changes, and even assigning genera to modes of diversification, tractable for the first time for a large proportion of Upper Guinean/African angiosperms. The estimation of a mega phylogeny for African angiosperms could be used to distinguish mechanisms that are likely to have been significant for the evolution of globally rare taxa, seeking correlates with their ecology, dispersal ability and distribution. This initiative is underway (Janssens et al. *in prep*.)

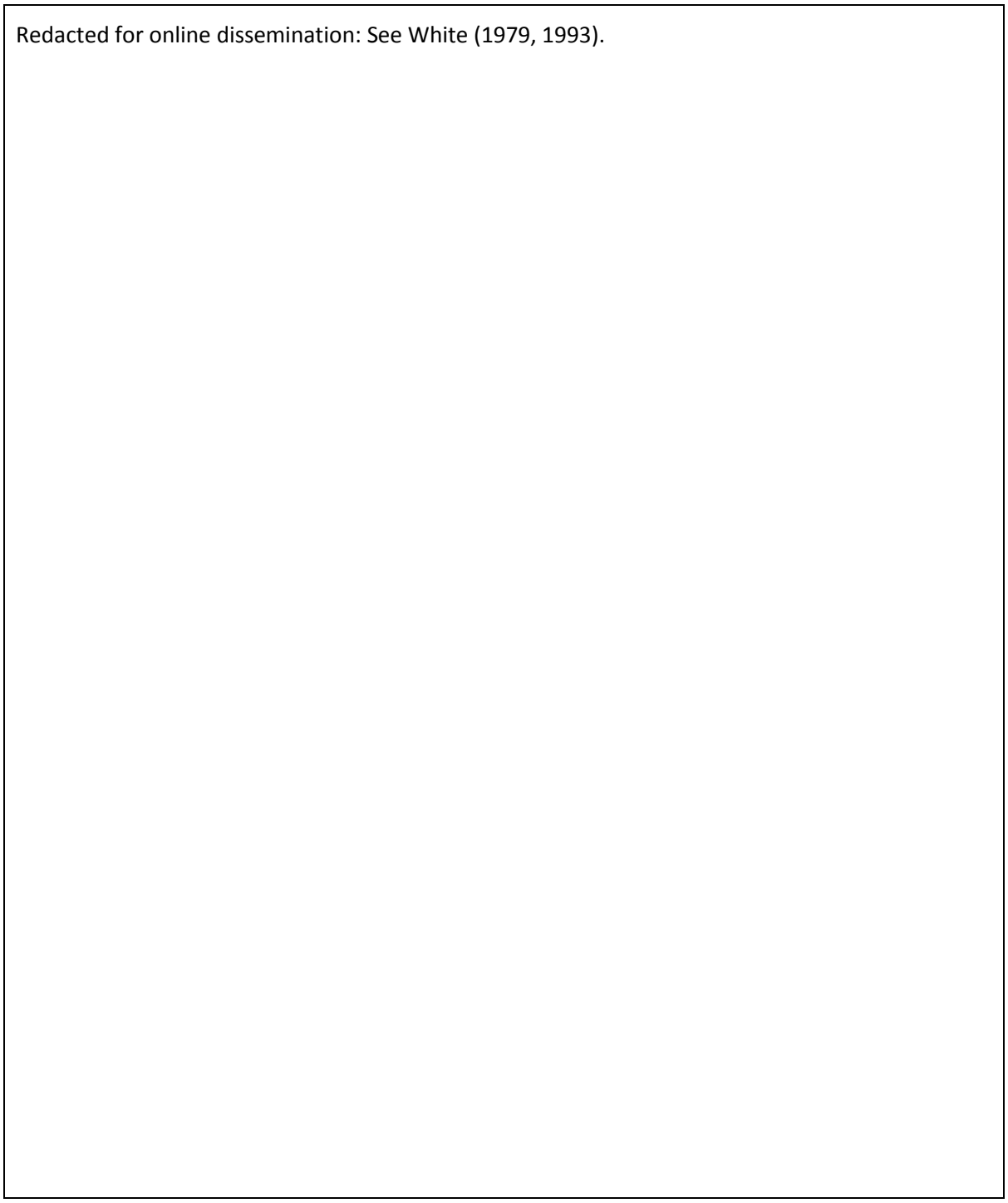
In Chapter 7 I explore the timing and modes of diversification for globally rare Upper Guinean species in two large clades: the Rubiaceae and the magnoliids. I estimate two dated phylogenies using the DNA barcode marker *matK*, for 25% of all African magnoliids, and 12% of African Rubiaceae species respectively, and compare the timing of species origins with respect to major historical climate changes in globally rare and globally widespread species. Details of species' distributions relative to areas of proposed climatic stability (refuges) are used to identify evolutionary mechanisms which may have been important for the diversification of the world's rarest species. Adding these few details to our understanding of the evolution of hotspots in Upper Guinea can benefit interpretations of other hotspots to build a global understanding of how biodiversity came to be distributed the way it is.

Some of the complexity and uncertainty associated with understanding and interpreting evolutionary patterns in species origins, particularly as concerns globally rare species, is the result of the complexity and uncertainty surrounding the identity and distributions of species. Much of this thesis has been concerned with laying the foundations to tackle this latter research question, and

was borne out of the difficulty I had in interpreting phylogenies without sufficient background data. It was also born out of frustration that patterns of endemism in Africa are almost always attributed to 'past climate changes' almost offhand, while more prosaic or at least more easily testable hypotheses have not yet been fully explored, and even the location of these areas of high endemism were subject to substantial vagueness. Meanwhile, that African species could not have diverged from each other within Pleistocene time frames seems to be another centrally held tenet of African botany (Moritz et al. 2000, Plana 2004), although molecular phylogenetic studies are very limited (Harris et al. 2000, Davis et al. 2002, Plana et al. 2004, Couvreur et al. 2008, Holstein & Renner 2011), and this assertion is challenged in Chapter 7.

In Chapter 8 I present the general conclusions from my thesis, answering the questions proposed in this introduction. A list of abbreviations follows Figure 1.1.

Figure 1.1. White (1993) regions of Africa, White (1979) subregions of Guineo-Congolia, and putative hotspots of Upper Guinea. 1: Upper Guinea subregion of the Guineo-Congolian Regional Centre of Endemism (RCE), 2: Lower Guinea subregion of Guineo-Congolia, 3: Congo subregion of Guineo-Congolia, A: Nimba Mountains putative hotspot, B: SW Ivory Coast/SE Liberia putative hotspot, C: SW Ghana/SE Ivory Coast putative hotspot, D: Western Liberia putative hotspot. II: Zambezian RCE, III: Sudanian RCE, IV: Somalia-Masai RCE, V: Cape RCE, VI: Karoo-Namib RCE, VII: Mediterranean RCE, VIII: Afromontane RCE; IX, Afroalpine (not shown). X: Guinea-Congolian/Zambezian regional transition zone (RTZ), XI: Guinea-Congolia/Sudanian RTZ, XII: Lake Victoria Regional Mosaic (RM), XIII: Zanzibar-Inhambane RM, XIV: The Kalahari-Highveld RTZ, XV: Tongaland-Pondoland RTZ, XVI: Sahel RTZ, XVII: Sahara RTZ, XVIII: Mediterranean/Sahara RTZ.



Abbreviations

AIC: Akaike information criterion
AM: Afromontane
AOO: Area of occupancy
APD: African Plants Database
APG: Angiosperm Phylogeny Group
Bio1-19: Climatic variables, see Table 6.1
BIOTA-BISAP: BIOTA project's Biogeographical Information System on African Plant Diversity
BK: Black Star
BRAHMS: Botanic research and herbarium management system
BU: Blue Star
C: Congolia
CART: Classification and regression tree
CC: Cyrille Chatelain
CEA: Congolian/E. Africa transition
CJB: Conservatoire et Jardin botaniques de la Ville de Genève
CM: Cicely Marshall
CVL: Cameroon Volcanic Line
DBH: Diameter at breast height
DCA: Decorana
DEM: Digital elevation model
DG: Dahomey Gap
DS: Degree square
DwCA: Darwin Core Format
EA: East African Herbarium
EIA, ESIA: Environmental (and Social) Impact Assessment
ESS: Effective sample size (phylogeny estimation)
FR: Forest Reserve
FS: Flora of Somalia
FTEA: Flora of Tropical East Africa
FWTA: Flora of Tropical West Africa
GADM: Global Administrative Areas database
GAM: Generalised additive model
GBIF: Global Biodiversity Information Facility
GC: Guineo-Congolia
GCM: General circulation model
GCSU: Guineo-Congolian/Sudanian transition zone
GCSUE: Guineo-Congolian/Sudanian transition zone (east)
GCSUS: Guineo-Congolian/Sudanian transition zone (south)
GCSUW: Guineo-Congolian/Sudanian transition zone (west)
GCZA: Guineo-Congolian/Zambezian transition zone
GD: Gold Star
GHI: Genetic Heat Index
GIS: Geographic Information System
GLM: Generalised linear model

GN: Green Star
GX: Green Star (anthropogenic distributions)
IPSL: one of the general circulation models
IUCN: International Union for the Conservation of Nature
JW: Jan Wieringa
KH: Kalahari-Highveld
KN: Karoo Namib
kya: Thousand years ago
LG: Lower Guinea
LG1: Lower Guinea (drier types)
LG2: Lower Guinea (wetter types)
LGM: Last Glacial Maximum
LS: Land surface (sensu Sayre et al. 2013)
LT: Lithology (sensu Sayre et al. 2013)
LV: Lake Victoria regional mosaic
MA: Madagascar (sensu APD)
masl: metres above sea level
matK: a chloroplast gene, used as a DNA barcode
MG: Vegetation macrogroup (sensu Sayre et al. 2013)
mHol: mid-Holocene
MRI: one of the general circulation models
MSE: Mean standard error
mya: Million years ago
NA: North Africa (sensu APD)
NG: Nigerian Guinea
NHN: Nationaal Herbarium Nederland
NMDS: Non-metric multidimensional scaling
NP: National Park
NPLD: Non-pioneer light demander
OOB: Out of bag
P: Pioneer
PAM: Partioning around medoids
PI: Pioneer Index
RAINBIO: Project to document and analyse species distributions in central Africa
rbcL: a chloroplast gene, used as a DNA barcode
RBS: Rapid Botanic Survey
RCE: Regional Centre of Endemism
RM: Regional mosaic
RMSE: Root mean standard error
RR: Resource Reserve
SA: Sahel
SA: Southern Africa (sensu APD)
SB: Shade-bearer
SM: Somalia-Masai region
SM1: Somalia-Masai transition zone
SU: Sudanian region
TA: Tropical Africa (sensu APD)
TM: Topographic moisture potential (sensu Sayre et al. 2013)

TO: Thermo/ombro types (sensu Sayre et al. 2013)
TRAFRICA: Database of tropical African species (Chapter 2)
TZ: Transition zone
UG: Upper Guinea
UG1: Upper Guinea (drier types)
UG2: Upper Guinea (wetter types)
UPGMA: unweighted pair group method with arithmetic mean
USGS: United States Geological Survey
WAFRICA: Database of western African species
WCMC: World Conservation Monitoring Centre
WCSP: World Checklist of selected plant families
WDPA: World Database on Protected Areas
WH: William Hawthorne
WR: Wildlife Reserve
ZA: Zambezian region
ZI: Zanzibar-Inhambane region

Chapter 2

TRAFRICA: A database of tropical African plants

Cicely A. M. Marshall[1], Jan J. Wieringa[2], and William D. Hawthorne[1]

1 Department of Plant Sciences, University of Oxford, South Parks Road, Oxford, OX1 3RB, UK.
2 Naturalis Biodiversity Center, National Herbarium of the Netherlands, Darwinweg 2, 2333 CR Leiden, The Netherlands.

Author contributions

Conceptualization: C.M. and W.H.; Methodology, Software and Analysis: C.M. and W.H.; Resources and Data Curation: W.H., J.W., and C.M.; Writing – Original Draft: C.M.; Review & Editing: C.M., SH., J.W., and W.H.

Acknowledgements

Cyrille Chatelain provided the African Plants Database Tropical Africa species list. Data capture, editing and broadcast of the African Plants Database is the product of a collaboration between the South African National Biodiversity Institute, the Conservatoire et Jardin botaniques de la Ville de Genève, Tela Botanica and the Missouri Botanical Garden. BISAP (Biogeographical Information System on African Plant Diversity) has been established at the Nees Institute, University of Bonn in the context of BIOTA Africa project together with several external partners, especially Jon Lovett (University of Leeds) and Peter Linder (Zurich). For more details about the dataset and a full acknowledgement of all contributors please refer to Linder et al. 2005 and Küper et al., 2006. C.M. acknowledges support from the Clarendon Fund and Merton College; W.H. acknowledges support from the James Martin 21st Century School and BP Biofuels.

Article status

Not intended for publication.

Abstract

Aim: To compile available species distribution records for tropical African vascular plant taxa.

Location: Mainland tropical Africa between 24°N and 24°S.

Methods: Records were compiled in BRAHMS, before being cleaned geographically in Geophila. The species backbone was drawn from the African Plants Database. The largest contributions by record are from the Global Biodiversity Information Facility (GBIF), the National Herbarium of the Netherlands, and a database assembled by W.H. for West African plants.

Results: The database, TRAFRICA, holds distribution records for 28,803 accepted species at one degree square resolution or finer within the mainland tropics of Africa. At one degree square resolution, there are 459,825 unique species x degree square records, and at half degree square resolution, there are 618,141 unique species x degree square records. The total number of records exceeds 3 million.

Main conclusions: The database is likely to be the largest yet compiled for tropical African species, in terms of species coverage and number of records. Nevertheless, at one degree square resolution, only 74% of mainland tropical African cells have >10 taxa recorded, and at half degree square, this figure drops to 45%. The exercise has revealed a number of challenges and surprises, including insights and developments in the way that we digitally geolocate records as a community; the high proportion of singleton species in the digital flora; the need to improve the inventory of plants even at country level and at relatively coarse resolution; the integration of different file formats and software into a coherent framework; and the challenges of sharing data, rights and responsibilities amongst the community of African botanists.

Introduction

The TRAFRICA (Tropical Africa) database is the largest and most complete inventory of tropical African species distributions that has been reported. The utility of a comprehensive, plant-specific, distribution database for tropical Africa is significant, allowing the analysis of floristic patterns across the continent, as well as having practical applications for plant conservation and land management. The database is central to this thesis; this chapter describes how TRAFRICA was assembled and the data it holds. I present some summary statistics describing the flora of tropical Africa, including richness at the specific, generic and family levels, and number of records and species at country, one degree square, and half degree square, resolution.

Conceptually, TRAFRICA is similar to RAINBIO (Dauby et al. 2016, Sosef et al. 2017), an effort to compile species distribution records for central Africa. As digital snapshots of the African flora, both TRAFRICA and RAINBIO are natural progressions from the BIOTA-BISAP dataset, which held 6,000 species at 1 degree square resolution. BIOTA-BISAP is included within TRAFRICA. RAINBIO and TRAFRICA were conceived and born independently and almost coevally, which is probably testament to the great advances in specimen digitisation over the last 10 years, as well as advancements in software and hardware capacity. The RAINBIO dataset has been published recently (Dauby et al. 2017, Sosef et al. 2017). The outputs of TRAFRICA and RAINBIO are complementary to each other, and each serves as a point of reference for the other. Together, they contribute to an exciting and productive period in African botany.

TRAFRICA would not have been compiled were it not for the African Plants Database (APD). In our opinion, this excellent resource is the most comprehensive and authoritative checklist of African plants, and the utility of a single taxon list, with a small number of curators making decisions over synonymy where necessary, is profound. We acknowledge that such a list will ever be as up-to-date and comprehensive as its maintainers may wish. For us, TRAFRICA is a logical progression from the

WAFRICA (West Africa) database maintained by W.H., which held some herbarium collections and many more records derived from vegetation surveys across West Africa. Our aim with TRAFRICA was to assemble the available data into a comprehensive distribution database for tropical African species in order to conduct a series of analyses, as published in this thesis and elsewhere.

Materials & Methods

Software

We first assembled TRAFRICA in BRAHMS (Filer, Dept Plant Sciences, 1985-2017). BRAHMS is written in Microsoft Visual FOXPRO and is a freely available database system dedicated to the management of botanical data. BRAHMS is the world-leading software for herbarium management and is used in diverse contexts, for example managing loans of herbarium material, or living collections. However, as much of our analysis is outside its scope of current activity, W.H. wrote a bespoke FOXPRO database to compile, clean, geolocated, analyse, and house, TRAFRICA. The database software is currently known as Geophila. Geophila was not written for public release, and such a task would be significant.

Geographic scope

The geographic scope of the database was defined as tropical Africa. The tropics of Cancer and Capricorn are close to 23.5 degrees north, and 23.5 degrees south, and we rounded to 24 degrees north and south where we had a choice. As is discussed below, the species list and distribution data do not respect these hard boundaries, and many South and North African species have been included because of the practicalities of assembling the database. However, the database should not be used outside of tropical Africa, where it is unrepresentative of the flora and available records. Although we did not aim to exclude tropical African islands, due to discrepancies in naming of territories, coastline boundaries, and search procedures, the database is also less reliable for non-mainland tropical Africa, and we have generally restricted our analyses to mainland territory. At the

northern limit, 24 degrees north includes all of the Sudanian region, all of the Sahel, and up to 3 degrees latitude of the Sahara (*sensu* White 1983). At the southern limit, 24 degrees south crosses (east to west) the Karoo-Namib, the Kalahari Highveld, and the Zambezian Regions. The first two of these are predominantly found further south. The Zambezian Region is predominantly tropical African, but extends to about 25 or 26 degrees south towards the east of mainland Africa to include the South African state of Limpopo. Zambezian Region species, which are part of the tropical African flora (*sensu* APD), will have had most their SA range extensions included, because any botanical records associated with their species name were submitted to the database. In the south, Namibia, Botswana, South Africa and Mozambique are not fully within the tropics. The APD TA area includes Mozambique, but excludes the first 3 countries.

Taxonomic scope

The taxonomic scope of the database was defined as any vascular plant collected on the mainland of Africa between the tropics of Cancer and Capricorn. The aim was to build a complete inventory. We limited the inventory to vascular plants because other plant groups are too poorly known to assess accurately their distribution and taxonomic status. We did not attempt to distinguish between native, naturalised, introduced or invasive species by any very thorough means. Species whose distributions we knew to be anthropogenic have been given the Star 'GX' (Chapter 4), which serves as an approximate estimate of the number and identity of introduced or cultivated taxa.

Species names & synonymy data

We compiled first a species list in BRAHMS. BRAHMS has one table for 'species' names, which holds the finest level of identification possible for a botanic record. This means that records in this table include: infraspecific taxa, generic/family-level classifications (e.g. *Psychotria* sp.; Rubiaceae sp.), unpublished taxa (e.g. *Hypoxis* sp. A of Flora Zambesiaca), *auct* names (where the name has been erroneously applied in a prominent reference work such as a regional flora), hybrids and cultivars, in

addition to validly published binomial species names. 'Mother species', defined as validly published binomial names excluding vague names, hybrids, cultivars and infraspecific taxa, were identified and marked in the database.

The initial tropical African species list was derived from the tropical African section of the African Plants Database (APD) and was transmitted to C.M. by Cyrille Chatelain in September 2014. This was the TA (tropical Africa) section of the African Plants Database, which was originally drawn from Lebrun & Stork's (1991-2012) checklist for tropical Africa, and is maintained by C.C. to reflect taxonomic changes. We reformatted the original MSExcel export for import into BRAHMS, with limited editing of the content except to add in missing synonyms (e.g. where the accepted name was found in the South African (SA) database). We codified the species list for WAFRICA against TRAFRICA, and added in any species records not present in TRAFRICA. These new species records included so-called vague names from survey work (e.g. '*Gaertnera sp.1 Ankasa*'), alternate spellings for species names (e.g. *Byttneria catalpifolia*/*Byttneria catalpaefolia*), fern species (which were not comprehensively treated for TA in APD at the time), and some synonyms not recorded in APD.

We followed APGIII (2009) for angiosperm family names. At the generic and specific level, synonymy follows APD 2014 as far as possible. Species of the Poaceae, Cyperaceae, and Orchidaceae were edited to reflect Kew's World Checklist for Selected Plant Families (WCSP, Govaerts et al. 2016), because it is more authoritative for these monocot groups than APD. For Pteridophytes we followed the taxonomy presented in the Flora of Tropical East Africa (FTEA) via JStor Global Plants as far as possible (RBG Kew 1948-2012). Accepted names and some synonyms were added from these sources where they were missing in TRAFRICA. Monographs and original publications were consulted where information was particularly current, or where a discrepancy between APD, WAFRICA, or other authoritative sources, such as the Plant List or WCSP, needed resolving.

Autonym subspecies were added where missing. All names were checked for cyclical synonyms (where both names point to the other as the accepted name) and were corrected. Designations of infraspecific taxa as varieties or subspecies follow the first source which donated that name (mostly APD), and subsequent variants of the same name were made synonyms or were overwritten.

Distribution data

Once a working species list had been established, we started populating the collections file of TRAFRICA. The collections file holds botanical records. A botanical record is any record of a species name in a place. Botanical records describe a species' distribution. Botanical records can be localised at any resolution, from GPS coordinates to 'Africa'. A botanical record may be associated with a vouchered specimen in a herbarium, but can also be an unvouchered observation.

Species names and synonyms were added to the database iteratively with distribution records. Before each new batch of botanic records was added, any new species names were either accepted, made synonyms of existing species entries, or merged to existing species entries, in order to keep the species list clean. 'Merging' moves all linked records from one record to another record before deleting the original entry (e.g. the species list entry '*Garcinia afzeli*' was merged to the entry '*Garcinia afzelii*': all botanic records stored under the first name were moved to the second name before the first name was deleted). A clean species list is essential, or botanic records for a species are split between name variants.

We began with the collections file of WAFRICA. These records came from survey work conducted in the region, and were enriched by the European Community-supported ECOSYN project (1996), which digitised many West African herbarium records. The Hall & Swaine database of Ghanaian records was included in WAFRICA (Hall & Swaine 1976, Hall & Swaine 1981).

C.C. sent C.M. records of species from Cote d'Ivoire, and country-level distribution data for the tropical African flora.

Jan Wieringa transmitted the Nationaal Herbarium Nederland (NHN) BRAHMS database in its entirety to C.M., once in October 2014, once in September 2015 after NHN's digitisation efforts were complete, and a final update in June 2016. NHN has a very strong African collection which is managed in BRAHMS. Herbarium records from the NHN collections are available online (NHN 2016). Distribution data for tropical African species were extracted from the database by C.M. We extracted from the NHN herbarium database the botanical records located in Africa, and codified the collections against TRAFRICA's species list. Species that were not codified at this stage were removed from the file, and were not imported. The result of this was that any record of a tropical African species from anywhere in Africa was maintained. Vague records, non-vascular plant records, and records only localised to 'Africa' were also removed from the file and not imported, to reduce the number of only marginally helpful records. In October 2015, the NHN's herbarium digitisation project was complete, and we were given an update of their latest database. This included another 1.6 million records in addition to the 2.4 million records from 2014 (not all African), as well as edits to the earlier records. The new data were imported, and edits to previously imported records were incorporated.

The BIOTA-BISAP plant distribution dataset was transmitted to C.M. in October 2014. This database was assembled by (W. Kuper of) Bonn University, and included selected families which had been recently monographed, and survey data from willing collaborators (including W.H.'s RBS data for Ghana). The BISAP dataset covers mainland Africa, south of 19 degrees north, and is gridded to one degree square. We imported to TRAFRICA the botanic records from all the grid cells of tropical Africa, omitting cells south of 24.5 degrees. At this stage the botanic records were assigned point coordinates in the middle of their degree square, although they have subsequently been geolocated as degree square polygons.

Distribution records for African conifers were extracted from the BRAHMS database maintained by Aljos Farjon by C.M. with permission, and for Mount Mulanje, Malawi, created by Alison Strugnell

(Strugnell 2006). Four published checklists, all from the journal CheckList, were included by formatting the pdf text to dbf tables for import (Mbayngone et al. 2008, Ouédraogo et al. 2011, Assédé et al. 2012, Ismail et al. 2015). There are many more published checklists, and floras, which could have been included, and this activity is ongoing.

GBIF

Distribution data from the Global Biodiversity Information Facility (GBIF) were incorporated into the database. GBIF is the world's largest repository for species distribution data. Data download was broken up geographically and taxonomically. Only vouchered records were downloaded, and only records with 'no known geographic issues' for records with coordinates. Downloads in Darwin Core Format (DwCA) were formatted for BRAHMS by C.M.

Using GBIF's browse occurrences website option, we applied the following filters to all occurrence data: Magnoliophyta (angiosperms); Pteridophyta (ferns and fern allies); Specimen; no known coordinate issues, and georeferenced only. These filters were designed to select the most reliably geolocated records for taxa of interest. We omitted observation data (records not supported by a herbarium voucher) to increase the reliability of the data and reduce the overall number of records to handle. Because the number of records left after these filters were applied was large (>1 million), it was necessary to split up the record downloads both taxonomically and geographically. For the Pteridophyta, we applied a bounding box at N23.5, -24; E-20, 55, to cover tropical Africa, resulting in 44,822 records. For Magnoliophyta, we applied two bounding boxes covering the northern portion of tropical Africa (24 to 1 degrees latitude, 758,773 records), and the southern portion of tropical Africa (-24 to 1 degrees latitude, 563,587 records). These records were downloaded in Darwin core format and extracted. We did not extract gymnosperm records from GBIF (an oversight), but gymnosperm records were not excluded from any other data source, and the conifer dataset of A.

Farjon was included. Archived links corresponding to the GBIF searches follow Table 2.1, and include the contributors to the dataset for each search, and terms of reuse.

The occurrence.txt download files were converted to .dbf tables suitable for import as BRAHMS botanical records by C.M. Because the size of a database file is a function of its width as well as its length, the wide Darwin Core Format files are poorly suited to the manipulation of long datasets, and splitting files into many separate shorter files is error prone. In the end, the only fields maintained in the database were the taxon name fields, the GBIF reference number, the geolocation fields, collector & institution fields, and date fields.

The structures of these .dbf tables were edited to use standard BRAHMS field names. The records were codified against the species list, after cleaning the name fields to maximise the chance of matches. Name cleaning included bringing all name information together in one appropriate field (e.g. all GBIF fields related to generic names were merged under the field genus), editing all name fields to fit our species list (e.g. replace all rank1 with “subsp.” For taxonrank=”SUBSPECIES”), and adding in missing “sp.”. We removed at this stage all records not identified to family level or below.

For geolocation fields, country codes were converted to full country names and cleaned, and records from countries outside mainland Africa were deleted. All fields related to textual geolocation were amalgamated into a Locnotes memo field. Any records with major or minor areas not consistent with the including region were deleted (e.g. a batch of records with US States as their major areas, likely a reversed lat/long issue from the data provider). Coordinate data were collated and converted to BRAHMS formats. Collection date fields were converted to BRAHMS fields, but other date fields were not. Fields related to collectors, collection numbers, institutes, and people were summarised into one memo field, the collector set to GBIF, and collection number set to the GBIF ID. Whilst this is not ideal (better to save the herbarium collector and collection number), it was the only feasible option given the plethora of fields in which this information was stored, and the number of records

involved. However, all original information was maintained in a Memo field, and any original field can be reimported if necessary using the GBIF ID. All other fields were deleted to reduce file size and corruption risk (any field can be reimported relating on the ID).

From these prepared files, records which could not be matched to a TRAFRICA species name were moved to a separate file. Although this will have resulted in some records being unfairly dismissed (e.g. typing errors in the genus/species names which would still leave the name recognisable to a human), it was not feasible to check these records by hand. In the end, around 132,000 records were dropped from the original download, because of issues with names, geolocation, or missing data, and approximately 1.23 million records imported.

We later searched GBIF for all records from tropical Africa without coordinate georeferences. This batch of around 500,000 records was surprisingly rich, because all the JStor scanned type specimens for Africa have been submitted to GBIF, and many of these are old and not stored with coordinates. We also had to go back to GBIF for all plant records from the Democratic Republic of Congo, as DRC was not included in GBIF's list of countries until 2015.

Out of Africa data

Records for tropical African species from outside of tropical Africa were gridded at the one degree square resolution for the purpose of revealing global distributions (Star rating, see Chapter 4) and were not submitted to the main database (c. 700,000 records). We assembled these non-African records from two sources only; the NHN database and GBIF. The NHN database was queried for collection records from our tropical African species list where the continent did not equal 'Africa'. We queried GBIF through the R package rgbif (Chamberlain 2016) for any species from our tropical African species list that did not already have a range recorded as >50 degree squares in TRAFRICA, where Continent equalled all continents except Africa, or equal to Africa excluding the original bounding box search area. The NHN and GBIF results were merged in a single .dbf, gridded at one

degree square resolution, and made unique to give only one record of a species for one degree square. For each species, the number of degree squares occupied (outside of Africa) was summarised. Because these search terms exactly omitted our original search procedures for the 'inside Africa' distribution data, for each species we added the number of degree squares occupied outside Africa to the total degree square occupancy inside Africa.

TRAFRICA records by provider

Table 2.1. Sources of distribution data included in the TRAFRICA database. Readers must defer to the data providers for terms of reuse.

Dataset	Geographic coverage	Number of records in TRAFRICA	Provided by
GBIF records	Tropical Africa, Global	1,659,027	Various, see below
WAFRICA	West African	514,912	University of Oxford Dept of Plant Sciences
NHN database	Global	493,729	NHN
WCSP country level	Global	80,314	WCSP
BIOTA-BISAP	Africa	73,759	Uni. Bonn
CJB Ivory Coast	Ivory Coast	64,054	CJB
CJB country level	Africa	62,236	CJB
Nimba survey data	Local: Nimba County	31,869 8,005	ArcelorMittal Liberia Euronimba Liberia Ltd (as curated by WH & CM, University of Oxford Dept of Plant Sciences)
Hall & Swaine 1981	Ghana	19,132	M.D.Swaine *pers.comm*. to W.H. (as curated by WH & CM, University of Oxford Dept of Plant Sciences)
Aljos Farjon, Conifers	Global	1086	Alijos Farjon
Alison Strugnell	Mount Mulanje, Malawi	1378	Alison Strugnell
Ouédraogo et al. 2011	Arly National Park, Tapoa, Burkina Faso	465	Ouédraogo et al. 2011
Mbayngone et al. 2008	Pama, Burkina Faso	448	Mbayngone et al. 2008
Assédé et al. 2012	Pendjari, Atacora Province, Benin	682	Assédé et al. 2012
Ismail et al. 2015	Eastern Nuba Mountains, South Kordofan, Sudan	246	Ismail et al. 2015

GBIF.org (3rd July 2015) GBIF Occurrence Download http://doi.org/10.15468/dl.cw8ol3
GBIF.org (3rd July 2015) GBIF Occurrence Download http://doi.org/10.15468/dl.wc8wzw
GBIF.org (21st July 2015) GBIF Occurrence Download http://doi.org/10.15468/dl.jbojyk
GBIF.org (19th January 2016) GBIF Occurrence Download http://doi.org/10.15468/dl.wjnsjs
GBIF.org (17th March 2016) GBIF Occurrence Download http://doi.org/10.15468/dl.yuylfv
GBIF.org (17th March 2016) GBIF Occurrence Download http://doi.org/10.15468/dl.1l8wgx

Cleaning geographic data

At this stage it was clear that the raw collections file in TRAFRICA would not be clean enough to produce meaningful results. Visualising the locations of the distribution records at this stage revealed many records in the sea, a harbinger for less obvious coordinate mistakes on land. We embarked on a phase of data cleaning, which was led by, and mostly conducted by, W.H. This was the stage at which TRAFRICA left BRAHMS. Geophila was written largely out of the need to clean and store geographic locality information, and in particular to manage bounding box or polygon-based geolocation.

Place name dictionaries were imported into Foxpro and Manifold GIS from public domain gazetteers. Core spatial dictionary reference files were compiled in Manifold GIS version 8. Polygons for African administrative regions were downloaded from the Global Administrative Areas database (GADM). These were processed by GIS, assigned unique geocodes and split into six layers depending on the administrative hierarchy. WDPA Protected Areas polygons were also processed, as defined by the World Conservation Union and UNEP-World Conservation Monitoring Centre's World Database on Protected Areas (2007, WCMC, Cambridge, UK, 2007). This dataset is available through the Global Land Cover Facility. Individual polygons in each layer were coded with the geocode of all polygons at all coarser scales in which they were entirely enclosed. African place names and point localities were downloaded from the Geonames database. These were allocated unique geocodes and also assigned geocodes of all polygons in which they were enclosed at various levels.

Distribution database records were divided into those with specified coordinates, and those without. About 500,000 of the original distribution records did not have coordinates: the text locality information was compared against the core gazetteer dictionary files in a geocoding procedure to assign point-with-radius coordinates, or polygon coordinates where the specimen was not finely localised (e.g. only localised to a district, or country). Using these procedures we were able to assign polygon or point with radius coordinates to around 98% of them, although in many cases this was

for larger administrative region only which do not fit into single degree square cells. For records with coordinates, the coordinates were reverse geocoded to assign all hierarchical place names to the record, and the text locality information was also geocoded to assign point or area coordinates. Where the two sources of information for a distribution record conflicted, records were checked and corrected by hand where the problem was detectable, with the remainder being omitted from the analyses (see next paragraph). All subsequent analysis and mapping was conducted by asking whether the record's perimeter (or point with radius) fitted entirely inside the sampling area of interest, rather than by assigning points to the middle of the bounding box as is commonly done, so that the geographic resolution at which the record was originally collected could be respected.

We manually checked 2,033,783 of the records (c. 67% of all records) for geographic accuracy. We achieved this by first making a file of unique locations (text locality-coordinate combinations). We checked: each unique gazetteer with more than 25 records; localities which are typically problematic (for example country centre points, frequently used in lieu of country polygons); and localities where stated coordinates conflicted with stated text locality information. Also included in these 2 million checked records are records for which we trusted the collectors (e.g. our own samples located by GPS, or BIOTA-BISAP dataset). Although we checked the locality records with the greatest impact on results, this still leaves a significant number of original records unchecked. For many localities with few botanical records and no supplied textual locality information, we are, effectively, trusting the original collectors and data providers, having taken reasonable precautions to identify situations where mistakes would have been made. 18,294 records that were otherwise usable were lost because they straddled a sampling (one degree grid cell) border. 5684 records were unable to be geolocated, and many others were reduced to country or top administration level areas (e.g. Western Region).

Table 2.2. Summary of taxon totals in TRAFRICA.

	Number of records	**Calculated from HTRAF on 27 July 2017:**
All species.dbf entries	97,309	Tag all
Hybrids	78	Untag Hybrid="*"
Cultivars	142	Untag cultivar="*"
Vague names	2,039	Untag Vague="*"
Uncertain names	1,772	Untag taxstat="unc"
Invalid names	190	Untag taxstat="inv"
Synonyms	52,399	Untag taxstat="syn"
Accepted taxa, excluding hybrids, cultivars, vague, uncertain, invalid names and synonyms	**40,783**	Remaining tagged (not the same as taking numbers away from total as some spp in >1 category) (tag profile acc40783)
Of which not introduced or cultivated	38,694	Star=GX and tag="*" (2089 GX)
Of which angiosperms	39,687	Fa.group <> to non-angiosperm groups
With data at 1 deg square	c. 32,802	Unique spnumbers From file biogeog/spdsistallrecsnotcleantemp.dbf
Accepted infraspecific taxa with distribution records at 1 deg square in tropics	c. 30,916	Unique spnumbers From file biogeog/spdsistallrecsnotcleantemp.dbf, lat<24.1 and lat>-24.1
Accepted infraspecific taxa with distribution records at 1 deg square in tropics, no known cultivated	c. 30,361	(from biogeog analysis which excluded Star=GX, but also cells with <3% species sampled)
Accepted mother species	**32,653**	Taxa tag profile (acc40783), untag for sp2<>" ". Copied to /redfiles/trafmo.dbf
Of which introduced or cultivated	1842	Star=GX
Number of genera	3793	Sort unique on gecode in trafmo.dbf
Number of families	311	Sort unique on facode in trafmo.dbf
Angiosperms	31,706	Add phylum from family file, sum
Pteridophytes	787	Add phylum from family file, sum
Lycopods	39	Add phylum from family file, sum
Gymnosperms	109	Add phylum from family file, sum
Non-vascular	12	Add phylum from family file, sum
Accepted mother species with distribution records at country level	**31,987**	Country dst in thesisdraft2/chapter2
Accepted mother species with distribution records at 1 degree square	29,547	R/TRAFstats/Slimdst.dbf (mother sp), spp with recs at one deg sq.
Accepted mother species with distribution records **at 1 degree square** resolution **in tropical Africa**	**28,803**	Rde/mothersp25julytaonly.dbf, sort unique on spnumber.
Accepted mother species with distribution records at 1 degree square resolution in tropical Africa, **angiosperm only**	**c. 27,967**	Assumes non-angiosperm taxa have same proportional representation in the dataset (likely to be less, conservative estimate).
Accepted mother species with distribution records at 1 degree square resolution in tropical Africa, **excluding known cultivated**	**27,690**	Rde/mothersp25julytaonly.dbf, sort unique on spnumber, star<>GX

Our species list must be close to complete as regards the current collective understanding of the tropical African flora (Table 2.2). We started from the African Plants Database for Tropical Africa, and added to our species list any validly published species name which was missing from the c. 3,000,000 records located between 24°N and 24°S tropical Africa (mostly ferns and species from northern Namibia, Botswana and South Africa, and recent publications). However, some newly published species, and synonyms, will not be found here.

In the following paragraphs species totals follow species+infraspecific taxa totals in brackets.

We can compare our totals to the Klopper et al. 2007 summary of the African Plants Database, which cited 50,136 (44,830) names for tropical and southern Africa combined, and 32,424 (c. 29,181) names for tropical Africa only. Our equivalent (angiosperm only) figures are 39,687 (31,706). Our totals are around 9% higher than the APD tropical-only estimates because we included species of Namibia, Botswana and South Africa north of 24°S, where the tropical Africa area of APD excludes these countries.

We can compare our species totals to the recently published RAINBIO dataset (Dauby et al. 2017, Sosef et al. 2017). For comparison of TRAFRICA to RAINBIO, we excluded names in our dataset for which we have no distribution data, and names we have recorded as introduced or cultivated (as in RAINBIO), and from the RAINBIO dataset, all records from outside 24°N to -24°S (as in TRAFRICA). This leaves RAINBIO with 26,547 (22,219) names. Our equivalent figures are 30,361 (27,690), which makes TRAFRICA approximately 15% bigger than RAINBIO. We have a larger number of cultivated species recorded as such than the RAINBIO dataset. Sosef et al. 2017 cite 22,577 species for the RAINBIO dataset, using their own definition of tropical Africa.

Of 26,547 RAINBIO taxon names recorded from tropical Africa 24°N to -24°S, 777 names were not matched in TRAFRICA. We checked the first 77 taxa, sorted alphabetically by genus name. We would

add eight of them to the database: four new publications, and four South African species which extend just north of 24°S. The remainder were name variants (33 species), synonyms, vague names, strictly southern African taxa (sensu APD maps), uncertain, North African endemics, Madagascan endemics, or cultivated. At least by this method of reckoning, we have probably missed around 80 tropical African species that we regret.

Dauby et al. 2017 suggested that RAINBIO is 89% complete, which they estimated by comparing the taxon total for TA only given by Klopper et al. 2007 to RAINBIO's full taxon dataset including north and south Africa. If we exclude taxa restricted to northern and southern Africa from the RAINBIO dataset, the estimate is 82%. Using our taxa/species totals, which are probably the most accurate estimates available for 24°N to 24°S, our estimate of RAINBIO completeness is 70% at the species level, and 68% at the infraspecific level. If we use Sosef et al.'s total species estimate of 22,577, and Klopper et al.'s estimate at the species level (29,181), completeness would be 77%. Sosef et al. (2017) suggested the discrepancy in their species total and that of Klopper et al. or Govaerts et al. (2001) (29,887) was the result of a high rate of synonymisation since 2007, and a high incidence of dubious names present in African checklists that do not have specimens linked to them. TRAFRICA has distribution data for 28,803 species, around 99% of the Klopper et al. estimate, and 96% of the Govaerts et al. estimate; it seems more likely that RAINBIO is not a particularly appropriate dataset from which to make estimates of taxon totals from, given the large number of available records not included in RAINBIO, and the inclusion of names with specimen records within the dataset only.

Using our own estimates of the number of taxa/species in tropical Africa (24° to -24°), which will be more accurate than those of Klopper or Govaerts or RAINBIO, we have distribution data for 88% of TA species at one degree square resolution or finer; if we exclude from our species total the number of species we have only recorded outside of TA (which probably should not be in the species list),

our completeness is 90% with distribution data at the species level, and 79% at the infraspecific level.

At country level resolution, 98% of accepted mother species have distribution data associated with them (Table 2.3). That still leaves 666 accepted mother species names in our database without a record at country level or below. Some will be names we added before cleaning the distribution record out of the database, some will be names for which the geolocation data is above the country level (i.e. older circumscriptions of countries or territories e.g. 'West Afrika'; some will be poorly known species for which the type has been lost or destroyed, or otherwise not digitised, or where we have not yet got around to resolving any uncertainty in the geolocation of the type, or just mistakes.

Our estimates of species richness for tropical African countries are higher in all cases than the RAINBIO estimate, and the species richness we have documented is for most countries higher than the RAINBIO modelled maximum estimates of species richness (Table 2.3). Congo, Gabon and Burundi have a higher modelled estimate of maximum species richness than we have observed; at least for Congo and Gabon, the RAINBIO dataset is well sampled, thanks to personal collections and the BR herbarium dataset. Mozambique extends a little south of 24 degrees; we have included it as our richness estimate will be reasonable, although a bit low. Adding in new country level records from the FTEA, quarter of the Flora of Somalia (FS), and some checklists of Kenya coastal forests (all digitised recently by W.H.), increases Tanzania and Kenya's total richness by about 600 species each, and Uganda's by c. 700. Tanzania has the highest species richness, presumably as a result of its large size, diverse landscapes, and relatively good collection. South Sudan has the lowest species richness; it is amongst the least well sampled countries of tropical Africa, and its recent name change means many records geolocated at country level only will not have been ascribed to it. Combining the estimates of South Sudan and Sudan together would be more accurate. The estimates given by us in Table 2.3 are not likely to decrease; some species will be sunk but this loss will likely be balanced by

new species descriptions, at least for the next few decades. Many countries are still very under collected, with even common species not recorded (or at least not digitised and publically available) at country level. Digitising FTEA and part of FS has contributed another 1976 unique records of species presence even at country level; the extra information to be gleaned from adding flora and checklist information to the database is likely to be quite significant.

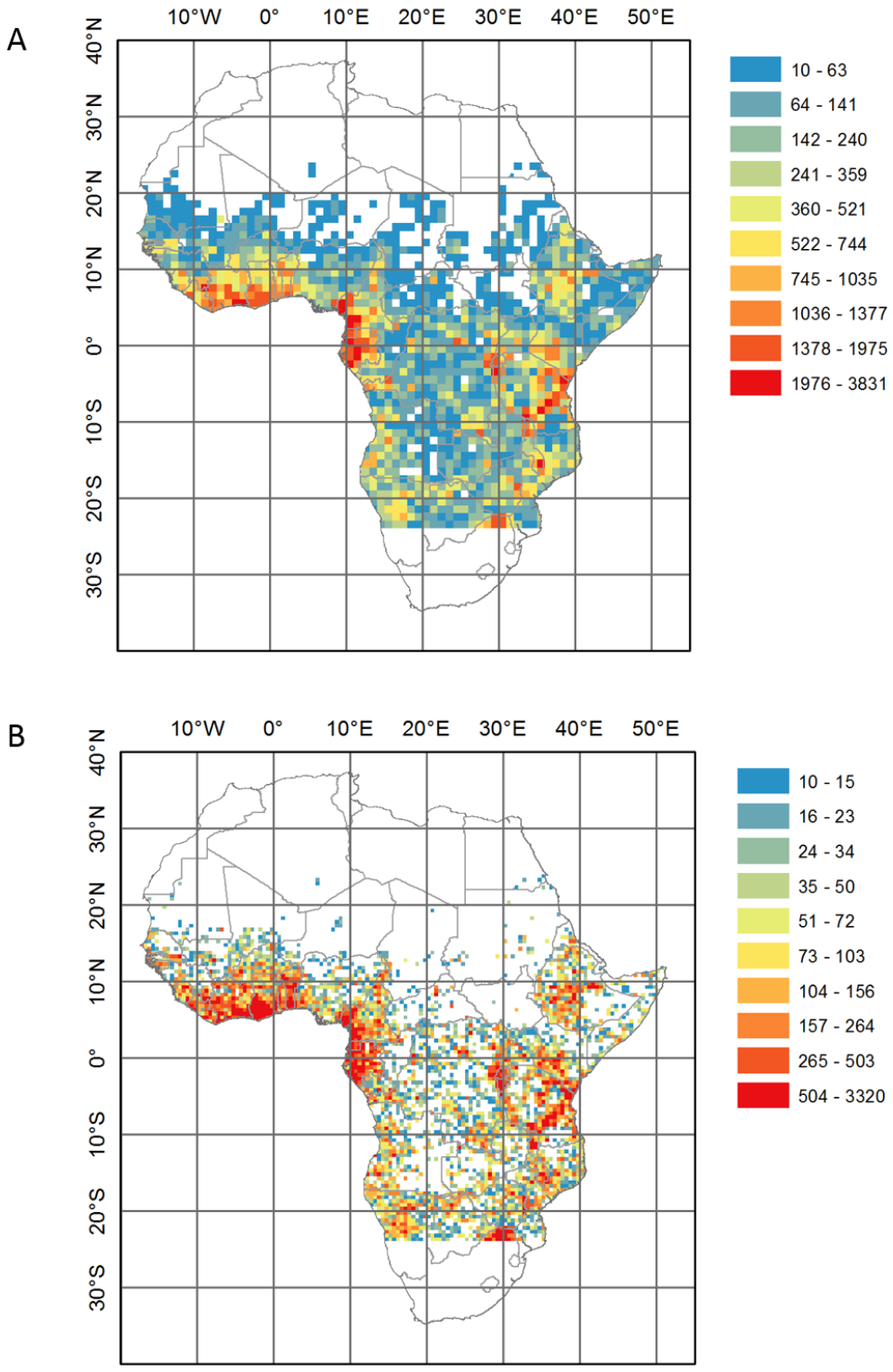
Table 2.3. Estimated species richness of countries completely within the TRAFRICA geographic area. Mozambique is included, although southern Mozambique is outside our area. RAINBIO estimates follow Sosef et al. (2017).

Country	TRAFRICA taxon richness	TRAFRICA species richness	Species richness with flora additions	RAINBIO species	RAINBIO Chao1 max estimate
(Mozambique)	6955	5926		4095	5309
Angola	8328	7398		2262	4388
Benin	3707	3275		2462	2889
Burkina Faso	2253	2019		879	1338
Burundi	3833	3310		2788	3591
Cameroon	9675	8488		6883	8057
Central African Rep.	4110	3780		2560	3503
Chad	2121	1970		N/A	N/A
Côte D'Ivoire	5516	4834		3689	4377
Dem. Rep. Congo	11844	10418		8860	10931
Djibouti	1050	988	994	N/A	N/A
Equatorial Guinea	4333	3988		3049	3856
Eritrea	2109	1931		N/A	N/A
Ethiopia	7070	6089		4481	5672
Gabon	6794	5973		5236	6144
Gambia	765	746		N/A	N/A
Ghana	5024	4488		2971	3667
Guinea	4140	3787		2533	3630
Guinea-Bissau	1682	1540		841	1360
Kenya	8364	7029	7607	4759	5991
Liberia	3617	3233		2403	2830
Malawi	6128	5256		3340	4413
Mali	2004	1864		903	1336
Niger	1398	1308		N/A	N/A
Nigeria	5947	5373		3378	4530
Republic of Congo	4911	4580		2403	3848
Rwanda	3037	2669		1883	2644
Senegal	2715	2444		1342	1953
Sierra Leone	3592	3284		1883	2545
Somalia	3488	3172	3251	1267	2381
South Sudan	282	275		N/A	N/A
Sudan	3868	3540		N/A	N/A
Tanzania	13031	10960	11613	8727	10605
Togo	2890	2623		N/A	N/A
Uganda	5115	4568	5307	2258	3348
Zambia	6708	5960		2807	4217
Zimbabwe	6431	5677		2807	4217

Our completeness rate of 90-98% at the species level does not mean that Africa is well sampled botanically. At degree square resolution, only 74% of mainland tropical African cells have >10 taxa recorded, and at half degree square, this figure drops to 45% (Figure 2.1). Numbers of taxa and species in each of White's major regions, and for a regional framework defined from the dataset, are summarised in Chapter 3. In Chapter 4, we map estimated taxon sampling completeness per cell. Figure 2.1 shows recorded taxon richness at half and one degree square resolution in the TRAFRICA dataset.

The number of records of unique species x sampling unit is the most important number of records in a dataset to quote, as this figure represents the true information content in the database for analysis. For TRAFRICA, these figures are 459,825 at the one degree square resolution, excluding GX species, and 618,141 at the half degree square resolution. This compares to 65,534 unique species x one degree square records in the RAINBIO dataset, making TRAFRICA about 7 times bigger than RAINBIO in terms of informative records.

Figure 2.1. Taxon richness in TRAFRICA. (A) Recorded taxon richness at one degree square resolution (total=459,825). (B) Recorded taxon richness at half degree square resolution (total=618,141). Cells with <10 species recorded are not shown. Legend gives number of recorded taxa; the colour scheme represents 10% quantiles.



Discussion

The database should not be used outside tropical Africa, nor on non-mainland Africa, although it includes some records from there. TRAFRICA seeks to be taxonomically and geographically representative of our current collective knowledge of the tropical African flora. Of course, not all plant species have been described from Africa, some species that have been described will be sunk or re-circumscribed; not all known plants have been collected throughout their ranges, not all available records have been digitised, and of those digitised records, we have not been able to include them all in TRAFRICA, not all digitised records are publically available, and we have not been able to digitise many sources of additional information such as floras and checklists.

Because the GBIF data could not be sensibly formatted to give collector and voucher, some of the GBIF records are certainly duplications of other records in the database. This is most apparent for the NHN collections, because NHN regularly contribute records to GBIF. The 400,000 or so African records submitted from the 2014 NHN database were also on GBIF, so the GBIF version of these records would have been imported again. Although not ideal, for our purposes the implications of having duplicated records in the database are modest, as all records are made unique by species name within a defined sampling unit before analysis. It does mean that our TRAFRICA record total (c. 3,000,000) is inflated relative to their information content. The RAINBIO group made a concerted effort to remove duplicate records, which is at least partly responsible for the very much larger number of records apparently in the TRAFRICA database. Although, if we compare informative records (unique species x unique sampling unit), TRAFRICA actually collapses less than RAINBIO does: TRAFRICA collapses to 15% of its original length at one degree square resolution, and RAINBIO to 11%.

Duplicate records (as in TRAFRICA) are only a problem in the analyses presented in this thesis if duplicate records have different, non-synonymous names, as then the outdated determination will

feature in the species list of the sample. For example, the same herbarium specimen could be in the database twice, once from NHN, with an up to date name, and once from GBIF, with a name which was accurate at the time of submission to GBIF. This problem will also be present within GBIF records, where duplicates of herbarium specimens have been independently curated and submitted to GBIF. Whilst this problem certainly exists, it probably does not affect a very large proportion of TRAFRICA records, because all data downloading was done within a relatively short time frame.

A large number of herbarium collections are very robustly named and are not prone to non-synonymous name changes (Goodwin et al. 2015). Goodwin et al. (2015) showed that even taking one of the worst case genera in Africa (*Aframomum;* not recently monographed, hard to collect, poorly preserved, generally overlooked, taxonomically challenging), only 15% or so of the collections going back to the 1850s were truly misnamed, with the remainder being synonymous name changes or vague names changing to specific names. The latter categories of name change pose no problem to a database such as TRAFRICA (or indeed any well curated herbarium database), because synonyms are maintained in the database and are easily handled, and a vague name is always superseded by a specific name. Of the specimens which might be misnamed, only a small proportion of these are likely to have had their names changed and digitised differently by different herbaria, for the following reasons: (i) most collections are modern and have not had a long time to diverge substantially, (ii) even in a big herbarium the number of specimens which have their names changed each year is not high, and (iii) of the specimen names which do get updated, we must assume that independent botanists would be capable of picking the correct name independently at least sometimes. Whilst some TRAFRICA botanic records are likely to be contradictory, the number is likely to be <15% of the total, and let us also remember that all of taxonomy is somewhat subjective, with the exception of type specimens. Differences of opinion are an important part of the taxonomic process. It remains the case, though, that there is an abundance of poor quality botanic data freely available for download.

Another issue with TRAFRICA, because it is an issue in the contributing datasets, and an issue with almost all botanic records save modern survey data, is that some records have been 'falsely' geolocated. Records collected before the GPS was used widely are typically localised to nearby communities, reserves or other landmarks. Conscientious curators or data users commonly give retrospective coordinates to localities, so that modern software can use the data. This logically impure practice means that batches of records tend to get given coordinates in the centres of well-known reserves, states, degree squares, and countries, to name a handful of cases apparent in the database. For an analysis conducted at a coarse resolution, this solution is much more useful than it is a hindrance, because at that resolution the information is very likely to be correct. At a finer resolution, this can become misleading, because collections from all over a reserve appear to e.g. come only from the centre of the reserve. Where we identified likely retrospective coordinates, we removed them in favour of the textual locality description on the specimen and an administrative region polygon or coordinate place name with a radius. The decision over whether we trusted more the supplied coordinates or the textual locality information for the c. 2,000,000 records we checked is subjective: we will not have caught all the retrospective coordinates, and some records will have been unfairly unlocalised. At least with a large number of data contributors all being slightly random with their coordinate assignation, taking the spread of coordinates together probably more or less aligns with the original collection localities. Never the less the conclusion is that TRAFRICA should not be analysed at too fine a grain, or to do very sensitive analyses, for which sample based datasets such as those analysed in Chapter 5 are more appropriate.

26% of degree squares across Africa have 10 or fewer species known from them, and at finer resolution than this, the holes exceed the fabric. Improvements to this situation can still be expected from further herbarium digitisation, perhaps in particular the digitisation of the East African herbarium (EA). Beyond this, a concerted effort to collect (Rapid Botanic Survey-type) data across the continent will be needed. The role of herbaria traditionally has been to facilitate specimen

identification, curate type specimens, and to accept significant other collections such as new country records. The care with which herbarium specimens are curated precludes every plant being accessed from every degree square in which it is found. Many herbarium records do not reflect the modern distribution of species, due to landscape changes in their collection localities, and a large influx of modern specimens would help to correct this. Although such records need not necessarily be supported by specimens, in reality many of them must be because of the difficulty of naming plants without a specimen; specimens also preserve the relevance of a name in the future, and help to define or refine species names themselves. In my opinion, specimens derived from sample-based vegetation surveys will need a new sort of repository into which they may be accessed, analogous to the establishment of seed banks once a need is recognised, or the establishment of Genbank once sequencing became commonplace.

In the course of examining the global range of 40,000 named taxa from Africa (Chapter 4), it became clear that a large number of species are known from their types only. Around 20% of TRAFRICA species with distribution data have only one record in the database (6202 species). The true number of species known only from their type specimen is much harder to calculate, but was estimated at around 10-20% by Cyrille Chatelain (*pers. comm.*). This unexpectedly high proportion of singleton species must have consequences for our understanding of the floristics of tropical Africa, particularly for many modelled analyses, where minimum numbers of records or a more even distribution of records would usually be expected. This feature of the digital African flora seems to have gone generally unrecognised, except by curators of large databases. It is hard to estimate the impact this has on our understanding of the tropical African flora, for example in this thesis, and at the current moment there is no way to solve this 'problem' anyway.

The collection, integration and curation of much of tropical Africa's plant data into a single database framework is of significant utility to those working in tropical Africa. We are happy to facilitate the work of others by sharing extracts from TRAFRICA, although I will raise four points on the subject.

Firstly, although our curation efforts are ours to give, this cannot be separated from the records themselves, which are largely not ours to give, and the rights of the original data providers must be respected. Secondly, restrictions on use have been specified differently by the donating datasets. All allow for scientific use with citation, but they do not all allow for commercial use, for example. Thirdly, the publication of the RAINBIO database in the public domain (CC4, use with citation) has alleviated the immediate pressing need to release TRAFRICA in order to allow African botany to progress. I hope that TRAFRICA and RAINBIO will be united ultimately, but this is not a small undertaking. Fourthly, TRAFRICA is stored within Geophila, software which is not written for immediate, supported distribution. We have geolocated by polygons, rather than coordinates, so the collections cannot be easily formatted in a table, and are not interpretable by BRAHMS, although obviously a solution could be reached. RAINBIO, the African Plants Database, and the NHN database between them constitute excellent sources of publically available data, and with the exception of our recent survey data, all the data we compiled are in any case in the public domain. All synthetic results from analysis of the input distribution data have been made public, either here or in related publications, for example the Star ratings, GHI scores, biogeographic framework, and richness summaries.

Chapter 3

Where is Upper Guinea?

Cicely A. M. Marshall[1*], Jan J. Wieringa[2], Stephen A. Harris[1], and William D. Hawthorne[1]

1 Department of Plant Sciences, University of Oxford, South Parks Road, Oxford, OX1 3RB, UK.

2 Naturalis Biodiversity Center, National Herbarium of the Netherlands, Darwinweg 2, 2333 CR Leiden, Netherlands.

* Corresponding author

Author contributions

Conceptualization: C.M. and W.H.; Methodology, Software and Analysis: C.M. and W.H.; Resources and Data Curation: W.H., J.W., and C.M.; Writing – Original Draft: C.M.; Review & Editing: C.M., S.H., J.W., and W.H.

Acknowledgements

Cyrille Chatelain provided the African Plants Database Tropical Africa species list. Data capture, editing and broadcast of the African Plants Database is the product of a collaboration between the South African National Biodiversity Institute, the Conservatoire et Jardin botaniques de la Ville de Genève, Tela Botanica and the Missouri Botanical Garden. BISAP (Biogeographical Information System on African Plant Diversity) has been established at the Nees Institute, University of Bonn in the context of BIOTA Africa project together with several external partners, especially Jon Lovett (University of Leeds) and Peter Linder (Zurich). For more details about the dataset and a full acknowledgement of all contributors please refer to Linder et al. 2005 and Küper et al., 2006. C.M. acknowledges support from the Clarendon Fund and Merton College; W.H. acknowledges support from the James Martin 21st Century School and BP Biofuels. J.W. acknowledges support from the Dutch Economic Enhancement Fund (FES) facilitating data capture of the Dutch herbaria.

Article status

In revision at the Journal of Biogeography.

Abstract

Aim: We investigate how multivariate statistical methods and large datasets have called into question the definition of 'Upper Guinea'. Upper Guinea cannot be defined without considering the definition of Guineo-Congolia, which must be defined in the context of its adjacent floras; we therefore propose a set of chorological regions for tropical Africa.

Location: Tropical Africa between 24°N and 24°S, Guineo-Congolia, Upper Guinea.

Taxon: 30,361 tropical African vascular plant taxa.

Methods: We applied Decorana and UPGMA clustering at one degree square resolution to define biogeographic regions, seeking concordance with previous classifications.

Results: We found that White's regional framework fitted our tropical African dataset well, but slightly different boundaries increased total endemism. To the north of Guineo-Congolia, the Jos Plateau and Cameroon Volcanic Line, between 8 and 10°E, are important east/west divides. The Sassandra river is supported as an east/west divide within Upper Guinea.

Main conclusions: Upper Guinea is recovered as part of Guineo-Congolia, distinct from the Sudanian region. All forest zones west of the Dahomey Gap (Upper Guinea) can be recovered as distinct from the Nigerian forests, giving Upper Guinea a Dahomey Gap termination in the east. Upper Guinean forests are only weakly distinct from Nigerian forests east of the Dahomey Gap to Cross River. Within this Nigerian forest zone, western cells are most similar to Upper Guinea, and eastern forest cells are most similar to Lower Guinea. Therefore, naming the Nigerian forests from Togo to the Cross River as 'Nigerian Guinea' and thinking of them as an Upper-to-Lower Guinea transition zone reflects the chorological patterns most closely. Although combining this transition zone with Upper Guinea to produce a Guineo-Congolian sub-region from the west coast of Sierra Leone to Cross River is supportable, the sub-region should be called Upper Guinea + Nigerian Guinea to avoid ambiguity.

Introduction

In this Chapter the definition of Upper Guinea and its botanic affinities is investigated, making use of the TRAFRICA dataset.

Biogeographic regionalisation maximises taxon similarity and endemism within defined regions, creating regions that are predictive of species distributions, and supporting the interpretation of patterns of species assemblage and distribution. Guineo-Congolia occupies a central geographic position in Africa, and holds most of Africa's rainforest. It is amongst the most species rich, and record rich, regions of tropical Africa, and has one of the highest rates of endemism (White 1983, Beentje et al. 1994, Linder et al. 2005, Stropp et al. 2016, Sosef et al. 2017). The definition of Guineo-Congolia relative to Africa's other major floras, and the subdivision of Guineo-Congolia, has received considerable scientific attention. The most widely used framework of African plant biogeographic regions was defined and mapped by White (1979, 1983, 1993). White used species distributions of a subset of the flora to define phytochoria: floristic units of any size, defined by their endemic taxa. According to White, a Regional Centre of Endemism (RCE), like Guineo-Congolia, holds >1000 endemic species, and at least 50% of the flora must be endemic, although he also accepted broad transition zones between RCEs with low endemism, and regional mosaics which he could not resolve into the fine-scale patterning they likely warrant.

Most authors working on African phytochoria have recognised 'Guineo-Congolia' with more or less consistent boundaries (Eig 1931, Lebrun 1961, Engler 1964, Good 1974, Takhtajan 1986), although White's subdivision of Guineo-Congolia is controversial. White (1979) subdivided Guineo-Congolia into three Subcentres of Endemism: Upper Guinea, Lower Guinea and Congolia. Upper Guinea was defined as a 100-350 km-wide strip of forest along the West African coast from Sierra Leone in the west to Ghana in the east. Lower Guinea was defined from southern Nigeria, to the flood plain of the Sangha River in south west Central African Republic, whilst Congolia corresponded to the eastern

part of the central African rain forest, and included most of the rain forest in the Congo River basin. Upper Guinea was separated from Lower Guinea by the Dahomey Gap, a 250 km-wide savanna corridor spanning Benin and Togo.

Today, the Dahomey Gap physically isolates the main forest blocks of White's Upper and Lower Guinea. However, 8400-4500 years ago, semi-evergreen forest prevailed in the Gap (Salzmann & Hoelzmann 2005), and even during dry phases, gallery forests could have acted as stepping stones between Upper and Lower Guinea, resulting in a relatively homogenous flora. The Dahomey Gap has been supported as the Upper Guinea/Lower Guinea border by authors other than White (Mangenot 1955, Lebrun 1961, Aubréville 1962, Hall & Swaine 1981, Takhtajan 1986). Others have drawn the Upper/Lower Guinea boundary at: (i) the Cross River on the Nigerian/Cameroon border (Léonard, 1965; Clayton & Hepper, 1974 (for forest species); Wickens 1976, Linder et al. 2005, Droissart et al. 2017); (ii) c. 5^{o} East, west of the Niger Delta (Fayolle et al. 2014); and (iii) the Cameroon Volcanic Line (CVL) (Clayton & Hepper, 1974; for grass species). Linder et al. (2012) provided a fourth interpretation of the botanical affinities of the forests of West Africa, where Upper Guinea *sensu* White was subsumed within the Sudanian region. Their Congolian region (which included Lower Guinea) terminated in the west around 4° East, comparable to definition (ii), and when the plant dataset was combined with animal datasets, Linder et al. drew the Upper Guinea boundary at the Nigeria-Cameroon border, including the area of the Cameroon Volcanic Line in Upper Guinea, comparable to definition (iii), at least from the Lower Guinea perspective.

Within Upper Guinea, the Sassandra River in SW Ivory Coast may act to structure species distributions. The wet evergreen forests of SW Ghana and SE Ivory Coast fall to the east, with Liberian evergreen and semi-deciduous forests on the west (e.g., the Western Guinean Lowland Forest and the Eastern Guinean Lowland Forest ecoregions of Olson & Dinerstein 1998, Bongers et al. 2004). Some genera have species restricted to either the east or the west, for example *Cola*

umbratilis and *C. buntingii* (Hall & Swaine 1981). Other species show variation in their phenotypes east to west, e.g. *Chassalia kolly* and *Diospyros vignei* (Hawthorne & Jongkind 2006).

White's regional framework was produced more than 30 years ago, from limited datasets, and without the aid of multivariate statistics. It was superimposed on a detailed map of physiognomic vegetation types, and his regional chorological boundaries may never have been much more than continental-scale sketches, drawn for 'cartographic convenience' (White 1983). Quantitative regionalisations have shown moderate concordance with White's framework (Denys 1980, Linder et al. 2005, 2012, Fayolle et al. 2014). We investigate the extent to which quantitative methods and new datasets have called into question White's regional framework, particularly the definition of Guineo-Congolia, and most specifically, the boundaries of Upper Guinea to the north and the east.

Materials & Methods

A summary of the TRAFRICA dataset was used (Marshall et al. 2016, Chapter 2). Analyses were conducted with a one-degree square grid, a grain size commonly employed as a compromise for analyses of plant distribution data, which balances the benefits of fine-grain resolution and large numbers of records per cell. Analyses were conducted at the lowest named taxonomic rank, after excluding vague names, hybrids, cultivars, and taxa we knew to be introduced or cultivated. The lowest named taxonomic rank was preferred to species rank, because biogeographic signal may be found at the infraspecific rank. Where values are given at the specific level in Tables 3.1 and 3.2, all subspecies were collapsed into their mother species and recalculated. Any degree square was excluded from analysis if it: had a predominantly azonal flora (*sensu* White 1983); had fewer than 10 taxon records; had <3% of its estimated total richness sampled; was outside of the tropics (24°N to 24°S); or was associated with islands only. This resulted in a dataset with 443,993 records for 30,361 tropical African taxa, from 1230 degree squares. White's major regions were digitised from his 1983 accompanying memoir, and one degree grid cells were classified to their major region by summing

the number of records belonging to each region within the cell (Table 3.1). The Sahara, the Karoo-Namib and the Kalahari-Highveld (*sensu* White 1983) have most of their area outside this study's remit, which applies to their tropical African portions only. Climatic variables were extracted from the WorldClim resource for degree square mid points.

There is inherent subjectivity in the classification of regions, at least when cluster structure is weak. To define the regions, we conducted a series of detrended correspondence analysis (DCA) ordinations on subsets of the dataset. These datasets comprised: (i) the whole tropical African dataset (Fig. A1.1); (ii) The cells assigned to the Zambezian, Guineo-Congolian/Zambezian transition zone (GCZA), and Afromontane regions in the analysis from (i) (Fig A1.2); (iii) The cells assigned to the Guineo-Congolian, Guineo-Congolian/Sudanian transition zone (GCSU), and Sudanian regions in the analysis from (i) (Fig. 3.4); (iv) The cells assigned to Upper and Lower Guinea in the analysis from (iii) (Fig. 3.5); (v) The cells assigned to Upper Guinea in the analysis from (iv) (Fig. 3.5).

DCA provides a quick, unconstrained ordination of sample and species data in multidimensional space. As the tropical African dataset is large and species rich, the speed of analysis of DCA is particularly advantageous. Ordinating data subsets allows gradients to be drawn out that might be obscured by stronger gradients elsewhere in the total dataset. DCA axes scores were visualised initially by transforming axis 1 scores to red values, axis 2 scores to blue, and axis 3 scores to green values of the RGB colour space.

We tested White's (1983) regional framework by annotating DCA ordination space with his regions. If White's framework can be supported by the dataset, the clustering of cells into regions in ordination space will be coherent, enclosed in volumes of ordination space which do not overlap with other regions. It is debatable whether the enclosing polygons or hulls in 2d perspective must be convex, or whether for instance some 'indentations' or concavity would be admissible, though convex volumes would be more convincing.

We subsequently clustered axes 1-3 of the DCA ordinations using UPGMA. UPGMA portrays the similarity between cells hierarchically, and represents this structure in a dendrogram. UPGMA joins cells together, and later groups of joined cells together, at a node height proportional to their similarity. UPGMA defines clusters by collapsing together all 'descendant' cells of a branch, below a (user-specified) node height, to produce clusters which are equally dissimilar to each other. UPGMA will therefore define a 'cluster' for a single cell if the cell is very dissimilar to its most similar cells. A cell can appear very different from other cells if it truly has a very distinctive species composition, or if it is unrepresentatively sampled.

We specified a node height that produced regions we could relate to White's regions. We subsequently collapsed some of the UPGMA-defined clusters back to higher nodes where there was no strong biogeographic rationale for maintaining them. In the case of the very small clusters, this was because we suspected the dissimilarity of the clusters was an artefact of sampling biases. In the case of larger clusters, this was because maintaining the cluster would result in an over-split framework for some parts of Africa relative to others, because the dataset is richer for some regions than others. Consequently, the regions defined here are not equally dissimilar to each other. UPGMA regions which are equally dissimilar to each other are shown in Figures A1.1 and A1.2.

Three small regions from different branches of the UPGMA dendrogram were combined to recover the Zanzibar-Inhambane region. We clustered DCA axis scores, so we can visualise the clusters defined by UPGMA in ordination space, as well as on the dendrogram. We can see that the placement of dendrogram branches is dependent on the smallest of differences in ordination distance (species dissimilarity) between clusters (Fig. A1.2). White's Zanzibar Inhambane Regional Mosaic has been particularly challenging to recover at degree square resolution. It is a long, narrow region, where the species composition is very finely structured geographically, and at degree square resolution, the relevant cells have various admixtures from the flora of the Afromontane, Somalia-Masai and Zambezian regions.

To investigate the extent to which Guineo-Congolia definition might be the result of methodology or subjectivity, the beta-sim matrix (without ordinating) was clustered using UPGMA (Figure A1.3) and PAM (Figure A1.4), and DCA axes scores were clustered divisively (Figure A1.5). The PAM 7-cluster solution for tropical Africa (Figure A1.6) and the PAM 12-cluster solution were compared with the final framework (Figure A1.7). The beta-sim matrix was ordinated using NMDS (Figure A1.8), and the NMDS ordination space was clustered with UPGMA (Figure A1.9).

The success of the overall floristic regionalisations was evaluated by calculating total endemism in the framework, i.e., the number of taxa endemic to a single region, divided by the total number of taxa in the dataset. Total endemism is one when tropical Africa is not broken into any regions, but decreases as the number of regions increases. Consequently, only frameworks with the same number of regions were compared. White and others assume that gross patterns of floristic similarity are strongly aligned with areas of endemism, even if not all distinctive areas are endemic rich or some centres of endemism include many vegetation zones. This overlap is key to optimising the phytochoria for both criteria. Consequently, we relied on values of total endemism to assess our frameworks. Clustering algorithms define regions using the gross patterns of floristic similarity.

We drew polygons around the one degree cell classifications to produce regional polygons, available as a shape file (data supplement, S3.1, Figure 3.7). We cut the Sahel north polygon using White's (1983) vegetation map boundary to the Sahara. All cell-level attribute data (sampling levels, raw and final regional definitions, lat/long) are available in Appendix S3.2.

Clustering was performed within the R package cluster (Maechler et al. 2016), specifically the functions *PAM*, *agnes* and *diana*. Ordinations were performed using the R package vegan (Oksanen et al. 2016), specifically the functions *decorana* and *metaMDS*. We made use of vegan3d (Oksanen et al. 2016b), rgl (Adler et al. 2016) and scatterplot3d (Ligges & Mächler 2003) for 3d visualisation of ordination space, plotrix (Lemon 2006) for transforming axis scores to RGB colour space, and

betapart (Baselga et al. 2013) for calculating a beta-sim matrix. Mapping was performed in QGIS 2.10.1 and ArcMap 10.4.1.

Results

Testing White's Tropical African regions

It is clear there is geographic structure in the dataset (Fig. 3.1b), but also that species turnover is continuous, as evidenced by a limited number of 'gaps' in species composition between cells in ordination space (Fig. 3.1a). For effecting a regionalisation, the absence disjunctions in ordination space means a diversity of possible divisions.

Overall, White's regions shown in Fig. 3.2a do occupy distinct convex volumes of the ordination space shown in Fig. 3.2b. In Fig. 3.2b, only 2 dimensions are visible, but even from this perspective there is minimal overlap, and these overlaps are largely resolved in the third dimension. Guineo-Congolian (GC) cells form a distinct cluster, although some GC cells are located within the cluster of White's Guineo-Congolian/Sudanian Regional Transition Zone (GCSU) and White's Guineo-Congolian/Zambezian Regional Transition Zone (GCZA). Single degree squares in mapped transition zones between two regions may include degree squares of either extreme, in a mosaic, or may be truly intermediate in species composition. However, the alteration of zone boundaries to apportion cells to either extreme zone may usefully allow the boundary of the non-transition regions to be extended.

We estimate that the proportion of species endemic to only one of White's regions is 40% (Table 3.1). With the exception of the Karoo-Namib, Kalahari-Highveld and Sahel, our estimates for the number of species present in each region are considerably higher than White's. The low richness reported for the Karoo-Namib and Kalahari-Highveld is because most of the area of these regions occurs further south than our 24°S limit. Although the richness of the Sahel is similar to White's, the

absence of data for many Sahelian cells means our estimate is likely to be an underestimate of its richness. Endemism estimates are much lower than those of White for all Regional Centres of Endemism, except the Afromontane Centre.

Figure 3.1. DCA ordination of tropical African cells. (a) DCA ordination scores coloured in RGB, with red for axis 1, green for axis 2, and blue for axis 3 (third axis not shown). Latitude, Longitude, Bio12 (annual mean precipitation), Bio1 (annual mean temperature), Bio4 (temperature seasonality) and Bio15 (rainfall seasonality) are correlated with ordination axes and are fitted as vectors. (b) Geographic location of cells, coloured according to the same scheme.

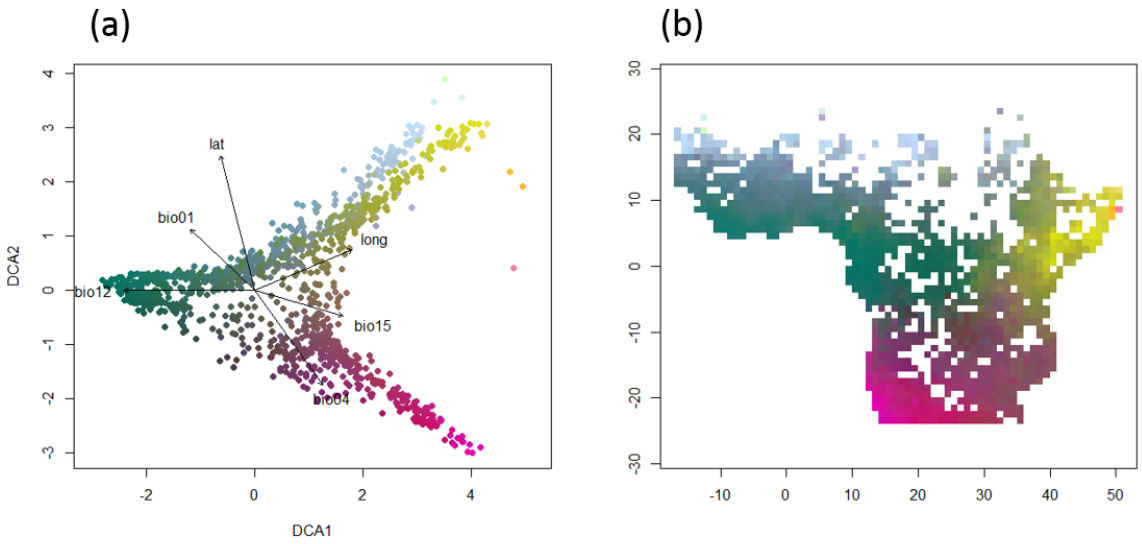


Figure 3.2. DCA ordination of tropical African cells, ordinated by White's (1983) regional phytochoria. (a) Map showing polygons of White's (1983) regional delimitation. One degree grid cells are coloured by their major region. Country polygons are shown in grey. Codes follow Table 1. (b) Axes 1 and 2 of DCA ordination space, coloured by White's regions.

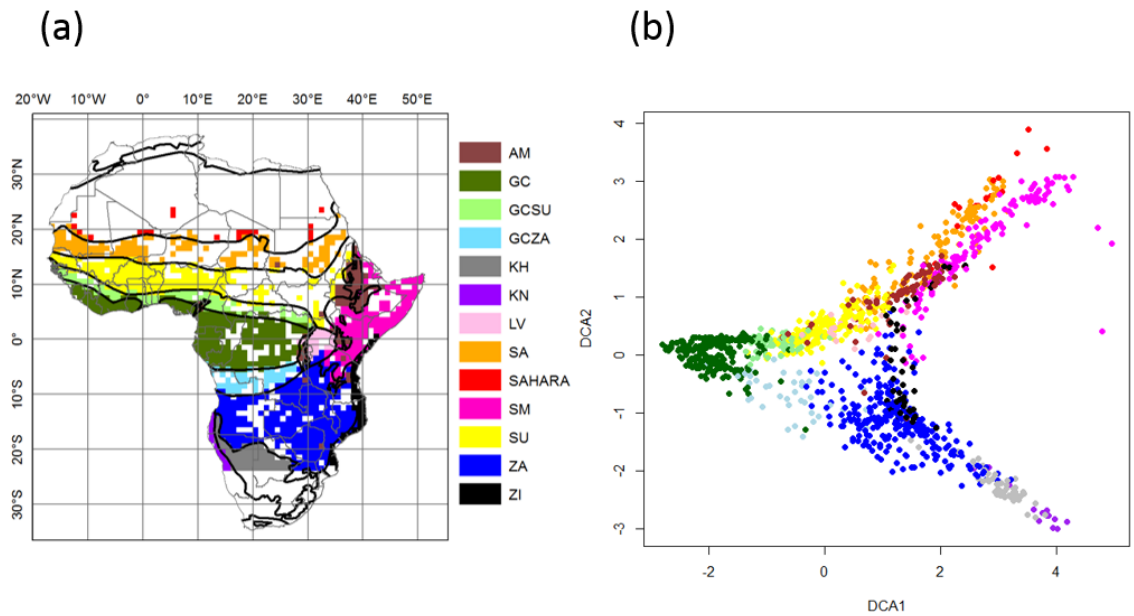


Table 3.1. Endemism in the dataset at one degree square resolution, for White's (1983) regions defined as in Fig. 3.2a. The 'Total' row refers to the total number of taxa in the dataset, and the number of taxa endemic to any one region: Total percentage endemism in the framework is calculated from these two numbers. 'BIOTA' refers to % endemism in White's regions using the BIOTA-BISAP dataset (Linder et al. 2005). 'White No.' and 'White %' are White's (1983) estimates for the (polygon) regions. Taxa includes infraspecific taxa. RTZ=Regional Transition Zone. RCE=Regional Centre of Endemism. RM=Regional Mosaic.

Code	Region	No. taxa (species)	No. endemic taxa (species)	% endemic taxa (species)	BIOTA: % endemic	White: No. species	White: % endemic
AM	Afromontane RCE (taxa)	11,633 (10,160)	1182 (892)	10% (9%)	5%	4000	7%
GC	Guineo-Congolia RCE	12,252 (10,804)	3897 (3201)	32% (30%)	29%	8000	80%
GCSU	GCSU RTZ	5929 (5480)	145 (104)	2% (2%)	1%	2000	very few
GCZA	GCZA RTZ	4580 (4331)	222 (202)	5% (5%)	5%	2000	very few
SA	Sahel RTZ	1484 (1419)	125 (111)	8% (8%)	4%	1200	3%
SM	Somali-Masai RCE	8213 (7234)	1468 (1260)	18% (17%)	11%	>2500	>50%
SU	Sudanian RCE	5534 (5048)	240 (182)	4% (4%)	1%	2750	33%
LV	Lake Victoria RM	4476 (4091)	102 (79)	2% (2%)	1%	3000	very few
ZI	Zanzibar-Inhambane RM	6185 (5470)	593 (448)	10% (8%)	8%	3000	c. 10%
ZA	Zambezian RCE	14,729 (12,637)	3655 (2902)	25% (23%)	22%	8500	c. 54%
SAHARA	Sahara RTZ (tropical only)	244 (240)	29 (26)	12% (11%)	12%	1620	12%
KH	Kalahari-Highveld RTZ (tropical only)	3619 (3276)	374 (292)	10% (9%)	2%	3000	very few
KN	Karoo-Namib RCE (tropical only)	1550	107	7%	19%	3500	>50%
	Total	30,361 (25,710)	12,139 (9795)	40% (38%)			

Deriving a regional framework

In Figure 3, we present a set of regional boundaries defined as a result of a series of DCA ordinations and UPGMA clustering analyses of the tropical African grid cells. In the resulting biogeographic framework, total endemism is increased from 40% using White's boundaries, to 47% using the new boundaries (Fig. 3.3, Table 3.2).

The regional framework derived here using multivariate methods is remarkably similar to that of White (1983). Within White's Somalia-Masai region, two regions are suggested by our analysis (Table 3.2). The SM has a high (32%) proportion of endemic species, whilst SM1 has low (4%) endemism, and includes species otherwise found in the Zambezian, AM, and ZI regions, to which it is geographically adjacent (suggesting SM1 is more like a transition zone, an effect exaggerated by our one degree square unit of analysis). SM1 and SM could be combined to make a region geographically similar to White's Somalia-Masai. Similarly, within White's Sahel our analysis suggests two regions: 'Sahel north' (SAN) and 'Sahel south' (SAS). As SAN has a much higher proportion of endemic species (20%) than the Sahel south (4%), a boundary between them is maintained. There is still higher endemism in SAN than SAS when the Saharan cells are excluded from the dataset. The Afromontane region (AM) detected by the analysis corresponds to the Ethiopian highlands and the mountains of south western Kenya and north eastern Tanzania, and covers White's Lake Victoria regional mosaic. There are outposts associated with Jebel Marra (Sudan), and with the CVL (Cameroon). However, the isolated Afromontane patches within the Zambezian cluster have not been resolved in this analysis. There is no Guineo-Congolia- Zambezian transition zone (GCZA). Zambezian cells are structured into eastern, central and southern portions, but even within these divisions cells located in White's GCZA lack distinctive species composition (Fig. A2.2). There is however a zone of overlap, where some 'Congolian' cells fall into the Zambezian (ZA) region, and some 'Zambezian' cells fall into the GC region. The cluster most closely aligned with the Kalahari-Highveld is larger than White's Kalahari-Highveld, with the inclusion of nine complete additional degree square cells from ZA. The eastern

part of Guineo-Congolia-Sudanian transition zone extends further north than that of White, although the data are limited here.

Figure 3.3. Major regions of tropical Africa, derived using a series of DCA ordinations and UPGMA cluster analysis. Regions were nested in the Zanzibar-Inhambane, whilst some regions were collapsed to maintain biogeographic integrity. No cells have been reassigned. Black polygon boundary lines delimit White's 1983 regional phytochoria, and grey polygons show countries. Legend codes follow Table 3.2.

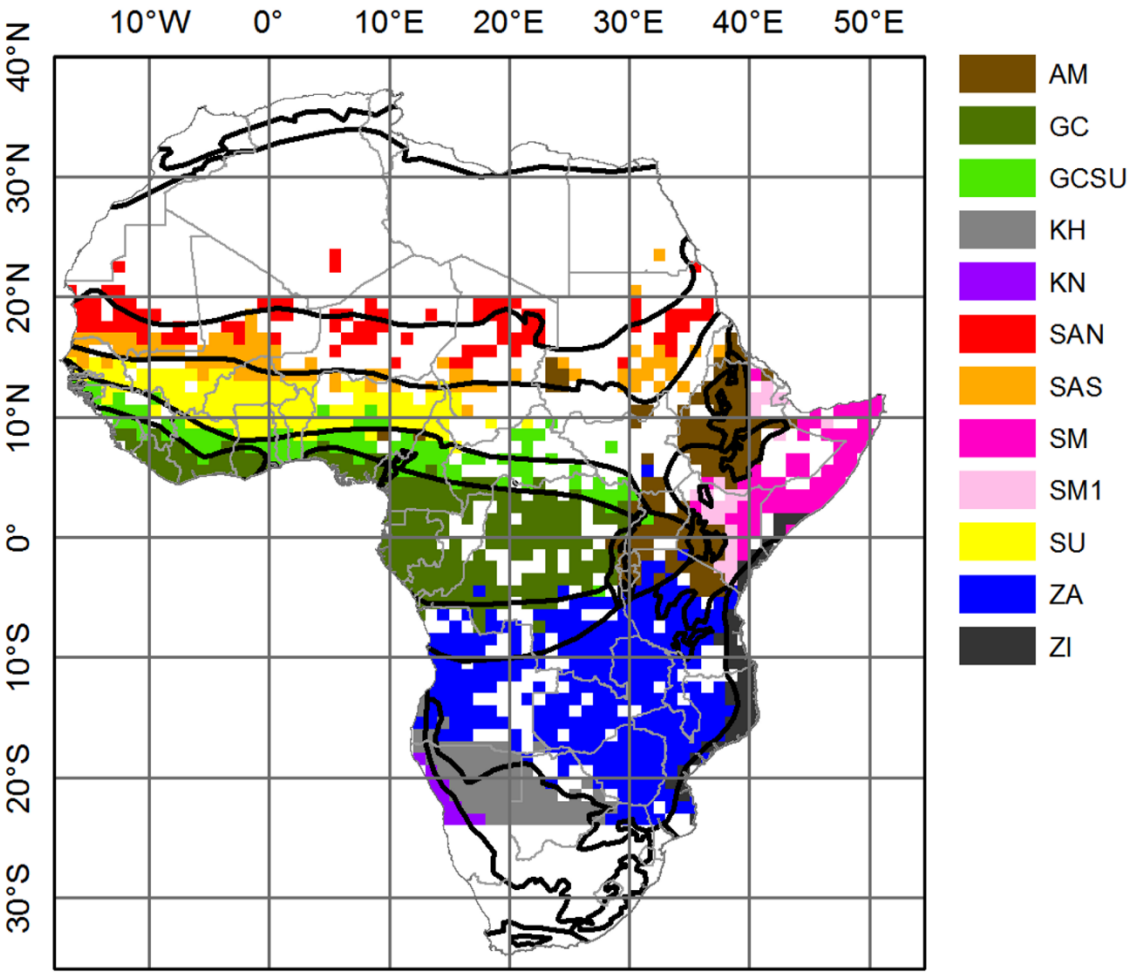


Table 3.2. Endemism for the tropical African regions, defined here, as shown in Figure 3. Region names refer to regions as redefined here, and are named following the White regions to which they have most geographic overlap. The 'Total' row refers to the total number of taxa in the dataset, and the number of taxa endemic to any one region: Total percentage endemism in the framework is calculated from these two numbers.

Code	Name	No. taxa (species)	No. endemic taxa (species)	% endemic taxa (species)
AM	Afromontane	10,075 (8613)	1667 (1249)	17% (15%)
GC	Guineo-Congolia	12,816 (11,321)	4370 (3616)	34% (32%)
GCSU	GC-Sudanian transition zone	7177 (6517)	309 (240)	4% (4%)
KH	Kalahari-Highveld (tropical)	3299 (2974)	373 (297)	11% (10%)
KN	Karoo-Namib (tropical)	1070 (1035)	58 (54)	5% (5%)
SAN	Sahel north + tropical Sahara	721 (707)	145 (127)	20% (18%)
SAS	Sahel south	1495 (1433)	56 (45)	4% (3%)
SM	Somalia-Masai	2711 (2595)	872 (824)	32% (32%)
SM1	SM w. montane elements	2635 (2498)	113 (82)	4% (3%)
SU	Sudanian	4043 (3670)	148 (110)	4% (3%)
ZA	Zambezian	16,545 (14,144)	5591 (4563)	34% (32%)
ZI	Zanzibar-Inhambane	5262 (4679)	426 (323)	8% (7%)
Total		**30,361 (25,187)**	**14,128 (11,530)**	**47% (46%)**

Subdividing Guineo-Congolia

In Fig. 3.4, cells of our Guineo-Congolian and Sudanian region (*sensu* Fig. 3.3) are reordinated and clustered. Lower Guinea does not extend east of 18°E, which coincides with White's eastern limit of his Sangha River Interval (SRI) (Fig. 3.4). The cells of White's SRI are assigned mostly to Lower Guinea in the north, and Congolia in the south, although data are lacking from many of White's SRI cells, so confidence in this boundary is low. The Congolian forest sub-region has a distinct cluster of cells towards the east, where Congolian forest transitions to the Afromontane region.

The main division (Fig. 3.4c) within Lower Guinea (LG) is into a northern portion, which includes drier types of Lower Guinean forest and the forests of Nigeria (LG1), and a wetter southern portion (LG2). The main division within Upper Guinea (UG) is into a northern, drier portion, and a wetter, southern portion, as for Lower Guinea. The northern type of UG forest (UG1) extends to include two cells of Nigerian forest, i.e. it is not delimited by the Dahomey Gap, but the Niger Delta. Wetter UG (UG2) is delimited approximately by the Dahomey Gap, at 2°W. Wet UG and wet LG are very distinct from each other in ordination space, but there is no clear disjunction between drier UG and drier LG.

To the north of the GC region there are three regions transitional to the Sudanian. In the west there is the western Guineo-Congolian/Sudanian transition (GCSUW), which terminates in the east at 9°E. To the east of this in the same latitudinal band there is the eastern Guineo-Congolian/Sudanian transition (GCSUE), which takes over from the GCSUW at the Jos Plateau. South of the GCSUE is another transition zone, called here GCSUS, which lies in ordination space between those two regions and terminates in the west at 10°E, at the CVL.

Delimiting Upper Guinea

In Figure 3.5, cells of our Upper Guinea and Lower Guinea (*sensu* Fig. 3.4) are reordinated. Our results show that the Nigerian forest cells, which are the contentious cells as regards the UG/LG border, have a species composition which is intermediate between the forests of Upper Guinea west

of the Dahomey Gap, and the forests of Lower Guinea (Fig. 3.5b), reflecting their geographically intermediate position (Fig. 3.5a).

In the UPGMA 2-cluster solution of this ordination space, the Nigerian forest cells cluster with our Upper Guinean cells west of the Dahomey Gap, giving 'Upper Guinea' a Cross River termination. Whereas in Figure 3.4, most of these Nigerian forest cells were clustered with our Lower Guinea subregion, promoting a mid-Nigeria termination line between UG *sensu lato* and LG. However, within the UPGMA cluster corresponding to Nigerian forest cells + UG west of the Dahomey Gap, the Nigerian forest cells form a cluster distinct from UG west of the Dahomey Gap, giving our Upper Guinea a Dahomey Gap termination, as in White's Upper Guinea, and the Nigerian forest cells their own subregional identity.

The Nigerian forest cells circled in red in Fig. 3.5b are the cells responsible for the mid-Nigerian termination points of Upper Guinea. Including the first in Upper Guinea gives a termination point at 4°E, as in Linder et al. (2012); the next cell is the Nigerian delta cell which clusters with Upper Guinea in Fig. 3.5b; and including the next two in Upper Guinea gives a termination point at 5^{o}E as in Fayolle et al. (2014). The cells of the CVL itself, and Cross River itself, cluster with the Lower Guinean cells.

The main structure within our Upper Guinea is east-west, with additional structure from north to south (Fig. 3.6a). Axis 1 divides western Upper Guinea from eastern Upper Guinea. Mean annual rainfall is higher in the west. Axis 2 separates the cells with more seasonal rainfall from the cells with less seasonal rainfall. Both latitude and longitude explain the position of cells in ordination space better than mean rainfall (bio12) or rainfall seasonality (bio15). Eastern Upper Guinea (sometimes called Lower Upper Guinea) includes, and is terminated in the west by, the Sassandra River, just visible on Fig. 3.5d as an inland intrusion of the SW Ivory Coast coastline in pale grey.

Figure 3.4. UPGMA cluster solution of DCA axis scores for Guineo-Conglia. (a) UPGMA cluster solution of DCA ordination space. C, GCSUS, CEA and LG1 are separated in ordination space by the third axis (not shown). Latitude, Longitude, Bio12 (annual mean precipitation) and Bio1 (annual mean temperature) are correlated with ordination axes 1 and 2, and are fitted as vectors. (b) Clusters mapped in geographic space. Black polygons show White's regions. The Cross River in Nigeria is shown in purple, as is the Sangha River in southern Central African Republic, the CVL is represented by White's polygon of Afromontane vegetation on the border between Nigeria and Cameroon.

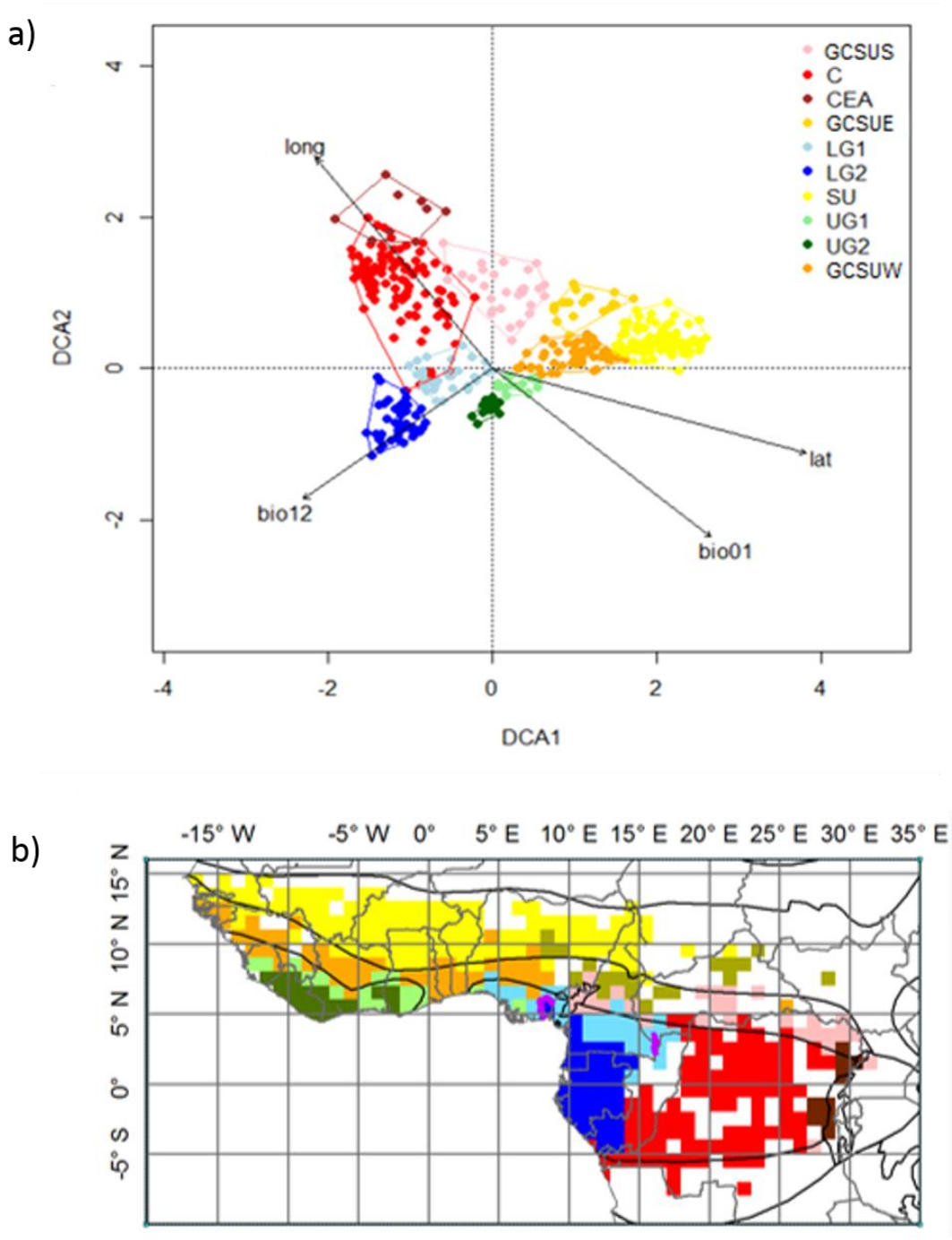


Figure 3.5. DCA ordination of Upper and Lower Guinea. (a) Cells coloured by proposed disjunctions between Upper and Lower Guinea. (b) Position of cells in DCA ordination space, coloured as in (a). Red circle shows the four cells responsible for mid-Nigerian termination points of 'Upper Guinea'.

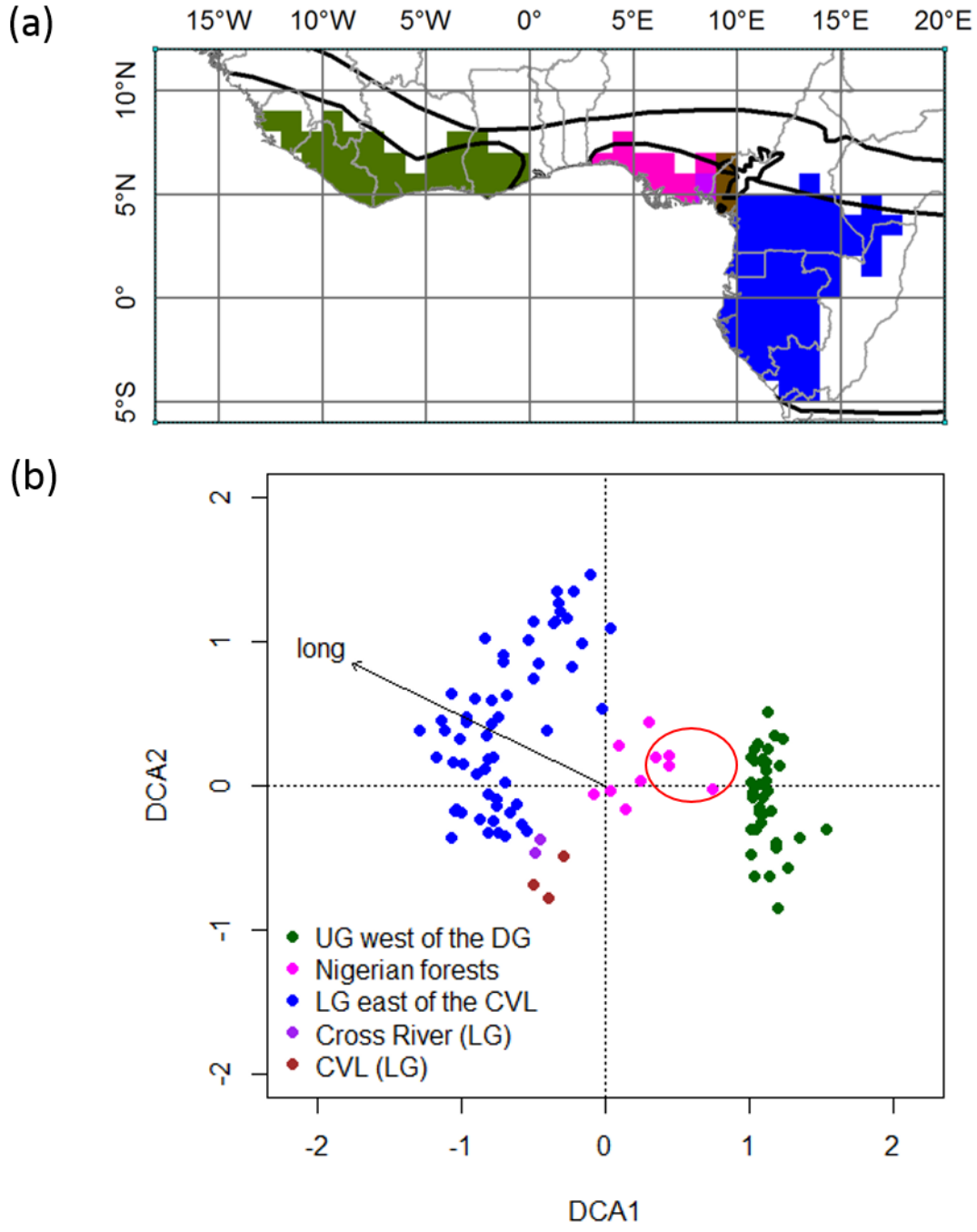


Figure 3.6. DCA ordination of Upper Guinea, defined as in Figure 3.5. (a) Cells of Upper Guinea west of the Dahomey Gap ordinated with DCA and divided into four groups using the origin lines of axis 1 and axis 2. (b) Geographic position of cells, coloured as in (a).

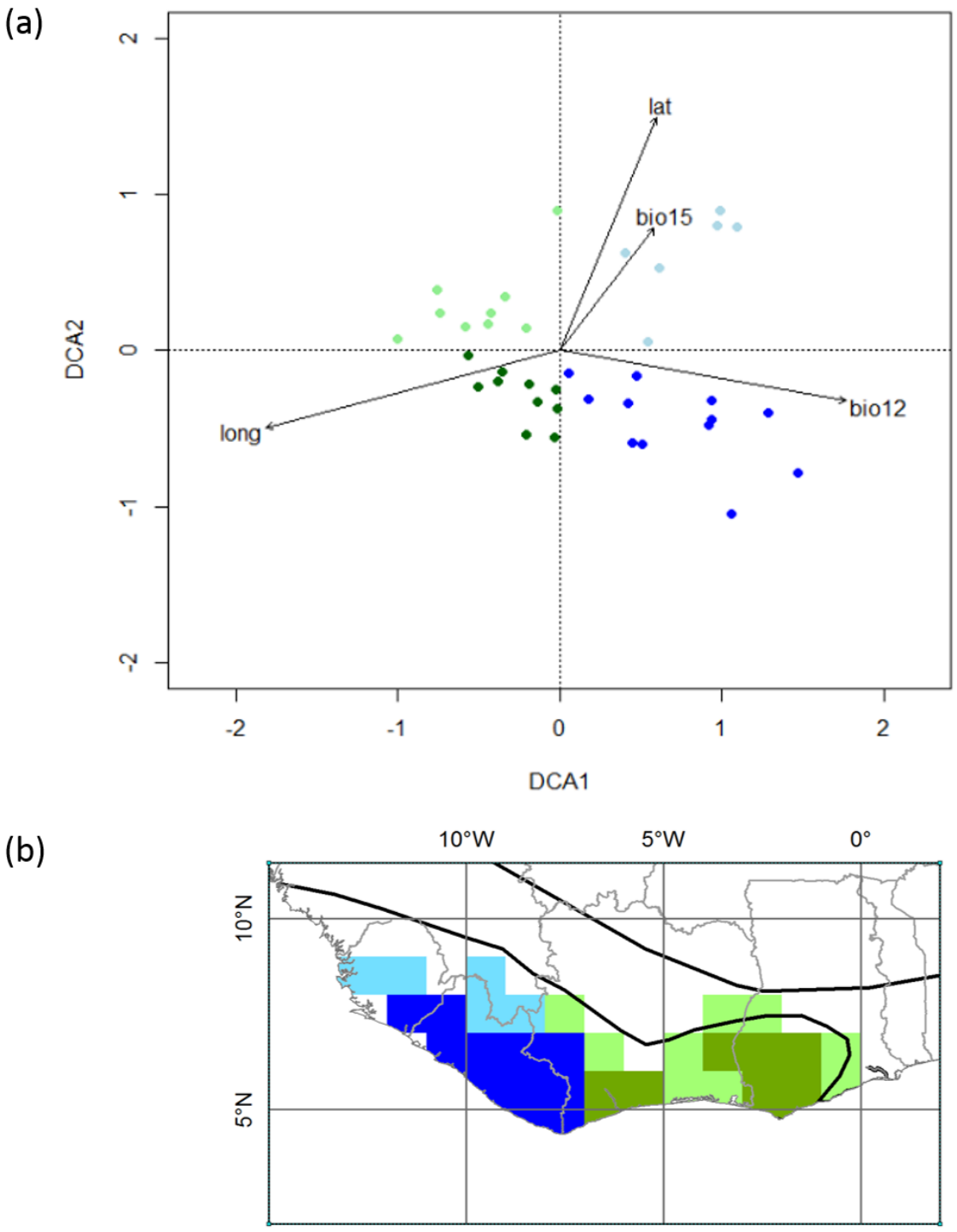
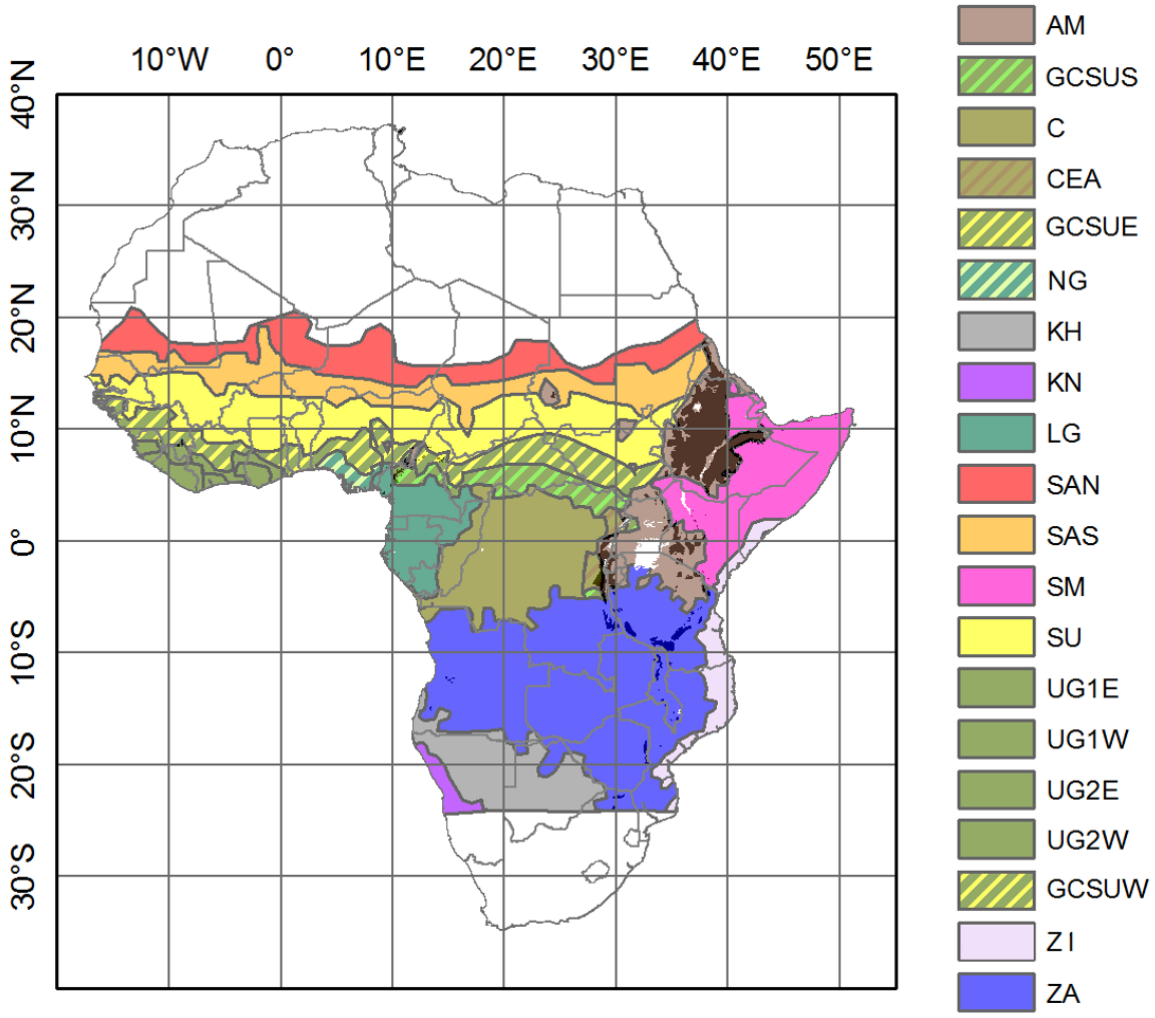


Figure 3.7 Regions of tropical Africa, as in Appendix S3.2



AM	Afromontane
GCSUS	Guineo-Congolian/Sudanian transition (south)
C	Congo
CEA	Congo transitional to the east
GCSUE	Guineo-Congolian/Sudanian transition (east)
NG	Nigerian Guinea
KH	Kalahari-Highveld
KN	Karoo-Namib
LG	Lower Guinea
SAN	Sahel north
SAS	Sahel south
SM	Somali-Masai
SU	Sudanian
UG1E	Upper Guinea: eastern dry
UG1W	Upper Guinea: western dry
UG2E	Upper Guinea: eastern wet & hyperwet
UG2W	Upper Guinea: western wet & hyperwet
GCSUW	Guineo-Congolian/Sudanian transition (west)
ZA	Zambezian
ZI	Zanzibar-Inhambane

Discussion

Distribution of the tropical African flora is characterised by gradual species turnover, at least at one degree square resolution. The exception is the disjunction between the Congolian flora and the geographically adjacent East African flora, which are distinct in ordination space. This east-west disjunction at the Albertine Rift was identified by Fayolle et al. (2014). In spite of a lack of disjunctions, clusters defined in ordination space are geographically coherent, given the large number of species evaluated, large grain size, and limited distribution records (27% of cells have 3-5% of their estimated species richness recorded).

White's (1983) regional framework fitted our dataset remarkably well (Fig. 3.2, Table 3.1). With a total endemism rate of 40%, White's 13-region framework fits the dataset slightly better than the PAM framework for 12 regions (39%, Fig A2.7). By successively reordinating and clustering subsets of the data, a set of regions have been defined that reflects White's framework, but improves the total endemism rate to 47%. Derivation of this regional framework was not independent of White's (1983) work, since we collapsed and occasionally combined clusters to recover his regions. It is therefore perhaps unsurprising that our final regions should be so close to White's. What is more surprising is that our final 12 region framework, which is deferential to White's, has higher total endemism than the PAM 12-cluster solution, which is derived entirely independently of White's framework. The PAM 12-cluster solution is also very reminiscent of White's framework, although the Kalahari-Highveld is not distinct from the Karoo-Namib, and the Zanzibar-Inhambane region was not recovered within it.

White defined Regional Centres of Endemism if they held >1000 endemic species and had >50% of their species restricted to them, describing these thresholds as 'convenient but not sacred'. In our dataset, none of his regions have >50% of species endemic to them, but all his regions have >1000 endemic taxa (892 endemic species for the Afromontane), excepting the Sudanian, the definition of

which as a RCE White was uncertain about, and discounting the Karoo-Namib, for which only the tropical portion has been investigated. Endemism estimates for each region are more similar to those made with the BIOTA-BISAP dataset (Linder et al. 2005) than they are to White's 1983 estimates. The endemism rate in White's GC is only 38% if his transition zones are included in the definition. White appears to have underestimated the number of 'chorological transgressors'; his richness estimates for each chorion are too low, and his endemism rates are too high. However, in the present regionalisation (Fig. 3.3), 47% of taxa are endemic to only one region, which is close to his 50% rate, although no single region can achieve this.

An analysis at much finer resolution, respecting the results so far but with refined boundaries following vegetation boundaries rather than degree lines, would be expected to increase the numbers and proportions of endemic species in each region significantly; this is an ambitious topic for further work, and would need to deal with the often transitory status and hectare-patterned mosaic evident in transition zones and regional mosaics.

As regards Guineo-Congolia, different regionalisations offer different latitudinal boundaries to the north. Whilst it is clear that Upper Guinea abuts the GCSU floristically, it has not been possible to recover a result as extreme as in Linder et al.'s (2012) analysis, where Guineo-Congolia excluded every cell of White's Upper Guinea. It was unclear to the authors whether many false absences in the plant dataset led to poor resolution of the cells into clusters, or whether the communities truly did not have strong intrinsic clusters due to continuous environmental gradients. The results reported here suggest both factors are operating. The turnover in species composition from southern Upper Guinea to the Sudanian is characterised by a lack of disjunctions (although presumably this would be reduced with a reanalysis of non-gridded data), and Upper Guinea is closer geographically and floristically to the Sudanian Region than Lower Guinea is. However, the cells of evergreen and hyperwet Upper Guinea are very remote in ordination space from the cells of the Sudanian region (Fig. 3.4a). It has not been possible to define Guineo-Congolia to exclude the

forest region of West Africa using multivariate methods with the present dataset, which is well sampled for Upper Guinea.

A definition of Upper Guinea that terminates at the Dahomey Gap can be justified by our analysis (Fig. 3.5). Subdivision of the cluster of Upper and Lower Guinean cells shows that wet UG west of the Dahomey Gap is most distinct from wet LG, from which it is also the most geographically distant. However, the wetter, southern cells of Upper Guinea merge gradually into the more seasonal UG forests further north. These more seasonal UG forests in turn merge into the drier, more western Nigerian forests (from which they were probably not isolated c. 4,500 years ago; Salzmann & Hoelzmann 2005), which are in turn similar to the northern LG forests, which are similar to the rest of LG. This similarity in species composition through the drier forest belt of Guineo-Congolia means that Upper Guinea may cluster with some of the western Nigerian forest cells (Niger Delta termination, or at 4 or 5°E in Nigeria), or all of the Nigerian forest cells (Cross River termination). The distinct flora of the CVL and Cross River cells themselves make it difficult to interpret patterns of species turnover through Nigerian forests to central Cameroon; it also means these two forest types may indeed present a real barrier for many species.

UG also merges into the GCSU flora, and in turn into the SU. The GCSU has a disjunction between 8°E and 10°E, at the Jos Plateau further north, or the CVL further south, which helps to give 'Upper Guinea' a Cross River termination when some or all of the GCSU cells are included in its definition. In the GC+GCSU+SU analysis (Fig. 3.4) it was possible to separate the wettest Upper Guinean forest zone from the more seasonal Upper Guinean forest zone, and thus from any flora east of the Dahomey Gap, which would give Upper Guinea a narrow distribution no more than 300 km inland from the coast. When only the UG and LG forest zones were ordinated, the division between Upper Guinea and Nigerian forests was clear, resulting in a definition of Upper Guinea which includes drier types of forest, and terminates at the Dahomey Gap, as in White's definition. Relative to White's UG,

the current definition, which still includes the drier fragments of forest in Sierra Leone, terminates on the west coast of Africa at 9°N, whereas White's (1983) subregion terminated at 11°N.

The main gradient within Upper Guinea (excluding the transition zone to the SU) is between western and eastern Upper Guinea, with a secondary gradient north/south (Figure 3.6). Restricted range taxa, such as *Gaertnera luteocarpa*, respect this divide, with one subspecies in south west Ghana and south west Ivory Coast (Sassandra), and the other restricted to Liberia (Jongkind 2015). We have divided Western from Eastern Upper Guinea using the 0 point of Axis 1, which coincides with the Sassandra River. The forests associated with NW, SW, NE and SE Upper Guinea west of the Dahomey Gap together constitute the present definition of Upper Guinea. The two western clusters represent 'Upper Upper Guinea', and the two eastern clusters 'Lower Upper Guinea', the two northern forest clusters are more seasonal than the two southern forest clusters (Figure 3.7).

The forests of Nigeria have been maintained as a transition zone between Upper and Lower Guinea, called here 'Nigerian Guinea'. In the west it is more similar to Upper Guinea, whilst in the east it is more similar to Lower Guinea. The term 'Middle Guinea' has been used elsewhere either for Fouta Djallon (in Guinea) or the Dahomey Gap flora, and should therefore be avoided for referring to Nigerian Guinea. Nigerian Guinea often is grouped with Upper Guinea, giving an Upper Guinea + Nigerian Guinea forest block with a Cross River termination. In this case we suggest that the resulting sub-region should be named explicitly as Upper Guinea + Nigerian Guinea. A Lower Guinea sub-region east of the Cross River, a Congo sub-region east of the SRI, and a Congo-Afromontane transitional region, have been retained (Figure 3.7).

The Guineo-Congolian region proposed here is therefore subdivided into Upper Guinea, Lower Guinea and Congo sub-regions, and one (sub-regional) transitional zone, Nigerian Guinea. We have maintained one western and two eastern Guineo-Congolian/Sudanian regional transition zones, which are divided longitudinally by the Jos Plateau and the CVL (8-10°E), and together correspond to

our Guineo-Congolian/Sudanian regional transition zone. We have maintained one regional transition zone to the Afromontane in the east.

The absence of strong disjunctions makes subdivision of Guineo-Congolia subjective, as is the case for tropical Africa. We suggest, though, that the evergreen and hyperwet forest zone of Upper Guinea will always be recovered as part of Guineo-Congolia, and that there will always be a way to separate these forests from anything east of the Dahomey Gap, unless the analysis is overstretched or limited to drier extremes. We have focussed here on the subdivision of Guineo-Congolia, and in particular Upper Guinea. In our opinion, each of the broad regions we have defined warrant similar treatment, as would the subdivision of Lower Guinea and Congo.

Supporting Information

Appendix 1

A document describing 8 additional analyses concerning the definition of the Guineo-Congolian region (bound with thesis).

Appendix S3.1

A shape file of our regional framework for tropical Africa.

Appendix S3.2

All results for each degree square, including sampling levels.

Chapter 4

Bioquality hotspots in the tropical African flora

Cicely A. M. Marshall[1*], Jan J. Wieringa[2] and William D. Hawthorne[1]

1 Department of Plant Sciences, University of Oxford, South Parks Road, Oxford, OX1 3RB, UK.

2 Naturalis Biodiversity Center, National Herbarium of the Netherlands, Darwinweg 2, 2333 CR Leiden, Netherlands.

* Corresponding author

Author contributions

Conceptualisation: C.M. and W.H.; Methodology and Software: C.M and W.H.; Resources and Data Curation: C.M. J.W, W.H.; Writing – Original Draft: C.M.; Writing – Review & Editing: C.M, J.W., W.H.

Acknowledgments

Cyrille Chatelain provided the African Plants Database Tropical Africa species list. Data capture, editing and broadcast of the African Plants Database is the product of a collaboration between the South African National Biodiversity Institute, the Conservatoire et Jardin botaniques de la Ville de Genève, Tela Botanica and the Missouri Botanical Garden. BISAP (Biogeographical Information System on African Plant Diversity) has been established at the Nees Institute, University of Bonn in the context of BIOTA Africa project together with several external partners, especially Jon Lovett (University of Leeds) and Peter Linder (Zurich). For more details about the dataset and a full acknowledgement of all contributors please refer to Linder et al. 2005 and Küper et al. 2006.

Article status

Published as C. A. M. Marshall, J. J. Wieringa & W. D. Hawthorne (2016). Bioquality hotspots in the tropical African flora. *Current Biology* 26: 3214–3219.

Abstract

Identifying areas of high biodiversity is an established way to prioritize areas for conservation (Stattersfield et al. 1988, Myers et al. 2000, Olson et al. 2002), but global approaches have been criticized for failing to render global biodiversity value at a suitable scale for local management (da Fonseca et al. 2000, Mace et al. 2000, Brummitt & Lughada 2003). We assembled 3.1 million species distribution records for 40,583 vascular plant species of tropical Africa from sources including plot data, herbarium databases, checklists, and GBIF, and cleaned the records for geographic accuracy and taxonomic consistency. We summarised the global ranges of tropical African plant species into four, weighted, categories of global rarity called Stars. We applied the Star weights to summaries of species distribution data at fine resolutions to map the bioquality (range restricted global endemism) of areas (Hawthorne 1996). We generated confidence intervals around bioquality scores to account for the remaining uncertainty in the species inventory. We confirm the broad significance of the Horn of Africa, Guinean forests, coastal forests of east Africa, and Afromontane regions for plant biodiversity, but reveal also the variation in bioquality within these broad regions and others, particularly at local scales. Our framework offers practitioners a quantitative, scalable and replicable approach for measuring the irreplaceability of particular local areas for global biodiversity conservation, and comparing those areas within their global and regional context.

Introduction

In this Chapter, bioquality is calculated and mapped across mainland tropical Africa. Bioquality assessment is not new (Hawthorne 1996). GHI has been measured and mapped across Japan (Nakamura 2013), Trinidad & Tobago (Baksh-Comeau et al. 2014), the UK (unpublished), the forest zone of Ghana (Hawthorne 1996), and locally at numerous sites across Africa: Liberia (Putu Hills, Nimba County), Guinea (Nimba Mountains), Malawi (Shire River Basin), Ethiopia (Wondo Genet, Sof

Omar), Mali (Komana), Sierra Leone (Bumbuna), and elsewhere around the world, e.g. in Brunei (Kuala Belalong), and Chile (Maule).

However, this Chapter reports the first time bioquality has been measured and reported at near continental scales. This was made possible by the assembly of a large species distribution database for tropical African plants (Chapter 2), which was necessary for both Star rating the tropical African flora, and bioquality mapping across this scope. The number of plant distribution records available digitally has soared, but the development of systems and measures to extract simple but accurate meaning for non-specialists for practical applications has not kept pace. Here, we present a simple system for extracting only one, but an important, biologically relevant measure, from big data, for practical applications. We argue can that this system can become best practice in land use planning globally.

Materials & Methods

Distribution data

The data for this analysis are drawn from TRAFRICA, the assembly of which is documented in Chapter 2. Species were summarised uniquely by one degree square (Figure 4.1) for the analysis. Excluding straddling records (where location is defined by a polygon and that polygon straddles a sampling grid line), unlocalisable records, duplicate records (same species same degree square), records not identified at or below species level, and species with uncertain Stars or taxonomic status, resulted in 498,701 unique species records for the one degree square summary.

Star rating

Tropical African species were assigned a Star rating based on their global degree square occupancy, and the grid was defined with its origin at the meridian/equator intersect.

Star rating followed the principle that Black Star (BK) species occupy on average 2.7 degree squares globally and carry a weight of 27; Gold Star (GD) species occupy on average 8 degree squares and carry a weight of 9; Blue Star species (BU) occupy on average 24 degree squares and carry a weight of 3; and Green Star species (GN) occupy on average 72 degree squares and carry a weight of 0. Species which were not believed to regenerate in natural vegetation in tropical Africa – widely cultivated species – were assigned the Star 'GX' and were excluded from analysis.

We initially assigned mother species to a tentative Star by summing the number of degree squares occupied by each species globally. We assigned any species with >50 degree squares occupancy to Green Star. We assigned any species with 16 to 30 records to Blue Star. We assigned any species with 6 to 10 records to Gold Star, and any species with 1 to 3 records only to Black Star. Species with intermediate occupancy were assigned to intermediate Stars for checking by hand. 6528 species had been Star rated by W.H. before the thesis; if the degree square occupancy of one of these species had an intermediate Star and the original Star was one of them, we accepted the original Star rating.

We assigned to Star by hand all species with these 'intermediate' distributions, and all Black star species. To assign a species to Star by hand, we reviewed the number of degree squares in our database, and consulted other online resources for additional records, including the Plant List, the JStor Global Plants Initiative, the African Plants Database, WCSP, IUCN Red List, Tropicos, and original publications, particularly for recently described species. Where we found a higher degree square occupancy recorded in one of those sources than in our database, we recorded the new number, and used this to assign a likely Star. We also rapidly reviewed tentative Gold and Blue Stars, confirming or reviewing their status. The review process has been ongoing, and will be opened to broader review on a planned Star Server website, but meanwhile we believe the current suite of African Stars to be a reasonable first review, suitable for yielding robust bioquality scores in the face of future minor revisions.

We initially assigned infraspecific taxa to the same Star as their mother species. Where infraspecific taxa inherited an intermediate Star rating they were assigned by hand. Refinement of Star ratings for infraspecific taxa is ongoing, although Nakamura (2013) demonstrated that such adjustments had very limited impact on bioquality scoring for the Flora of Japan, where there an unusually high proportion of species have infraspecific variants. Infraspecific taxa may take a Star category that is intermediate between the Star they would be assigned if they were a species, and the Star category of their mother species. This down weighting for infraspecific taxa relative to their range is justified by their implied reduced genetic distinctiveness relative to two species.

Digitised distribution records are not yet sufficiently comprehensive to reflect accurately all plant species' true distributions globally, and this is more pronounced for infraspecific taxa, as in many digitised herbaria, such determinations lag behind species level determinations. It is known that certain groups of species are under recorded in herbaria, because of their geography or ecology. Fortunately, additional information about species' plausible distributions is available in formats other than gridded distribution data, for example range descriptions in a flora. A categorical system is better able to take advantage of this information than a continuous system because it is possible for a knowledgeable person to estimate which broad category of global rarity a species is likely to belong to, but not to assign a precise number of degree squares of occupancy. Although this introduces a degree of subjectivity to the system, the results are more accurate and better reflect the true breadth of knowledge regarding species' distributions than a strict reliance on digitised records would, and allows biases in a species' recorded collections to be compensated for. The shortfall of point data is particularly acute for more widespread species, yet statements like "widespread in the tropics" can be interpreted as Green Star with some confidence. A summary of the rules used to manage this subjectivity is included in Table 4.1.

Occasionally, species are too poorly known to allow a Star to be assigned, and they are designated with '?'. Synonyms carry the Star rating '!' in the database. Infraspecific taxa have their Star rating in lowercase. Vague names, invalid names, hybrids and species outside our remit carry the rating '-'.

We estimated (post-hoc) the mean degree square occupancy of each Star category to be 52.7 (GN), 16.11 (BU), 5.63 (GD), and 2.11 (BK) from records with point or small (<1 degree diameter) polygon localities held in our database. These values are around one third lower than the standardised degree square occupancy values of 72, 24, 8 and 2.7. This means that in our subjective Star checking against other botanic resources, we have tended to overrule the preliminary Star assignments which were based on degree square occupancy alone, towards lower weight (commoner) Star ratings. Our post-hoc weights mean that we have judged that we hold in our database (at degree square resolution or lower) only two-thirds of the known range of species, on average. As we digitise more data into our database, recorded degree square occupancy for each Star category for African species will rise to values closer to the ideal.

Star ratings are updated following available information. The latest Star ratings for species can be found as a data supplement associated with this Chapter (S4.1). A detailed discussion of Star Rating in general can be found in the RBS manual (Hawthorne & Marshall 2016).

Table 1. Key to define Stars for tropical Africa.

Global range	Interpretation	Star
Very local endemic; 1-5 degree square occupancy.	Known from the type only, or from 3 or fewer collections, or from a handful of populations. Known from up to 5 degree squares, if squares are adjacent to each other and recording is likely to be good. E.g. Endemic to Gulf of Guinea Islands; endemic to a single mountain or known from a single reserve, or 1 or 2 Flora regions in FTEA/FS.	Black
Local endemic; 4-12 degree square occupancy.	Known from more than 3 collections, but usually restricted to part of a region in tropical Africa. E.g. endemic to several mountains, or 3-5 Flora regions in FTEA/FS, or upper Upper Guinea (see Chapter 3) OR Apparently a Black Star species, but found in common habitats within tropical Africa and likely to be more widely distributed, or likely to have been under recorded (e.g. some ephemeral aquatics). OR apparently a Black Star species, whose specific distinctness is questionable.	Gold
Fairly widespread across one tropical African region, or sparsely recorded from more than one region; 8-30 degree square occupancy.	Sparsely distributed or otherwise specialised, for example, scattered on mountain tops within the eastern Arc mountains of East Africa, or only on lateritic outcrops.	Gold
	Could be known from as few as 8 collections from 8 degree squares, but where those degree squares are scattered and likely to imply occupancy between the known localities; or where recorded localities are common, yet sparsely collected, habitats in the region.	Blue
>30 degree square occupancy. Fairly or very widespread across more than one tropical African region; global species; pantropical species.	Widespread in the tropics, or across tropical Africa. Generalist species.	Green
Exotics	Not native to tropical Africa: exotics and garden or crop plants, or otherwise commonly planted.	Green (GX)
Insufficiently known	Where the type record is dubious, or where not mentioned in regional floras in spite of type recorded from the area, or where the species name is not recognised by multiple online resources, or where the taxonomic status of the species is very much in question (but not yet formally synonymised). Distribution insufficiently know, i.e. no reliable information available about where this species lives.	?
Uncategorisable	Species is found outside tropical Africa (e.g. South Africa, or North Africa, or Madagascar) but whose name appears in the database for some reason. Vague species, e.g. 'Begonia sp.; hybrid species. Invalidly published names.	-
Synonym	Synonymous taxa names.	!

Calculating GHI, GHI confidence intervals, and sample completeness

GHI is calculated using the following formula, where N_{BK}, N_{GD}, N_{BU} and N_{GN} are the number of Black, Gold, Blue and Green Star species, and W_{BK}, W_{GD} and W_{BU} are the respective weights.

$$GHI = 100 * (N_{BK} \times W_{BK} + N_{GD} \times W_{GD} + N_{BU} \times W_{BU})/(N_{BK} + N_{GD} + N_{BU} + N_{GN})$$

GX species are invisible in this GHI calculation. GX species are species whose distributions are largely if not entirely anthropogenic in tropical Africa (species like cassava or *Ixora coccinea)*, and incidentally are not regularly collected by botanists. These records were excluded from GHI calculations so that the GHI of a cell would not be unfairly brought down by the presence of a city, agricultural research station, or botanic garden (for example) in the cell. Although these species are plants, there are treated here like a statue would be – invisible – because we are concerned with documenting bioquality in natural vegetation, and not the bioquality of gardens (which is an interesting but separate matter). We have run the same analyses including GX species, and find there is almost no perceptible difference in the overall patterns, but cells containing well documented botanic gardens (e.g. Limbe botanic garden, Cameroon) have a slightly lowered GHI when GX species are included.

Confidence intervals were generated using a resampling procedure. Ideally, GHI is based on all or at least >90% of the flora of a place, or failing that on a subsample that is believed to have a balance of Stars close to the true balance. For local samples, this is achievable. At the one degree resolution, we know that we have <90% of the species recorded for most cells, and we do not know whether we have a biased (with respect to Star) sample of the flora. The bootstrapping approach described here does not test for these biases. Rather, it was used to create confidence intervals around the GHI scores, such that if there were a bias in either direction for a cell, the true value would still be within the confidence interval.

We estimated that 7 degree squares in tropical Africa had >90% of their species recorded, by comparing cell richness to richness estimates for degree cells based on Barthlott et al. (2005) (see below). For each such 'sub-complete' cell, we calculated the 'true' GHI (set GHI), using all the species present. Then, we drew 2.5% of those species and recalculated the GHI of the small sample. We repeated this 10,000 times, and from this batch of 10,000 sample GHIs we calculated the 1st and 99th percentiles of the GHIs. We plotted these min and max points against the true set GHI; we did this for all 7 cells; and then we modelled linearly the min and max values across the range of the GHI. This allows set GHI to be estimated from any sample GHI; each cell has a different balance of Stars and number of species, and this influences the maxGHI-minGHI ranges; the linear models are an approximate fit to this variation. We repeated this, drawing a random subset of 5%, 10%, 20%, 30%, 40%, 50%, 75%, 90% of the species from each cell; as a greater proportion of species are drawn, the range of values that sampled GHIs can take becomes more constrained by the true value of the cell (Figure 4.1). Having established these relationships, we treated each cell as a sample GHI with a particular sampling proportion, and asked what values the true set GHI could take, given the sample GHI and the sampling proportion (sampling proportion defined as in Figure 4.1A). For example, a cell with 50 species recorded from it where 2000 species might be expected has an estimated sampling intensity of 2.5%. If the GHI calculated from those 50 species is 500 (the sample GHI), then the 'true' GHI of the cell is likely to fall between 276 and 920. 276 is the minimum likely true GHI for the cell, given the sample GHI of the 50 species present. MinGHI was not mapped for any cell where <2.5% of the species were estimated to have been recorded.

Total species richness for each one degree cell was estimated using Barthlott et al. (2005). The map was digitised, and each degree square was assigned to the maximum Diversity Zone which it overlapped, corresponding to an estimated species richness at approximately this scale. This map was created using a much smaller dataset, and reports species richness estimates at 10,000 km^2 resolution rather than one degree square, but it serves as a reasonable approximation for this

purpose. We estimated sampling completeness by comparing our recorded richness against these estimates and presented the results as a percentage. To estimate total richness for areas of different sizes (for example in the quarter degree squares, some of which are visible in Figure 4.3), we did the following: (i) assigned each degree square to a simplified version of the African phytochoria, (ii) took estimates of total species richness for each phytochorion from White (1983), (iii) averaged species richness per one degree cell for each phytochorion (from Barthlott et al. (2005), (iv) created a dummy species list of equal length to the total richness values for each phytochorion, (v) resampled this pool by drawing the average estimated number of species per one degree square per phytochorion, (vi) calculated the number of unique dummy species in each 2, 3, 4 and 5 degree square area, (vii) fitted a species area curve to the values, to create theoretical species area curves for each phytochorion. This method is crude, as it does not account for any variables affecting species richness beyond the size of area and the phytochorion, but produces plausible richness estimates for the purpose of constructing confidence intervals around GHI at scales other than one degree square.

GHI results at one degree square, including sampling levels and lat/long, are supplied as a data supplement (S4.2).

Mapping

Mapping was conducted using QGIS 2.10.1. Geocoding and reverse geocoding analyses were conducted in part using Manifold.

Fieldwork

The survey data shown in Figure 4.4 were collected using the RBS method (Hawthorne & Marshall 2016). RBS has been refined over 25 years to fill the distribution data gap which exists between herbarium records and formal (tree-) plot records. Survey took place between 2010 and 2012 as part of ArcelorMittal Liberia's (AML) Environmental and Social Impact Assessment (ESIA) (URS 2013) and

an ESIA of the rail corridor for Euronimba Liberia Ltd. 31,775 records were made of plants across 310 RBS samples. Species were identified by W.H. and C.M, Pierre Poilecot and Ouo-ouo Haba, Patrick Ekpe, Carel Jongkind, James Kpadehyea, David Bilivogui, Steven Heathcote, Wing-Yunn Crawley and Daniel Dorbor. Specimens were deposited at FHO (Oxford) and with AML pending the resurrection of Liberia's National Herbarium.

Figure 4.1. Relationship between set and sample GHI at different sampling intensities. Linear models are fitted to seven, c. 100% sampled cells found in the TRAFRICA one degree dataset. Set GHI refers to the observed GHI of the cell. Sample GHI refers to the mean GHI of 10,000 bootstrap replicates drawn from each cell, when 2.5% to 90% (labelled 0.025 to 0.9) of species are drawn.

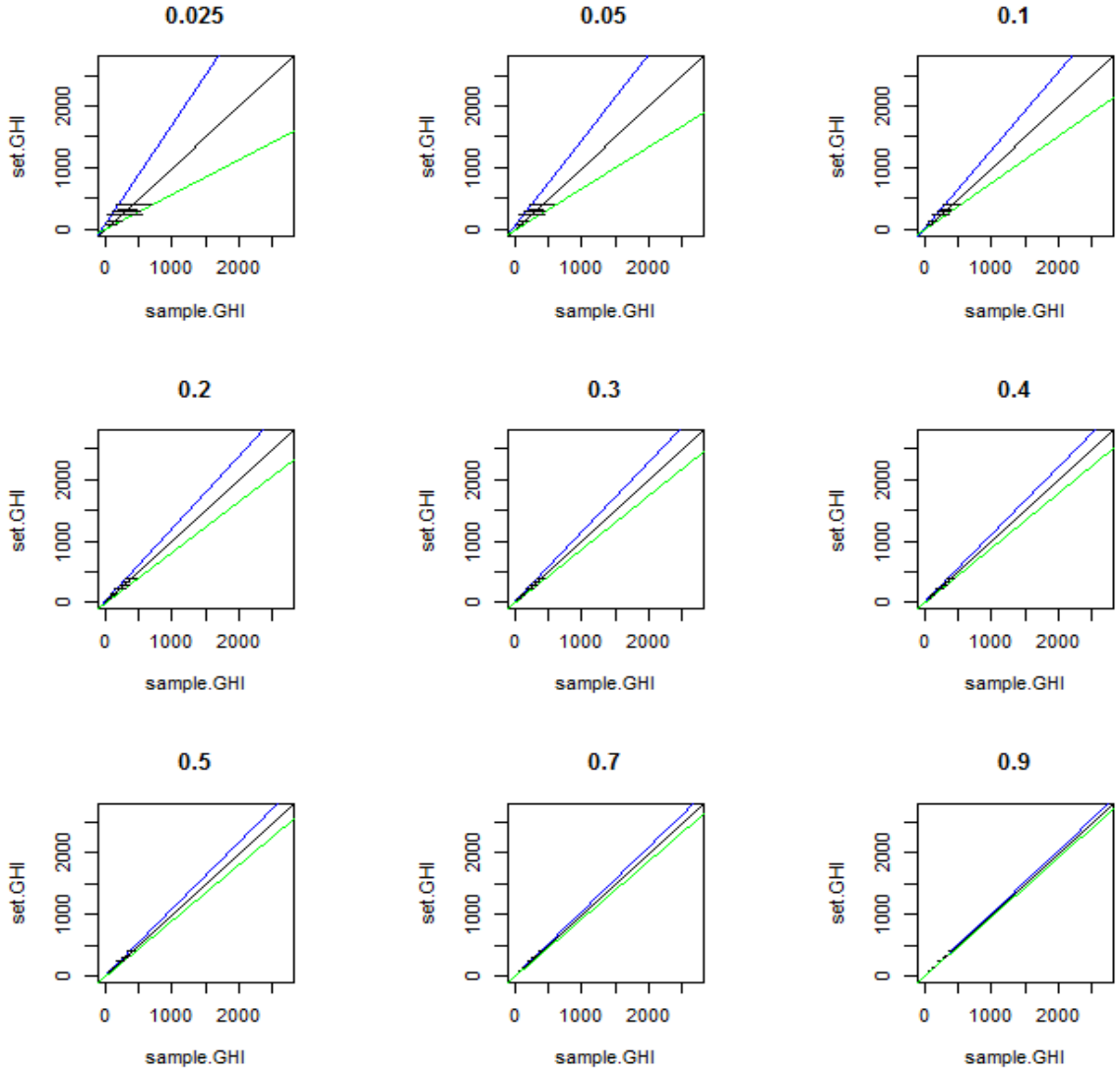


Figure 4.2. Example distribution patterns for a species of each Star. Black Star species occupy on average 2.7 degree squares globally. Gold Star species occupy 8, Blue Star species occupy 24, and Green Star species occupy 72 degree squares globally (or 100 x 100 km, whichever is the larger). Mapped distribution for *Allophylus africanus* includes distribution data for named formas and varieties.

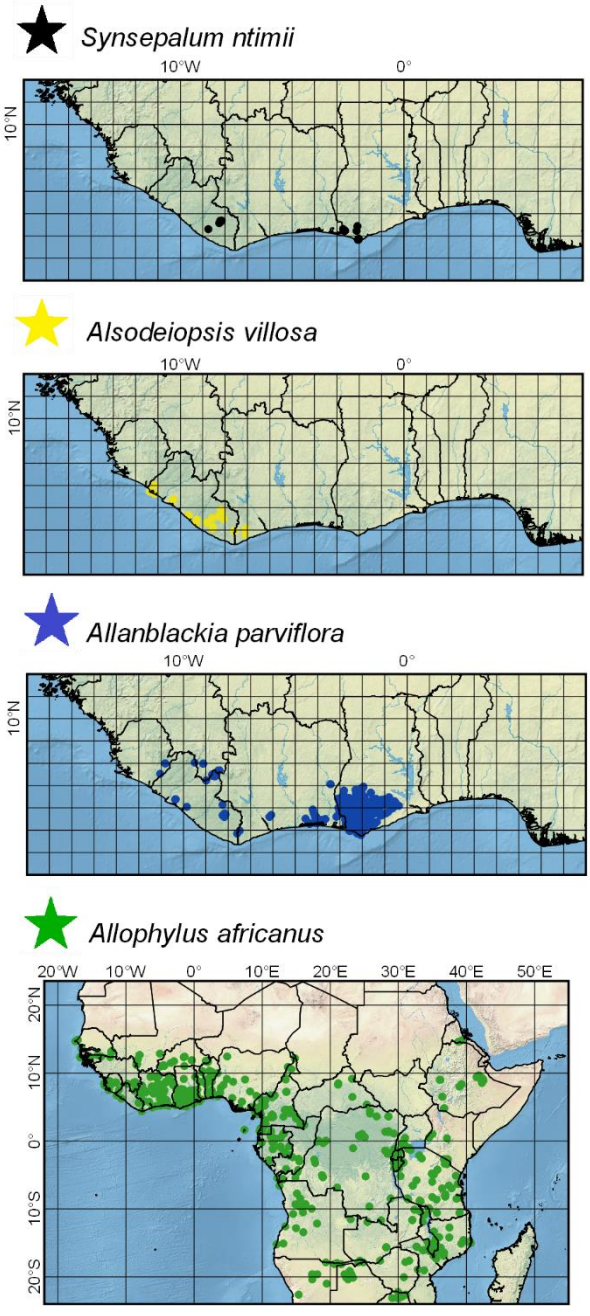


Figure 4.3. Bioquality hotspots in the tropical African flora. A: Ratio of species richness in our database to total species richness estimated using Barthlott et al. (2005). B: Bioquality mapped at one degree square resolution using minGHI, a reliable minimum estimate of GHI; minGHI is a conservative GHI estimate expected to be closer to the true GHI if collections are currently biased towards globally rare species. C: GHI values for bioquality (assumes no species sampling bias with respect to Star). D: maxGHI, maximum likely GHI assuming species sampling is currently biased towards the globally commonest species (probably the least likely scenario). Confidence intervals (minGHI to maxGHI) are larger where species sampling is poorer (compare panels A, B, C, D). 'True' GHI values, assuming complete inventory, would fall between minGHI and maxGHI estimates for each cell.

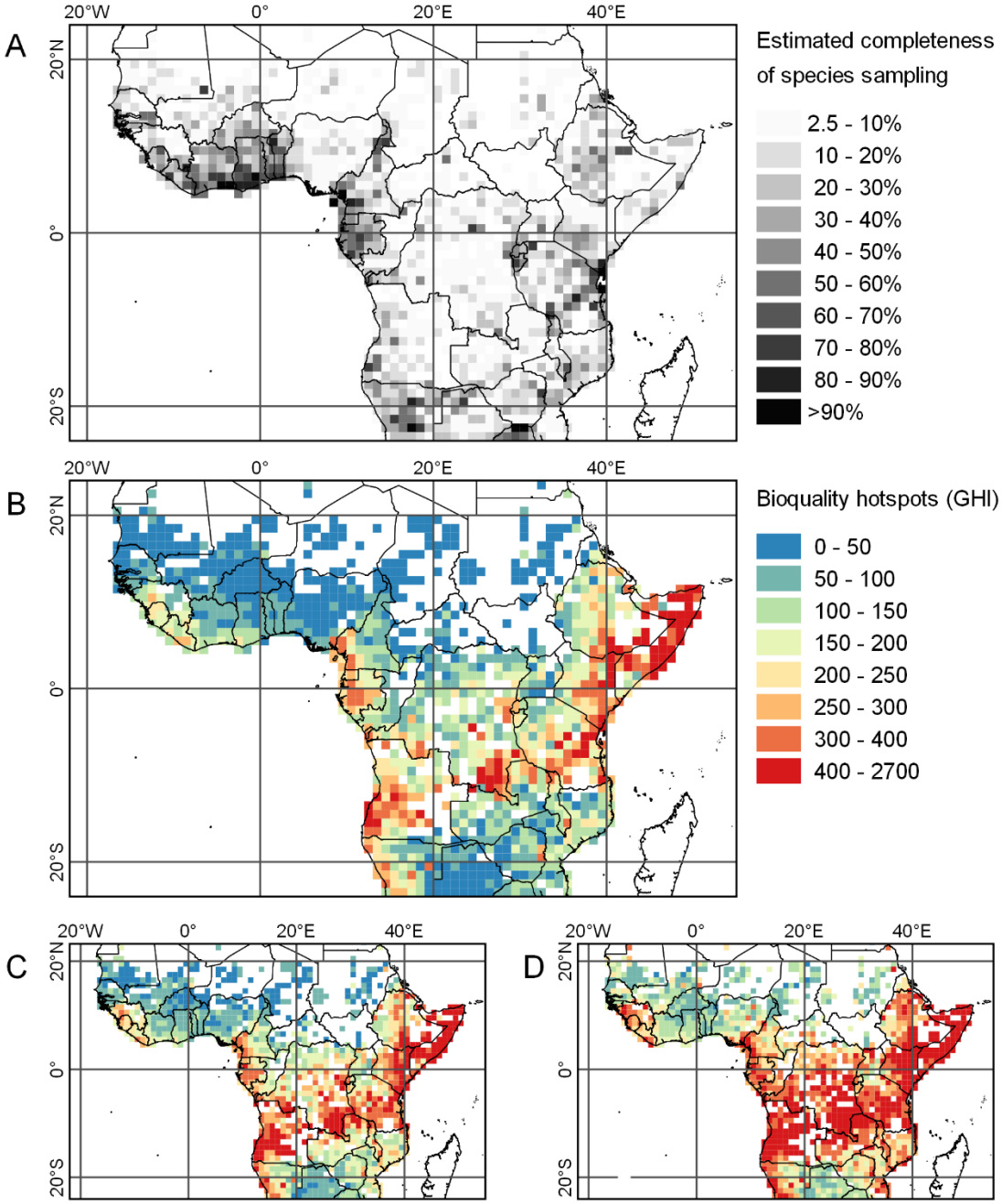
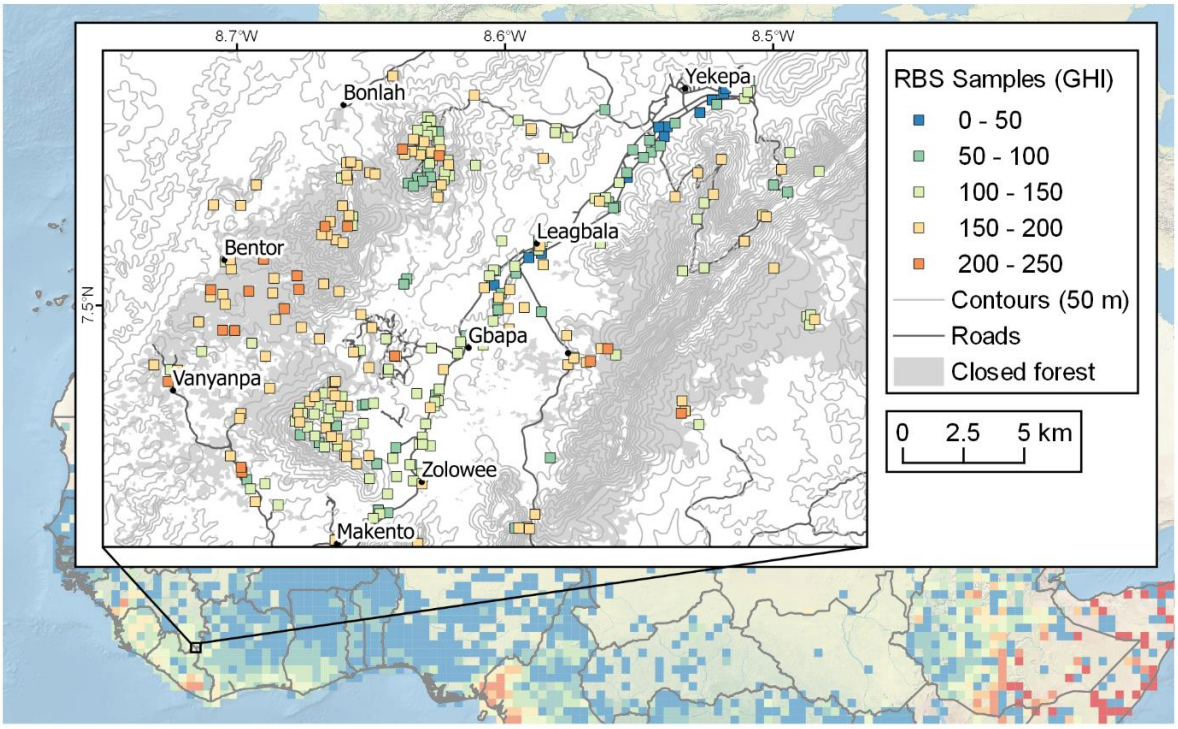


Figure 4.4. Bioquality at local scales. GHI calculated from 310 Rapid Botanic Survey (RBS) samples across northern Nimba County, Liberia. The hotter GHI (>200) scores equivalent to the minGHI estimate for the degree square as a whole were recorded in forest in this region, although not all the forest had such a high GHI. Background map shows minGHI for 0.5 x 0.5 degree squares.



Discussion

Distribution data for tropical African plants

Biodiversity hotspots were originally identified using the richness of species endemic to large, biogeographic realms (of very varied size) which had been significantly degraded (Myers et al. 2000), largely because species distribution data were available only at this coarse resolution (Brooks et al. 2006). This situation has improved rapidly as online public repositories (e.g. GBIF), collection digitisation efforts (e.g. JStor's Global Plants Initiative), and data journals (e.g. Check List) were established, increasing the available number of geolocated species records.

We assembled 3.1 million global species distribution records for tropical African vascular plants, from plot data, herbarium databases, checklists, and the Global Biodiversity Information Facility (GBIF). We limited our GBIF search to records supported by herbarium specimens and those without reported geographic issues. Our tropical African species list was derived from the African Plants Database, and includes 40,583 accepted species or infraspecific names which were checked for synonymy and comprehensiveness against other resources. We refer to species for simplicity, but all analysis was conducted on the lowest named taxonomic unit at or below the species level.

Of the 3.1 million distribution records, 0.5 million were without coordinates. We geolocated these records by comparing the text locality information provided in the collectors' notes to standardised gazetteer dictionary files, and assigned to the records either point-with-radius, or polygon, coordinates, depending on the detail available in the notes. Records assigned polygons were included in the following analyses if they fitted inside the sampling units in question. We used similar geolocation methods to detect records for which the supplied coordinates and supplied text locality information conflicted; such records were checked and corrected by hand or were omitted from analysis. Many older specimens, often including types, were collected without coordinates. Using these novel methods we were able to compile records for almost all vascular plant taxa present in

tropical Africa, ensure taxonomic consistency and geographic accuracy, and respect the geographic resolution of the original collection in the analysis.

We estimated the completeness of our species distribution data by comparing our species sampling levels against published estimates of species richness (Barthlott et al. 2005) (Figure 4.1A). There are many areas for which species sampling is far from complete, particularly for central Africa (Stropp et al. 2016). We must continue our efforts to fill these data gaps, but we cannot afford to ignore the biogeographic signal present in existing data or the plants we seek to record will be gone.

Star rating: Species-level conservation assessment

We summarised the global range for all plant species in tropical Africa into 4 categories of global range, called Stars (Hawthorne 1996) (Figure 4.2). Globally rare species are the important elements of biodiversity to conserve locally, in order to conserve species richness globally. Black Star species have the narrowest global ranges (c. 2.7 degree squares occupancy on average), Green Star species are the globally commonest (c. 72 degree squares), and Gold and Blue Star species are intermediate. Star ratings are taxon-specific, mutually exclusive and globally applicable, so that each species or infraspecific taxon in the world can have only one Star. Global ranges were categorized, rather than using a continuous occupancy metric, to produce a memorable framework that is robust in the face of moderate future range extensions and which retains the necessary subtlety to reveal robust biogeographic patterns. Given that the full degree square occupancy of all species globally is not yet known (Figure 4.1A), the categorical system also allows for interpretation of the appropriate Star rating for species which are inadequately represented in herbaria, for example due to geographic or ecological biases in collections. We reviewed each species' Star in light of the best available information from online floras and other botanic resources, unless it was already a Green Star species (globally widespread). Although this introduces a degree of subjectivity to the system, the results better reflect the true breadth of knowledge regarding species' distributions than a strict reliance on digitised records would. Each Star category carries a weight which is inverse to the mean

range (measured as degree square occupancy) for all the included species of that Star category, so that rarer species and Stars have a higher weight.

Star rating can be compared with the IUCN Red Listing approach when criterion B2 (AOO) is invoked, but Star rating requires no explicit measure of population change, regional ratings are not necessary or allowed, and the grid size for AOO calculations is standardised to one degree square (or 100 x 100 km, whichever is larger), for all plant species. Globally, three times as many vascular plant species have a Star rating compared with a Red List Category (62,868 cf. 20,147; 100% cf. 8% tropical African plant species assessed). Star rating offers a biologically pure assessment of a species' range, free from difficult projections of population change, which is relatively fast to conduct, and is useful for scientific analyses of distribution patterns as well as conservation assessment. As a consequence of this study, all tropical African vascular plant species have a Star rating, so the system can now be used to support or prioritise conservation anywhere in tropical Africa, and could be extended to other taxa.

Bioquality hotspots in tropical Africa

We used the Star ratings and species distribution summary tables to produce a quantitative measure of plant biodiversity value for areas across tropical Africa. Indices respecting species global ranges reflect a particular component of what specialists tend to recognise as the biodiversity value of a place. We refer to this attribute of plant biodiversity as *bioquality*, and the particular index used to measure bioquality is the Genetic Heat Index (GHI) (Hawthorne 1996). GHI is calculated for a unique species list for an area, by averaging over the weights of the Stars for those species found in the area. An area with a high proportion of globally rarer species in its flora achieves a high GHI and a high bioquality hotspot score.

This is similar to calculating range-size rarity (Küper et al. 2004, Williams et al. 1996), except we measured ranges globally rather than within the study area, to produce scores which are

comparable globally. Range size rarity uses the continuous degree square occupancy of species, whereas we have binned ranges into the four Star categories to produce results which are not artificially precise, given that the full degree square occupancy of all species is not yet known (Figure 4.1A). The biggest difference is that the GHI divides by the number of species present (to produce a weighted average), which means that the GHI is independent of richness or diversity. This has the possible disadvantage that areas with high absolute numbers of rare species achieve lower GHI scores if they also include many common species, but a number of significant advantages: areas are not downgraded if their species inventory is not complete, making the measure robust to missing data. GHI scores decrease where vegetation is invaded by globalized species. Species richness increases with the size of area under consideration: ignoring richness means that GHI scores can be calculated and meaningfully compared for areas of any shape or size, including the very local.

To conserve species globally, it is not important to prioritise individual areas with high species richness. Rather, it is important to protect areas where a high proportion of the individuals belong to globally rare species, otherwise those species would be lost from the global species pool (Mittermeier et al. 2011). In fact, the number of species in an area, whether rare, threatened or simply present (richness), is now generally recognised as a poor metric for identifying conservation priorities, because richness alone reveals little more than the availability of data, the size and shape of the area under consideration, and the biome type (Brummitt & Lughadha 2003).

When GHI is calculated from an essentially complete species list for an area, then confidence intervals are not necessary. Neither would they be necessary if species were sampled incompletely but representatively with respect to the true balance of Stars in the full flora, because as a weighted average the GHI includes is independent of richness. However, we cannot tell whether the recorded flora is currently biased towards the globally rarest (or commonest) species. We therefore estimated bootstrapped confidence intervals for the GHI for each degree square, given the apparent GHI (Figure 4.3C) and current estimated species sampling completeness (Figure 4.3A), to produce a

confidence interval within which the true GHI value of each area is expected fall, even if sampling is currently biased with respect to Star (Figure 4.3B and 4.3D). This is one way in which uncertainty can be quantified and reliable conclusions drawn, whilst the species inventory is incomplete.

Figure 4.3B reveals tropical Africa's biodiversity patterns in their most complete, repeatable, and intimate detail yet. On the whole, the results fit comfortably with previous studies of the distribution of Africa's plant biodiversity (White 1983, Lovett 1998, Lovett et al. 2000, Linder 2001, Küper et al. 2004, Wieringa & Poorter 2004), by highlighting the generally rather low endemism in the Sahara, Sahel and Sudanian regions, and medium to high endemism for the Guineo-Congolian, Zambezian, Somalia-Masai, Karoo-Namib, Zanzibar-Inhambane and Afromontane regions. The Somalia-Masai (Horn of Africa) flora comes out as one of the highest bioquality floras in tropical Africa. While the large number of endemic species has been recognised (Thulin 2003), Somalia's high bioquality has perhaps been underappreciated relative to Africa's wetter and montane forest regions (Küper et al. 2004), most likely due to undersampling and relatively lower species richness.

Smaller bioquality hotspots are visible around Mount Cameroon, Mount Mulanje and Mount Chimanimani. In Guineo-Congolia, bioquality peaks in the high rainfall forests of Cameroon and Gabon towards the coasts; is higher for western Upper Guinea than in the east; and bioquality is somewhat lower but comparable for Congolia (though data are sparser). Bioquality peaks in the Zambezian region in south eastern Democratic Republic of Congo, and in southwest central Angola. For the Karoo-Namib, the coastline of southern Angola is particularly hot; the flora of the eastern coast of Africa (Zanzibar-Inhambane regional mosaic) is particularly hot in south east Tanzania.

Bioquality at local scales

Our bioquality metric (GHI) is based on a weighted average of globally rare plants, and as proportions scale meaningfully with richness and area, the scale (grain) and shape of sampling units for an analysis can be matched to its application. The data for such fine-scale bioquality analyses can

be derived for a project area by on-the-ground sampling. In particular, Rapid Botanic Survey is a botanical survey technique specifically designed to collect this information with the minimum possible effort, although a meaningful GHI score can be calculated from any reasonably taxonomically complete survey data e.g. relevés (Moore 1962) or all-species transects.

Figure 4.4 reveals the local variation in bioquality found by local sampling within one of these degree squares, around Yepeka (Nimba mountains, northern Liberia), and across different vegetation types and altitudes. Such local-scale information is particularly useful for land management planning. Bioquality around the Nimba Mountains is lower for the more populated, lowland area around the central road corridor, and peaks in the closed canopy slope forests at higher elevations, with some variation apparent even within this forest type. It is clear that this 'hotspot' at the one degree square scale is a patchwork of hot and cold spots at a finer level. It is useful to be able to measure how hot an area is at this rather local scale, because it is at this scale where decisions impacting biodiversity are often taken.

The background map shows minGHI at 0.5 x 0.5 degree square resolution, and reveals bioquality patterns in greater detail than the one degree map of Figure 4.3B, although fewer data points can be resolved to this higher resolution grid.

Bioquality as a conservation framework

Bioquality is measured using the global range of plant species. Vascular plants are often used as an indicator taxon for biodiversity measurements because they are relatively well known taxonomically and geographically, and define the terrestrial habitats in which other taxa live. If high bioquality is used to define priorities for conservation, or to inform local land management, it makes sense to consider many other aspects of an area (Hunter & Hutchinson 1994), including species other than plants (Grenyer et al. 2006), social factors (Cincotta et al. 2000), economic cost/benefit analyses (O'Connor et al. 2003), ecosystem-wide benefits (Loreau 2001), phylogenetic diversity and

evolutionary processes (Vane-Wright et al. 1991), and rates or risk of habitat loss (Myers et al. 2000). We keep such measures out of our plant bioquality analysis, and promote viewing them as independent GIS layers, because mixing criteria in a single metric makes results harder to interpret and to make globally consistent. We accept that the proportion of globally rare plant species in a flora is by no means the only important factor when designing a land management plan, but it is a critical one.

As a consequence of this study, all mainland tropical African plant taxa have a Star rating and GHIs can now be calculated easily anywhere in tropical Africa where the species composition is at least partly known. This should prove useful in the context of Environmental Impact Assessment, or Protected Area planning, because a local scale hotspot map and database can: describe a baseline; inform the positioning of infrastructure or protected areas; identify appropriate offset areas; allow precise monitoring of impacts and changes through time (with resurvey); and help devise management plans for the globally rarest species. We accept as a premise of the system that the data are never complete, and that taxonomic boundaries also shift, so the system is built to be robust in light of new information.

As much as 79% of Earth's land surface has now been prioritized for conservation under one system or another (Brooks et al. 2006), and we do not wish to define yet another set of broad areas of conservation importance. Instead, our framework offers conservationists and land managers a quantitative and replicable approach for measuring the irreplaceability of particular local areas for global biodiversity conservation, and comparing those areas within their global and regional context.

Supporting information

Appendix S4.1

Star ratings for tropical African species.

Appendix S4.2

Sampling levels and GHI scores for each cell in the analysis.

Chapter 5

Anatomy of a hotspot: gradients in species composition and bioquality in Upper Guinean hotspots

Cicely A. M. Marshall[1*], Jonathan Dabo[2], Markfred Mensah[2], Patrick Ekpe[3] and William D. Hawthorne[1]

1 Department of Plant Sciences, University of Oxford, South Parks Road, Oxford, OX1 3RB, UK.
2 Forestry Research Institute of Ghana, P. O. Box UP 63 KNUST, Kumasi, Ghana
3 Ghana Herbarium, Botany Department, P.O. Box 55, Legon, Ghana

* Corresponding author

Author contributions

C.M., M.M., J.D. and W.H. collected the data, C.M., M.M., J.D., P.E. and W.H. identified the specimens, C.M. conducted the analysis, C.M. wrote the text with input from S.H. and W.H.

Acknowledgements

We thank the Oxford University Expeditions Council; Wilson Fund, Department of Plant Sciences; and Merton College, University of Oxford for funding the new field work in SW Ghana. We thank Samartex, FORIG and the Ghana Herbarium for facilitating the new field work in SW Ghana. We acknowledge the support of the ODA (now DfID) to W.H. for previous field work in SW Ghana. We thank John Howell and Jamison Suter of ArcelorMittal Liberia and BHP Billiton for supporting the field work in Liberian Nimba. We thank James Kpadehyea, Ouo-Haba, David Bilivogui, Pierre Poilecot, Carel Jongkind, Daniel Dorbor, Wing-Yunn Crawley, Steven Heathcote, Hannah Greene, Lincy Barlon, Eric Bontor and Rosemarie Wonbenyakeh for fieldwork and identification in Liberian Nimba. We thank Carel Jongkind, Patrick Ekpe, Jan Reitsma, Elze Hesse, Danae Maniatis, Malika Virah-Sawmy, David Bilivogui, Paquilé Cherif, Moussa Diabaté, Ouo-Ouo Haba[9], Jean-Louis Holié, Delamou Londiah, John Amponsah, Nèma Soua Loua, Gononan Traoré, Guila Gbamou and M. Zojbelema for fieldwork and identification in Guinean Nimba.

Article status

Intended for submission to Biological Conservation.

Abstract

Aim: The recognition of an area as a 'hotspot' of endemism for globally rare plants does not imply that all vegetation within it is 'hot'. We investigate patterns of floristic global endemism within two of Upper Guinea's hotspots at the local scale (sample level). These patterns are interpreted in the context of local environmental gradients and disturbance regimes. We investigate the bias that may arise in our interpretation of hotspots as a result of changes in taxonomic and range conceptions of species over time.

Location: 114 sample locations within forest reserves of south west Ghana (SW Ghana), and 454 sample locations within the Nimba Mountains (Liberia and Guinea).

Taxon: 1,042 vascular plant taxa (species and infraspecific) of SW Ghana, and 1735 vascular plant taxa of the Nimba Mountains.

Methods: Linear modelling of bioquality scores, where bioquality is calculated as a weighted average of species' global ranges for each sample. Modern climatic variables, latitude and longitude, altitude, disturbance (Pioneer Index), local landscape position and swampiness were explored as potential correlates.

Results: Increasing disturbance, measured with the Pioneer Index, reduces bioquality significantly in both datasets. Disturbance reduces bioquality more rapidly in areas that would otherwise be hot than in areas that are otherwise 'cold'. Latitude is well correlated with climatic gradients in SW Ghana, and annual rainfall in particular is a significant predictor of the high global endemism associated with the far south. In the Nimba Mountains, where the altitudinal range is large, higher global endemism is found at higher altitudes, particularly over 1000 m. Bioquality scores have not changed overall in SW Ghana over 20 years, although the species responsible for this perception have changed.

Main conclusions: Bioquality scores are much more stable than our understanding of species' identity and ranges, which allows for some consistency in results and conservation priority setting over time and in the face of partial information. Areas of high rainfall and high altitude are geographically restricted within Upper Guinea, which must contribute to the high global endemism of the species found in these areas. The Pioneer Index is the only significant term retained in both the SW Ghana and the Nimba models, and suggests a degree of generalisability of the negative impact of disturbance on the distribution of globally rare species.

Introduction

In this Chapter, bioquality patterns within two of Upper Guinea's hotspots are modelled within a linear framework, at the sample level.

Upper Guinea is a chorological term for the forest zone of west Africa, running from Sierra Leone in the west to the Dahomey Gap (Ghana) in the east, from the coast up to 350 km inland (White 1979, Chapter 3). The recognition of Upper Guinea as a sub-region of Guineo-Congolia implies a relatively high degree of vascular plant endemism, and a relatively high degree of species homogeneity. Guineo-Congolia has the highest rate of endemism in tropical Africa (34%, equal with the Zambezian region); while Upper Guinea has 7% taxon endemism, the lowest of the Guineo-Congolian sub-regions (Chapter 3).

Three studies identified hotspots of endemism within Upper Guinea at 1-degree square resolution or lower (Marshall et al. 2016 (Chapter 4), Küper et al. 2004, Wieringa & Poorter 2004). Küper et al. (2004) mapped range size rarity across Africa using the BIOTA dataset at one degree square resolution. Range-size rarity combines the number and range sizes of species in each cell within the study area (Gaston 1994). In Upper Guinea, highest range size rarity was identified in the Nimba Mountains, Monrovia, SE Liberia/SW Ivory coast, and Abidjan, with slightly lower range size rarity in SW Ghana. This technique was sensitive to the different collection intensities for each cell, which

probably accounts for the high range size rarity associated with Monrovia and Abidjan. Wieringa & Poorter (2004) mapped the range size rarity of species across Upper Guinea at 0.5 degree square resolution, controlling for collection intensity, and identified high values in the Nimba Mountains, SE Liberia, SW Ivory Coast, and SW Ghana. Marshall et al. (2016) found high bioquality within Upper Guinea in Nimba, SE Liberia, SW Ivory Coast and SW Ghana, as well as generally elevated scores for western Liberia and Sierra Leone, confirming the results of Wieringa & Poorter (2004). Bioquality is similar to range size rarity, except that species' ranges are measured globally at one degree square resolution rather than within the study area at any resolution, and species ranges are categorised into four groups, called Stars, rather than being treated continuously (Chapter 4).

These three studies used herbarium databases, which have the advantage that they summarise existing records, generally with taxonomic accuracy, but have the disadvantage that records are likely to have a geographic or taxonomic bias that can be problematic for ecological applications. An alternative to using herbarium databases is to conduct ground-based surveys at a series of sample points. One method that has been used widely in West Africa is Rapid Botanic Survey (RBS), where samples are located in homogeneous vegetation units, and the intention is to collect or record every species of vascular plant present locally (Hawthorne & Marshall 2016). Sample data representing plant communities is particularly valuable, because herbarium collections rarely reveal much of the floristic associations of a species.

Globally, we have a limited understanding of the processes by which taxa characteristic of hotspots have come to be spatially aggregated, or are maintained in aggregation (Simon et al. 2009). This is particularly the case at the sample level, where information about specific global endemism is less likely to be known or considered. There is the possibility that hotspots defined at international scale are a 'false' result due to collecting bias (Nelson et al. 1990). An inadequate knowledge of taxon distributions has been called the Wallacean shortfall, and of taxon identity the Linnaean shortfall (Brown & Lomolino 1998, Lomolino 2004, Whittaker et al. 2005). Low to moderate levels of sampling

can act to increase artificially endemism scores for a locality, particularly if the rest of the region is poorly collected, because herbarium collectors tend to collect rare, novel or unusual species. Controlling for collection intensity when calculating endemism for a degree cell only partly corrects for this issue: our understanding of a species' range is still affected by the collection bias. When a new species is described it is commonly known from only one or a few localities and becomes a local endemic; this can result in artificial rarity for newly described species. Conversely, species may be sunk based on new collections that reveal a single, widespread variable species concept is more appropriate than separate, less variable, less widespread species.

Upper Guinean hotspots have been interpreted as the result of species' adaptation to present day environmental determinants, such as altitude and topography (Linder 2001, Mutke et al. 2001, Barthlott et al. 2007) and rainfall (Wieringa & Poorter 2004). For high endemism to occur, it is necessary that these environmental determinants be restricted geographically, and for species to be poor dispersers relative to the scale of geographic restriction. Most authors have also supported the interpretation of hotspots in Africa as the result of historic changes in climate, particularly during the Pleistocene (2.8-0.01 Ma) (Hamilton 1976, Lovett et al. 1996, Fjeldså & Lovett 1997, Linder 2001). SW Ghana was likely to have remained forested during the Pleistocene, on the evidence of palynological studies of a deep-sea marine core off the Dahomey Gap (Maley 1996, Morley 2000). The Nimba Mountains could have provided a range of conditions suitable for species adapted to a variety of habitats during periods of climate change, due to their altitudinal range (Robbrecht 1996, Maley 1996).

A bioquality hotspot at a relatively coarse resolution does not necessarily imply that all areas within it will be rich in globally rare species. We present an analysis of the gradients in species composition and bioquality at the sample level, within two of Upper Guinea's hotspots, SW Ghana and the Nimba Mountains, to examine the 'anatomy' of these hotspots. We investigate the extent to which modern environmental and geographic variables can predict high bioquality, at very local resolution, within

two of Upper Guinea's bioquality hotspots (putative refugia). SW Ghana is regarded as a local hotspot of plant endemism within Upper Guinea mostly due to the relatively higher numbers of range restricted species known from the Ankasa Resource Reserve (RR) and neighbouring Wet evergreen forests (Cape Three Points, Draw River, Neung and Ndumfri Forest Reserves). Less is known about the species composition, or bioquality, of the forest reserves immediately to the north of Ankasa. Using new data collected from these more northern reserves, we aim to establish how far north the hotspot of SW Ghana extends, and whether or not the rarest species are confined to particular vegetation types with the area, such as swamps, high elevation, or undisturbed forest. Within the Nimba Mountains hotspot, the altitudinal gradient is significant, with the highest peak c. 1600 m above sea level. We investigate the extent to which the Nimba hotspot is confined to these high altitudes, using a dataset of 454 samples from a range of altitudes and vegetation types.

We also investigate the extent to which observed high bioquality has been stable in SW Ghana, in the face of changing conceptions of species' range and identity over 20 years.

Materials & Methods

Botanic Survey data

27 samples were enumerated in SW Ghana, which were compiled with 87 botanic samples from the region from different surveys (Table 5.1). We compiled 454 samples from the Nimba Mountains, from three previous survey efforts (Table 5.2). The SW Ghana and Nimba datasets were analysed separately from each other.

Records identified to genus or family level only were dropped from the datasets. Analysis was conducted at the lowest named taxonomic level (species or named infraspecific taxa). Any record carrying a synonymous name was updated to the accepted name before analysis, using the TRAFRICA taxonomic framework. Specimens were collected of all plants for which identification was

not absolutely certain. Specimens are housed currently at the Daubeny Herbarium in Oxford (FHO), with AML in Liberia, and with Euronimba in Liberia. For SW Ghana, identification and fieldwork was carried out by J.D., M.M., C.M., and W.H., with identification assistance at the Ghana herbarium from P.E. For Nimba, identification and fieldwork was carried out by W.H., P.E., C.M., and others named in the acknowledgements.

With the exception of the Hall & Swaine 1981 A plots, which were 25 x 25 m samples of the vascular flora, and B plots (the first 30-60 vascular plant species encountered), all samples were conducted using the Rapid Botanic Survey (RBS) method (Hawthorne 1996, Hawthorne & Marshall 2016). For every sample enumerated, we aimed to record every vascular plant species present in the understorey, as well as the identity of canopy trees. An RBS sample is defined as a sample of a local, homogeneous, vegetation of interest to the surveyors, and is bounded in two ways; each sample must include at least 40 canopy tree species, and a sample should not take longer than 4 hours to complete. No specialist tree climbing equipment was used, so epiphytes were probably under recorded.

27 new RBS samples were conducted in SW Ghana for this publication, arranged along a latitudinal gradient from 5.21 to 5.69 °N. The area is generally low lying (55-185 m), but vegetation was sampled to include local hill tops, valley bottoms and riversides in the dataset. Samples were conducted mainly in undisturbed forest, although four samples showed signs of moderate disturbance from logging or footpaths. Latitude, longitude (decimal degrees), altitude (m asl), reserve name, landscape notes and vegetation notes were recorded for samples in the field.

The abundance of understorey species in each sample was coded as 1 for present, 2 for common and 3 for dominant. Canopy trees were counted. For compatibility of canopy data with understory data during analysis, we subsequently and additionally coded canopy tree species ‘1’ for abundance

if they accounted for less than one third of the counted stems in their sample, '2' for one to two thirds, and '3' for > two thirds of stems counted in the sample.

Permission to collect and export plants in SW Ghana was obtained from the Ministry of Lands and Natural Resources, Samartex, and Wildlife Division for Ankasa RR, and in Nimba, from the concession owners (AML, BHP, Euronimba) and the CITES offices of Liberia and Guinea.

Calculating GHI

Bioquality scores were calculated for each sample using the Genetic Heat Index, GHI (Hawthorne 1996). Star rating follows Marshall et al. 2016 (Chapter 4). Each Star rating carries a weight (BK=27, GD=9, BU=3, GN=0). The GHI for each sample was calculated from the species present, using the following formula, where N_{BK}, N_{GD}, N_{BU} and N_{GN} are the number of Black, Gold, Blue and Green Star species in a sample, and W_{BK}, W_{GD} and W_{BU} are the respective weights.

$$GHI = 100 \times (N_{BK} + W_{BK} + N_{GD} \times W_{GD} + N_{BU} \times W_{BU}) \div (N_{BK} + N_{GD} + N_{BU} + N_{GN})$$

Environmental data

Bioclimatic variables (Bio1 to 19) were extracted from the WorldClim resource for the sample locations at 30 arc-second resolution (c. 1 km^2) using the packages raster (Hijmans 2016) and sp (Pebesma & Bivand 2005, Bivand et al. 2013) in R version 3.3.1.

Pioneer Index & Swamp index

Disturbance scores were calculated for each sample using the Pioneer Index, PI (Hawthorne 1995, 1996). Guilds follow Hawthorne (1996). PI is calculated as:

$$\text{Pioneer Index (PI)} = ((N_P \times 2) + N_{NPLD} \times 100) \div (N_{PI} + N_{NPLD} + N_{SB})$$

Pioneer species (P) are defined as species which regenerate in gaps, and tend to dominate early successional forest; crowns of all sizes of pioneer trees tend to be more exposed than average for

the size class. Seedlings of non-pioneer light demanders (NPLD) are more commonly found in shade than gaps, although saplings are soon very light demanding; the crowns of medium and large NPLD trees tend to be more exposed than average for the size class. All sizes of species of the shade-bearer guild (SB) can be common in undisturbed forest, especially the understorey, and all sizes of shade-bearer tree have crowns which tend to be more shaded than average for the size class. Species of the swamp guild are largely restricted to riversides or periodically inundated or permanent swamps and are not included in the pioneer index.

Analysis

Samples were ordinated using NMDS with a Bray-Curtis dissimilarity matrix on abundance data with the Vegan package (Oksanen et al. 2016) in R version 3.3.1. Variables were fitted to the ordination space and visualised. Linear modelling was conducted using base R version 3.3.1. The statistical population of GHI scores has a normal distribution, except that some values cannot be achieved on account of the weights being multiples of three, and the index being bounded at 0 and 2700. Species range size distributions typically follow a log normal distribution (Gaston 1996). The GHI of ecological samples does not always follow a (log) normal distribution, depending on the location and vegetation types sampled. In the SW Ghana dataset, GHI scores were log transformed, rendering the dataset normal (Shapiro-Wilk test $p=0.36$ for log(GHI), $p=0.79$ for residuals). In the Nimba dataset, the transformation rendering the dataset most normal (natural logarithm) produced GHI scores which remained right-skewed and non-normal (Shapiro-Wilk test, $p<0.0005$). The dataset can be made normal by removing the 23 samples with GHI over 300, but this would remove much of the interesting signal from the dataset. The linear model presented has residuals which are not normally distributed (the model has worse fit at higher GHIs). The model's general conclusions were checked by fitting local regressions to the dataset, and by analysis of the normal subset of the dataset. Mapping was conducted in QGIS version 2.16.1.

Table 5.1. Datasets compiled for SW Ghana. The location of the five reserves, and the samples, is shown in Figure 5.1. The dataset contains 13,357 records of 1,042 taxa identified to species level (or infraspecific level where applicable), from 114 samples. The header file (sampling levels, sample attributes) for the 114 RBS samples can be found in Appendix 5.1.

Dataset	Reference/Ownership	No. samples: Ankasa	Boi Tano	Tano Nimri	Jema Assemkron	Nini Suhien	Total
2015RBS	Marshall et al. (this publication)	2	16	9	0	0	27
TSP	Marshall & Hawthorne unpublished	3	0	0	0	0	3
HANDS	Hall & Swaine 1979, 1981	2		2	1	0	5
RBS1991	Hawthorne & Abu Juam 1995; Hawthorne 1996	7	4	4	3	0	18
ANK1998	Hawthorne 1998	45	0	0	0	10	55
GSBA	Hawthorne et al. 2002	0	5	0	1	0	6
		59	**25**	**15**	**5**	**10**	**114**

Table 5.2. Datasets compiled for Nimba Mountains. The location of these samples is shown in Figure 5.4. The dataset contains 40,259 records of 1735 taxa (species or named infraspecific ranks) from 454 samples. The header file for the 454 RBS samples can be found in Appendix 5.2.

Dataset	Date range	Reference/Ownership	Country	No. samples
AML	2010-2011	URS 2013. AML; Hawthorne & Marshall	Liberia	236
BHP	2007-2008	BHP Billiton; Hawthorne & Marshall	Guinea	139
Euronimba	2012	Euronimba; Hawthorne & Marshall	Liberia	79
				454

Results

Stability of hotspots

If species are upgraded or added to the Black, Gold and Blue Star list at more or less the same rate as they are downgraded or sunk, then GHI scores will be stable, even if the identity of the species responsible for them shift. The stability of Star ratings and GHI scores is explored for Ghana, where Star ratings from 1996 are available (Hawthorne & Abu-Juam 1995, Hawthorne 1996).

Many species have changed Star over 20 years (Table 5.3). For each Star category, more species have moved into a lower Star category than a higher one. New collections, and an increase in the information available digitally about species' ranges, has resulted in larger ranges for Ghanaian species recorded in 1996.

However, GHI scores for SW Ghana have remained consistent: GHI scores were calculated for the 114 samples of the SW Ghana dataset using the 2017 Star ratings, and using the 1996 Star ratings. A paired t-test shows no significant difference between the mean GHI scores ($p=0.47$, $n=114$). For 22 samples, we have a record of the GHI scores calculated at the time (Hawthorne & Abu-Juam 1995). The GHI scores as reported in 1995 are also not significantly different from our calculation of the GHI of those samples in 2017 ($p=0.24$, $n=22$).

This is because new Black, Gold and Blue Star ratings have been made since 1996, which on average, have counteracted the increasing range of previously rated species over time. Newly rated species are the result of new records of species in Ghana ('new record'), new species descriptions ('new species'), and new Star ratings for species which were too data poor to be assessed in 1996 ('new rating'). This process is explored for Black Star species of SW Ghana only, for brevity, in Table 5.4. Black Star species have the greatest weight in GHI calculations.

Of the 23 Black star species of SW Ghana known in 1996, only 8 have remained on the current list, with the other 15 species either synonymised (1 species) or downgraded. But 15 Black star species have been added to the list, through new species publications (10 names), new records (4 species), and new ratings (1 species), with the result that there are still 23 Black star species known from SW Ghana.

Table 5.3. Movement of species between Star categories, comparing Star ratings of 1996 to Star ratings of 2017. 250 species names from 1996 were synonyms; these were converted to their currently accepted name for analysis. Nine species were then redundant, as they had been sunk into names already listed (1 BK, 2 GD, 2 BU, and 4 GN); these were removed from the calculations. Of 46 Ghanaian species recorded as Black Star in 1996, 17 remain Black in 2017, 22 have been downgraded to Gold in light of new data, and 7 have been downgraded to Blue. Of the species with Stars in 1996, 24 species are considered Black in 2017, compared with 46 which were considered Black in 1996.

Star 1996	BK (2017)	GD (2017)	BU (2017)	GN inc. GX (2017)	? (2017)	**Totals 1996**	**Totals 2017**
BK	17	22	7	0	0	**46**	**24**
GD	3	29	132	13	0	**177**	**87**
BU	2	19	176	109	1	**307**	**602**
GN (+ Reddish)	0	3	236	635	1	**875**	**1056**
X	0	2	30	285	2	**319**	**-**
Z	0	1	9	8	0	**18**	**-**
?	2	11	12	6	2	**33**	**6**
Total						1775	1775

Table 5.4. 23 Black star taxa are currently accepted as present in SW Ghana. In 1996, there were also 23 Black Star species known from these reserves. 14 of those species considered Black Star in 1996 have been downgraded, and 1 has been sunk (all downgraded species are shown in grey at the bottom of the table). 15 new Black Star species have been added to the list: 10 by new publications, 1 by new rating, and 4 by new records in the region ('Notes'). 8 Black star species are stable: recorded as BK Star species of the reserves in 1996, and still BK Star species of the reserves in 2017.

Species	Star 2017	Star 1996	Notes
Calycobolus hallianus	BK		New species (previously *Calycobolus* sp. A with no Star)
Chytranthus verecundus	BK	BK	Stable
Cola umbratilis	BK	BK	Stable
Dorstenia embergeri	BK	?	New rating
Ficus pachyneura	BK		New record 2015
Gaertnera luteocarpa subsp. luteocarpa	bk		New species
Hymenostegia gracilipes	BK	BK	Stable
Leucomphalos discolor	BK	BK	Stable
Momordica sylvatica	BK		New species
Monocyclanthus vignei	BK	BK	Stable
Pavetta abujuamii	BK		New species
Pavetta ankasensis	BK		New species
Pavetta sonjae	BK		New species
Pavetta subglabra	BK		New record 1998
Psychotria ankasensis	BK	BK	Stable
Psychotria hawthornei	BK		New species
Psychotria longituba	BK	BK	Stable
Psychotria nigrostellata	BK		New species
Schefflerodendron ekpei	BK		New species
Suregada ivorensis	BK		New record 2015
Synsepalum ntimii	BK		New species
Tapura ivorensis	BK	BK	Stable
Tricalysia parva	BK		New record 2015
Chrysophyllum azaguieanum	GD	BK	Downgraded
Dactyladenia hirsuta	GD	BK	Downgraded
Dasylepis blackii	BU	BK	Downgraded (was *D. assinensis*)
Leptoderris cyclocarpa	GD	BK	Downgraded
Leptoderris miegei	GD	BK	Downgraded
Neolemonniera clitandrifolia	BU	BK	Downgraded
Nephthytis swainei	GD	BK	Downgraded
Pierreodendron kerstingii	GD	BK	Downgraded
Placodiscus bancoensis	GD	BK	Downgraded
Psychotria brachyanthoides	BU	BK	Downgraded
Psychotria subglabra	GD	BK	Downgraded
Sclerosperma mannii	BU	BK	Downgraded
Shirakiopsis aubrevillei	BU	BK	Downgraded
Synsepalum aubrevillei	GD	BK	Downgraded
Trichoscypha chevalieri	-	BK	Sunk to *T. lucens*

Gradients in bioquality

In SW Ghana, GHI scores are higher in the south than in the north, although not all southern samples are hot (Figure 5.1). Annual mean rainfall increases to the south, and annual mean temperature decreases.

The position of samples in NMDS ordination space is related to latitude, i.e. climatic changes over the latitudinal range of the samples. Samples from northern reserves adopt higher axis 2 positions than samples from further south (Figure 5.2). Low lying samples from all reserves have higher axis 1 scores than samples from higher positions in the local landscape. Axis 2 represents the latitudinal rainfall/temperature gradient, while axis 1 represents local water availability: Swamp species have higher median axis 1 scores than do non-pioneer species. Two parallel gradients of species turnover within SW Ghana can be detected: one for hill top and flat land from Tano Nimri south to Ankasa, and one for low lying samples from Tano Nimri to Ankasa. GHI increases with lower axis 2 scores (towards the south), and with lower axis 1 scores (towards higher, less disturbed landscape positions). GHI is correlated with both axes, suggesting that both latitudinal and local variation is important to account for, in order to make realistic GHI predictions at the sample scale.

We can formalise these relationships by modelling the relationship between GHI and latitude, climatic variables, disturbance, and local landscape position. A linear model is fitted to the dataset from the variables latitude, mean annual rainfall, landscape position, and Pioneer Index (LAT, BIO12, LAND, PI). The adjusted R squared for the model is 59%, and the model has the equation:

$GHI_{hilltop}$ = exp (-1.2625989 + 0.0036085*BIO12 + -0.1499967*PI + 0.0267400*LAT*PI)

$GHI_{non\text{-}hill}$ = exp (-1.2625989 + 0.0036085*BIO12 + -0.1499967*PI + 0.0267400*LAT*PI - 0.0892025)

The highest GHI scores (>300) are associated with relatively undisturbed samples in the south (Figure 5.3). The interaction between LAT and PI means that the slope of PI is different depending on

LAT, and vice versa. When PI is low (samples are undisturbed), GHIs are around 125 points higher in the south (GHI c. 275) than for samples in the north (GHI c. 150), and the slope of latitude on GHI is strongly negative (Fig. 5.3A, solid line). When PI is high (samples are disturbed), the slope of latitude on GHI is almost flat, meaning that disturbed samples always have the same (low) GHI, regardless of their latitude (Fig. 5.3A, dashed line). GHI decreases linearly with increasing disturbance (PI), but much more steeply for samples in the south compared with samples in the north (Fig. 5.3B). For the less disturbed samples, GHIs are much higher in the south than in the north. In the more disturbed samples, the north and the south converge on generally low GHIs. In the south, when disturbance is low, being on a hill top adds 25 points to the GHI. Mean rainfall (Bio12) is positively correlated with both GHI, and negatively with latitude.

Figure 5.1. Sample locations of 114 samples in SW Ghana, coloured by GHI. Reserves are named.

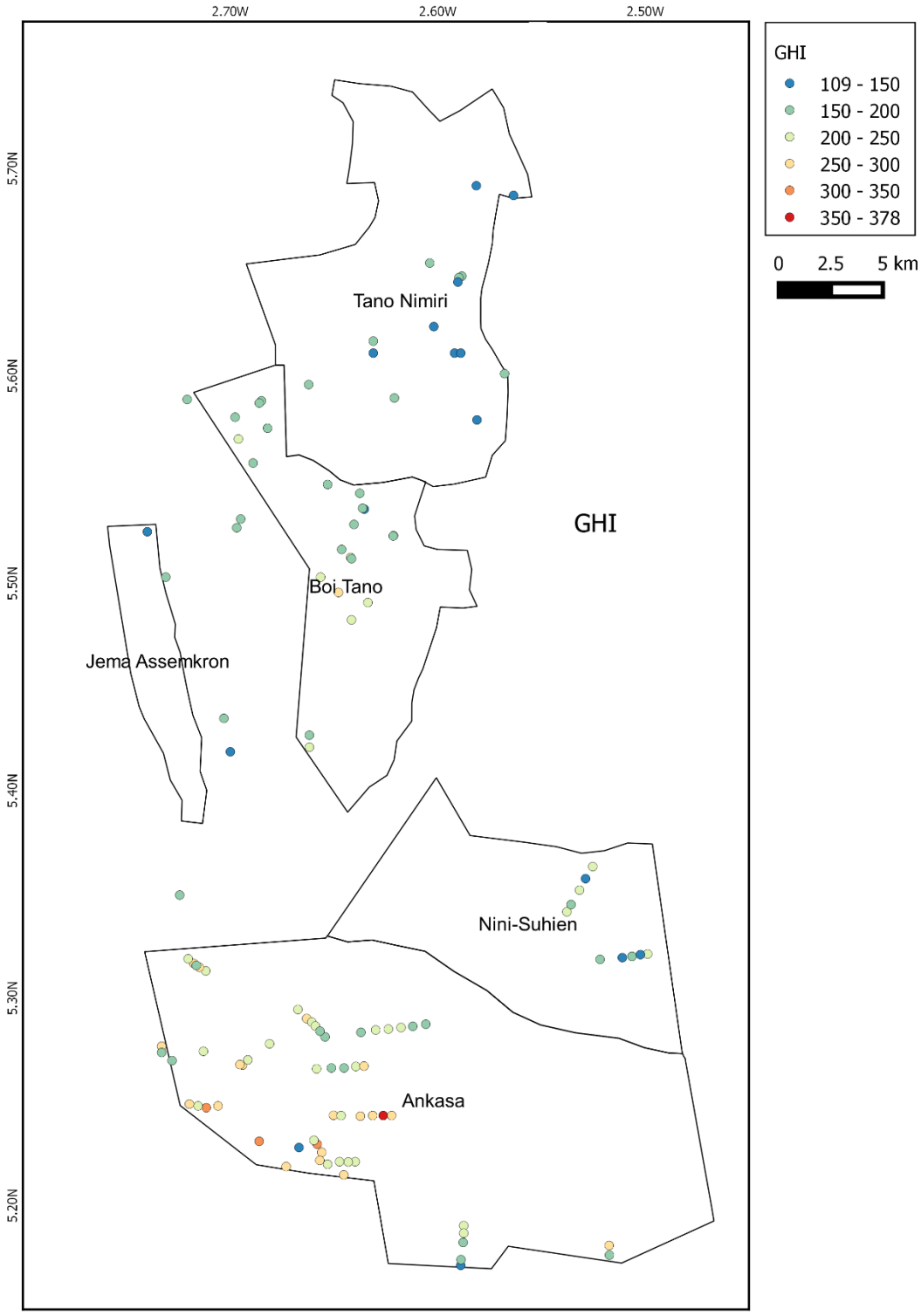


Figure 5.2. NMDS ordination of 114 samples of SW Ghana. Samples are coloured by reserve and sized by landscape position. Disturbed samples shown as squares. Significant environmental vectors are fitted (LAT = Latitude, Bio12 = annual mean rainfall).

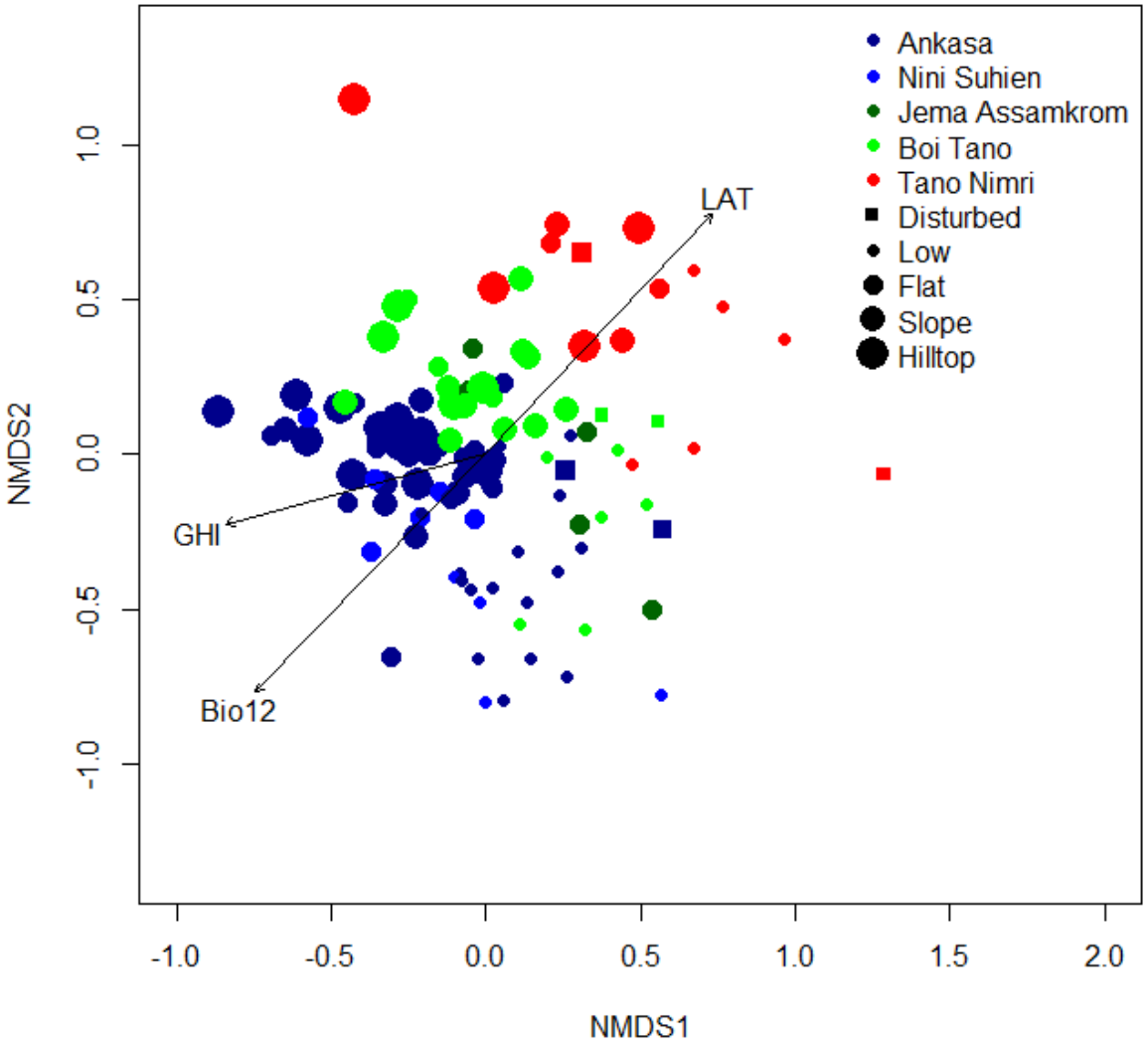
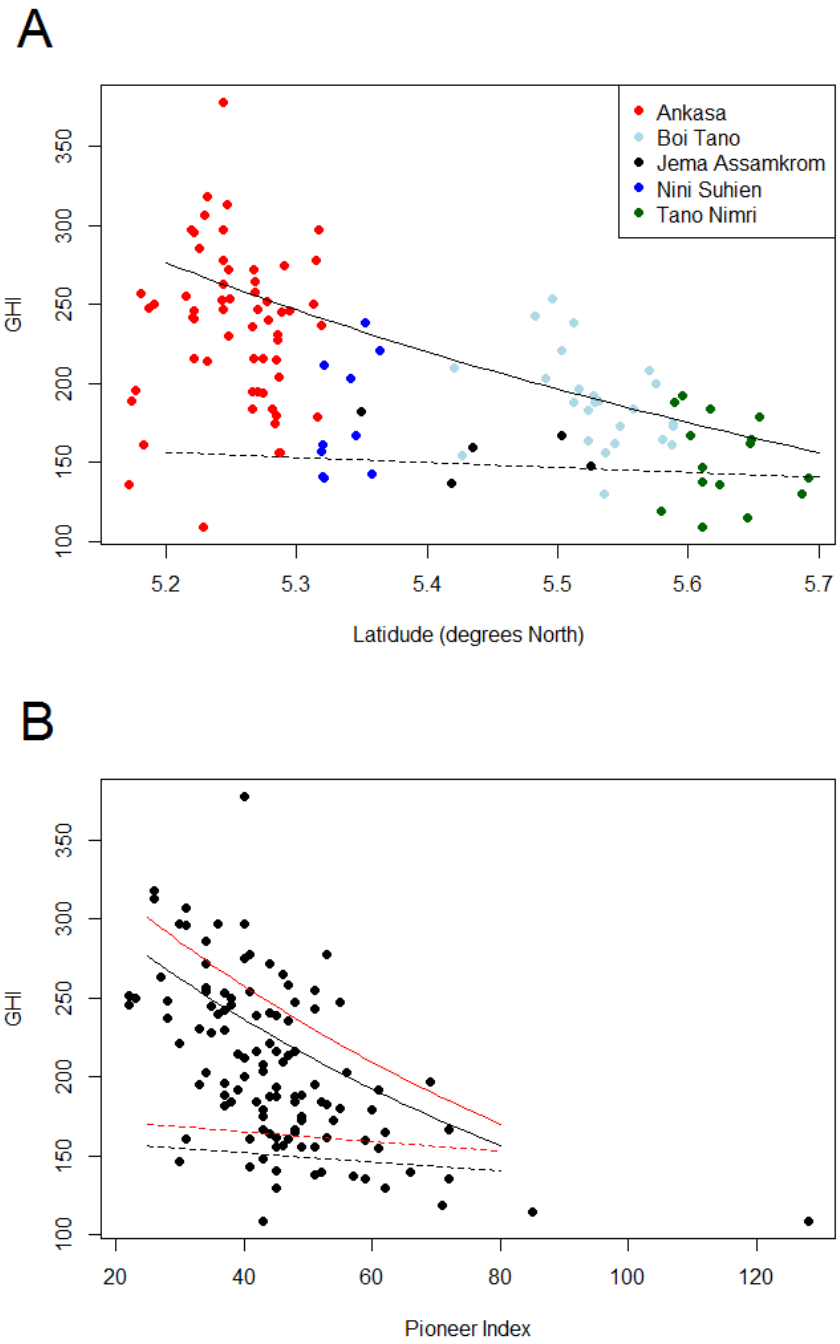


Figure 5.3. Linear model for GHI in SW Ghana. GHI = exp (-1.2625989 + 0.0036085*BIO12 + -0.1499967*PI + 0.0267400*LAT*PI (- 0.0892025 for non-hill)), Adj. R2 = 59%, n = 114, all terms are significant (p<0.05). (A) Samples are coloured by Reserve. The dotted line shows model fit with high PI (PI=80), the solid line at low PI (PI=25); Bio12 (annual mean rainfall) is set to its mean value in the dataset (1910 mm), LAND is set to Non-hill. (B) Model fit for high latitude (Lat=5.7, dotted lines) and low latitude (Lat=5.2, solid lines); hill top samples are shown in red, non-hilltop samples in black. Bio12 is set to its mean value (1910 mm).



The major feature of the Nimba survey area is the Nimba Mountains range, and compared with the SWGhana dataset the altitudinal range is dramatic (400-1660 masl, cf. 85-115 masl). The highest GHI scores (>300) are found at high altitude in Guinea, where some GHIs exceed 600 (Figure 5.4). Disturbance decreases GHI, and altitude increases it (Figure 5.5). The model has the formula GHI= exp(5.8 + -0.000871*ALT + -0.0104*PI + 0.000000818*ALT^2); adjusted R^2 is 57%. Although the application of the PI, which was designed for the Ghana forest zone, is open to some caution in the context of savanna in Liberia and Guinea, these habitats do yield a high PI and can be considered highly disturbed, due to their open nature and high frequency of fires. Disturbance decreases GHI scores linearly: holding Altitude constant, increasing the PI by 10 points decreases GHI by 10 points. GHI increases non-linearly with altitude: GHI increases more for every 100 m ascent at higher altitudes than at lower altitudes, particularly over c. 1000 m. High GHIs (over 300) are only found in areas > 1000 masl, and in samples where the Pioneer Index is less than 100. Areas above 1000 m can also have low GHIs, particularly if they are disturbed. Altitude increases GHI, but only in relatively undisturbed vegetation.

There is some circularity when the Pioneer Index is used as an explanatory variable in a model of GHI, because it is an index derived from the same sample species data, and more importantly, because Pioneer and NPLD species as guilds have lower GHI than shade bearers (on average, shade-bearers are globally rarer than pioneers and NPLDs). In Nimba, GHI(pioneers)=81, GHI(NPLD)=107, GHI(Shade-bearer)=174. However, an independent check on the relationship between disturbance history and GHI can be made for Nimba. For the Nimba dataset, we recorded the age of the vegetation by asking local residents when the land was last cleared. The model of GHI ~ AGEYEARS tells us that 1 year of age adds 1.2 to the GHI of a sample, between 1 and 44 years of age (the range of our data); n=63 (Figure 5.6).

Figure 5.4. Sample locations of 454 samples of the Nimba Mountains, coloured by GHI. Contour lines are spaced at 50 m. The thicker black line indicates rail (and road) corridor from Yekepa south westwards.

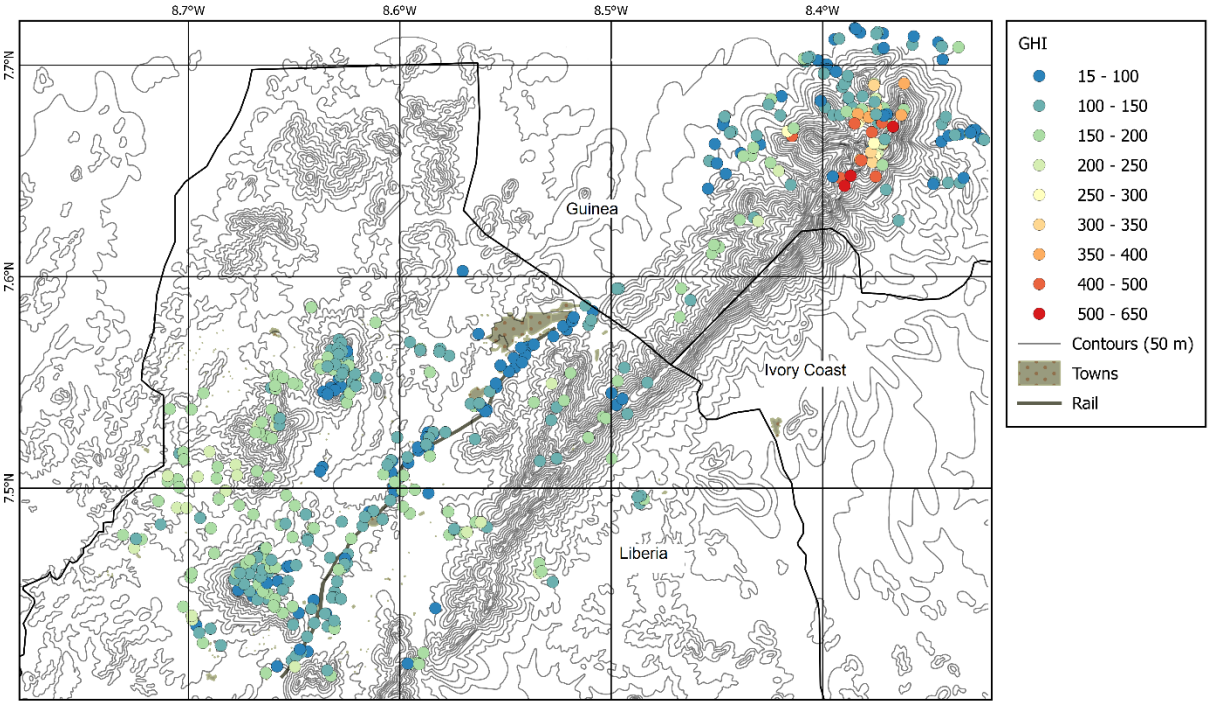
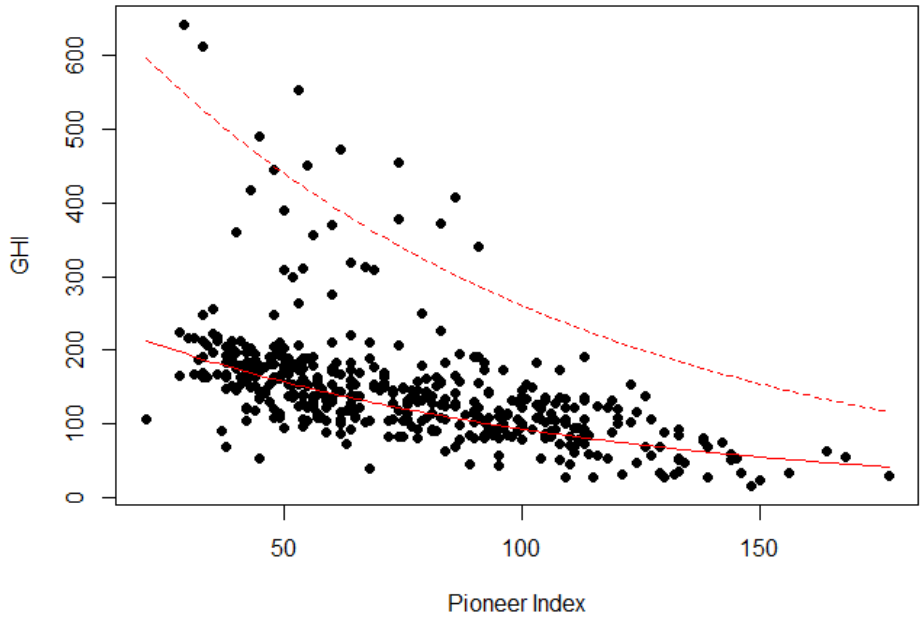


Figure 5.5. Linear model for GHI in Nimba. GHI= exp(5.8 + -0.000871*ALT + -0.0104*PI + 0.000000818*ALT2); Adj. R2 is 57%, n=454, all terms are significant (p<0.05). (A) Dotted line shows model fit at high altitude (1660 m), solid line shows model fit at low altitude (420 m). (B) Dotted line shows high Pioneer Index (PI=177), solid line shows low Pioneer Index (PI=22).



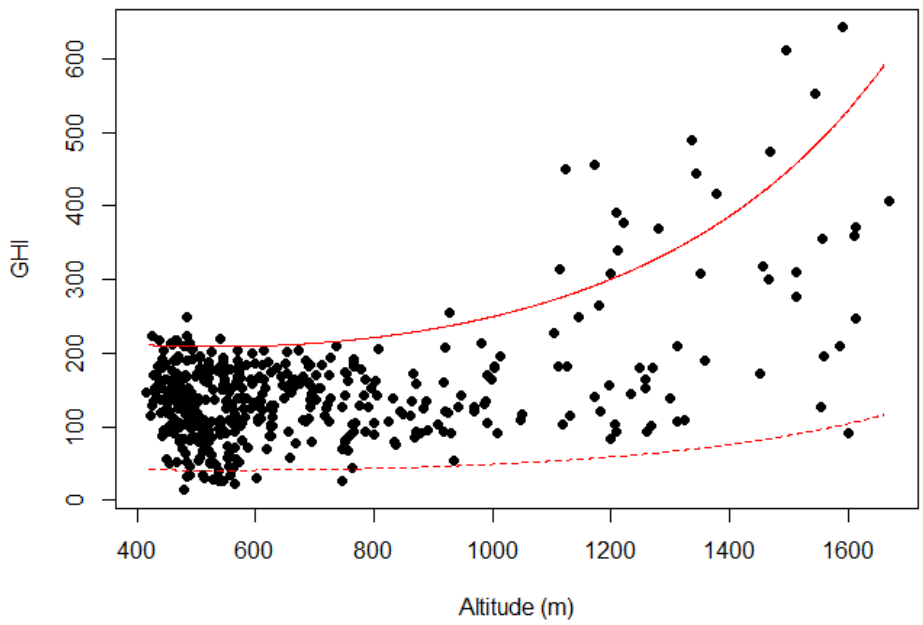
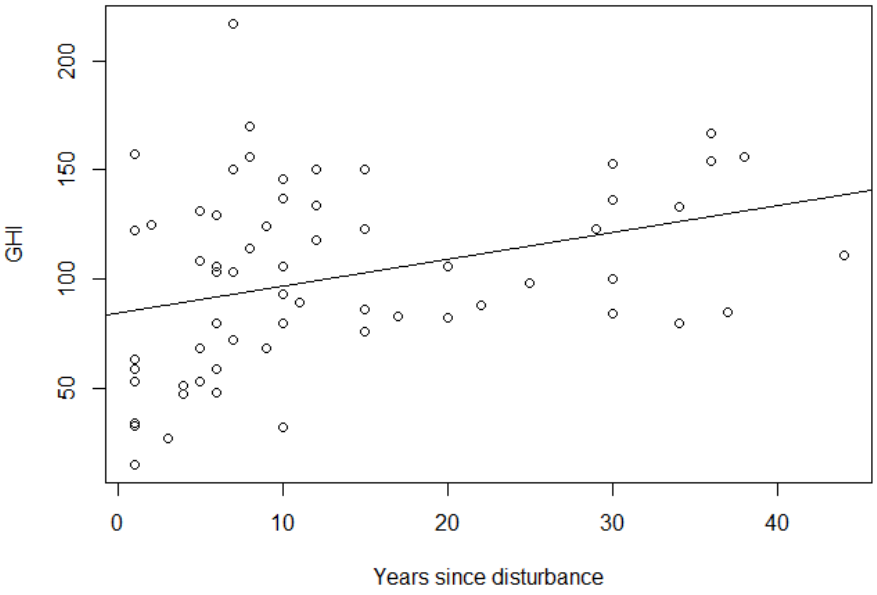


Figure 5.6. Linear model of GHI by number of years since disturbance, for 63 samples of the Nimba dataset. GHI = 84 + 1.24 × AGE, p=0.006, R2 = 10%.



Discussion

The hotspot of SW Ghana is confined to relatively undisturbed samples, decays linearly to the north, and peaks locally on hill tops. Boi Tano FR has GHI scores intermediate between Tano Nimri FR and Ankasa RR, and is home to at least one globally rare Black Star species not yet found in Ankasa (*Momordica sylvatica* Jongkind). The strong influence of disturbance on GHI scores in the south is sobering, because the great majority of samples in this dataset are relatively undisturbed (PI<70) compared with land outside the forest reserves. The theoretical maximum for PI is 200, where every species is a pioneer. The vast majority of the samples in this dataset are from closed or slightly open forest inside forest reserves, and even within these vegetation types the signature of disturbance on GHI can be observed. Disturbance in these forest reserves is largely due to logging (except in mostly unlogged Ankasa), flowing water in the wet season, natural tree falls and (in Ankasa) elephant activity, and moderate human activity such as trails. Undisturbed forests further north already have much lower GHIs, and here disturbance has little impact on the GHI.

The influence of disturbance on SW Ghana's bioquality is mirrored in the Nimba survey area, where the hotspot of the Nimba Mountains is confined to relatively undisturbed samples over 1000 m. In Nimba, GHI decreases more steeply at high altitude, where GHIs are otherwise high, than at lower altitudes, where GHI is generally lower. Disturbance in the Nimba survey, which incorporates more samples from non-forest than the SW Ghana surveys, is the result of farming and iron ore mining operations, localised fires, as well as natural tree falls: the range of PIs (c. 0-150) is wider than in the SW Ghana dataset. The Pioneer Index is the only significant term retained in both the SW Ghana and the Nimba models, and indicates a degree of generalisability of the impact of disturbance on the distribution of globally rare species. GHI increases with the number of years since vegetation was last cleared (and the Pioneer Index decreases), giving an indication of the recovery potential of the vegetation of hotspots if left undisturbed.

Maximum altitude was significantly positively correlated with species endemism at broad scales across Africa, but explained only 4% of the variation observed (Linder et al. 2001). Wieringa & Poorter (2004) found no correlation between species endemism and altitude in Upper Guinea. Topographic complexity, rather than altitude, has been linked to high biological diversity (Plana 2004, Barthlott et al. 2007). Topodiversity is relatively low around SW Ghana, but higher for the Nimba Mountains (Mutke et al. 2001). Our results show that altitude is an important predictor of bioquality in the Nimba Mountains. Even in the SW Ghana dataset, where the altitudinal range is small, local landscape position (hill tops vs slopes, flat land or swamps/riversides) has an important influence on the GHI, probably mediated by disturbance.

Environmental variables (rainfall, soil water holding capacity, altitude, soil fertility, soil pH and cation exchange capacity) were all significantly correlated with endemic richness in Upper Guinea, and rainfall in particular explained a large amount of the variation in rare species richness, with the most drought intolerant species having the smallest distributions (Wieringa & Poorter 2004). In SW Ghana, our results confirm the importance of annual rainfall for predicting the occurrence of globally rare species, with higher rainfall associated with higher bioquality locally.

Brenan (1978) cited Ghana as having 43 species endemic to the country, while Hall & Swaine (1981) cited 18; Brenan (1978) cited 59 Liberian endemic species, while Jongkind (2012) made an estimate of 50. Our results show that there has been considerable flux in our understanding of both the taxonomic identities and geographic distribution of plants found in SW Ghana since 1996. For instance, local endemics *Uvariopsis globiflora* & *U. guineensis* of Hall & Swaine (1981) were sunk by Hawthorne & Jongkind (2006), and the Black Star species *Trichoscypha chevalieri* of SW Ghana has been sunk. Our results show a trend towards downgrading of species once considered Black Star with time. However, this has been entirely offset by the equal number of new Black Star species described or recorded from the area over the same time period, with the consequence that the number of Black Star species, and indeed the GHI scores and patterns within the area, have

remained remarkably stable. This balancing of the trends between the downgrading of species through the Star ranks, and adding new species at the top, makes hotspot scores much more stable than our understanding of species' ranges themselves, and allows for some consistency in results and conservation priority setting over time and in the face of partial information.

40% of species assigned to a Star in 1996 have changed Star since, in light of new collections and digitisation of herbarium records that were unavailable for analysis in 1996. This high level of flux means it would be wise to consider our current taxonomic framework and interpretation of species' ranges to be a hypothesis, pending new information. 60% of species have been stable over 20 years, and we predict that this proportion will increase with time, although we may still be many years from the asymptote.

Our shifting conception of taxa and their ranges, and the possibility of overlooking species even within these relatively well studied reserves, makes it hard to measure 'true' changes in the conservation value of a reserve, for example if we wanted to know whether or not a particular reserve had been effective at maintaining high endemism. At least we can say that GHI scores have not decreased, on average, over 20 years. Even if GHIs generally do decline due to apparent range increases in the coming decades, the ranked scores are expected to be more stable. From the point of view of the conservation of globally rare plants in SW Ghana, it is important to keep additional disturbance in the south, particularly Ankasa, to a minimum, as is the current practice. Boi Tano, which is actively logged, has GHI scores high enough to be brought down by disturbance, and the importance of leaving enough time for forest to regenerate between felling cycles, and for maintaining the integrity of the unloggable Globally Significant Biodiversity Areas, is highlighted by our results. In Boi Tano we found *Dalbergia* sp. 'trifoliate' and *Keetia tenuiflora* 'type b', both likely to be new, as-yet undescribed, species, as well as 8 Black Star species, including *Momordica sylvatica,* as yet not found outside of Boi Tano. We found seven Black star species in Tano Nimri, including *Ficus pachyneura* on the Tano Nimri/Boi Tano border. This speaks to the high conservation

value of Boi Tano forest reserve, but with the exception of *F. pachyneura*, *M. sylvatica* and the trifoliate *Dalbergia*, all these species are also found in Ankasa.

Both Nimba and SW Ghana have been suggested as climatic refuges over paleo timescales. Paleo-climate modelling, and by extension paleo-vegetation modelling, is subject to great uncertainty at geographic resolutions finer than 0.5 degree square resolution (Hardy 2013, Lima-Ribeiro et al. 2015, Varela et al. 2015), so it has not been possible to test explicitly the importance of e.g. Pleistocene annual rainfall on modern bioquality at the sample level. However, we can explain 57% and 59% of the variation in GHI scores within the Nimba and SW Ghana hotspots without invoking past climate changes. The high rainfall regime in the far south of Ghana is geographically isolated from other high rainfall areas. This means that species adapted to these modern day, very high rainfall, conditions in the SW of Ghana must disperse a long way, over unsuitable intervening habitat, if they are to be globally common (or at least relatively widespread) today. This modern day climatic isolation with respect to rainfall, must be at least in part be responsible for maintaining the higher concentration of range-restricted species in SW Ghana. The Nimba high altitude savanna is very different floristically to the lower altitude forests. Afromontane savannas are not at all common in the region, and this geographic and climatic isolation must be at least partly responsible for restricting the range of the species found there.

In both datasets, accounting for disturbance has a marked effect on the GHI, at least within areas which would otherwise be hot, and for the SW Ghana dataset, local landscape position is retained as significant variable. Such local influences have a high impact on bioquality and are lost to broadscale analysis of herbarium data. The generality of the patterns presented here may be determined by analysing RBS sample datasets from other Upper Guinean hotspots, particularly in SE Liberia, where we might expect the relationships to hold, in trend if not in absolute terms. In colder and drier areas there is less scope for local variation in bioquality.

Supporting Information

Appendix S5.1

Sample header data for SW Ghana cells

Appendix S5.2

Sample header data for Nimba cells

Chapter 6

Modelling GHI across Africa

Article status

In prep.

Abstract

Aim: To predict bioquality patterns within tropical Africa at 0.5 degree square resolution, in order to fill the evident gaps across the continent at this resolution. The model is used to extrapolate GHI scores across northern and southern Africa, generating testable predictions from the model. The relative contributions of the many variables posited as important for predicting high endemism are explored.

Location: Mainland tropical Africa, mainland Africa.

Taxon: c. 40,000 vascular plant taxa (species and infraspecific taxa)

Methods: Conditional forest inference trees.

Results: Our tropical GHI model can predict known GHI values with an error of around 100 GHI points. Transferability tests suggest that the model can predict GHI scores of Ghanaian cells with acceptable accuracy, when no training data are given for Ghana. Predicted bioquality across tropical Africa and mainland Africa approximate previously established patterns (Chapter 4). All included terms held some predictive power over GHI.

Main conclusions: The biogeographic frameworks of White (1983) and Sayre et al. (2013) were the most important predictors of GHI across tropical Africa. Climatically, maximum temperature of the warmest month was the most important variable, at least in the IPSL model. The stability of climate parameters over the last 21,000 years held predictive power for GHI, and in particular the stability of annual rainfall. Surficial lithology, longitude and latitude were amongst the top predictors; altitude and slope were of middling importance. Predicting bioquality patterns for areas where distribution data is lacking allows inferences to be drawn before species collection is sufficiently robust.

Introduction

In this Chapter bioquality is modelled at 0.5 degree square resolution, across tropical Africa. Modelling bioquality using environmental and geophysical variables is desirable for three reasons. Firstly, the model can be used to fill in data gaps across tropical Africa at half degree square resolution, allowing GHI to be mapped for areas lacking species distribution data (c. 60% of tropical Africa at 0.5 degree square resolution). Secondly, it can be used to extrapolate from tropical Africa to northern and southern Africa, although this is subject to greater uncertainty. And thirdly, the model can be used to identify particular variables which are important for predicting high global endemism. Understanding the relationship between environmental and geophysical factors and high endemism is of fundamental research interest, for example in understanding the response of global endemism to climate change.

Choice of explanatory variables

Explanatory variables used in the models are summarised in Table 6.1. The rationale for the inclusion of each follows in this section.

Latitude, longitude and area

Latitude was significantly correlated with corrected weighted endemism for sub-Saharan Africa and explained 15% of the variation (Linder 2001). A Spearman's rank correlation revealed significant correlation between richness of rare species and latitude within Upper Guinea, although the range of latitudes in Upper Guinea is rather small and this result could have more to do with high rainfall or putative refugia on the coast (Wieringa & Poorter 2004). The latitudinal gradient in species richness is a global phenomenon (Brown 1988, Rosenzweig 1995, Kaufman & Willig 1998) although the underlying causes and mechanisms are heavily debated (Rohde 1992, Hillebrand 2004, Barthlott et al. 2007), and the relevance of this to patterns of endemism are unclear. Longitude does not

correlate with species richness or endemism in Africa in the southern hemisphere (Linder 2001) or in Upper Guinea (Wieringa & Poorter 2004). As a weighted average, GHI scores should be relatively impervious to absolute numbers of species, which are related to area. However, as the size of sampled area increases, more Black Star species are encountered whilst Green Star species tend to stay the same, so GHI scores probably do inflate somewhat as the sampling area increases.

Altitude, topography and slope

Hotspots of endemism within Africa have been found to correlate with high altitude. Maximum altitude was found to be significantly positively correlated with richness and endemism across sub-Saharan Africa, but explained only 4% of the variation (Linder 2001). Wieringa & Poorter (2004) reported high endemic richness within Upper Guinea on Mt Nimba, Mt. Ziama and Mt. Tonkoui, but no significant relationship was found between altitude and endemic richness after high orographic rainfall associated with high elevations was accounted for. Lovett et al. (1998) report locally endemic species from across the altitudinal range in E. Africa, but with a peak in numbers at submontane elevations, and noted the smaller area (thus high concentrations of locally endemic plants) at montane elevations. Doubtless, African mountains do harbour globally rare species, isolated and adapted to high altitude conditions and daily low temperatures. Topographic complexity, rather than altitude *per se*, has been linked to high species richness (Mutke et al. 2001, Barthlott et al. 2007), and may be related to high endemic richness.

Lithology and soil

Soil water holding capacity, soil fertility, soil pH and cation exchange capacity were all significantly correlated with endemic richness in Upper Guinea (Wieringa & Poorter 2004). Soil type, and lithology, influences which species can grow in a place, for example the mineral-rich bed rock found in parts of Katanga (e.g. Meerts 2016, an example of soil based vegetation classification).

Biogeography

White (1983) divided the surface of the African continent into 18 regions (excluding Madagascar), azonal vegetation (particularly mangrove), and water bodies (e.g. Lake Victoria). As each region, or chorion, shares a flora, we expect chorion to be predictive of GHI scores. Although some of the choria (regional centres of endemism, except for Sudanian) are defined by having >1000 endemic species, and > 50% endemism, others, such as the transition zones, have much lower percentage endemism. For example, White estimated the endemism of the Somali-Masai RCE at 50%, the GC RCE at 80%, the Sahel at 3%, and the Sahara at 12%, and we might expect these differences to result in quite different GHI scores between the choria. The biogeographic framework (Chapter 3) created for tropical Africa from this same dataset may better predict GHIs than one created from a different dataset (White 1983, Sayre et al. 2013). However, our own framework is constrained to tropical African regions, meaning we do not have predictions for regions in northern or southern Africa, and in any case, it is similar to White's (1983).

The terrestrial ecosystems and vegetation of Africa were classified and mapped as part of a larger effort and global protocol (GEOSS – the Global Earth Observation System of Systems) to map the terrestrial ecosystems of the earth in a standardized, robust, and practical manner (Sayer et al. 2013). The Sayre et al. vegetation classification is hierarchical, and was developed by a panel of African ecosystem scientists and vegetation geographers. The finest vegetation units in the classification are the 126 macrogroups, defined using diagnostic plant species and growth form, and each occupying an area of 100-10,000s hectares. Macrogroup was predicted for the continent at 90 m resolution using a CART method, predicting from a training set comprised of a series of sampling points of known vegetation class. This regionalisation is much more constrained by geographic factors, and less by species distributions, than the other two systems, because of the predictive variables used in the analysis, such as lithology and climate types.

Modern climate

Factors like mean annual precipitation, evapotranspiration or the length of the dry season show a close relationship with species richness in the tropics (Barthlott et al. 2007, Francis & Currie 2003, Mutke & Barthlott 2005, Mutke et al. 2001). Rainfall was significantly correlated with endemic richness in Upper Guinea, and the most drought intolerant species had the smallest distributions (Wieringa & Poorter 2004). Within south west Ghana, higher bioquality scores were associated with increasing rainfall (Chapter 5). In contrast, Linder et al. 2001 found that rainfall explained 43% of variation in species richness across Africa, but less than 1% of variation in corrected endemism.

Paleo-climate

Local endemism in tropical Africa has been explained by historical variations in the climate (Hamilton, 1976; Brenan, 1978; Lovett et al., 1996; Fjeldså et al., 1997, Linder et al. 2001), rather than by the modern climate. Lonnberg (1929) was the first to suggest that lowland rain forest refuges could have occurred as the result of drying and cooling conditions, particularly during the Pleistocene (c. 2.6 Ma – 12 Ka). Lowland rain forest is proposed to have shrunk to small enclaves where climatic conditions remained suitable, such as along rivers, at higher elevations, or in particular lowland areas with favourable edaphic conditions. The high endemism of refugia has been attributed to high speciation, and/or low extinction. The refugial speciation model argued that climate change led to fragmentation of species with isolation of conspecific populations in refuges. The isolates became genetically differentiated and eventually speciated (Hamilton 1976, Prance 1982). An alternative model suggested that refuges act as 'museums', accumulating species slowly over time due to climatic stability and low extinction rates (Stebbins 1974, Lovett & Friis 1996), predicting a gradual accumulation of species of all ages. Here, we include climate stability in our bioquality modelling, which tests the importance of both stability and instability of climate over the last 21,000 years for predicting bioquality.

Climate layers, both modern and paleo, are created by interpolating data from c. 50,000 weather stations worldwide, incorporating latitude, longitude and altitude using splines (e.g. WorldClim Hijmans et al. 2005; see New et al. 2002, Kriticos 2012 for other efforts), and for paleoclimates, by incorporating c. 100 fossil pollen records from the Last Glacial Maximum (LGM, Bartlein et al. 2011; Harrison et al. 2014) and mid-Pliocene (Dowsett et al. 2012). The geographic and taxonomic resolution of palynological cores remains limited. Climate and fossil observations are modelled using simulations of atmospheric and oceanic circulations: a general circulation model (GCM). Climatologists tune GCMs based on boundary conditions such as orbital parameters, solar forcing, greenhouse gas concentrations, CO_2 emissions, land-use, and ice coverage, coupled with atmospheric, vegetation, and ocean dynamics. Climatologists have predicted global climates for past (Miocene, Pliocene, late Quaternary), present (pre- and post-industrial), and future conditions (end of 21st century). The two main GCM working groups are the Coupled Model Inter-comparison Project (CMIP5; Taylor et al., 2012), and the Paleoclimate Modeling Intercomparison Project (PMIP3) (Braconnot et al., 2007), which have reported the results of 9 general circulation models between them (CCSM, CNRM, COSMOS, GISS, FGOALS, IPSL, MIROC, MPI, MRI) at up to 4 time periods (after Lima-Ribeiro et al. 2015). The LGM (21 ka) and mid-Holocene (mHol, 6 ka) represent glacial and interglacial phases related to the last ice age, while the Pliocene (Plio) time period represents the mid-Pliocene warm period (mPWP, ~3.3-3.0 Ma). Modern climates represent average conditions 1950-1999.

The mPWP is known as the 'Golden Age' for Africa, characterised by consistently warm, moist climates, and the expansion and diversification of rain forests, with cores rich in pollen from *Klaineanthus, Uapaca, Iodes, Petersianthus, Pycnanthus, Elaeis* and *Pandanus*. This was followed by very pronounced drying and cooling across the African continent. Glaciation of the northern hemisphere at 2.7 Ma saw a gymnosperm (*Afrocarpus falcatus* type pollen*)* appearing in the west African fossil record for the first time since the Cretaceous, and desert establishing for the first time

in the Sahara 2.6 Ma (Morley 2000). The most recent glacial period occurred 70,000-12,000 years ago with the last glacial maximum (LGM) at c. 21,000 years ago. Glacial periods were associated with cooler, drier conditions than interglacial periods. Interglacial periods were probably characterised by a greater rainforest extent than today, as during the mid-Holocene c. 6000 years ago when forest covered the Dahomey Gap. Lake Bosumtwi was unforested during the last glacial maximum, until 9000 years ago. Lake Barombi Mbo remained forested, as did SW Ghana, during the Pleistocene (Morley 2000).

Materials & Methods

Data processing

A grid at half degree square (0.5 × 0.5 DS) grain originating at the equator/meridian was created and clipped against the mainland of the African countries, resulting in 10,355 cells across mainland Africa, and 8117 cells for tropical Africa. Geographic processing was performed using ArcGIS. The latitude and longitude were calculated as the centroid values for each cell before clipping, and the area of each cell was calculated after clipping.

Altitude and slope data were derived from the GMTED2010 digital elevation model (DEM) at 30 arc second resolution. GMTED2010 replaces GTOPO30 as the elevation dataset of choice for global and continental scale applications. Nine 30-arc-second GMTED2010 tiles for Africa were downloaded from the USGS and were merged into a single layer. A slope raster layer was calculated from the DEM layer using the Slope tool of ArcGIS. Mean and standard deviation of altitude and slope were calculated for each cell at the new half degree square grain and appended to the grid. The land surface form (LS) layer from Sayer et al. 2013 was derived from the 90 m NASA/NGA shuttle radar topographic mission DEM, and data were characterised into seven landform classes following an analysis of slope and local relief (Sayre et al. 2013, Sayre et al. 2009, True 2002). The land surface

raster layer was summarised using the majority and variety of class occurrence within each cell, and appended to the grid.

The surficial lithology raster layer (Sayer et al. 2013) was summarised using the majority and variety of class occurrence within each cell, and appended to the grid. The layer was compiled from twelve geology, soil and lithology datasets (FAO 1998-2008) and data were reclassified into one of nine bedrock types, or one of ten unconsolidated surficial materials (Sayre et al. 2013).

Cells were assigned to one of the 18 major choria, water, mangrove or missing, of White (1983). The major chorion of each cell was determined using the largest overlap addin tool for ArcGIS, the number of choria per cell was calculated using the base spatial join tool; both fields were spatially joined to the grid. Cells were also assigned to the biogeographic framework of Sayre et al. (2013). Majority and variety Macrogroup were calculated for each cell. The higher level classifications class, subclass, formation and division for each cell were looked up from the layer attributes table, relating on majority macrogroup. Macrogroup was split along class boundaries to produce four attributes, where MG1 covers macrogroups from forest to woodland, MG2 from Grassland, savanna and shrubland; MG3 from Desert and semi-desert, and MG6 rock vegetation, to which the water and unknown macrogroups were appended.

Majority Country was calculated for each cell from the African countries file using the largest overlap addin tool in order to facilitate transferability analyses.

BIOCLIM, one of the first species distribution modelling approaches (Busby 1991a, 1991b, after Booth et al. 2014), gave rise to a series of bioclimatic envelope variables, now well established for example in the CliMond (Kriticos 2012) and WorldClim database (Hijmans et al. 2005) and supporting R packages (dismo, Hijmans et al. 2013). The 19 core Bioclim variables (Bio1-19) are a series of variables able to describe the climatic componenent of the niche of species whilst being calculable from a small number of raw monthly input variables: Pr, precipitation flux; tas, mean temperature;

tasmax, maximum temperature; tasmin, minimum temperature. There are others (Bio20-35), calculated from solar radiation and evaporation, but these data are harder to calculate at African extent so were not pursued.

We used the downscaled climate data from the ecoClimate dataset at 0.5 degree square grain (Lima-Ribeiro et al. 2015), using the modern climate as baseline and two available paleo time periods. The ecoClimate downgraining from the native resolution of the GCMs (c. 1-4 degree square) is described in Lima-Ribeiro et al. 2015. The 9 GCMs make substantially different predictions of modern climate, and paleo climates. Varela et al. 2015 identified groups of models with the most similar predictions, and recommended that one model from each group should be used to test the robustness of any conclusions drawn from environmental data: COSMOS (the model with the most discrepent predictions); one model from the group MPI and MRI; one model from the group CCSM3, FGOALS and MIROC; and the last model from the group GISS, CNRM and IPSL. The authors also recommended excluding Bio14 and Bio15, as these variables show the highest discrepancy between GCMs, but as we use the predictions of only one GCM at a time, we have not excluded these variables, but are happy to accept that these variables are especially uncertain.

Mean values for the variables Bio1-Bio19, estimated under the IPSL and MRI general circulation models, estimated at Modern, mHol and LGM time periods, were extracted from the EcoClimate database and were related to the grid using the centroid lat/long values. Paleo-climate estimates were taken using the Modern climate (1950-1990) as a reference point. We estimated the models presented here using the IPSL dataset, and repeated all analyses using the MRI dataset (Appendix 2).

Climate stability variables were calculated by us in R for each Bioclim variable, separately for each GCM. A stability measure was calculated as the sum of the absolute difference between modern & mHol, and mHol and LGM for each variable; the min value was subtracted from these values and divided by the range so that each cell has a normalised measure of stability, with 0 representing max stability and 1 the greatest change.

The two climate-derived raster files from Sayre et al. (2013): thermotypes & ombrotype (TO) and topo-moisture potential (TM) were summarised as majority and variety for each cell. TM holds little information when upgrained to half degree square resolution (binary wet or dry at 90 m resolution), but was maintained for completeness.

Not all cells have data from all sources, due to lack of species distribution data in the case of response variables, and discrepancies in coastline boundaries between the other data sources. Cells missing particular datasets were tracked using an ID number for each data source, with an ID of 0 indicating missing data. Cells missing explanatory variables were excluded from the data set during analysis. Putative explanatory variables used in the models are summarised in Table 6.1.

The two response variables, GHI and minGHI, were calculated from GeoPhila and linked to the grid table via centroid lat/long. Only cells estimated to have at least 2.5% of their total species richness sampled were included in the training set (3426 cells). Islands were excluded as the input GHI data are not representative for them, and there is a wealth of evidence to suggest that endemism scores are not comparable for islands vs. contiguous mainland (e.g. Kier et al. 2009).

The grain of the layers (90 m) from Sayre et al. (2013) and DEM data layer (30 arc second) is much finer than can be representatively produced for the GHI (response) layers, or indeed for the EcoClimate data (0.5 degree square). At a finer grain than half degree square, most cells would lack reliable GHI scores, as most cells would lack botanic records, so the distribution data limit the resolution at which the GHI models can be estimated. Technically, any grain could have been chosen for prediction (e.g. by using the 1 km resolution WorldClim database); however, the finer the grain, the more intensive computation is, and the greater the uncertainty – if the model is not given information about the relationship between GHI and these geospatial variables at finer resolutions, then we are assuming that the same relationships from half degree squares will hold at finer scales. We therefore chose not to push the model beyond the resolution of the GHI input layers, half

degree square. Half degree square resolution is sufficient to reveal the broad biogeographic variation across the continent, as reported largely by herbarium databases; and at fine scales very local factors, particularly disturbance, become important anyway (Chapter 5).

Summary of variables

Table 6.1. Summary of environmental and geophysical variables used in the GHI models.

Variable	Description	Unit	Source	Summary attributes
BIO1	Annual Mean Temperature	°C	EcoClimate, 0.5 D.S. resolution	Mean
BIO2	Mean Diurnal Range (Mean of monthly (max temp - min temp))	°C	EcoClimate, 0.5 D.S. resolution	Mean
BIO3	Isothermality % (BIO2/BIO7) (× 100)	°C	EcoClimate, 0.5 D.S. resolution	Mean
BIO4	Temperature Seasonality (standard deviation × 100)	°C	EcoClimate, 0.5 D.S. resolution	Mean
BIO5	Max Temperature of Warmest Month	°C	EcoClimate, 0.5 D.S. resolution	Mean
BIO6	Min Temperature of Coldest Month	°C	EcoClimate, 0.5 D.S. resolution	Mean
BIO7	Temperature Annual Range (BIO5-BIO6)	°C	EcoClimate, 0.5 D.S. resolution	Mean
BIO8	Mean Temperature of Wettest Quarter	°C	EcoClimate, 0.5 D.S. resolution	Mean
BIO9	Mean Temperature of Driest Quarter	°C	EcoClimate, 0.5 D.S. resolution	Mean
BIO10	Mean Temperature of Warmest Quarter	°C	EcoClimate, 0.5 D.S. resolution	Mean
BIO11	Mean Temperature of Coldest Quarter	°C	EcoClimate, 0.5 D.S. resolution	Mean
BIO12	Annual Precipitation	mm/m^2	EcoClimate, 0.5 D.S. resolution	Mean
BIO13	Precipitation of Wettest Month	mm/m^2	EcoClimate, 0.5 D.S. resolution	Mean
BIO14	Precipitation of Driest Month	mm/m^2	EcoClimate, 0.5 D.S. resolution	Mean
BIO15	Precipitation Seasonality (Coefficient of Variation)	cm	EcoClimate, 0.5 D.S. resolution	Mean
BIO16	Precipitation of Wettest Quarter	mm/m^2	EcoClimate, 0.5 D.S. resolution	Mean
BIO17	Precipitation of Driest Quarter	mm/m^2	EcoClimate, 0.5 D.S. resolution	Mean
BIO18	Precipitation of Warmest Quarter	mm/m^2	EcoClimate, 0.5 D.S. resolution	Mean
BIO19	Precipitation of Coldest Quarter	mm/m^2	EcoClimate, 0.5 D.S. resolution	Mean
Stability 1-19	Sum of the absolute difference between	Normalised	This publication	Value

Variable	Description	Unit	Source	Summary attributes
	modern & mHol, and mHol and LGM for each variable; min value subtracted and divided by the range			
LS	Land surface form	7 categories	Sayre et al. 2013	Variety, majority
LT	Lithology	20 categories	Sayre et al. 2013	Variety, majority
MG6	Vegetation macrogroups of class 6	4 categories	Sayre et al. 2013	Variety, majority
MG3	Vegetation macrogroups of class 3	28 categories	Sayre et al. 2013	Value
MG2	Vegetation macrogroups of class 2	40 categories	Sayre et al. 2013	Value
MG1	Vegetation macrogroups of class 1	27 categories	Sayre et al. 2013	Value
Class	Vegetation class (hierarchy level 2)	6 categories	Sayre et al. 2013	Value
Subclass	Vegetation subclass (hierarchy level 3)	9 categories	Sayre et al. 2013	Value
Formation	Vegetation formation (hierarchy level 4)	18 categories	Sayre et al. 2013	Value
Division	Vegetation division (hierarchy level 45)	44 categories	Sayre et al. 2013	Value
TO	Thermotypes/ Ombrotypes	157 categories	Sayre et al. 2013	Variety, majority
TM	Topographic moisture potential	2 categories	Sayre et al. 2013	Variety, majority
Alt	Altitude	Masl	GMTED2010, 30 arc second resolution	Mean, STD
Lat	Latitude	Degrees north	-	Value
Long	Longitude	Degrees east	-	Value
Area	Cell area on land	Degrees square	-	Value
Slope	Derived from Altitude	Decimal degrees	ArcGIS processing of GMTED2010	Mean, STD
Chorion	Regional phytochoria	19 categories	White 1983	Variety, Majority
Country	Country	Named categories	Countries shape file, contiguous mainland only	Value

Analysis

Conditional inference decision trees

A conditional inference decision tree (Hothorn et al. 2006) was built to predict GHI across Africa. The main difference, and here advantage, of such an ensemble decision tree prediction method (CART analysis) over modelling techniques such as GLMs or GAMs is that no assumptions need be made about the underlying distributions of variables and their relationships. CART techniques are suitable for analyses where hundreds of variables are potentially predictive of a response, where a GLM would run out of degrees of freedom. The result is that CART analysis tends to produce rather reliable predictions, at least within the training data. The disadvantage is that they can be relatively poorer at inference outside out the training data precisely because relationships are not abstracted, and that the significance, 'importance' or relationship between a particular variable and the response is not entirely transparent.

Random Forest (Breiman & Cutler) is perhaps the most established of the machine-learning methods of classification or regression tree analysis (CART analysis). The predominant implementation in R is found in the package randomForest (Liaw & Wiener, 2002). In general, a decision tree is made by taking a training data set with a response, say GHI, and some explanatory variables, say altitude. If GHI>500 is always associated with altitude >2000 m, then a single decision tree could be built by taking all the data, and creating just one decision node: "Altitude > 2000 m?" to predict whether the GHI was less than or more than 500. When more than one explanatory variable is available, a measure of information gain is used to decide which variable to use for each successive node in the tree, such that the entropy in the response is reduced as far as possible by selecting a particular explanatory variable for the split; and if variables are continuous, where to split each variable. Internal nodes are added to the tree until the tree is fully grown, although leaf nodes must contain > 5 samples (by default) to reduce overfitting and allow for learning.

A ‘forest’ of such decision trees is created by ‘bagging’ (bootstrap aggregating) rows, and columns (‘feature bagging’). Bagging is used to improve model performance, and provides a way to estimate model fit. For a training data set with X explanatory variables and Y samples, a bag is created of width X and length Y, where variables and samples are chosen randomly with or without replacement from all the available data, and a decision tree is trained on that subset of the data. The procedure is repeated n.tree times to produce n.tree decision trees in the forest. The predictions of a single tree may be sensitive to noise in the original training set, but the average predictions of many trees are less sensitive, as long as the trees are different. Bagging produces different trees in the forest by showing them different training sets. For prediction, an average of the predictions of the single trees has been shown to highly outperform the single trees (Breiman 1996, Breiman 1998, Bauer & Kohavi 1999, Dietterich 2000), by means of smoothing the hard cut decision boundaries created by splitting in single classification trees, which in turn reduces the variance of the prediction (Bühlmann & Yu 2002, after Strobl et al. 2008).

‘Out-of-bag’ error can be determined by calculating the mean squared prediction error for the samples, using only the trees that did not have that sample in their training set, providing a way to estimate overall model performance.

The importance of a particular variable in prediction of the response can be assessed by computing the out-of-bag prediction error for each sample, then permuting the values of that variable within the trees and calculating the out-of-bag prediction error for the samples again. The before and after difference in values is calculated and optionally normalised by the standard deviation of these differences to give the % increase in mean standard error (%IncMSE). Variables with large values are interpreted as more important than variables with small values.

randomForest has a number of drawbacks. Firstly, variable selection for nodes in the individual CART trees is biased in favour of continuous variables, and categorical variables with many levels, over smaller continuous variables and categorical variables with few levels. This is because the information gain metric is not penalised by the number of potential split points, so variables with more potential cutpoints are more likely to produce a good criterion value by chance, as in a multiple testing scenario (Strobl et al. 2007). This behaviour is undesirable because a variable with the same information can be rendered more or less informative by the way it is presented to the algorithm, for example by changing the scale of measurement or the way that factors are grouped into categories. Larger, continuous variables tend to be selected higher up in the tree, and more often, and therefore have a more profound influence on prediction than other variables. This is a further problem when variables are permuted to measure their permutation importance, because variables high up in the tree can affect the prediction accuracy of a large number of observations, compared with those occurring less often or lower down in the tree. Conditional inference trees (Hothorn et al. 2006), as used in the *cforest* function of the R package party (Strobl et al. 2007), do not have this bias in variable selection. Variable selection is conducted by minimising the p value of a conditional inference independence test, such that the number of categories of each variable is incorporated in the degrees of freedom.

Variable selection for a split, and by extension its importance for prediction, is further affected by the way that observations are sampled to create the bag for a tree (in either cforest or randomForest). An association between a response and an unrelated explanatory variable can be induced by bootstrap sampling of the original observations. This is because even theoretically unrelated variables usually have a small deviation from zero correlation (independence), and this deviation gets amplified when bootstrapping as some observations are repeatedly picked and others dropped. The effect is more pronounced for variables with more categories, because each category holds fewer observations, and is thus more vulnerable to stochastic deviations from independence.

The solution is to sample observations without replacement (rather like X-fold cross validation), so that the relationship between a variable and response is not unfairly inflated for some types of variable.

Variable selection for a split, and by extension the assessment of a variable's ability to inform the prediction of a response, is further challenged by the presence of collinear (or otherwise co-related) variables (Strobl et al. 2008). When a variable is perturbed to assess its importance, it is perturbed on its own without reference to other variables in the tree, so it is testing the independence of the variable from the response, but also assuming independence of the variable from every other variable. The resulting permutation importance scores are therefore marginal estimates of importance, and the effect is that a variable collinear with a predictive variable appears predictive, where it would appear independent of the response if considered conditional on the other variable. Feature bagging (where variables are dropped) aggravates this problem as collinear variables are not always assessed in the presence of each other. A tree can be constructed with only one of the collinear variables, and in this tree whichever feature is included will be deemed 'important'. When all the trees are averaged, the resulting importance value for a correlated variable tends to be inflated relative to uncorrelated variables, and it is unclear which variable is the more important.

party offers a conditional test of variable importance for its *cforest* function, where a variable of interest is permuted within levels of its correlated variables, maintaining the correlation structure between the variable of interest and its correlated predictor variables. For example, if places with higher mean temperatures also have higher maximum temperatures during the hottest month, a conditional importance test would permute maximum values only within places having similar mean temperatures, in order to assess the additional information added by maximum temperature. Although the implementation works well for theoretical datasets with relatively small numbers of

correlated variables, unfortunately it does not run on this data set except with combinations of two or three variables.

Cforest models were estimated using the package party in R, with GHI as the response, and separately a model with minGHI as the response. Categorical explanatory variables (such as the major lithology class of a cell) were modelled as factors, although their values are numeric. Bagging was conducted without replacement.

Variable importance was estimated using permutation tests in cforest. Because conditional tests of importance were not possible to run, we reduced collinearity in our input datasets by modelling only one of the GCM outputs at a time (see Appendix 3 for correlations, particularly between a single model's modern estimate and its own paleo estimates). All paleo-climate values were dropped in favour of the (less strongly correlated) stability metrics. Correlation coefficients for each continuous variable were calculated to help with interpretation.

Model validation

Model validation is an important step, and has two aspects. One is to ask how well a model captures variation within the input data, i.e. how much error is associated with predictions made by the model. We estimated this by calculating root mean square error (RMSE) for out of bag (OOB) predictions.

The second is whether the input data are appropriate and representative of the response being predicted (fitting a good model to poor or unrepresentative data will not give reliable results). By preserving the input resolution for our predicted values we avoid one source of additional uncertainty. However, it is still important that the GHI data should be representative with respect to the geospatial explanatory variables for accurate predictions to be made, and more profoundly, that the GHI scores themselves are reliable (Chapter 4). It is clear from Chapter 2 that the Sahel and Sahara are under sampled relative to central Africa, and the models may predict relatively poorly

here, although this is hard to test without an additional data source. Several of the phytochoria are unsampled, including the Mediterranean and the Cape floras. Without any training data from these regions we are expecting the relationships between environmental factors and GHI established for tropical Africa to hold once extrapolated to these new floras and climates, which is a big assumption.

Model transferability, *sensu* Wenger & Olden (2012), was assessed by dropping cells of Ghana and Cameroon from the training dataset and then predicting GHI scores for them. During OOB model validation, random cells are dropped from the training sets and the model is used to predict GHI for those missing cells. The differences are quantified. However, bioquality is strongly spatially ordered, with adjacent cells having more similar GHIs than distant cells. A strong test of model performance is therefore to drop whole contiguous blocks of cells from the training dataset, and ask whether the model can recover the known bioquality patterns over an area.

Specific research questions and their model approaches

As the minGHI model fits rather poorly compared with the GHI model, and because it is a function of sampling intensity which is always zero outside of the training dataset, the minGHI model is not pursued further, and all further results concern GHI only.

What is the predicted bioquality across tropical Africa?

For this aim, a *cforest* model was built using the *cforest* function of the party R package. We included all variables of Table 6.1, in order to make robust predictions, and to assess the importance of each variable. We repeated this analysis for the IPSL (main body) and MRI GCMs (Appendix 2).

Cforest has a number of parameters which can be tuned to improve model performance. These include *ntree*, the number of trees to be grown (with 500 as the default); *sampsize*, the number of cells to be randomly sampled for a single tree (two-thirds by default); *mtry,* the number of variables to be randomly sampled for a single tree; *nPerm*, the number of times a variable should be

permuted to calculate its importance value (1 by default); *nodesize*, the minimum number of cells at a terminal node (5 by default); *maxnodes*, the maximum number of terminal nodes (all trees are fully grown by default).

Nodesize and *maxnodes* were left at their default values, and represent a compromise between making the most predictive trees (i.e. every cell is predicted by its values), and allowing some 'learning' to happen, permitting for better model fit when unknown training values are encountered. If enough trees are grown, *sampsize* is not an important parameter in the regression implementation. Increasing *nPerm* makes importance estimates for variables 'slightly more stable, but [is] not very effective' (Liaw & Wiener 2015), and rapidly increases computation time. Following tuning, we set mtry to the default two-thirds. We increased *ntree* to 1500, and set sampsize to the default 0.632. We set *nPerm* to 10 (a compromise between stability and computation time), as a one off investment in this model.

Model validation was carried out by calculating RMSE from OOB error estimates. Predictions were made for every cell for which training data were available for every variable, which excludes, for example, the Mediterranean and Cape phytochoria. Mapping was conducted in ArcGIS. OOB predictions were mapped for cells in the training data, rather than input values, or the values obtained by running training data through the model, which are over fitted.

Which environmental variables best predict bioquality patterns? What is the relative contribution of climate, climate stability, soil, biogeography, and topology to bioquality?

The relative importance of individual variables in the bioquality model described above were calculated using party's *varimp* function and plotted. Models and variable importance were estimated using four different seeds, and a Friedman test was applied to test the null that that the rank of each variable was the same across model runs.

The marginal relationships between each variable of interest and GHI were visualised by plotting cell GHI predictions against the respective input variable data and fitting smooth splines with 10 knots for continuous variables; or boxplots for categorical variables. This is not the same as a partial dependence plot, because the values are not calculated based on the variable's position in the trees and conditionally on higher nodes, but gives a general view of the relationship. Partial dependence plots are excruciatingly slow to perform on this dataset.

Extrapolating from the tropics: what is the predicted bioquality across mainland Africa?

A response can be extrapolated from a model with continuous variables, but not from unordered categorical variables. To predict bioquality across mainland Africa, we dropped all categorical variables with factor levels missing from the (tropical) training set: TOMAJORITY, CHORIONMAJ, MG 6, MG3, MG2, MG1. Otherwise, the model and parameters are as before.

Model validation was carried out by calculating RMSE from OOB error estimates, visual inspection and comparison with our best model's estimates, and published estimates of endemism patterns across northern and southern Africa.

We assessed model transferability by dropping cells of Ghana, and cells of Cameroon, from the model. To test for model transferability, cells for Ghana were dropped out of the training dataset to see if the resulting model could predict observed bioquality patterns for Ghana as a whole. Ghana was chosen because species sampling is high here, so we have a reliable estimate of the true bioquality pattern for the country. Dropping one of the best sampled countries from the training dataset also constitutes a strong test for the model, because the model has considerably less information to work off. We separately dropped cells for Cameroon, another well sampled country, from the training set and then predicted bioquality across the continent.

Mapping was conducted in ArcGIS. OOB predictions were mapped for cells in the training data, rather than input values, or the values obtained by running training data through the model, which are over fitted.

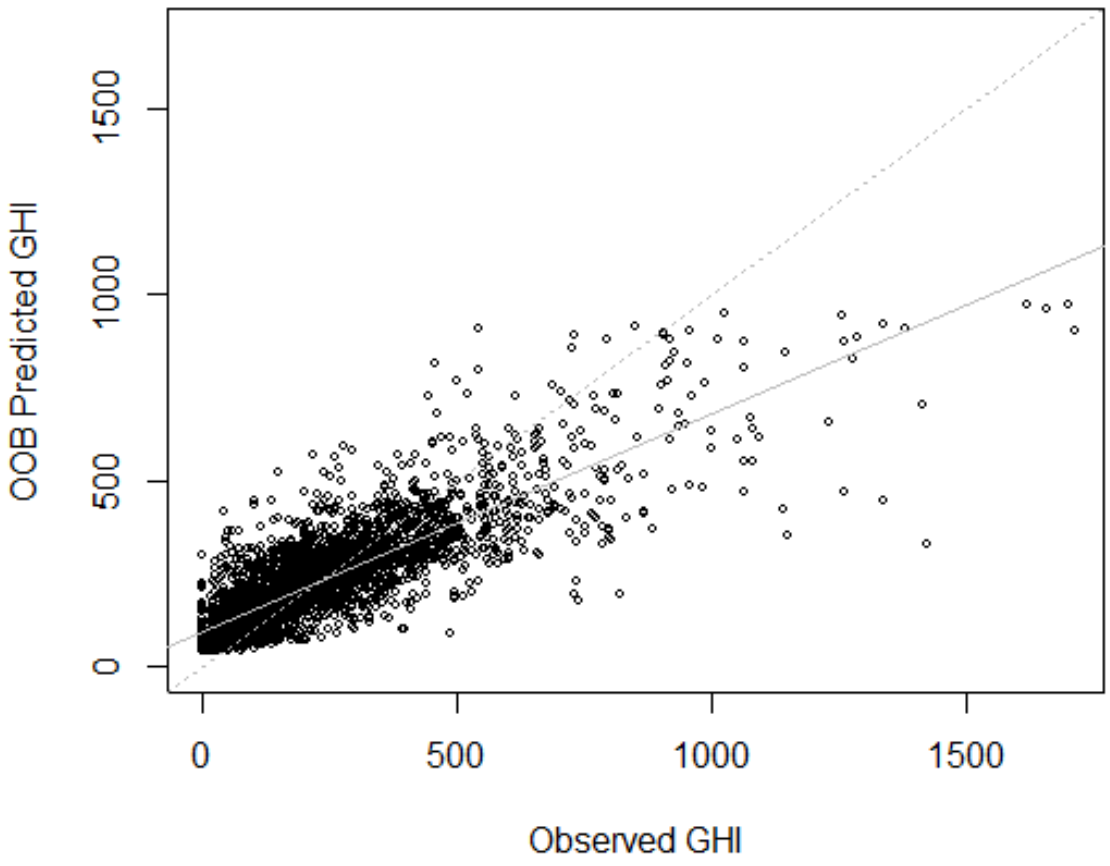
Results

Models presented here use the IPSL GCM. All input and output data associated with IPSL model can be found in S6.1. All input and output data associated with MRI model, presented in Appendix 2, can be found in S6.2.

Predicting bioquality across tropical Africa

Our tropical African model predicts GHI within 114 GHI points, comparing input GHI scores to out-of-bag GHI predictions. The OOB plot (Figure 1) shows that the model renders GHI values more average.

Figure 6.1. Plot of OOB predictions against input (observed) GHIs. Grey dotted line shows the 1:1 line, solid grey line shows the linear regression of the two datasets. Predicted values are more average than the input values. n=3426, root-mean-square error (RMSE)=113



Predicted GHI for tropical Africa is faithful to the training data, but is much smoother (Figure 6.2, Figure 6.3). The coldest areas are identified as the Sudanian, Sahel and southern Sahara; and the Kalahari-Highveld transition zone/southern Zambezian. Excluding the Sahara, these two savanna regions in northern and southern Africa are arid, smooth plains; unlike the other arid regions of Africa (northern tip of Somalia, Karoo-Namib and parts of the northern Sahara), which achieve high altitudes and slopes. The central 'sash' of high bioquality is maintained, and seems to relate best to mountains, peaking in the extreme east of Somalia. The longitudinal stratification of Somalia is helped by the inclusion of longitude in the model, which identifies Somali high scores in general.

Figure 6.2. Input GHI, for 3426 cells across tropical Africa.

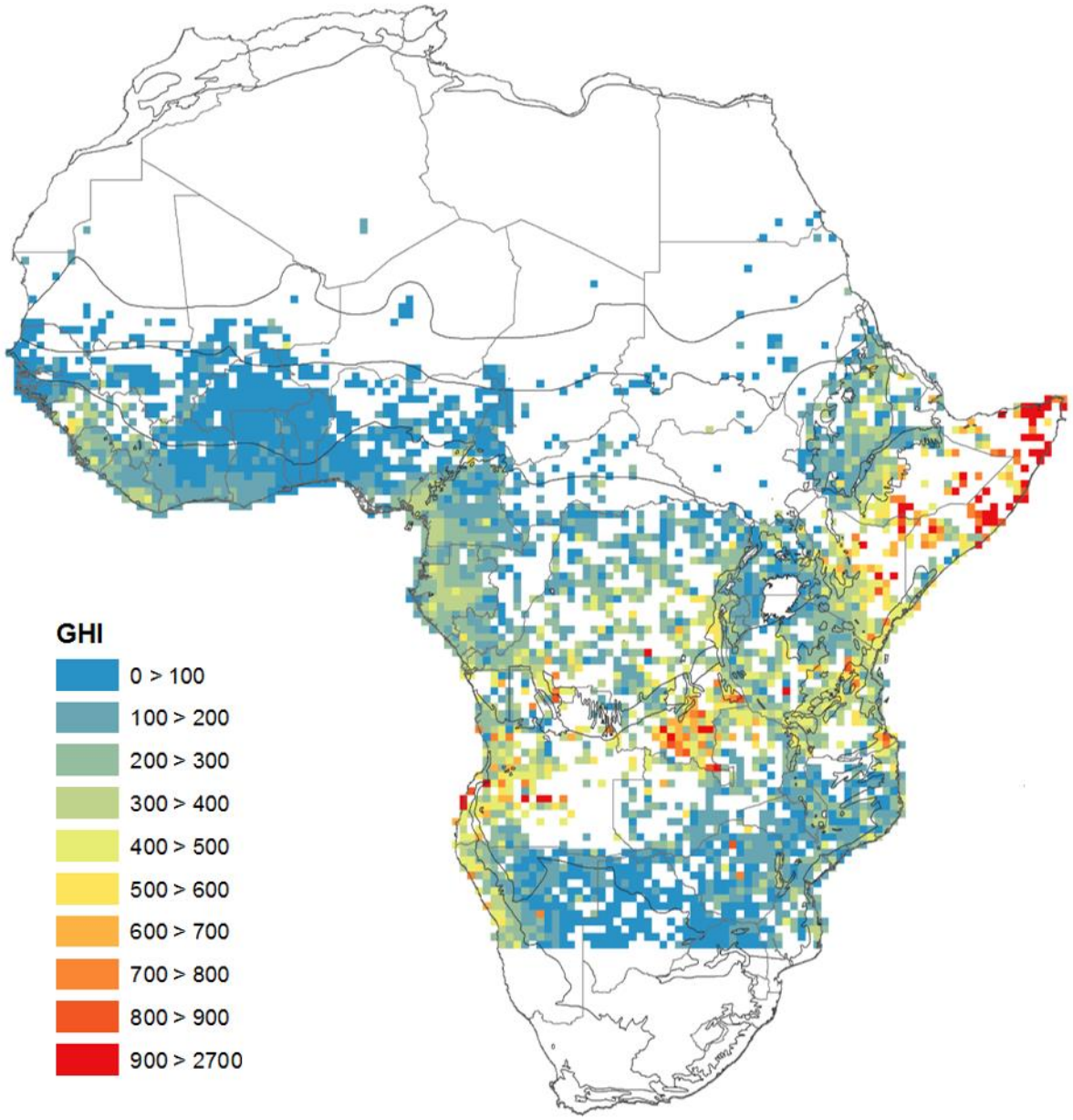
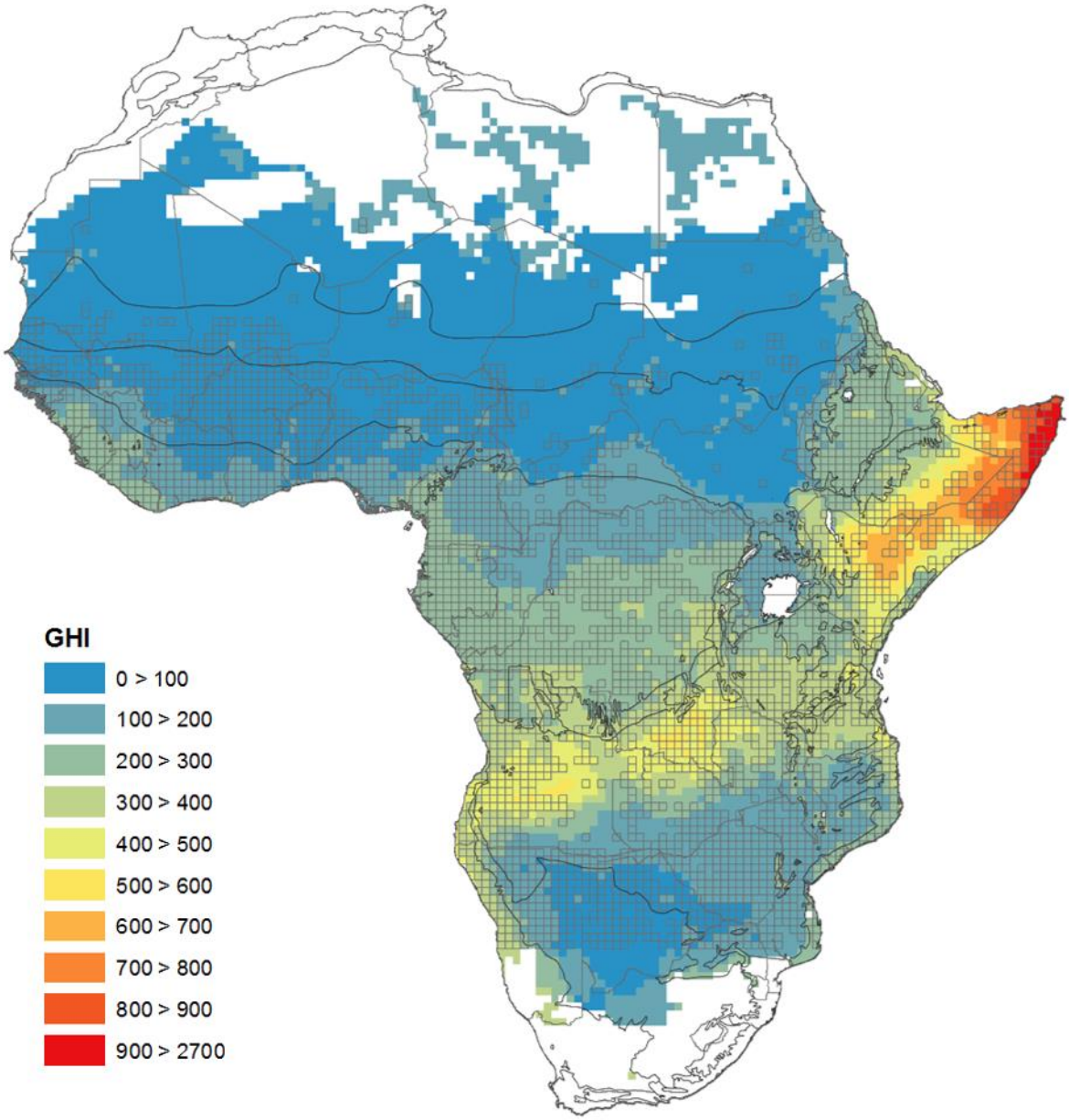


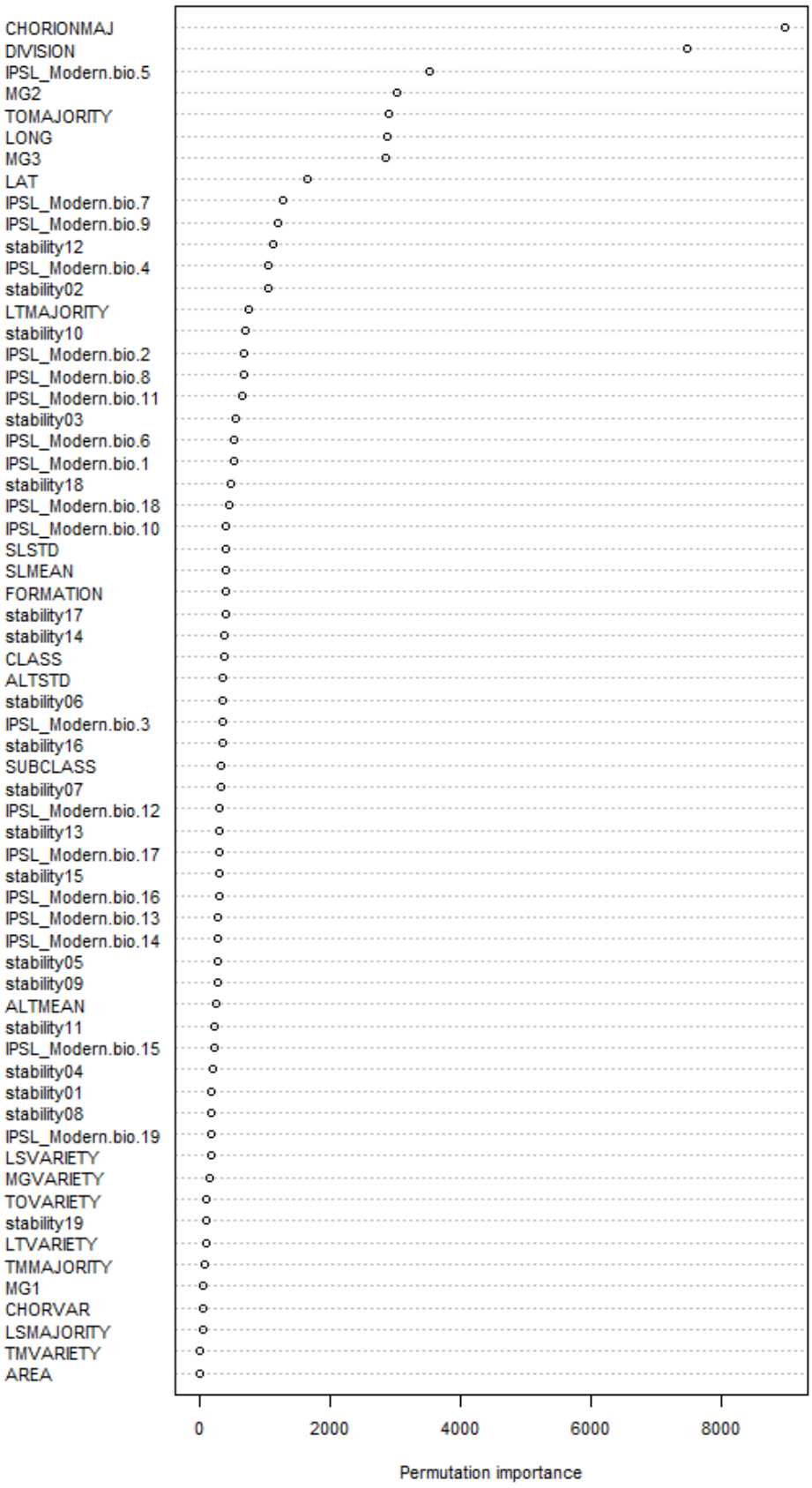
Figure 6.3. Modelled GHI for tropical Africa. OOB predicted values for cells in the training data are mapped with grey borders.



Variable importance and major relationships with GHI

Variables included in the model are ordered by their importance for the prediction of bioquality in Figure 6.4. All variables hold predictive power for GHI, except area. The relationship between each of these predictors and GHI are presented in Figures 6.5 to 6.10.

Figure 6.4. Importance scores for explanatory variables within the GHI model. Terms follow Table 6.1.



For rainfall, areas which have experienced the greatest change since the LGM (stability=1) tend not to have low GHIs, but the very highest GHI scores are associated with highest stability (Figure 6.6). This is generally true for temperature, but the relationship is weaker, and it does not apply to Bio2 (mean diurnal temperature range), Bio3 (isothermality), Bio6 (minimum temperature of the coldest month), or Bio11 (mean temperature of the coldest quarter), where the highest scores are associated with the greatest change (Figure 6.5). The stability of annual mean rainfall over 21,000 years is amongst the top predictors of GHI.

Figure 6.5. GHI plotted against stability measures for the temperature-related Bio1-11.

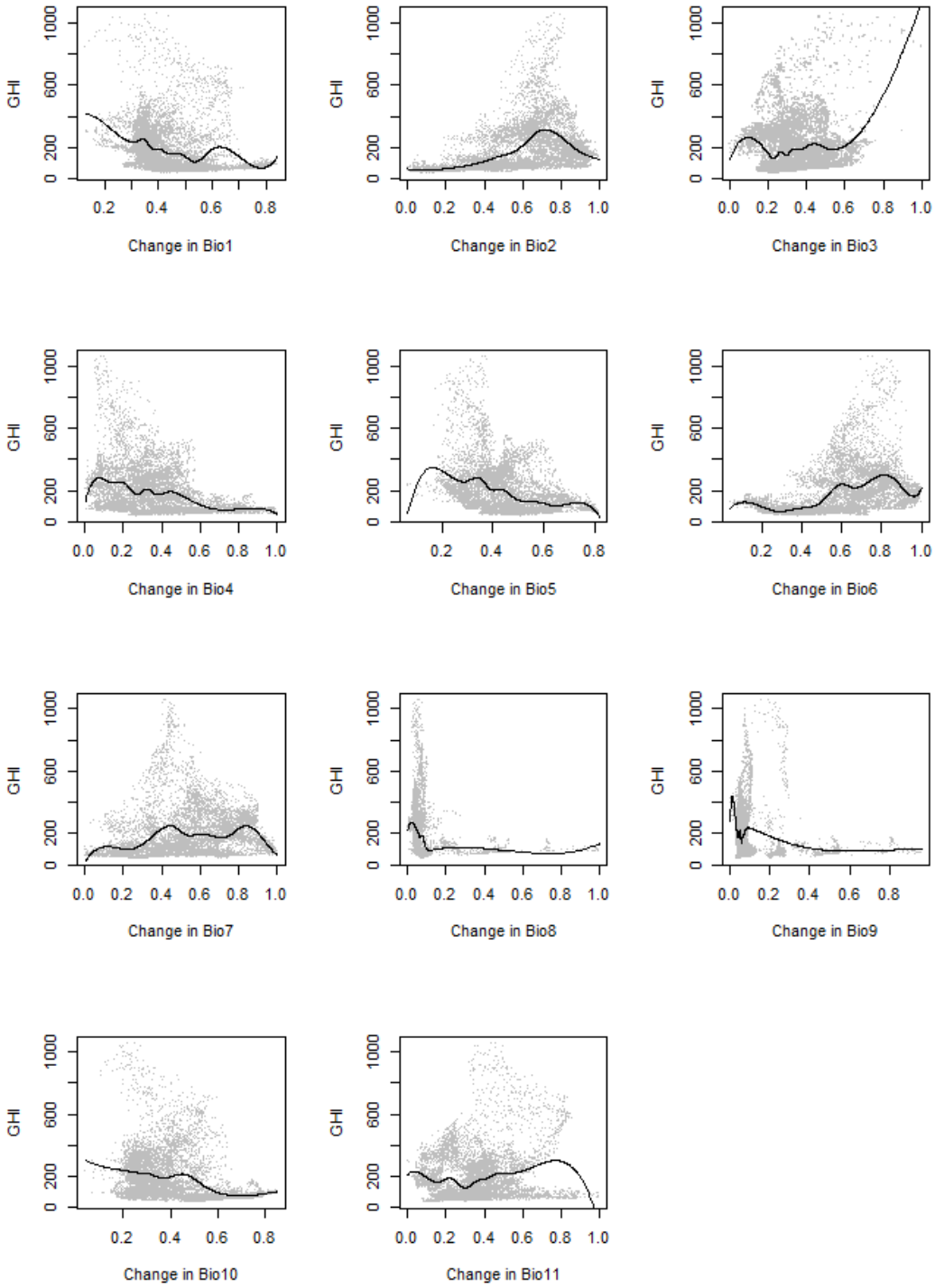
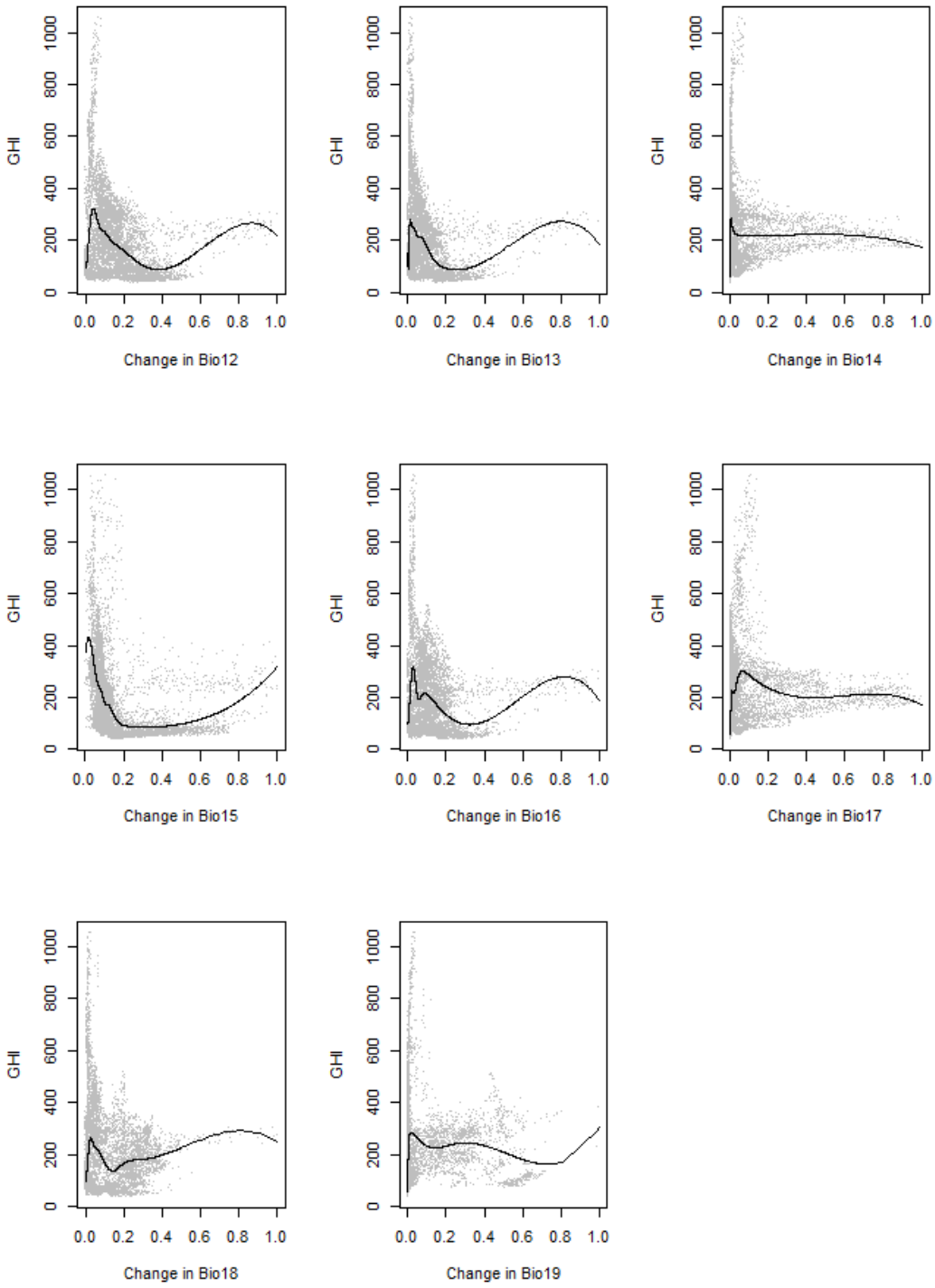


Figure 6.6. GHI plotted against stability measures for the rainfall-related Bio12-19.



For modern climate, we can divide a group of more important variables: Bio5, Bio7, Bio9, Bio4, Bio2, Bio8, Bio11, Bio6, Bio1, Bio18 and Bio10 from the less important Bio 3,12,17,16,13,14,15,19 (Figure 6.4). This is more or less a division between rainfall variables (less important) and temperature variables (more important), except Bio18, which is Precipitation of Warmest Quarter, and is interacting with temperature. This is supported by a t-test of the difference in importance values between Bio1-11 and Bio12-19 (Table 6.2). To summarise the rather interrelated Bio1-11, the pattern is for the highest GHI values to be associated with the least extreme temperatures (monthly Bio5, annually Bio7, seasonally Bio4, quarterly Bio9, diurnally Bio2), but high mean annual temperatures (Bio1) (Figure 6.7).

GHI generally increases with increasing annual mean rainfall, although the highest GHIs are associated with the lowest annual mean rainfall values (Figure 6.8). The Bio14/17 plots (driest month and quarter respectively) show particularly interesting relationships, with the very highest and the very lowest GHI scores associated with the most extreme dry climates, and every cell with monthly rainfall more than about 20 mm having moderate GHI scores.

Table 6.2. T-tests for (logged) variable importance. Friedman test on rank orders, Friedman chi-squared = 0.22, df = 2, p-value = 0.90; Accept null (rank importance is the same on each run). Results from seed 42 are shown here (arbitrarily).

Variable	Mean	*p*
Bio 1-11	998.7621	<0.0005. True difference in means is not equal to 0.
Bio12-19	302.9121	
Stability 1-11	403.094	0.80. True difference in means is equal to 0.
Stability 12-19	439.8394	
Bio 1-19	705.7726	0.08. True difference in means is equal to 0.
Stability 1-19	418.5657	

Figure 6.7. GHI plotted against mean modern measures for the temperature-related Bio1-11.

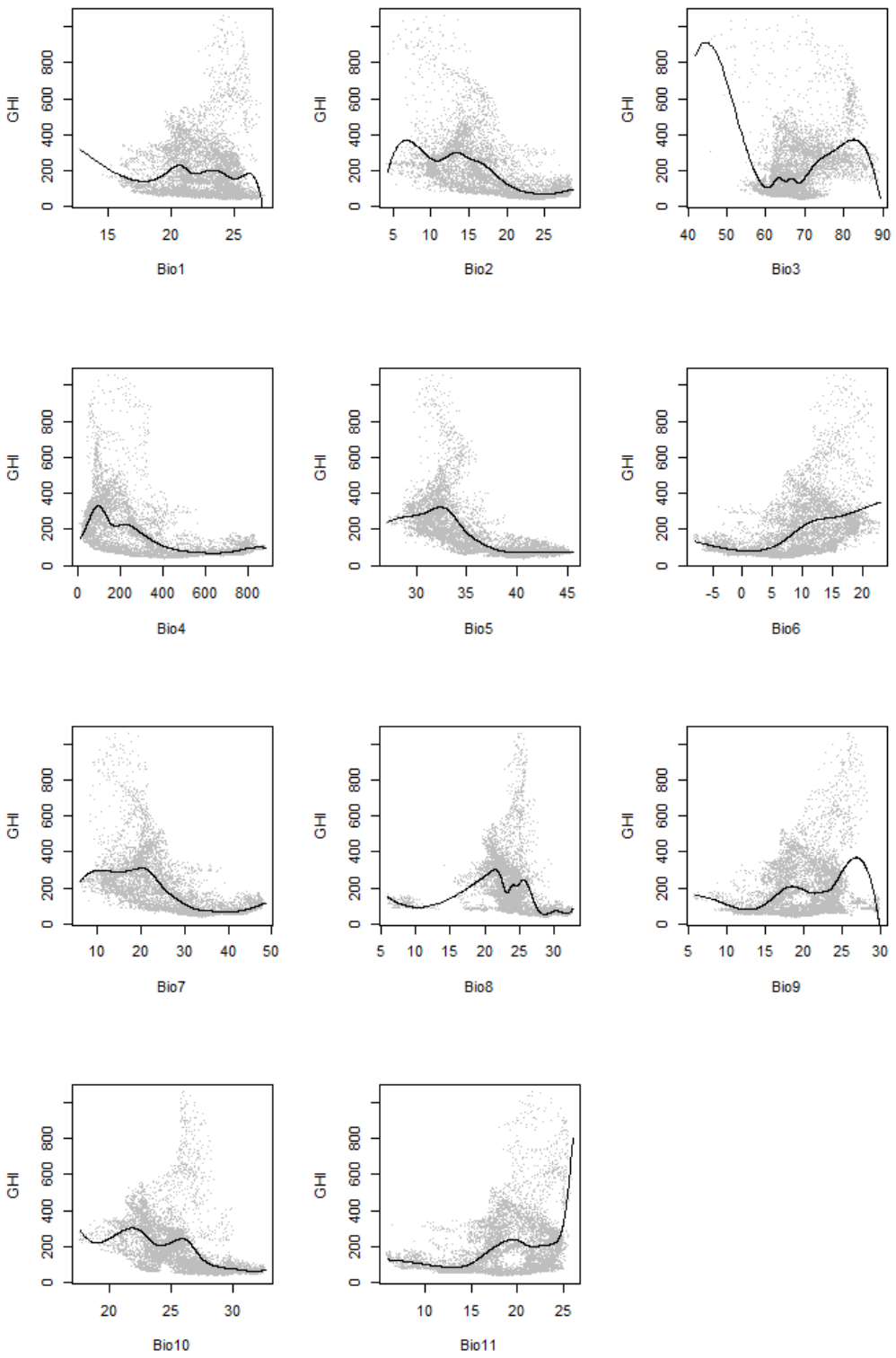
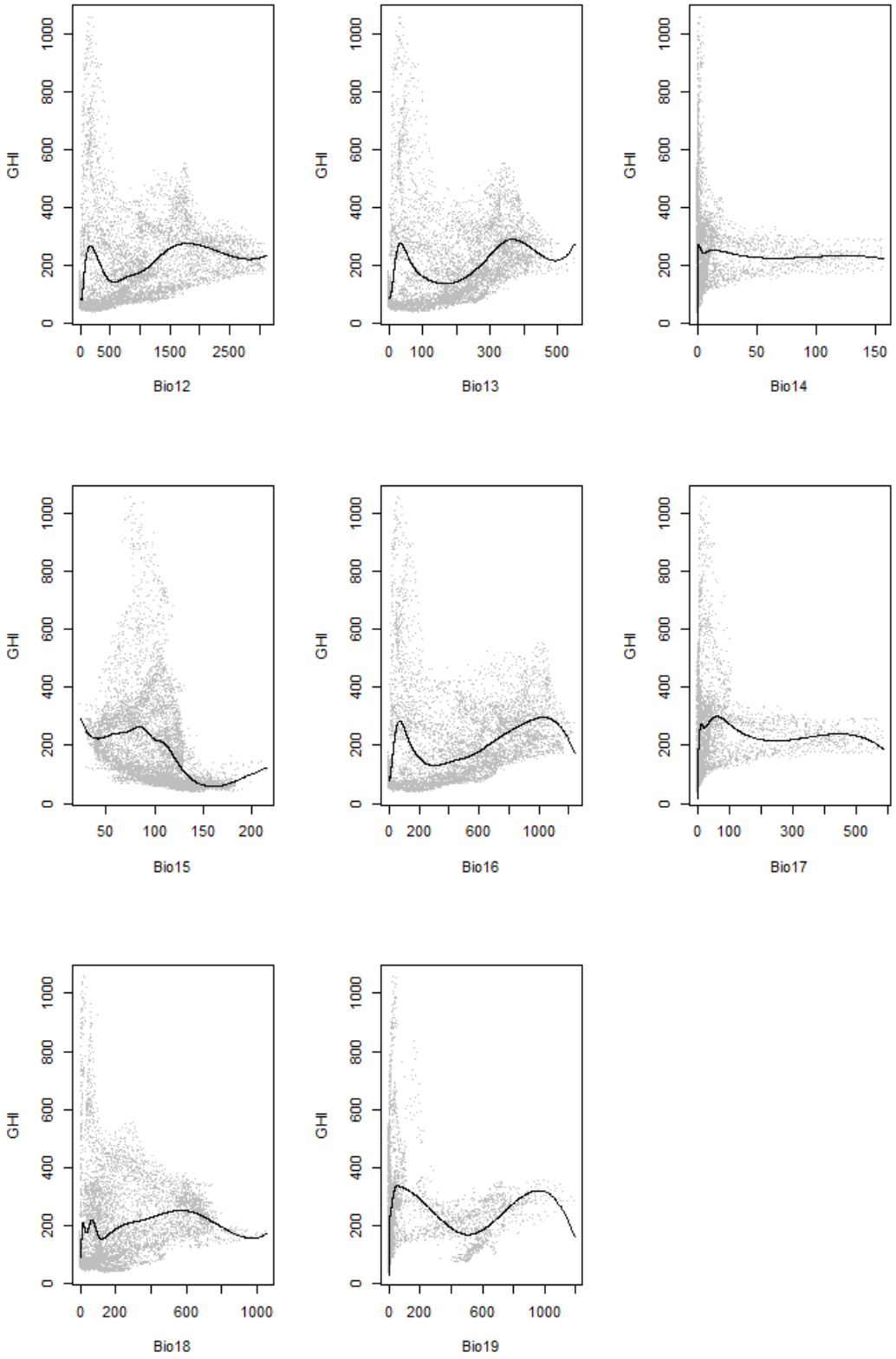


Figure 6.8. GHI plotted against mean modern measures for the rainfall-related Bio12-19.

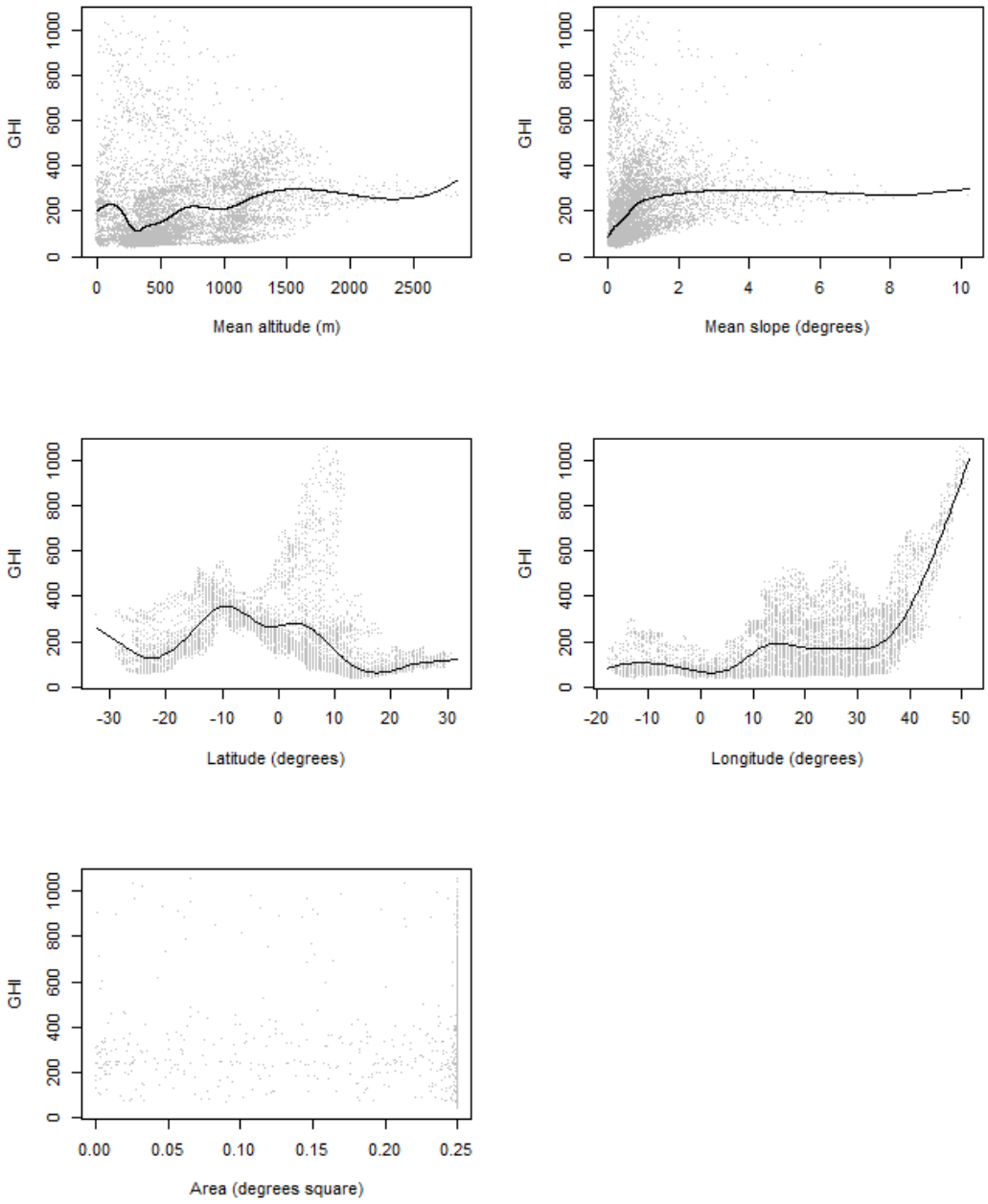


On average, slope is more important than altitude for the prediction of bioquality, and both are generally of middling importance (Figure 6.4, Figure 6.9). Land surface majority, a Sayre et al. 2013 classification, comes out as one of the least important predictors, and land surface variety (the number of different types of land surface within a cell) does little better. Mean altitude and mean slope are correlated, such that higher altitude areas generally have higher slopes, from the available evidence it seems that slope is the leading variable of these two. GHI increases with both altitude mean and slope mean, with greater variation at low altitudes and slope. Cells with altitude > c. 1750 m all have GHIs in the top quartile. With both plots we can see the 'wedge' pattern: high values of slope/altitude cannot achieve low GHIs, but GHI can be high or low at low values.

Longitude and latitude are important predictors of bioquality, ranking sixth and eighth respectively (Figure 6.4). The latitudinal gradient is high-low-high-low-high, with nadirs corresponding to the two main savanna ecosystems of the Sudanian and Zambezian regions, at mesic north and mesic south latitudes. The 'peak' at the equator actually peaks twice, once at 10 degrees north and again, more strongly, at 10 degrees south (Figure 6.9). The importance of longitude is due to the high scores of Somalia in the extreme east of the training dataset, beyond 40 degrees east (Figure 6.4).

Cells overlapping the coastline have a smaller area on land than half a degree square. Area (on land) of the cell was not an important predictor of bioquality at this resolution. This is an advantage for this analysis, but does not mean than bioquality and area would never interact over a more extreme range of resolutions (Figure 6.4).

Figure 6.9. GHI plotted against mean altitude, mean slope, latitude, longitude and area.



The most important predictive variable for the GHI model is the major phytochorion of the cell *sensu* White (1983). Unsurprisingly, the Somalia-Masai region (IV) has the highest median GHI scores; the lowest median scores are associated with the Sudanian (III), Sahel (XVI) and Sahara (XVII) choria (Figure 6.10). The surficial lithology type associated exclusively with Somalia has the highest GHI (Figure 6.10).

Divison, sensu Sayre et al. 2013, is the second more important predictor in the model (Figure 6.4). Division is one level up from macrogroup, the finest (Class>subclass>formation>division>MG). Division separates, for example within Guineo-Congolia, Guineo-Congolian Evergreen & Semi-Evergreen Rainforest from Guineo-Congolian Swamp Forest. MG2 and MG3 are the finest divisions of woodland ecosystems and desert ecosystems respectively, and both appear very high up in importance ranking. MG1 (types of forest) appears to hold near zero predictive value beyond Division, probably because the forests of Africa have rather similar GHI values to each other, compared with, for example, the radical differences in GHI between the Sahara and the Kalahari or Somali deserts.

Figure 6.10. GHI plotted against White (1983) regional phytochoria, and lithology classes. A) Boxplot showing median GHI for each chorion. Chorion names follow White (1983), MA is mangrove, WA is water. Chorion IV is Somalia-Masai. B) Boxplot showing median GHI for each lithology class. 17=Allvuim-Gypsum, found only in Somalia.

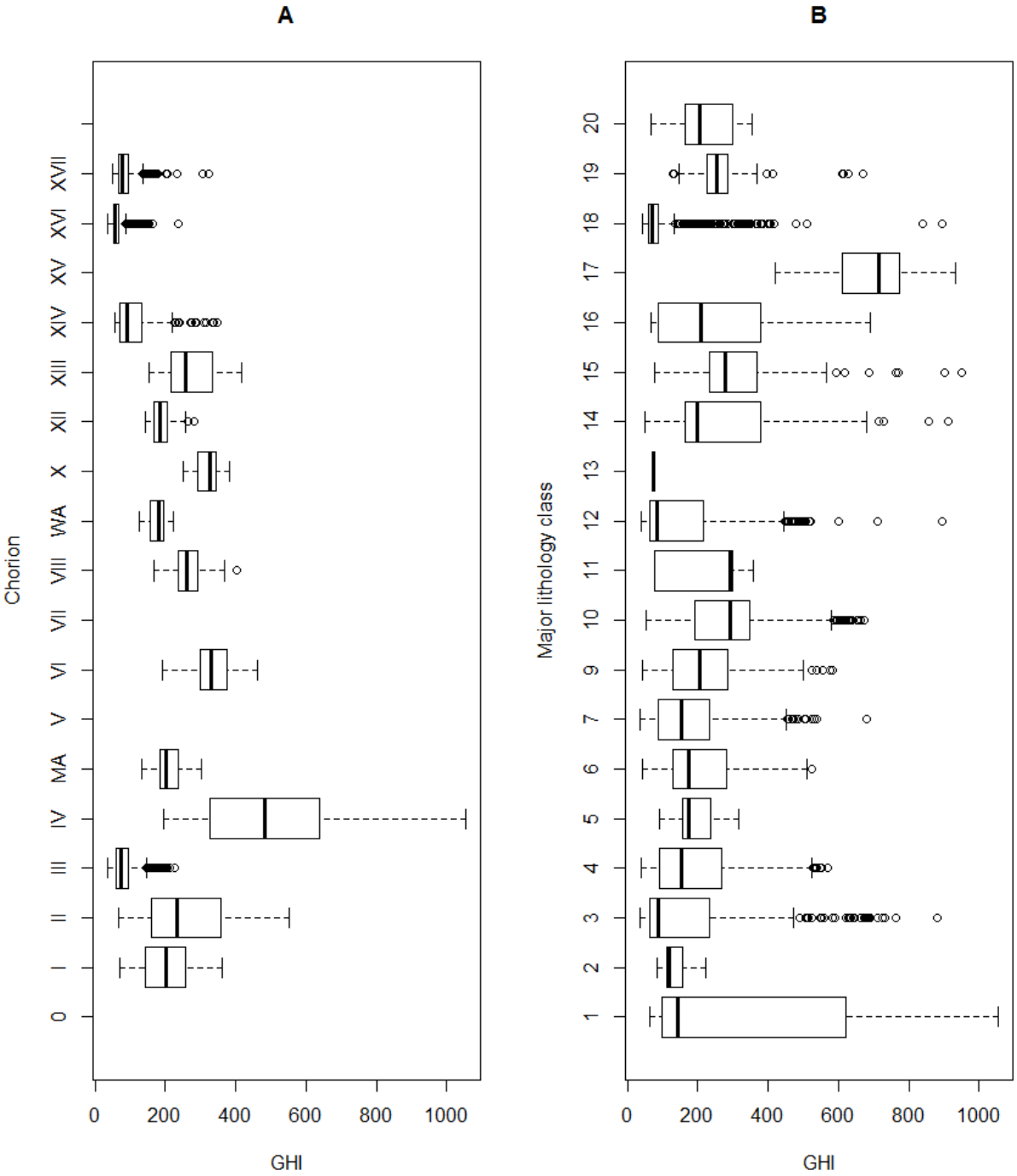
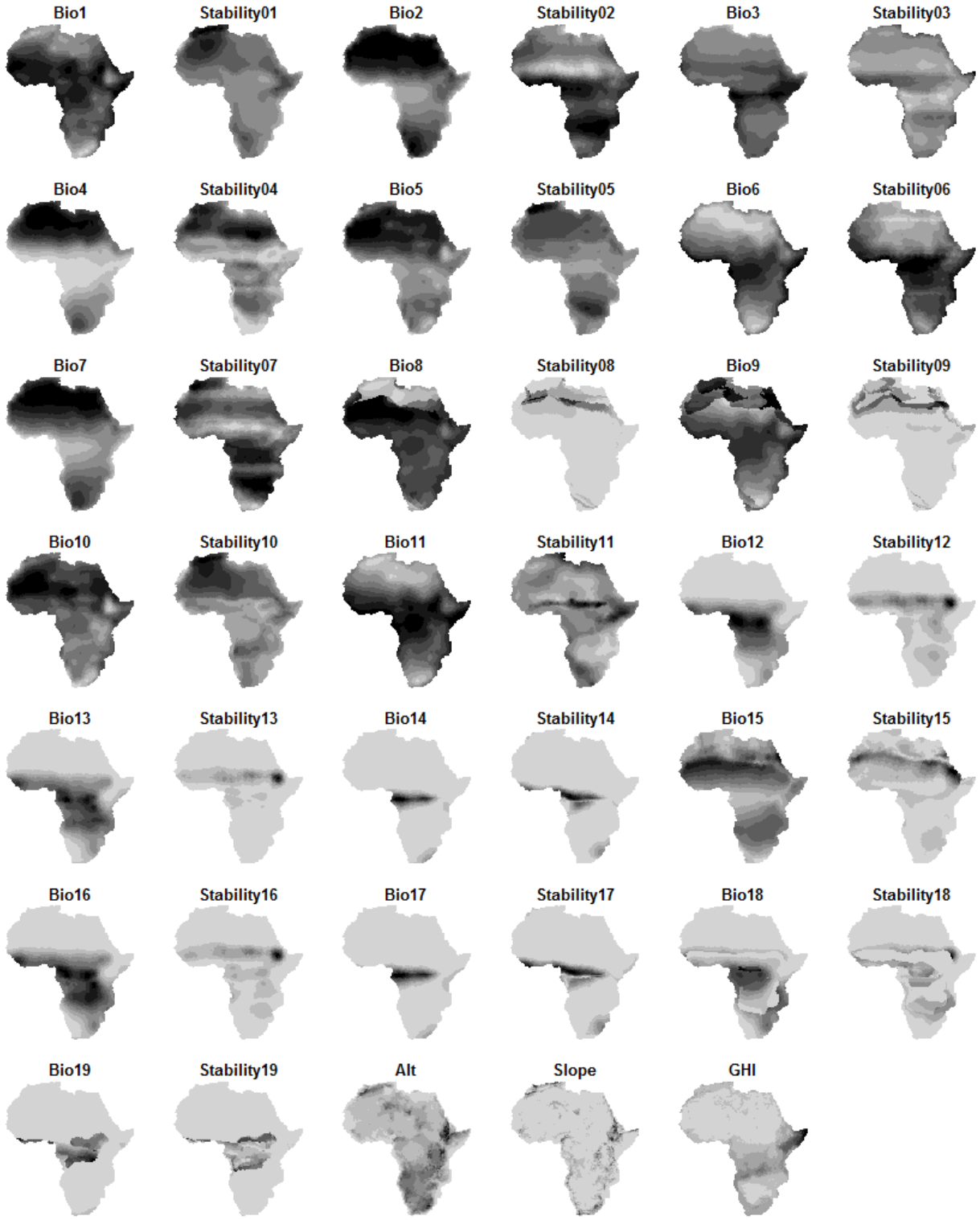


Figure 6.11. IPSL Modern Bioclim variables and stability measures, altitude, slope and GHI.
Dark=high values; 10 colours from light grey to black; equal intervals (not quantiles), e.g. for stability, 0-0.1, 0.1-0.2 etc. High values represent max change, low values represent max stability. Terms follow Table 1.



Extrapolating to mainland Africa

For Ghana, the results with and without Ghana are more similar to each other than either are to the input dataset (Figure 6.12). Both models predict a wider band for the SW Ghana hotspot, that is, Ghanaian bioquality is elevated in the model relative to observed GHIs, particularly without Ghana data in the training dataset. This is probably because Ghana, being well sampled vegetatively, has relatively more Green star records than the rest of Africa, so what the model is learning from other countries is inflated relative to scores it could learn from Ghana. The model also ascribes relatively higher scores to the border strip with Togo than in the input dataset, corresponding to Lake Volta and rift to the west. The Atewa Range achieves a high score of 197/205 when modelled without the input data, a higher score than it achieves in input data (140/139), presumably because the model picks up high altitude and slope there.

For Cameroon, unlike Ghana, all modelled values are depressed relative to the input dataset (Figure 6.13). In panel B, the odd cell with GHI of 206 corresponds to a very local peak in altitude. Both models maintain generally higher values towards the coast compared with inland, but in the model without Cameroon cells this is of reduced extent. The two input cells with high values are not maintained by either model (GHI=510 and 414).

Figure 6.12. Test of model transferability using Ghanaian cells. A) Input GHI values for Ghana; B) Output GHI values for Ghana; C) Output GHI values for Ghana, with Ghana cells omitted from training dataset. Friedman test on cell ranks, comparing model without Ghana to input: $p < 0.005$ (ranks are different). Paired t-test on mean of the differences between cells of the model without Ghana to input cell values: $p = 0.209$ (difference=0).

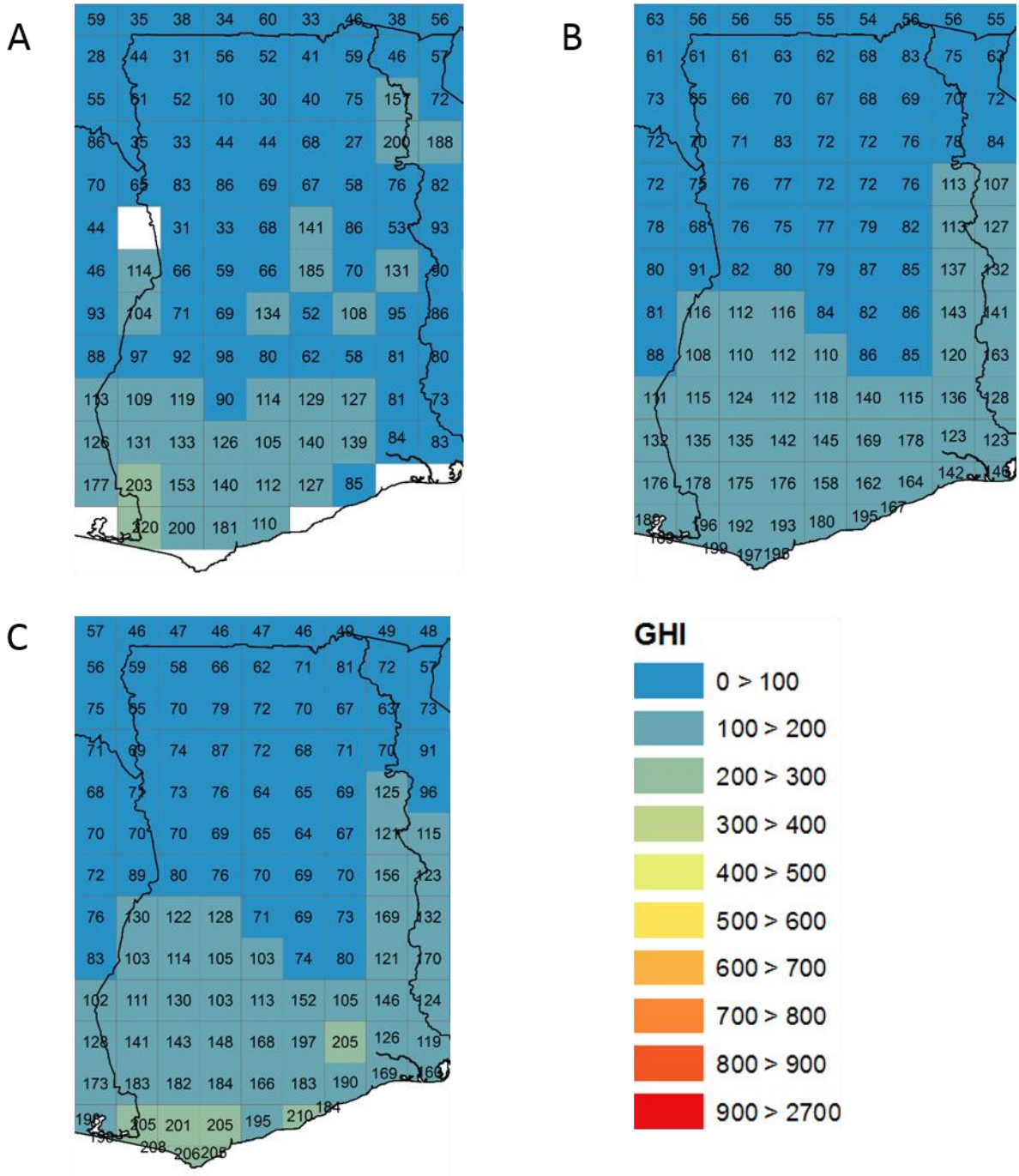
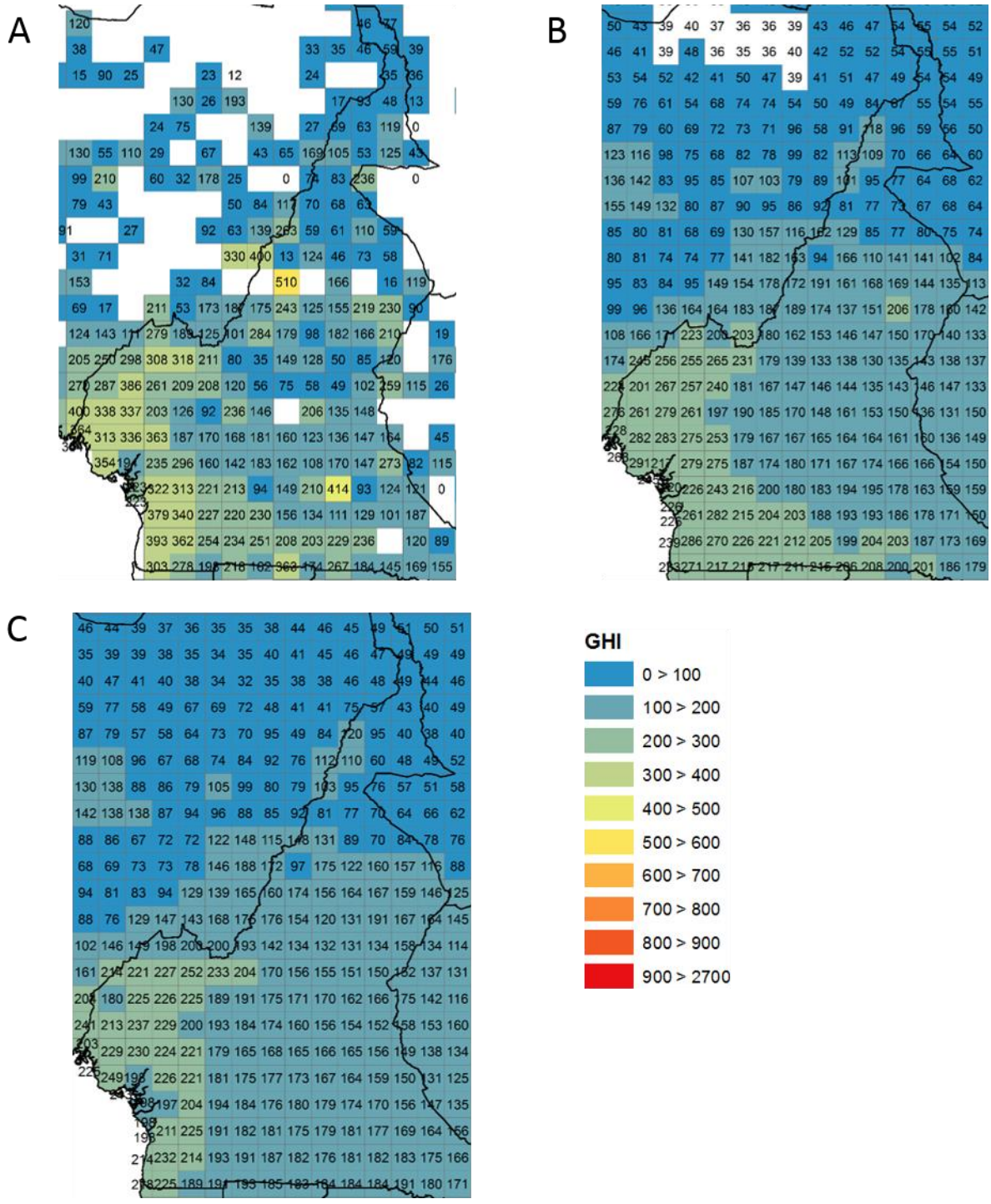


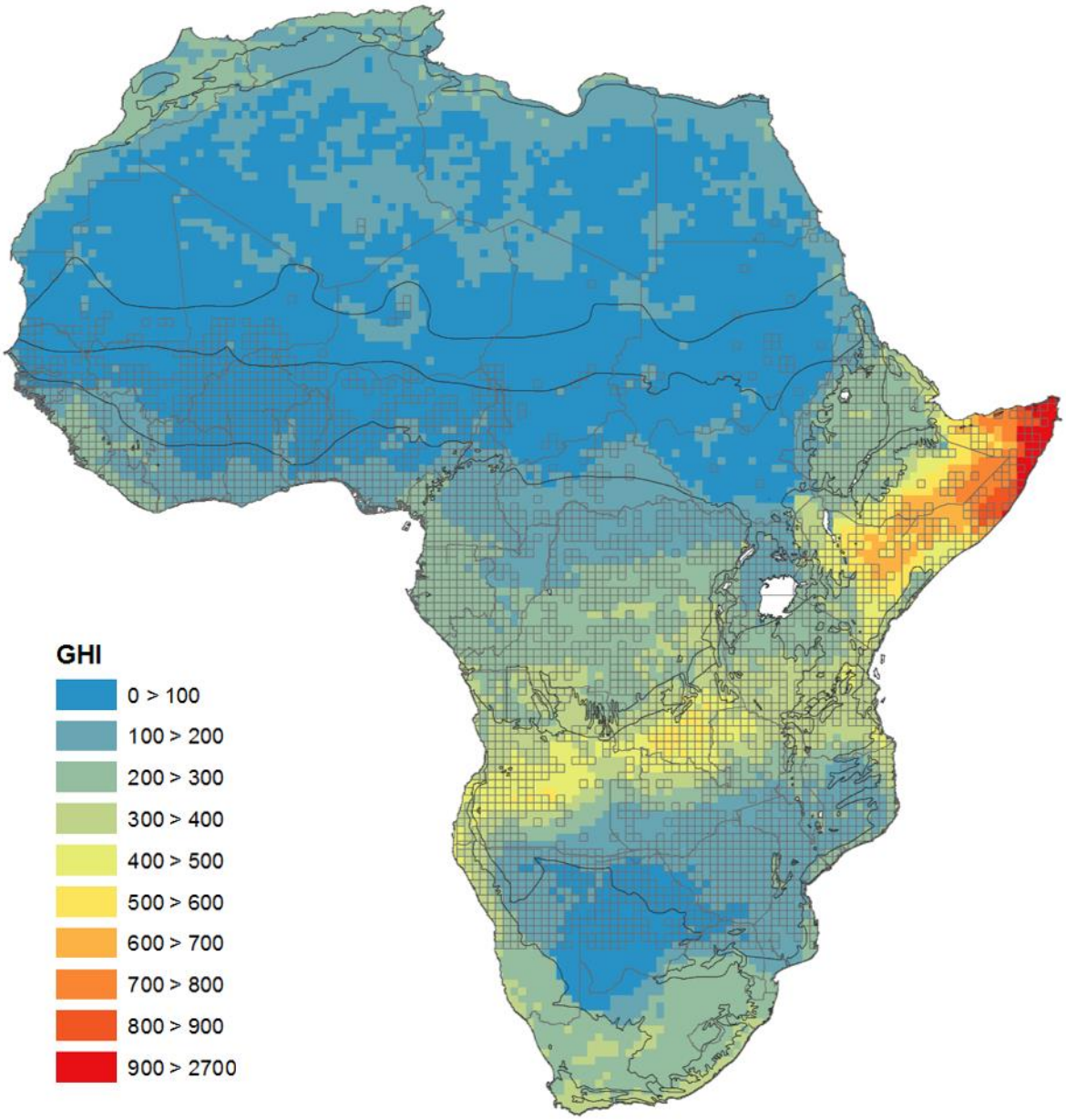
Figure 6.13. Test of model transferability using Cameroonian cells. A) Input GHI values for Cameroon; B) Output GHI values for Cameroon; C) Output GHI values for Cameroon, with Cameroon cells omitted from training dataset. Friedman test on cell ranks, comparing model without Cameroon to input: $p < 0.005$ (ranks are different). Paired t-test on mean of the differences between model without Cameroon to input: $p < 0.05$ (mean of the differences = 11.0).



For North Africa in our all Africa extrapolation, the model predicts the higher endemism associated with the Mediterranean flora compared with the Sahara/Sahel, without training data for this region, and even picks out the outpost of Mediterranean flora in Libya (Benghazi) (Figure 6.14). The Mediterranean region is wetter and has higher slopes than the Sahara. In general, predicted GHI scores are elevated in northern Sahara cf. the south and Sahel. Brito et al. (2016) mapped endemism of amphibians, reptiles, birds and mammals and shows a similar uplift in endemism for the Sahara relative to Sahel. Their patterns within the Sahara for high (vertebrate) endemism are not particularly congruent with ours for vascular plants.

The GHI model predicts the ongoing high bioquality of the Karoo-Namib further south, outside of the scope of the training data set (Figure 6.14). The Cape flora has higher predicted GHI than the Karoo-Namib surrounding it, and southern Kalahari Highveld, but does not achieve very high values. At 300-400, cells of the Cape approach the 90th percentile of all scores across Africa.

Figure 6.14. Modelled GHI map of mainland Africa. OOB predicted values for cells in the training data are mapped with grey borders.



Quantile breakdown of scores:

10%	20%	30%	40%	50%	60%	70%	80%	90%
59	70	80	99	136	177	226	283	372

Discussion

GHI was modelled across the African mainland at half degree square resolution. No information is derived about the species responsible for such modelled scores. An alternative method would be to build species distribution models for all species of tropical Africa at a given resolution, and then for a grid of the same resolution calculate the GHI from the probabilities of occurrence of the species in each cell. This method is not explored here, mostly because of the time it would take to parameterise and tune c. 30,000 species models. Also, for all but the best known species and areas, the resulting occurrence maps would doubtless be prone to substantial error, and the endeavour may well turn out to be more costly, less successful and less constructive than simply collecting plants from less well known regions. Although the gaps in our 'input' knowledge (uncertainty in most species' full distributions, uncertainty in climate conditions across the African continent) also affect the modelling of GHI, by modelling just one statistic (GHI) which averages over the known flora, we avoid multiplying our errors and we are better able to scrutinise and manage uncertainty.

That the chorological and structural divisions should be so highly favoured by the model is interesting. *cForest* should not have the bias towards categorical variables that randomForest suffers from, so the result is probably not artefactual. White's regionalisation was defined using estimated numbers of endemic species to areas, so the importance of this variable is perhaps less surprising, but the Sayre et al. (2013) regions are much more strongly influenced by vegetation physiognomy than vascular plant endemism. Both vegetation classifications in any case are defined at a much broader scale than half degree square, so the argument is not entirely circular.

The emphasis on temperature over rainfall shown here is unusual, with most previous analyses focussing instead on rainfall. Generally, climatically, GHI shows a trend to increase with increasing temperature, and is highest for areas with the most constant and *least extreme* temperature

regimes. GHI generally increases with rainfall, and is highest under the *most extreme* rainfall regimes, particularly at the dry end.

GHI generally increases with increasing *instability* of rainfall*,* but is highest for cells with both the most stable rainfall regimes. GHI generally increases with increasing stability of mean temperature, but this relationship is weak. This suggests that both hypotheses (stability promotes speciation, and endemism; change promotes speciation and endemism) may be at work, at least at this crude temporal and geographic resolution.

The MRI GCM results (Appendix 2) differ in fine detail, but the mapped GHI results are not substantially different to those with the IPSL model. For tropical Africa in particular, the broad patterns of bioquality are very similar. For the African mainland analysis, the IPSL model predicts somewhat higher GHI values in the high altitude regions of the Sahara. The stability of annual rainfall over paleo timescales remains the most important of the stability metrics, and is ranked as more important than all of the modern day climate variables except rainfall seasonality. Rainfall seasonality (Bio15), and temperature seasonality (Bio4), are the most important climatic drivers. Biogeographic-related variables remain the most important predictors of GHI. For the stability of rainfall, GHI increases with increasing instability, but the highest GHI values are associated with the most stable areas.

In general, the model results are smoother and more average than the input values, and dropping a whole country seems to not to have a big effect on model predictions. Had we lifted the minimum criterion up to 10% sampling the input GHI values would be more stable, but then many cells would have lacked a response variable.

The results presented here are tantalising, but need to be developed further before publication. The models should additionally incorporate the biogeographic framework of Chapter 3. I would also explore raising the 2.5% sampling threshold for GHI scores to 5% or 10% so the model can be trained

on a more reliable training set. I investigated the area and isolation of modern climate types as predictive variables for GHI, and found this to be a rather important variable in the prediction of GHI (not presented). The hypothesis here is that areas which are isolated climatically have higher GHIs, as species are less likely to be able to disperse to have wide distributions (as in SW Ghana and Nimba, where high rainfall and high altitude areas are rather distant from them and GHIs are high). My definition of climate types (divisive clustering of the bioclim variables) likely needs to be refined. I would backcast my definition to the available paleo-climate periods to define the area and isolation of climate types at the LGM, for example, as this is a more direct test of refuge theory. The models should also be run with the CCSM GCM, which includes a Pliocene (3 mya) climatic prediction. Splitting tropical Africa into its major regions and building separate models would also be illuminating. I suspect that the high GHIs of Somali in particular have, in these models, overruled much of the signal for other floras. Although it is interesting to draw very general conclusions about the impact of e.g. the stability of rainfall over 21,000 years on bioquality, this almost certainly belies interesting differences between the regions.

The fact that the broad trends in bioquality patterns in Ghana and Cameroon were predicted, albeit slightly more smoothed than reality, even without these countries in the training set, is very encouraging, and suggests these models have a lot to offer for filling in gaps on the map. However, the prediction of higher values for Ghana than recorded suggests that intensive local sampling across tropical Africa, yielding more accurate answers, would involve slightly lower scores in these other countries than modelled.

Acknowledgements

We acknowledge the World Climate Research Programme's Working Group on Coupled Modelling, which is responsible for CMIP, and we thank the climate modeling groups (listed in Table 6.1 of this paper) for producing and making available their model output. For CMIP the U.S. Department of

Energy's Program for Climate Model Diagnosis and Intercomparison provides coordinating support and led development of software infrastructure in partnership with the Global Organization for Earth System Science Portals. http://cmip-pcmdi.llnl.gov/cmip5/guide_to_cmip5.html

Supporting Information

Appendix 2

The analyses presented in Chapter 6 are rerun using the MRI climate model (bound with thesis).

Appendix 3

Analyses showing the correlation between paleo and modern climates, and between the different general circulation models.

Appendix S6.1

Header file for IPSL model.

Appendix S6.2

Header file for MRI model.

Chapter 7

Lineage diversification in globally rare species

Article status

Not for publication.

Abstract

Aim: To quantify divergence dates for Upper Guinean species, for sampled species of the magnoliids and Rubiaceae, and look for differences in the timing and tempo of lineage diversification between globally rare, and globally widespread species.

Location: Africa, Upper Guinea

Taxon: 103 species of magnoliid and 181 Rubiaceae species, mostly from Upper Guinea.

Methods: Dated phylogenies were estimated using the marker *matK* for 25% of all African magnoliids, and 12% of African Rubiaceae. For each species, the date of divergence from its closest sampled relatives was extracted. Divergence dates between the globally rarest species were compared with divergence date estimates of globally widespread species using a t-test.

Results: 11 species of magnoliid have median divergence dates inside of 2.8 Ma with very strong support. 14 Upper Guinean Rubiaceae species have 95% confidence intervals of their median age falling inside 2.8 Ma, while 28 species have median ages falling inside 2.8 Ma. 23 Black and Gold star species have well resolved sister relationships. Of these, 15 have median divergence dates in the Pleistocene (eight have 95% confidence intervals entirely within the Pleistocene), a further four diverged 2.8-5 Ma, and only four are older than 5 Ma.

Main conclusions: 26 species were shown to have Pleistocene origins with 95% confidence, of which 16 also had strong topological support. This is in contrast to the commonly held belief that 2.8 Ma is insufficient to observe speciation events in Africa. Globally rarer Black and Gold star species are 'younger' than widespread Green and Blue Star species (rarer species have closer sampled extant relatives). At least *Sericanthe toupetou & S. adamii* and *Psychotria brachyanthoides* and *P. linderi* would seem to support the lowland forest Pleistocene refuge hypothesis. *Monocyclanthus vignei,* belonging to a monotypic genus, has a divergence date from its closest relatives of 13.5 mya, and its

refugial distribution alongside other ‘neoendemics’ lends support to the museum hypothesis, as not all local endemics of Upper Guinea have Pleistocene origins.

Introduction

In this Chapter, mechanisms by which globally rare species of Upper Guinea may have diversified are explored. Many hypotheses have been put forward to explain the observed spatial aggregation of globally rare species (Table 7.1). Two broad classes of hypothesis can be distinguished: those that invoke past climatic or geographic changes, and those that rely on modern day ecogeographic variation.

Ecologists have sought to explain distribution patterns in terms of present day factors, such as rainfall or habitat heterogeneity, while historical biogeographers have sought to highlight evolutionary and paleo-climatic processes that have influenced speciation and extinction. The dichotomy between processes maintaining high species diversity in certain areas (ecological processes) and those that generate high diversity through time and across regions (historical evolutionary and biogeographic processes) is a blurred one; history has played a substantial role in shaping both regional and local species diversity and ecological factors can influence or drive speciation and extinction (Moritz et al. 2000). The mechanisms summarised in Table 7.1 are all likely to have played a role in shaping the angiosperm diversity, endemism and distributions that we observe today.

Table 7.1. Classes of evolutionary model for promoting spatially aggregated species endemism (after Moritz 2000).

Class	Sub-class	Mechanism	Barrier	Timing	Predictions	References
Pleistocene climate changes: increased allopatric speciation	Lowland forest (<1500 m) refuge theory	↑ vicariant speciation through isolation, drift, selection	Dry forest, savanna (local pockets of lowland forest as refuge during cooler, drier phases)	Pleistocene (< 2.8 Ma)	Sister taxa occur in different lowland refugia; congruence between species of similar ecological and dispersal traits.	Haffer 1969, Prance 1982, Hamilton 1976, Robbrecht 1996.
	Fluvial refuge theory	↑ vicariant speciation through isolation, drift, selection	Dry forest, savanna (gallery forest as refuge during cooler, drier phases)	Pleistocene (< 2.8 Ma)	Sister taxa occur in different riverine refugia; lowland forest species persist with differences in gallery forest; congruence in distributions of species which can tolerate periodic inundation.	Leal 2004, Colyn et al. 1991, Robbrecht 1996, Wieringa 1999.
	Montane (>1500 m) refuge theory	↑ vicariant speciation through isolation, drift, selection	Dry forest, savanna, lowland forest (higher elevations as refuge during warmer, wetter phase)	Pleistocene (< 2.8 Ma)	Sister taxa occur in different montane refugia; montane endemics are of recent origin.	Sosef 1994, 1996, Robbrecht 1996, Roy 1997.
	Disturbance theory	↑ speciation through competition, directional selection	Heterogeneous forest structure	Pleistocene (< 2.8 Ma)	Neoendemics are concentrated in areas which experienced maximal disturbance during climatic changes, rather than in stable refuges.	Bush 1994, Colinvaux 1993.
Pleistocene refugial: decreased extinction	Museum hypothesis	↓ extinction due to stability. Range contraction.	N/A; speciation by another mechanism	Before or after 2.8 Ma; gradual	Gradual accumulation of endemics of all ages; no Pleistocene 'burst'; sisters not necessarily in different refugia.	Stebbins 1974, Lovett & Friis 1996, Fjeldså & Lovett 1997.
Gradient: parapatric or allopatric speciation	N/A	Divergent selection with or without gene flow; Adaptation to present day ecogeographic conditions	None necessary; ecological gradients	Not correlated with climate changes.	Taxa have sisters in nearby different vegetation type; derived species are found at the edges of forest blocks.	Smith et al. 1997, 2001, Endler 1982, Fjeldså 1994, Davis et al. 2002, Harris et al. 2000, Möller & Cronk 2001.
Dispersal	N/A	Speciation in isolation	Distance	Not correlated with climate changes.	Well dispersed taxa have diversified	Wieringa & Poorter 2004.

Most authors who have interpreted hotspots in Africa have invoked historic changes in climate (e.g. Hamilton 1976; Brenan 1978; Lovett et al. 1996; Fjeldså et al. 1997, Linder et al. 2001). The extent of lowland forest in Africa has changed dramatically since the origin of the tropical angiosperm flora in the early Cretaceous. During the Eocene (39-36 mya), fossil evidence showed that tropical tree taxa were common as far north as Egypt (Morley 2000). Climatic changes in the late Miocene to Pliocene (10-2 mya) are thought to have had a particularly dramatic impact on the African flora. The late Pliocene (3.5-2.8 mya) was marked by several periods of very pronounced drying and cooling across the African continent, associated with glacial periods in the northern hemisphere, and rapid extinctions in the rain forest flora (Morley 2000). The Pleistocene (2.8-0.01 mya) was characterised by nine major and 12 minor glacial cycles with the most recent occurring 70,000-12,000 years ago (Hamilton 1982, Hamilton & Taylor 1991, Sosef 1994). Interglacial periods were probably characterised by a greater rain forest extent than today (Morley 2000).

Lonnberg (1929) was the first to suggest that lowland rain forest refuges could have occurred as the result of drying and cooling conditions. Lowland rain forest would have shrunk to small enclaves where climatic conditions remained suitable, such as along rivers, at higher elevations, or in particular lowland areas with favourable edaphic conditions. Geologic and palynologic data can be used to identify refugia by dating a sedimentary core and identifying the pollen. On land, the best (radiocarbon dated) palynological evidence for Guineo-Congolia comes from Lake Bosumtwi in Ghana (over 28,000 years), currently near the drier limit of the forest belt, and Lake Barombi Mbo in the (Lower Guinean) forest of Cameroon (over 30,000 years) (Morley 2000, reviewed by Maley 1996). Lake Bosumtwi, nowadays in the Moist Semideciduous forest belt (albeit denuded by human activity), was unforested during the last glacial maximum, until 9000 years ago. Lake Barombi Mbo remained surrounded by forest, confirming the Cameroon-Gabon refuge there. Palynological studies of a deep-sea marine core off the Dahomey Gap have confirmed a coastal rain forest flora in SW Ghana during the Pleistocene. However, the geographic and taxonomic resolution of these

palynological cores remains limited (Sosef 1996). Reconstructions of past vegetation based on paleo-vegetation modelling coupled to paleo-climate modelling are being developed, but are still subject to substantial uncertainty (e.g., Hély et al 2009, Anhuf et al. 2006, Hardy et al. 2013).

An alternative way to identify historical refugia is to look at modern species distributions, particularly narrowly endemic species. Under refuge theory, the rain forests surrounding former refugia are expected to contain some species derived from the local refugia, mixed with rapidly dispersing species from further afield; so we expect to find the highest diversity of rain forest species at locations of former refuges, particularly for poorly dispersing, rain forest obligate taxa (Sosef 1996). To the extent that refugia have been proposed using high species endemism as evidence, hotspots and refugia are one-and-the-same, allowing for some uncertainty over precise limits.

The high endemism of refugia has been attributed to high speciation, and/or low extinction. The refugial speciation model argued that climate change led to fragmentation of species with isolation of conspecific populations in refuges. The isolates became genetically differentiated and eventually speciated (Hamilton 1976, Prance 1982). This suggests an origin of these species during and after major climate changes. An alternative model suggested that refuges act as 'museums', accumulating species slowly over time due to climatic stability and low extinction rates (Stebbins 1974, Lovett & Friis 1996), predicting a gradual accumulation of species of all ages. These two models are not mutually exclusive, and the support for each can be tested using dated molecular phylogenies (Plana 2004).

A handful of authors have explicitly tested the hypotheses outlined in Table 7.1 for West African angiosperms using taxonomic, ecological and biogeographic methods (e.g. Sosef 1994, 1996, Robbrecht 1996, Fjeldså & Lovett 1997, Wieringa 1999); few have used molecular phylogenetic tests (Harris et al. 2000, Davis et al. 2002, Plana et al. 2004, Couvreur et al. 2008, Holstein & Renner 2011). 30% of *Begonia* species, a poorly dispersed rain-forest obligate genus, were estimated to have

diverged during the Plio-Pleistocene (5 mya) and all in the past 10 mya (Plana et al. 2004, Sosef 1994, 1996). The ITS sequences of African *Aframomum*, characteristic of marginal forest habitats, showed almost no variation and were estimated to have arisen in the Pleistocene (Harris et al. 2000). Upper Guinean species were shown to be sister to small groups of Lower Guinean species in *Bikinia*, *Aphanocalyx* and *Tetraberlinia*, all primary evergreen forest trees, and rapid Pleistocene speciation was found in *Tetraberlinia* and *Bikinia*, though slower rates in *Aphanocalyx* (Wieringa 1999, phylogeny reconstructed on morphological traits). In *Acridocarpus*, rain forest species hold a phylogenetically basal position (Miocene) with a pattern of increasing drought tolerance among the more recently diverged taxa, although not quite as recently as the Pleistocene (Davis et al. 2002). In contrast, Fjeldså & Lovett (1997) concluded that lowland areas, which have been postulated as Pleistocene refugia (in E. Africa), were dominated by species that represent lineages of pre-Pleistocene age. 25 species of African *Coccinia* were found to have diverged before the Pleistocene (Holstein & Renner 2011). Two influential reviews (Plana 2004, Moritz et al. 2000) suggest that "in most cases, the divergence of extant sister taxa predates the Pleistocene" (Moritz et al. 2000) or that "strong support for a Pleistocene model of recent speciation is exceptional and unusual" (Plana 2004). However, datasets for wet tropical African groups are seriously under-represented, and this impression appears to be largely speculative.

Dated phylogenies can be used to estimate divergence dates between putative sister species. Characteristics of these species can be 'mapped' onto phylogenies to show what geographic distributions and ecological characteristics are shared by closely related species. These observations can be related to the predictions of the various evolutionary models.

Here, dated *matK* phylogenies for two clades are presented: the magnoliids and the Rubiaceae. *matK* in combination with *rbcL* has emerged as the standard 2-locus barcode for plant species, on the recommendation of Hollingsworth et al. (2009). *rbcL* is more conserved among angiosperms than *matK,* and for this reason is particularly useful for discriminating families. *matK* is often used to

discriminate species, but is sometimes invariant within a genus or closely related genera. *rbcL* has not been included in either of the analyses presented here, but might be expected to help resolve order/family relationships within the magnoliids.

The magnoliids are an angiosperm clade comprising the orders Canellales, Laurales, Magnoliales and Piperales. Relationships between the orders are somewhat contentious, with the Piperales and putative outgroup Chloranthales being inconsistently placed. However, a preponderance of the morphological and molecular evidence suggests that the magnoliids ([CHLORANTHALES **[[MAGNOLIALES + LAURALES] [CANELLALES + PIPERALES]]**] [monocots [Ceratophyllaceae + eudicots]]) are monophyletic (Stevens 2001 onwards: APG website).

In Upper Guinea, the magnoliids are dominated by the Annonaceae, a pan-tropical family of the Magnoliales. There are 42 genera (c. 400 species) of African Annonaceae (Couvreur 2014). The family includes one of only three (monospecific) genera endemic to Upper Guinea, *Monocyclanthus.* The family has benefitted from recent taxonomic attention, making this family a good starting point. A molecular phylogeny for the family was presented which delimited the subfamilies, tribes and genera (Chatrou et al. 2012), and a dated phylogeny discussing the timings of the family's dispersal around the tropics has been presented (Couvreur et al. 2011). Pirie & Doyle (2012) presented a discussion of the dating issues in the family (stem group c. 91 Ma); and Couvreur et al. (2008) presented a dated phylogeny for African Annonaceae, showing the East African species to have had multiple periods of diversification (33, 16 and 8 mya) coinciding with known periods of aridification and geological activity. Several significant African genera have been revised recently (*Uvariopsis,* Kenfack et al. 2003; *Annickia,* Versteegh & Sosef 2007; *Isolona* & *Monodora*, Couvreur 2009; *Hexalobus*, Botermans et al. 2011; *Uvariastrum,* Couvreur 2014), although a few genera are still in need of revision: *Monanthotaxis* (in progress) and *Artabotrys.*

The Rubiaceae are a very large family in the order Gentianales, which also includes its sister family the Loganiaceae (dominated by the large genus *Strychnos* in Upper Guinea), and the Apocynaceae Gelsemiaceae and Gentianaceae (APG 2009). The Rubiaceae is one of the five most species-rich families in the world, with c. 13,000 species in 620 genera found on all continents (Goevarts et al. 2006). The Rubiaceae is the most species-rich family in Upper Guinea (c. 300 species) and account for > 10% of all angiosperm species found there. Three subfamilies are recognised, the Cinchonoideae, the Ixorideae and Rubioideae. The stem divergence time for the family has been estimated at 90.4 mya, with subfamily divergence times at 73.1, 73.1 and 84.4 mya (Bremer & Eriksson 2009).

The phylogenies are used to date divergence between sampled African species and their closest extant relatives overall, and for the globally rarest species in particular (Black and Gold Star). Some Upper Guinean Black and Gold Star species with Pleistocene origins are mapped relative to putative refugia of Upper Guinea, where species samping is reasonably representative.

Materials & Methods

Gene and taxon sampling

The chloroplast gene *matK* alone was used to estimate the phylogenies for the Rubiaceae and the magnoliids. Taxon sampling for the magnoliids is summarised in Table 7.1 and for the Rubiaceae in Table 7.2. 'Upper Guinean species' were defined as angiosperm species with at least part of their distribution in the forests of Upper Guinea. The checklist of 2673 Upper Guineaea species was based intitally on Jongkind (2004) and was updated for new species publications and synonymy (TRAFRICA). 'African species' were drawn from the African Plants Database 3.4.0, and include species of South and North Africa.

I selected and prepared 768 leaf fragments of Upper Guinean species for *matK* and *rbcL* sequencing in 2012, before my theis commenced. Sequencing was conducted by researchers in Canada using the Barcode of Life Data System (BOLD) (Ratnasingham & Herbert 2007). Sequencing was funded under an IDRC project (led by Olivier Hardy), the initial remit of which was to build a DNA barcode library for Ghanaian (timber) trees. Trees (including small trees down to 5 cm DBH) were therefore the initial focus of the species sampling, which was extended to include other woody and non-woody species.

These sequences were supplemented with publically available sequences held by GenBank (Benson et al. 2013) for Upper Guinea species not held in the BOLD project, and for the putative 'sister' species of Upper Guinea species. Putative sisters were identified by searching GenBank for African species in Upper Guinea genera, making the assumptions that: (1) Upper Guinean species are more closely related to a species found on mainland Africa than further afield; and (2) that Upper Guinean species will be more closely related to a species in its own genus than a different genus (i.e., genera are monophyletic). This method does not guarantee that the closest sister has been included in the phylogeny, but is an improvement on sampling only Upper Guinean species, and makes use of all available sequence data. Species are therefore at least as young as they appear in the phylogeny, but might become 'younger' with increased sampling. Sequences were collated in MEGA version 6 (Tamura et al. 2013) and were aligned manually.

Two species of Loganiaceae were chosen as the outgroup for the Rubiaceae. No outgroup was included for the magnoliids and the tree was mid-point rooted; although there is a preponderance of evidence that the Chloranthales are sister to the magnoliids, this is disputed, and the order is not present in Africa (Stevens 2001 onwards: APG website). The Chloranthaceae were therefore excluded from the phylogeny to reduce the parameters in the phylogeny and improve the resolution in the rest of the tree.

Phylogeny estimation

Time calibrated phylogenies were estimated in BEAST v. 1.8.0 (Drummond & Rambaut 2007). BEAST simultaneously estimates divergence times, topology and rates, and is preferable to previous relaxed clock methods (Couvreur et al. 2008). The best fitting evolutionary model was identified in each case by the Akaike information criterion (AIC, Sanderson 2002) as implemented by MEGA.

For the magnoliids, four independent runs of 15 million generations were sampled every 1000 generations. Log files were examined in Tracer to check for convergence and the default burn in of 10% was confirmed to be appropriate. Log and tree files were combined using LogCombiner, with resampling set to every 6000 generations and burn-in at 10%. The resulting 9000 samples were examined in Tracer to confirm good support (all parameters with ESS>200; BEAST manual); and the 9000 tree files were summarised as a maximum clade credibility tree using TreeAnnotator, with median values used to annotate the nodes. For the Rubiaceae, four independent runs of 20 million generations were sampled, resampled and combined using the same process.

Phylogenies may be calibrated in absolute time using either fossil evidence or secondary calibration points derived from previous phylogenies (which still rely ultimately on fossil evidence). For the magnoliids, a normal prior probability distribution with a mean of 91 mya and a standard deviation of 1.5 was put on the age of the Annonaceae stem (enclosing possible ages from c. 89 to 93 mya following Couvreur et al. (2008), and agrees well with other estimates derived from other approaches (Pirie & Doyle 2012 for a discussion). For the Rubiaceae, a normal prior probability distribution with a mean of 90 mya and a standard deviation of 3 (77 – 90 – 105 mya) was put on the age of the Rubiaceae stem following Bremer & Erikkson (2009). Divergence times for both phylogenies were estimated under a log normal relaxed clock method using the Yule model of speciation (Gernhard 2008).

Endemism

Species were annotated with their Star rating. Species with distributions entirely outside of Upper Guinea were coded as 'African' species. When prepared for publication, Stars will be updated for non-Upper Guinean species.

Results

59% of Upper Guinean magnoliid species were included in the phylogeny, and c. 14% of other African magnoliids. This brings the total sampling of African magnoliid species to 25% (Table 7.1). 38% of Upper Guinean Rubiaceae species and 6% of other African congeneric species were included in the phylogeny (Table 7.2).

Table 7.1. Magnoliids taxon sampling. Proportion of species by genera and families sampled within the magnoliids phylogeny (Figure 7.1). Upper Guinean species estimates follow Jongkind 2004, African species follow APD.

Family	**Genera in Upper Guinea**	**Upper Guinean species**		**Other African species**		**Overall**	
		Sampled	Total	Sampled	Total	Sampled	Total
Aristolochiaceae	*Aristolochia*	0	1	0	7	0	8
	Pararistolochia	2	6	1	5	3	11
Piperaceae	*Piper*	0	3	0	1	0	4
	Peperomia	2	2	0	10	2	12
Hernandiaceae	*Illigera*	0	2	0	2	0	4
Lauraceae	*Beilschmiedia*	3	5	2	70	5	75
Annonaceae	*Annickia*	1	1	3	8	4	8
	Anonidium	0	1	0	3	0	4
	Artabotrys	4	7	2	18	6	25
	Cleistopholis	1	1	1	2	2	3
	Duguetia	2	2	1	2	3	4
	Friesodielsia	3	4	1	3	4	7
	Greenwayodendron	1	1	0	1	1	2
	Hexalobus	2	2	1	3	3	5
	Isolona	3	5	0	10	3	15
	Mischogyne	1	1	1	1	2	2
	Monanthotaxis	4	13	6	30	10	43
	Monocyclanthus	1	1	0	0	1	1
	Monodora	4	4	2	11	6	15
	Neostenanthera	2	2	1	2	3	4
	Piptostigma	1	2	2	10	3	12
	Polyceratocarpus	1	1	0	6	1	7
	Uvaria	6	16	11	43	17	59
	Uvariastrum	2	2	1	5	3	7
	Uvariodendron	3	3	2	12	5	15
	Uvariopsis	2	2	4	14	6	16
	Xylopia	7	8	0	24	7	32
Myristicaceae	*Coelocaryon*	2	2	1	2	3	4
	Pycnanthus	1	2	0	1	1	3
Totals:		**60**	**102**	**43**	**306**	**103**	**407**
% sampled:			**59%**		**14%**		**25%**

Table 7.2. Rubiaceae taxon sampling. Proportion of species by genera and families sampled within the Rubiaceae phylogeny (Figure 7.2). Upper Guinean species estimates follow Jongkind 2004, African species follow APD.

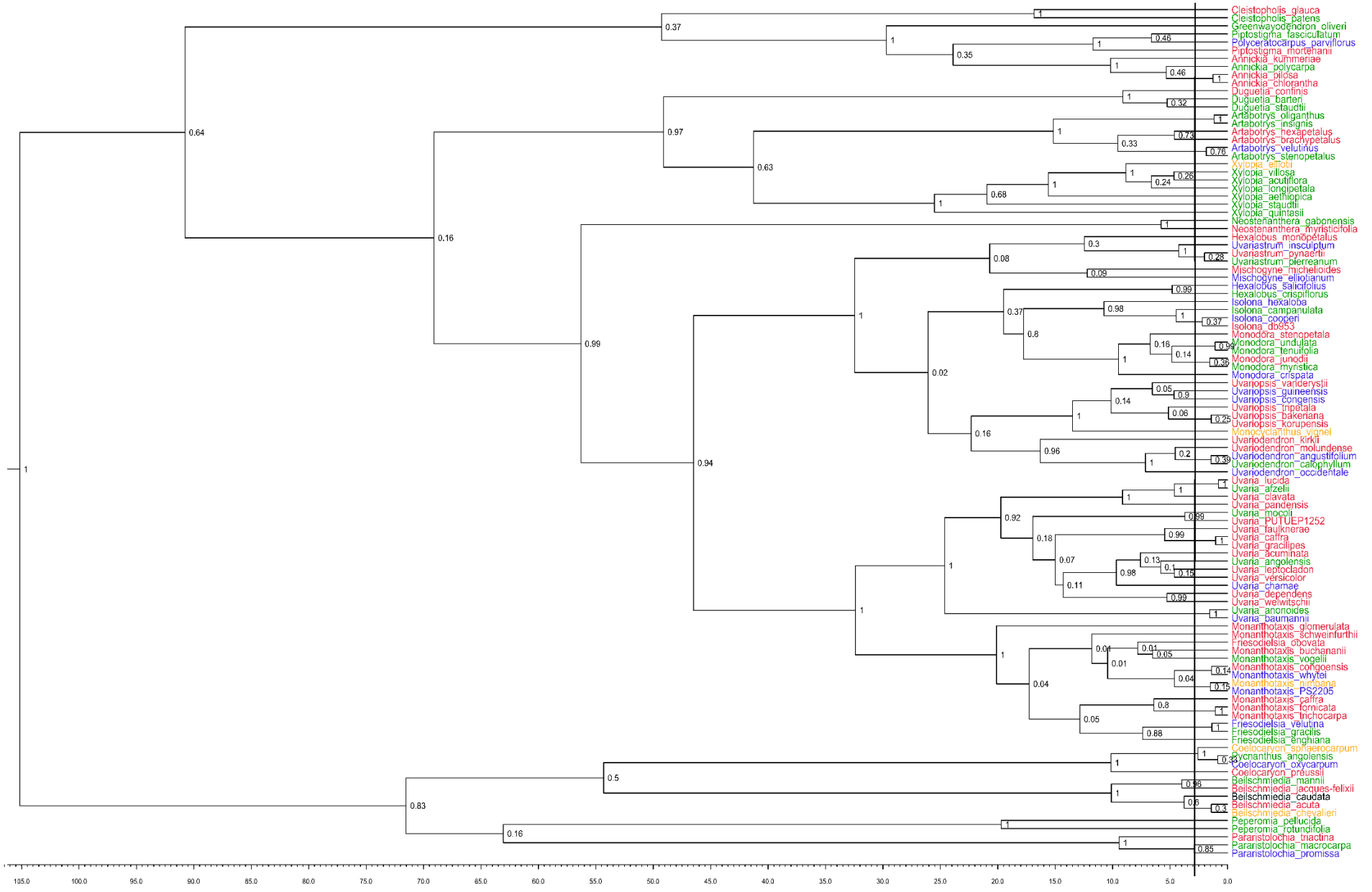
Genera in UG	UG species		Other African species		Overall	
	Sampled	**Total**	**Sampled**	**Total**	**Sampled**	**Total**
Aidia	1	1	0	8	1	9
Argocoffeopsis	1	4	0	5	1	9
Atractogyne	0	1	0	1	0	2
Aulacocalyx	2	2	2	8	4	10
Belonophora	1	1	2	4	3	5
Bertiera	3	8	1	29	4	37
Calycosiphonia	2	2	0	1	2	3
Chassalia	2	7	0	35	2	42
Chazaliella	2	6	1	14	3	20
Coffea	6	7	8	40	14	47
Corynanthe	1	1	0	2	1	3
Craterispermum	2	3	1	12	3	15
Cremaspora	0	1	0	1	0	2
Cuviera	3	3	1	18	4	21
Dictyandra	0	1	0	1	0	2
Didymosalpinx	0	1	0	3	0	4
Euclinia	1	1	0	1	1	2
Gaertnera	1	5	1	7	2	12
Gardenia	2	3	5	17	7	20
Geophila	1	3	0	7	1	10
Hallea	2	2	1	1	3	3
Hutchinsonia	1	2	0	0	1	2
Hymenocoleus	0	7	0	5	0	12
Hymenodictyon	1	2	1	2	2	4
Ixora	2	9	6	34	8	43
Keetia	1	13	1	19	2	32
Lasianthus	1	2	0	14	1	16
Leptactina	1	3	1	16	2	19
Macrosphyra	0	1	0	2	0	3
Massularia	1	1	0	0	1	1
Monosalpinx	0	1	0	0	0	1
Morelia	0	1	0	0	0	1
Morinda	2	4	0	5	2	9
Multidentia	0	1	0	9	0	10
Mussaenda	1	13	1	20	2	33
Nauclea	1	4	0	1	1	5
Nichallea	0	1	0	0	0	1
Otomeria	0	1	0	7	0	8
Oxyanthus	3	6	3	26	6	32
Parapentas	0	1	0	2	0	3
Pauridiantha	2	4	1	34	3	38
Pausinystalia	1	1	2	4	3	5
Pavetta	7	20	9	203	16	223

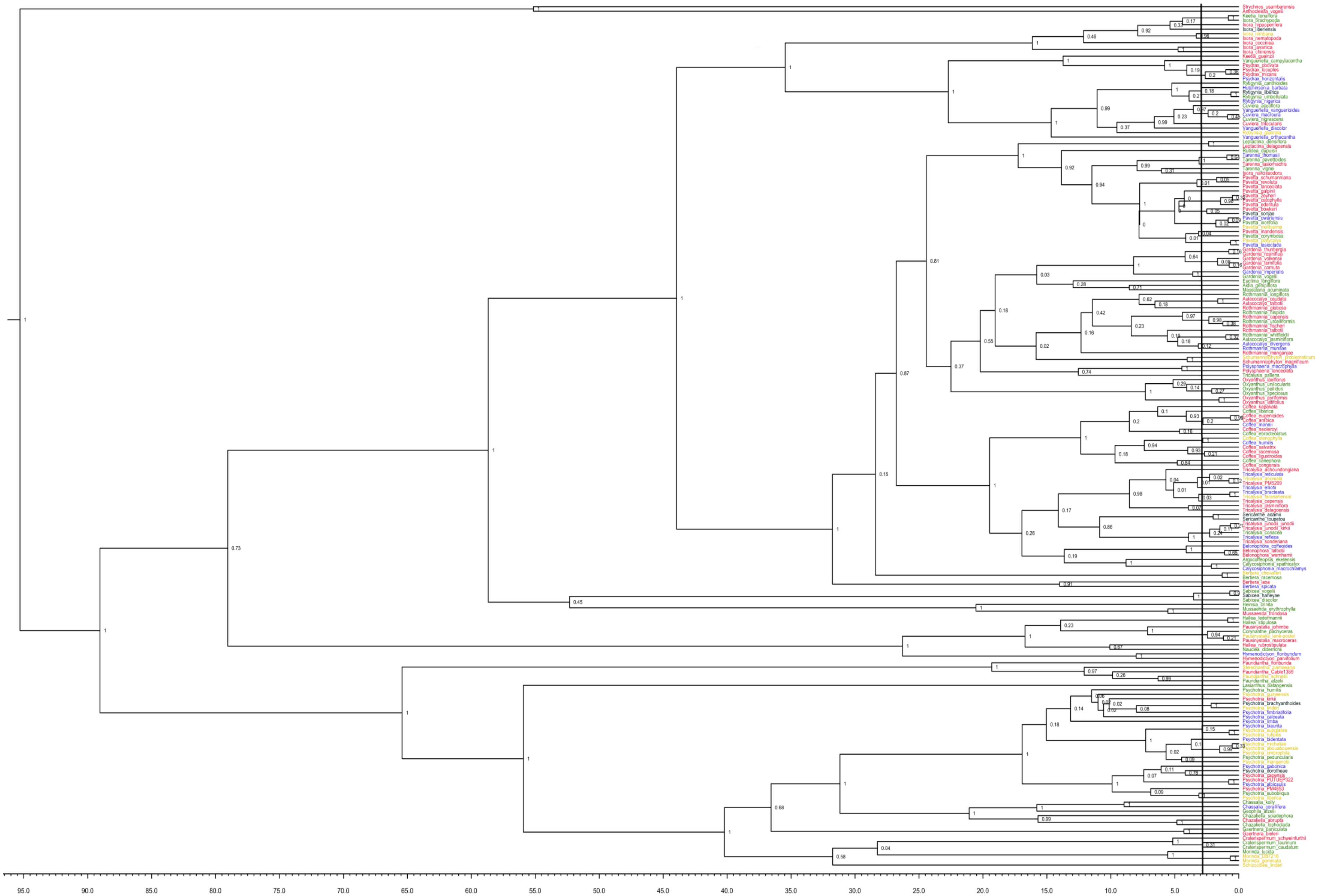
Genera in UG	UG species		Other African species		Overall	
	Sampled	**Total**	**Sampled**	**Total**	**Sampled**	**Total**
Pleiocoryne	0	1	0	0	0	1
Poecilocalyx	0	1	0	3	0	4
Polysphaeria	1	2	1	11	2	13
Pouchetia	0	3	0	2	0	5
Psychotria	19	43	4	176	23	219
Psydrax	1	7	3	27	4	34
Pyrostria	0	1	0	7	0	8
Robynsia	1	1	0	0	1	1
Rothmannia	5	5	5	17	10	23
Rutidea	1	5	0	15	1	20
Rytigynia	4	6	0	68	4	74
Sabicea	3	17	0	66	3	83
Sacosperma	0	2	0	0	0	2
Schizocolea	1	1	0	1	1	2
Schumanniophyton	1	1	1	2	2	3
Sericanthe	2	4	0	17	2	21
Sherbournia	0	2	0	11	0	13
Stelechantha	1	1	0	3	1	4
Tarenna	3	11	1	34	4	45
Tricalysia	8	11	5	56	13	67
Trichostachys	0	1	0	10	0	11
Uncaria	0	2	0	2	0	4
Vangueriella	4	6	0	11	4	17
Virectaria	0	2	0	6	0	8
Totals:	**113**	**297**	**68**	**1163**	**181**	**1461**
% sampled:		**38%**		**6%**		**12%**

Dated phylogenies from both analyses are shown in Figures 7.1 & 7.2.

Figure 7.1. Maximum clade credibility tree for African magnoliids. Node values are posterior probability values. Axis show time in millions of years; line designates the beginning of the Pleistocene (2.8 mya). Colours represent the global distribution of the species by Star (Green, Blue, Gold & Black), and red for African species not found in Upper Guinea. The phylogeny is available as supporting information (S7.1).

Figure 7.2. Maximum clade credibility tree for African Rubiaceae. Node values are posterior probability values. Axis show time in millions of years; line designates the beginning of the Pleistocene (2.8 mya). Colours represent the global distribution of the species by Star (Green, Blue, Gold & Black), and red for African species not found in Upper Guinea. The phylogeny is available as supporting information (S7.2).


Cleistopholis_glauca
Cleistopholis_patens
Greenwayodendron_oliveri
Piptostigma_fasciculatum
Polyceratocarpus_parviflorus
Piptostigma_mortehanii
Annickia_kummeriae
Annickia_polycarpa
Annickia_pilosa
Annickia_chlorantha
Duguetia_confinis
Duguetia_barteri
Duguetia_staudtii
Artabotrys_oliganthus
Artabotrys_insignis
Artabotrys_hexapetalus
Artabotrys_brachypetalus
Artabotrys_velutinus
Artabotrys_stenopetalus
Xylopia_elliotii
Xylopia_villosa
Xylopia_acutiflora
Xylopia_longipetala
Xylopia_aethiopica
Xylopia_staudtii
Xylopia_quintasii
Neostenanthera_gabonensis
Neostenanthera_myristicifolia
Hexalobus_monopetalus
Uvariastrum_insculptum
Uvariastrum_pynaertii
Uvariastrum_pierreanum
Mischogyne_michelioides
Mischogyne_elliotianum
Hexalobus_salicifolius
Hexalobus_crispiflorus
Isolona_hexaloba
Isolona_campanulata
Isolona_cooperi
Isolona_db953
Monodora_stenopetala
Monodora_undulata
Monodora_tenuifolia
Monodora_junodii
Monodora_myristica
Monodora_crispata
Uvariopsis_vanderystii
Uvariopsis_guineensis
Uvariopsis_congensis
Uvariopsis_tripetala
Uvariopsis_bakeriana
Uvariopsis_korupensis
Monocyclanthus_vignei
Uvariodendron_kirkii
Uvariodendron_molundense
Uvariodendron_angustifolium
Uvariodendron_calophyllum
Uvariodendron_occidentale
Uvaria_lucida
Uvaria_afzelii
Uvaria_clavata
Uvaria_pandensis
Uvaria_mocoli
Uvaria_PUTUEP1252
Uvaria_faulknerae
Uvaria_caffra
Uvaria_gracilipes
Uvaria_acuminata
Uvaria_angolensis
Uvaria_leptocladon
Uvaria_versicolor
Uvaria_chamae
Uvaria_dependens
Uvaria_welwitschii
Uvaria_anonoides
Uvaria_baumannii
Monanthotaxis_glomerulata
Monanthotaxis_schweinfurthii
Friesodielsia_obovata
Monanthotaxis_buchananii
Monanthotaxis_vogelii
Monanthotaxis_congoensis
Monanthotaxis_whytei
Monanthotaxis_nymphana
Monanthotaxis_PS2205
Monanthotaxis_caffra
Monanthotaxis_fornicata
Monanthotaxis_trichocarpa
Friesodielsia_velutina
Friesodielsia_gracilis
Friesodielsia_enghiana
Coelocaryon_sphaerocarpum
Sycnanthus_angolensis
Coelocaryon_oxycarpum
Coelocaryon_preussii
Beilschmiedia_mannii
Beilschmiedia_jacques-felixii
Beilschmiedia_caudata
Beilschmiedia_acuta
Beilschmiedia_chevalieri
Peperomia_pellucida
Peperomia_rotundifolia
Pararistolochia_triactina
Pararistolochia_macrocarpa
Pararistolochia_promissa
105.0
100.0
95.0
90.0
85.0
80.0
75.0
70.0
65.0
60.0
55.0
50.0
45.0
40.0
35.0
30.0
25.0
20.0
15.0
10.0
5.0
0.0



95.0
90.0
85.0
80.0
75.0
70.0
65.0
60.0
55.0
50.0
45.0
40.0
35.0
30.0
25.0
20.0
15.0
10.0
5.0
0.0

In the magnoliids phylogeny, families, tribes, and the majority of genera are monophyletic and have good support (Figure 7.1). The topology [[MAGNOLIALES + LAURALES] [CANELLALES + PIPERALES]] was not recovered; rather the Annonaceae were found as sister to the other families. Relationships between the orders of the magnoliids have been found by other researchers to be hard to reconstruct, and it is perhaps unsurprising that a different topology has been recovered here using only one gene and with unbalanced sampling across the orders. Within the Annonaceae, all African subfamilies and tribes as presented in Chatrou et al. (2012) have been recovered with good support, although the subfamilies Ambavioideae and Malmeoideae are placed as sister to each other, rather than Malmeoideae being sister to Annonoideae. Tribal divisions and topology within the Annonoideae are as expected. Within the Monodoreae, relationships between *Hexalobus*, *Uvariastrum* & *Mischogyne* are poorly resolved, as is their relationship to *Isolona*, *Monodora*, *Uvariopsis* & *Uvariodendron*, suggesting a lack of resolution in the marker *matK.* In the Uvarieae, *Monanthotaxis* & *Friesodielsia* are not well distinguished. The monophyly of *Friesodielsia* has been questioned and here the genus is paraphyletic with respect to *Monanthotaxis* in Chatrou et al. (2012).

The three subfamilies of Rubiaceae were recovered with maximum support and the stem ages broadly agree with the estimates of Bremer & Eriksson (2009) (Figure 7.2). As with the magnoliids phylogeny, support and monophyly for the genera is mostly strong, although the (closely related) genera *Robynsia, Vangueriella, Cuviera, Hutchinsonia, Rytigynia, Keetia* and *Psydrax* of the Ixorideae are not well resolved and are mostly paraphyletic with respect to each other.

Species with strong topological support show the same spread of ages as for all species. The following results, discussion and analyses use only species with strong topological support at their highest internal node. The great majority of species have estimated (maximum) median

divergence dates inside of the last 15 million years; 38% of Rubiaceae and 32% of magnoliid species have a median date of divergence from their nearest sampled relative during the Pleistocene (Figure 7.3, Figure 7.4).

Figure 7.3. Histograms showing median estimated ages of divergence of African magnoliid species. (a) All species in the magnoliid phylogeny. (b) Species with strong topological support. Species divergence ages are available as supporting information (S7.3).

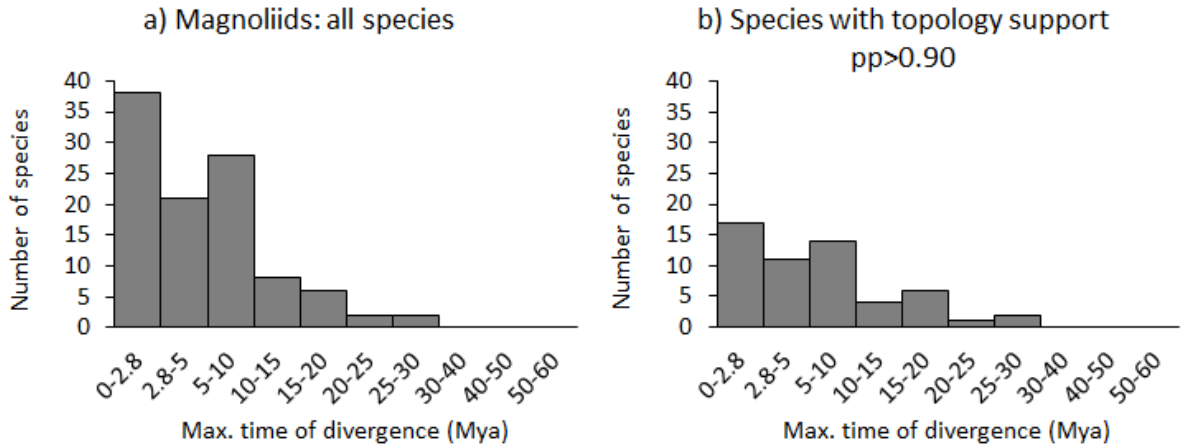
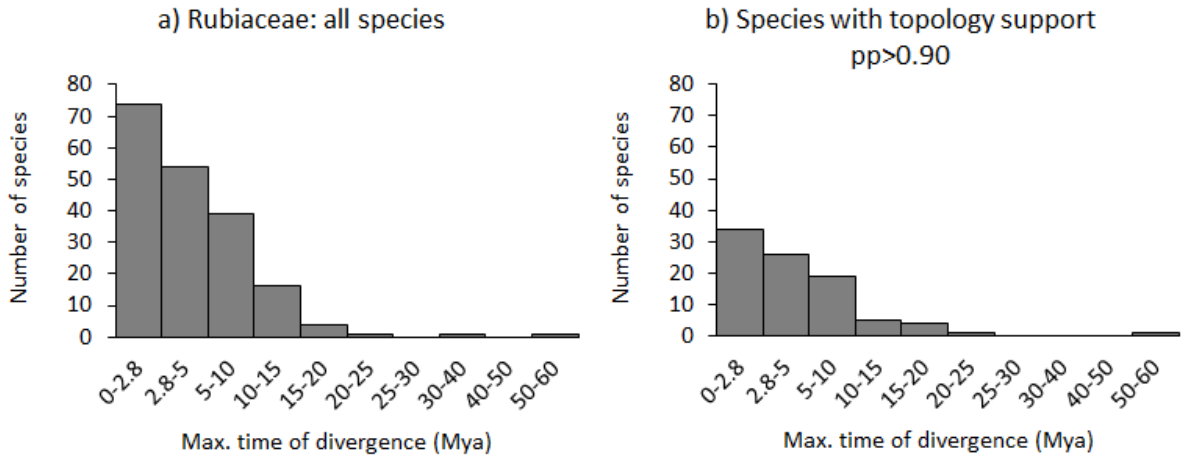
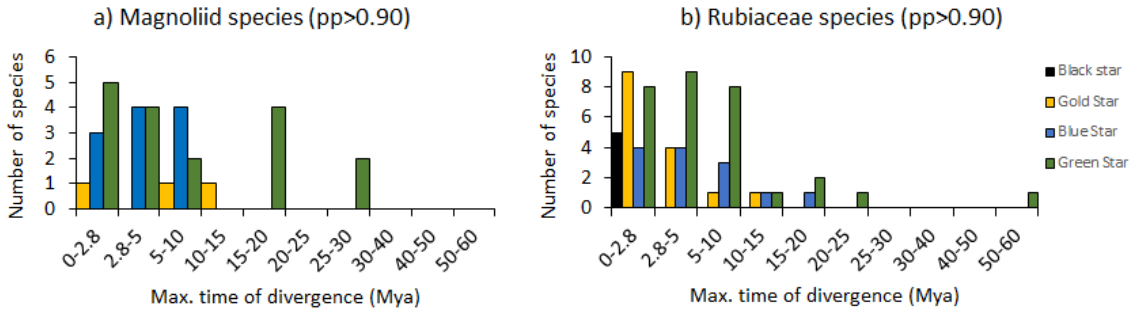


Figure 7.4. Histograms showing median estimated age of divergence of African Rubiaceae species. (a) All species in the ingroup of the Rubiaceae phylogeny. (b) Species with strong topological support. Species divergence ages are available as supporting information (S7.3).



A Welch two-sample t-test on the mean (median) ages of Black and Gold Star Rubiaceae species with posterior probability support greater than 0.9 (2.9 mya) vs. Blue and Green Star Rubiaceae species with posterior probability support greater than 0.9 (7.1 mya) shows the means to be significantly different from each other ($p<0.005$), $t(55)=-2.76$ with the globally rarer Black and Gold Star species younger than widespread Green and Blue Star species. Mean ages for Black and Gold Star species, and Blue and Green Star species, were calculated as the mean of the median species ages as summarised in the phylogenies. The number of Black and Gold Star species in the Magnoliids is too small to justify a statistical comparison of mean ages.

Figure 7.5. Species divergence dates by Star. (a) Magnoliids, (b) Rubiaceae. Colours represent categories of global rarity for Upper Guinean species coloured by Star (Black, Gold, Blue, Green).



For the magnoliids, no 95% age dates fall inside of 2.8 mya. 11 species have median divergence dates inside of 2.8 mya with very strong support (Table 7.3). For the Rubiaceae, 14 Upper Guinean species have 95% confidence intervals of their median age falling inside 2.8 mya. 28 species have median ages falling inside 2.8 mya (Table 7.4).

Table 7.3. Species sister groups in the magnoliid phylogeny with strong support at their joining node and divergence dates within 2.8 mya.

Species	Support	Age
Artabotrys oliganthus (GN) & *A. insignis* (GN)	1	1.18
Uvaria afzelii (GN) and *U. lucida* (E. Africa)	1	0.79
Uvaria anonoides (GN) & *U. baumannii* (BU)	1	1.56
Friesodielsia velutina (BU) & *F. gracilis* (GN)	1	1.36
Monodora tenuifolia (GN) & *M. undulata* (ex. *M. brevipes*) (GN)	0.99	1.08
Coelocaryon sphaerocarpum (GD)	1	2.58

Table 7.4. Species sister groups in the Rubiaceae phylogeny with strong support at their joining node and divergence dates within 2.8 mya.

Species	Support	95% inside?	Age
Pavetta platycalyx (GD) *& P. lasioclada* (BU)	1	Y	1.06
Rytigynia liberica (BK) *& R. umbellulata* (GN)	1	Y	1.08
Morinda geminata (GD) *& M. DB7216**	1	Y	1.13
Sabicea harleyae (BK) *& S. vogelii* (GN)	0.90	Y	1.17
Tricalysia faranahensis (GD) & *T. bracteata* (BU)	1	Y	1.17
Psychotria subglabra (GD) *& P. rufipilis* (GD)	1	Y	1.23
Psychotria albicaulis (BU) *& P. PUTUEP322*	1	Y	1.26
Keetia tenuiflora (GN) *& Ixora brachypoda* (GN)	1		1.29
Hallea stipulosa (GN) *& H. ledermannii* (GN)	1		1.33
Bertiera chevalieri (GD) & *B. racemosa* (GN)	1		1.77
Sericanthe toupetou (BK) *& S. adamii* (BK)	1		2.48
Calycosiphonia macrochlamys (BU) *& C. spathicalyx* (GN)	1		2.61
Psychotria brachyanthoides (BK) *& P. linderi* (GD)	1		2.63
Pavetta edulenta	0.98		1.42
Psychotria ombrophila	0.99		1.97

Figure 7.6. Maps from the African Plants Database (CJB 2014) showing the distributions of *Sericanthe toupetou* and *S. adamii.*

Redacted for online dissemination: See the African Plants Database online at http://www.ville-ge.ch/musinfo/bd/cjb/africa/recherche.php?langue=an

Figure 7.7. Maps from the African Plants Database (CJB) and Lachenaud 2013 showing the distributions of *Psychtoria linderi* and *P. brachyanthoides*.

Redacted for online dissemination: See Lachenaud (2013) and the African Plants Database online at http://www.ville-ge.ch/musinfo/bd/cjb/africa/recherche.php?langue=an

Discussion

Due to the impossibility of guaranteeing that sister species have been included in the phylogenies, species ages must be taken as estimates of their (median) maximal age. Even with this bias towards older age estimates, the greater majority of sampled African species have differentiated inside the last 15 million years, and around a third of sampled species are likely to have had Pleistocene origins. 26 species were shown to have Pleistocene origins with 95% confidence, of which 16 also had strong topological support. This is rather in contrast to the commonly held belief that 2.8 million years is insufficient to observe speciation events in Africa. However, elsewhere in the tropics, large radiations inside of 2 million years have been observed and accepted (e.g. 80 *Lupinus* species in the Andes, Hughes & Eastwood 2006). The degree to which the ages recovered here are artefactual as a result of the choices of secondary calibration points for example, or the topology recovered here, would need to be explored.

23 Black and Gold Star species appear in these phylogenies with well resolved sister relationships. Of these, 15 have median divergence dates in the Pleistocene (8 have 95% confidence intervals entirely within the Pleistocene), a further four diverged 2.8-5 mya, and only four are older than 5 mya. It would seem from this result that the majority of locally endemic species, characteristic of hotspots, are likely to be species with recent origins (or at least species with extant close relatives in the phylogenies), particularly as age estimates are more likely to get more recent with better sampling. Of the species that are older than 5 million years, one is *Monocyclanthus vignei* (13.48 mya), one of only three genera endemic to Upper Guinea (all are monotypic). By both morphological and phylogenetic analyses it is most closely related to *Uvariodendron,* a genus that is relatively well represented in this phylogeny (3/3 Upper Guinean species, 2/12 other African species). In this phylogeny, *Monocyclanthus vignei* is placed as sister to *Uvariopsis* on a long

branch in a clade including *Uvariodendron*, although topologies where *M. vignei* is nested within *Uvariopsis* have been recovered. Assuming the *matK* topology presented is to be trusted (Figure 7.1), it is interesting that this 'paleoendemic' is sufficiently different from its relatives to have been recognised as a distinct genus. Two of the other species > 5 mya, *Stelechantha ziameana* (12.56 mya) and *Schizocolea linderi* (32.2 mya), belong to species poor genera with only three and one other species respectively found in Lower Guinea, none of which were sampled. It seems likely that these 'old' age estimates are not plausible. *Schizocolea linderi* and *S. ochreata* have even been flagged as examples of young, vicariant species using morphological characteristics (Wieringa & Poorter 2004). The only other contender for a paleoendemic species from this sample is *Xylopia elliotii* (8.85 mya).

The Black star sister species *Sericanthe toupetou & S. adamii* show different distributions in the putative Upper Guinean refugia of Maley 1996: *S. toupetou* in SW Ghana and SE Liberia/SW Ivory Coast hotspots, and *S. adamii* in the Nimba mountains and further west in Liberia. According to the phylogeny, these two species are likely to have diverged from each other during the Pleistocene (Figure 7.6). The Black and Gold star sister species *P. brachyanthoides* and *P. linderi* have similar distributions, except that *P. brachyanthoides* is also found in SW Ghana. According to the phylogeny of Figure 7.2, these two species are likely to have diverged from each other during the Pleistocene (Figure 7.7). At least these two species pairs would seem to support the lowland forest Pleistocene refuge hypothesis of Table 7.1.

The level of species sampling within Upper Guinea is relatively good (c. 60% and 40%) and compares favourably with many published phylogenies. Sampling of other African species is much lower. Balancing the species sampling by taxon, region and Star would give more convincing results. If many Upper Guinean species have their closest relatives geographically close to them,

then many of the non-Upper Guinean Rubiaceae (e.g. those found in East Africa, or the Cape) might have little relevance to the sister relationships recovered in the phylogeny. However, one potential limitation associated with low sampling of African species is that more widespread species may be more likely to have a sister with a distribution outside of Upper Guinea, or even outside Africa. If the more endemic species are more likely to have their close relatives included in the phylogeny than globally widespread species are, then this will have systematically biased the age estimates of Black and Gold Star species towards younger ages relative to Green and Blue Star species in these current analyses. The younger ages of Black and Gold Star species relative to Blue and Green Star species presented here should therefore be taken as a tentative result worthy of deeper investigation.

The patterns presented here should be supported with investigation of more genera. The method employed here, of breaking the angiosperms into smaller clades, can be repeated to cover more of the angiosperm phylogeny, with each phylogeny being treated as a sample of the flora. Phylogenies with fewer species relative to the number of sequence characters can be estimated more robustly, particularly if species sampling is well balanced phylogenetically.

Although the phylogenies of Figure 7.1 and 7.2 are on the whole reasonable, and a good proportion (c. half) of species level relationships are well resolved, the lack of consistently good resolution within several genera makes inferring geographic patterns of species origination hard to discern with confidence, at least with *matk*. Relationships within a genus, even those with relatively low topological support, could be more confidently interpreted if the pattern is corroborated by an independent source (e.g. morphology).

Supporting Information

Appendix S7.1

Magnoliids phylogeny

Appendix S7.2

Rubiaceae phylogeny

Appendix S7.3

Species branch lengths

Chapter 8

General Conclusions

The earliest specimens recorded in the RAINBIO dataset date from 1780 (Sosef et al. 2017). A great deal of botany has been conducted across tropical Africa in the last 250 years, and in the 10 years since the work of the BIOTA-BISAP and ECOSYN working groups.

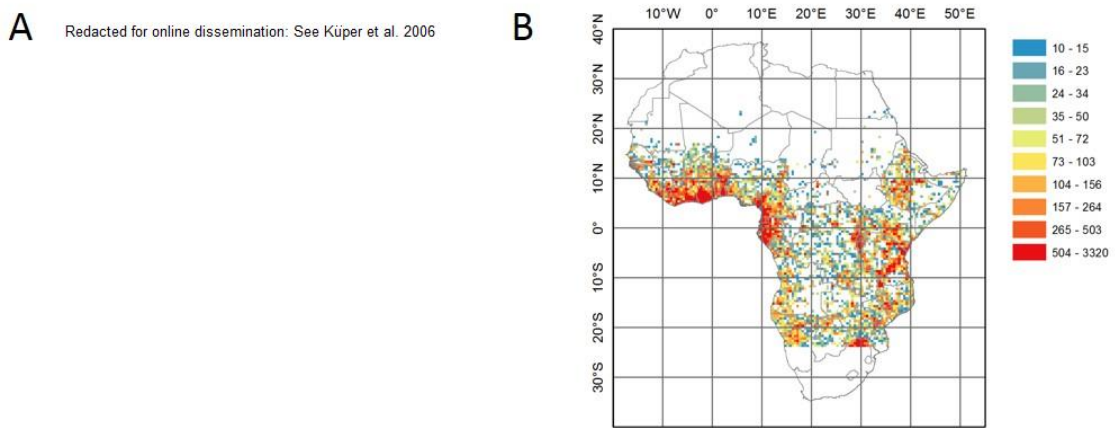
TRAFRICA, and the analyses it has facilitated, represents a significant development in tropical African botany. The exercise has revealed a number of challenges and surprises, including insights and developments in the way that we digitally geolocate records as a community; the high proportion of singleton species in the digital flora; the need to improve the inventory of plants even at country level and at relatively coarse resolution, the integration of different file formats and software into a coherent framework, and the challenges of sharing data, rights and responsibilities amongst the community of African botanists.

There has been considerable motivation worldwide to advance herbarium digitisation projects and normalise data sharing between scientists. This has resulted in a huge increase in digital distribution data availability. Reporting around 2006, the BIOTA-BISAP database was the most comprehensive digital account of sub-Saharan Africa's flora, holding 5438 unique species x degree square records. 10 years later, and excluding Southern Africa, TRAFRICA holds 498,701 unique taxa by degree square records for tropical Africa, and the RAINBIO dataset holds 600,000 unicate specimens records across central Africa. The most significant aspect of this 90-fold increase, in my opinion, has been the mobilisation of data across the taxonomic spectrum; the digitisation of all

African type specimens in particular has meant that hypotheses can be tested now across most described plant species.

In contrast, little has changed regarding the geographic coverage of species distribution data. Figure 1 shows collection intensity in 2006 (Fig. 8.1A), and in 2016 (Fig. 8.1B). Although the further digitisation of African herbaria, for example the East African Herbarium, may improve collection coverage across some areas, the major gains from digitisation projects are likely to have already been made.

Figure 8.1. Geographic patterns in species distribution data available in 2006 and 2016. A) Collection intensity for sub-Saharan vascular plants in 2006 (from Küper et al. 2006). B) Collection intensity at half degree square resolution for tropical African vascular plants in 2016 (Chapter 2).



The big data era has highlighted the depths of our data poverty as regards range descriptions for most plant species. 25% of degree squares across Africa have <2.5% of their estimated species richness recorded, and at finer grains than this, the holes exceed the fabric. Only with a concerted effort to collect survey type data across the continent are we likely to fill the data holes apparent from Figure 8.1.

Botanists collect plants neither systematically nor randomly, even as a collective. The sum of our knowledge of species distributions are the result of nearly 250 years of independent decisions taken over where to collect plants, considering accessibility and safety, the rewards of collecting in both very unknown and relatively well known localities, the chances of making new records of species or occurrences relative to the personal and financial expense of fieldwork, the convenience of funding, and personal ambitions. Some of the present day gaps have probably been considered 'boring' by centuries of botanists, and some areas will have always been inaccessible. It may be some time yet before western Angola, central Congo, much of Somalia and South Sudan are well sampled botanically. On the other hand, countries like Uganda and Guinea, and northwestern Tanzania, northern Kenya and northern Mozambique are quite undersampled even at the degree square resolution, but have good local botanists and strong botanic infrastructure and are generally safe and accessible. For these areas, a programme of collection across degree squares, targeting new species-by-degree-square records would be very informative, at least for ecological, biogeographic, floristic, and spatially explicit phylogenetic analyses. Such a programme of targeted collection may also shed light on the taxonomic validity and likely distribution of the c. 20% of tropical African species known only from a single record.

We intend to maintain TRAFRICA. Specifically, we continue to digitise floras and checklists to increase the coverage of distribution records, we will incorporate any new species x location records from RAINBIO not already in TRAFRICA, and at some stage the species names and synonymy will need to be updated following publications. And of course, we continue to carry out survey work in tropical Africa.

In Chapter 3 the extent to which multivariate statistical methods and large datasets have called into question the definition of 'Upper Guinea' was investigated. Upper Guinea cannot be defined

without considering the definition of Guineo-Congolia, which must be defined in the context of its adjacent floras: we therefore proposed a set of chorological regions for tropical Africa. Given the many false absences, and no doubt some false presences in the TRAFRICA dataset, it is perhaps surprising how strong and consistent the signals from the African flora have been. We found that White's regional framework fitted our tropical African dataset well, but slightly different boundaries increased total endemism. Having submitted the TRAFRICA dataset to a wide range of clustering techniques, we concluded that Upper Guinea is indeed part of Guineo-Congolia, and the result of Linder et al. 2012 was likely due to gaps in the BIOTA-BISAP dataset for Upper Guinea and an underpowered analysis. All forest zones west of the Dahomey Gap (Upper Guinea) could be recovered as distinct from Nigerian Guinea, giving Upper Guinea a Dahomey Gap termination in the east. However, we also showed that drier Upper Guinean forests are only weakly distinct from Nigerian forests east of the Dahomey Gap to Cross River. Within this Nigerian forest zone, western cells are most similar to Upper Guinea, and eastern forest cells are most similar to Lower Guinea. Therefore, naming the Nigerian forests from Togo to the Cross River as 'Nigerian Guinea' and thinking of them as an Upper-to-Lower Guinea transition zone reflects the chorological patterns most closely. Although combining this transition zone with Upper Guinea to produce a Guineo-Congolian sub-region from the west coast of Sierra Leone to Cross River is supportable, the sub-region should be called Upper Guinea + Nigerian Guinea to avoid ambiguity.

Identifying areas of high biodiversity is an established way to prioritize areas for conservation, but global approaches have been criticized for failing to render global biodiversity value at a suitable scale for local management. In Chapter 4 we presented the global range of 40,000 tropical African plant taxa in four categories of global range, and mapped bioquality across tropical Africa at one degree square resolution. We confirmed the broad significance of the Horn of Africa, Guinean forests, coastal forests of east Africa, and Afromontane regions for plant biodiversity. We found

the highest bioquality values to be associated with Somalia: this was the strongest deviation from previous maps of global endemism. The map of Linder (2001) did, however, show very high range size rarity measures for southern Somalia; as northern Somalia was not included in the analysis this result was considered likely artefactual, and was not picked up in the summary of hotspots.

We revealed the variation in bioquality within these broad regions and others, for example within Upper Guinea. Patterns of endemism within Upper Guinea were most congruent with those of Wieringa & Poorter (2004, considering rare species within Upper Guinea only), namely high global endemism associated with Nimba, SE Liberia, SW Ivory Coast and SW Ghana, as well as generally elevated scores for western Liberia and Sierra Leone. The framework presented in Chapter 4 offers practitioners a quantitative, scalable and replicable approach for measuring the irreplaceability of particular local areas for global biodiversity conservation, and comparing those areas within their global and regional context.

In Chapter 5, patterns of endemism within two of Upper Guinea's bioquality hotspots, areas which have been posited as climatic refugia, were explored at finer geographic resolution. We showed that increasing disturbance, measured with the Pioneer Index, reduced bioquality significantly in both datasets. Disturbance reduced bioquality more rapidly in areas that would otherwise be 'hot' than in areas that are otherwise 'cold'. Although disturbance can reduce bioquality, there is a separate need to explain why high bioquality in the absence of modern disturbance is restricted to the bioquality hotspots, and the history and isolation of particular environments must have a role here. Latitude is well correlated with climatic gradients in SW Ghana, and annual rainfall in particular is a significant predictor of the high global endemism associated with the far south. In the Nimba Mountains, where the altitudinal range is large, higher global endemism is found at higher altitudes, particularly over 1000 m. Areas of high rainfall and high altitude are

geographically restricted within Upper Guinea, which must contribute to the small global range of the species found in these areas. The Pioneer Index is the only significant term retained in both the SW Ghana and the Nimba models, and suggests a degree of generalisability of the negative impact of disturbance on the distribution of globally rare species. Overall, we could explain 59% and 57% of the variation in GHI at sample level within the two datasets, without invoking past climate changes.

Bioquality scores measured and reported for particular samples in SW Ghana have not changed overall over 20 years, although the suite of species responsible for this perception have changed. As much as 40% of species have changed Star over 20 years, even in the relatively well known reserves of south west Ghana. However, we showed that bioquality scores are much more stable than our understanding of species' identity and ranges, which allows for some consistency in results and conservation priority setting over time and in the face of partial information.

In Chapter 6, we built a model for global endemism within tropical Africa, which can predict known GHI values with an error of around 100 GHI units. Transferability tests suggested that, when no training data are given for Ghana, the model could predict GHI scores of Ghanaian cells with acceptable accuracy. This model is one way in which conclusions may be drawn about the likely conservation value of parts of tropical Africa for which distribution data may be lacking for some time.

All included terms held some predictive power over GHI. The biogeographic framework of White (1983) and Sayre et al. (2013) were the most important predictors of GHI across tropical Africa. Climatically, maximum temperature of the warmest month was the most important variable, at least in the IPSL model. The stability of climate parameters over the last 21,000 years all held predictive power for GHI, and in particular the stability of annual rainfall. Surficial lithology,

longitude and latitude were amongst the top predictors; altitude and slope were of middling importance at half degree square resolution. Both stability, and instability of rainfall since the LGM were related to higher bioquality scores.

In Chapter 7 we quantified divergence dates for Upper Guinean species, for sampled species of the magnoliids and Rubiaceae, and looked for differences in the timing and tempo of lineage diversification between globally rare, and globally widespread species. 26 species (of 284) were shown to have Pleistocene origins with 95% confidence, of which 16 also had strong topological support; around one third of species included in the phylogenies had estimated median divergence dates within 2.8 million years. This is in contrast to the assertion that 2.8 million years is insufficient to observe speciation events in Africa. Due to the fact that many potential sister species were unsampled, these estimates must be minimum values.

Globally rarer Black and Gold Star species were shown to be 'younger' than widespread Green and Blue Star species (rarer species had closer sampled extant relatives). At least *Sericanthe toupetou* and *S. adamii* and *Psychotria brachyanthoides* and *P. linderi* would seem to support the lowland forest Pleistocene refuge hypothesis, due to their sister relationship and current distribution in different putative Pleistocene refugia, and Pleistocene origins. *Monocyclanthus vignei,* belonging to a monotypic genus, had a divergence date from its closest relatives of 13.5 mya, and its refugial distribution alongside other 'neoendemics' lends support to the museum hypothesis, as not all local endemics of Upper Guinea have Pleistocene origins. However, these paleondemics can only be shown to be more recently evolved as the botanical exploration continues – we can wonder how long we need wait before a new *Monocyclanthus* species is discovered in Lower Guinea.

Globally rare species are aggregated at all scales, not least from one degree square to the sample level. Given the demonstrated importance of both climatic stability and change for high GHIs, at

least over 21,000 years; and the tentative evidence that speciation has both been substantial over Pleistocene time frames; and given that some globally rare species are from more phylogenetically distinct lineages, we can find evidence for at least two mechanisms by which globally rare species have been introduced, and localised within, the species pool. Nevertheless, at least in the two Upper Guinean hotspots studied in detail, high global endemism must at least be maintained by processes acting on more recent timescales. The localisation of modern climate and altitude are strong predictors of high local bioquality, and the signal from all of these can be effectively removed by the introduction of a more severe disturbance regime.

We have shown how conclusions can be reasonably drawn from incomplete data, so that conservation action may be taken before what we seek to document is gone. For conservation, the negative impact of disturbance on bioquality is sobering, but the consistency of GHI scores in SW Ghana is at least encouraging, and gives us confidence in the performance of the metric over time, and the possibility of maintaining high bioquality in reserves over decades. As I have done to the work of others, there will no doubt be refinements to the studies presented here over the next decades, as new datasets and methods emerge. Without assuming that our current moment in time is somehow universally correct, I take some confidence from my conclusion that little of our basic framework for tropical African floristics has been violated by my thesis, and the signals emerging from our current perspective on the tropical African flora are strong.

Concluding remarks

By combining methods from the major technical disciplines in botany, I hope that I have contributed substantially to both the fundamental unit of our science – species distribution data – while marshalling the amassed data to useful synthetic and analytic effect. I hope that the strength of my approach has been to integrate the wealth of data from both traditional and emerging sources, to have analysed it using a range of traditional and newer ecological, biogeographic, taxonomic and phylogenetic methods, and to have considered a significantly larger proportion of the angiosperms than has been tackled before. I hope to have illuminated the floristic patterns in Upper Guinea, taken in its tropical African context, and to have presented a richer interpretation of the occurrence of globally rare species within tropical Africa.

References

Adler, D., Murdoch, D., Nenadic, O., Urbanek, S., Chen, M., Gebhardt, A. … Senger, A. 2016. rgl: 3D Visualization Using OpenGL. R package version 0.95.1441. https://CRAN.R-project.org/package=rgl

African Plants Database (APD) (version 3.4.0). Conservatoire et Jardin botaniques de la Ville de Genève and South African National Biodiversity Institute, Pretoria, Retrieved June 2014, from http://www.ville-ge.ch/musinfo/bd/cjb/africa/.

Angiosperm Phylogeny Group. 2009. An update of the Angiosperm Phylogeny Group classification for the orders and families of the flowering plants: APG III. Botanical Journal of the Linnean Society 161: 105–121.

Anhuf, D., Ledru, M. -P., Behling, H., Da Cruz Jr., F. W., Cordeiro, R. C., Van der Hammen, T., Karmann, I., Marengo, J. A., De Oliveira, P.E., Pessenda, L., Siffedine, A., Albuquerque, A. L., Da Silva Dias, P. L. 2006. Paleo-environmental change in Amazonian and African rain forest during the LGM. Palaeogeography, Palaeoclimatology & Palaeoecology 39: 510–527.

Assédé, E. P. S., Adomou, A. C., and Sinsin, B. 2012. Magnoliophyta, Biosphere Reserve of Pendjari, Atacora Province, Benin. Check List 8, 642–666.

Aubréville, A. 1962. Position chorologique du Gabon. Flore du Gabon, 3, 3-11.

Baksh-Comeau, Y. S., Maharaj, S. S., Adams, C. D., Harris, S. A., Filer, D. L., and Hawthorne, W. D. 2016. An annotated checklist of the vascular plants of Trinidad and Tobago with analysis of vegetation types and botanical “hotspots.” Phytotaxa 250, 1–431.

Barthlott, W., Mutke, J., Rafiqpoor, D., Kier, G., and Kreft, H. 2005. Global Centers of Vascular Plant Diversity. Nova Acta Leopoldina NF 92 342, 61–83.

Barthlott, W., Hostert, A., Kier, G., Küper, W., Kreft, H., Mutke, J., Rafiqpoor, M. D. & Sommer, J. H. 2007. Geographic patterns of vascular plant diversity at continental to global scales. Erdkunde 61: 305–315.

Bartlein, P. J., Harrison, S. P., Brewer, S., Connor, S., Davis, B. A. S., Gajewski, K., Guiot, J., Harrison-Prentice, T., Henderson, A., Peyron, O., Prentice, I. C., Scholze, M., Seppa, H., Shuman, B., Sugita, S., Thompson, R. S., Viau, A. E., Williams, J., and Wu, H. 2011. Pollen based continental climate reconstructions at 6 and 21 ka: a global synthesis. Climate Dynamics 37: 775-802.

Baselga, A., Orme, D., Villeger, S., De Bortoli, J. & Leprieur, F. 2013. betapart: Partitioning beta diversity into turnover and nestedness components. R package version 1.3. https://CRAN.R-project.org/package=betapart

Bauer, E., Kohavi, R. 1999. An Empirical Comparison of Voting Classification Algorithms: Bagging, Boosting, and Variants. Machine Learning 36: 105-139.

Beentje, H. J., Adams, B., Davis, S. D. 1994. Regional Overview: Africa. In S. D. Davis, V. H. Heywood, & A. C. Hamilton (Eds.) Centres of Plant Diversity. A guide for their strategic conservation. Vol 1. Cambridge, UK: IUCN Publications Unit.

Benson, D. A., Cavanaugh, M., Clark, K., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J. & Sayers, E. W. 2013 Jan; Nucleic Acids Research 41 (Database issue): D36-42.

BIOTA-BISAP. BISAP (Biogeographical Information System on African Plant Diversity) has been established at the Nees Institute, University of Bonn in the context of BIOTA Africa project together with several external partners, especially Jon Lovett (University of Leeds) and Peter Linder (Zurich). For more details about the dataset and a full acknowledgement of all contributors please refer to Linder et al. 2005 and Küper et al. 2006.

Bivand, R. S., E. Pebesma, V. Gomez-Rubio. 2013. Applied spatial data analysis with R, Second edition. Springer, NY. http://www.asdar-book.org/

Bongers, F., Poorter, L. & Hawthorne, W. D. 2004. The forests of Upper Guinea: gradients in large species composition. In Poorter et al. 2004, Biodiversity of West African forests. 2004. CABI Publishing, UK.

Booth, T. H., Nix, H. A., Busby, J. R., & Hutchinson, M. F. 2014. bioclim: the first species distribution modelling package, its early applications and relevance to most current MaxEnt studies. Diversity and Distributions 20: 1–9.

Botermans, M., Sosef, M. S. M., Chatrou, L. W. & Couvreur, T. L. P. 2011. Revision of the African genus Hexalobus (Annonaceae). Systematic Botany 36: 33-48.

Braconnot, P., B. Otto-Bliesner, S. Harrison, S. Joussaume, J.-Y. Peterchmitt, A. Abe-Ouchi, M. Crucifix, E. Driesschaert, Th. Fichefet, C. D. Hewitt, M. Kageyama, A. Kitoh, A. Laîné, M.-F. Loutre, O. Marti, U. Merkel, G. Ramstein, P. Valdes, L. Weber, Y. Yu and Y. Zhao. 2007. Results of PMIP2 coupled simulations of the Mid-Holocene and Last Glacial Maximum – Part 1: experiments and large-scale features. Climate of the Past 3: 261-277.

Breiman L. 1998. Arcing Classifiers. The Annals of Statistics, 26: 801-849.

Breiman L. 1996. Bagging Predictors. Machine Learning 24: 123-140.

Breiman & Cutler's Random Forest: https://www.stat.berkeley.edu/~breiman/RandomForests/

Bremer, B. & Eriksson, T. 2009. Time tree of Rubiaceae: phylogeny and dating the family, subfamilies and tribes. International Journal of Plant Science 170: 766–793.

Brenan, J. P. M. 1978. Some aspects of the phytogeography of tropical Africa. Annals of the Missouri Botanical Garden 65: 437-378.

Brito, J. C., Tarroso, P., Vale, C. G., Martínez-Freiría, F., Boratyński, Z., Campos, J. C., Ferreira, S., Godinho, R., Gonçalves, D. V., Leite, J. V., Lima, V. O., Pereira, P., Santos, X., da Silva, M. J. F., Silva, T. L., Velo-Antón, G., Veríssimo, J., Crochet, P. -A., Pleguezuelos, J. M., & Carvalho, S. B. 2016. Conservation Biogeography of the Sahara-Sahel: additional protected areas are needed to secure unique biodiversity. Diversity and Distributions 22: 371–384.

Brooks, T. M., Mittermeier, R. A., Fonseca, G. A. B. da, Gerlach, J., Hoffmann, M., Lamoreux, J. F., Mittermeier, C. G., Pilgrim, J. D., and Rodrigues, A. S. L. 2006. Global Biodiversity Conservation Priorities. Science 313: 58–61.

Brown, J. H. 1988. Species diversity. Analytical biogeography (eds A.A. Myers and P.S. Giller), pp. 57-89. Chapman and Hall, London.

Brown, J. H. & Lomolino, M. V. 1998. Biogeography, 2nd edn. Sinauer Associates, Sunderland, MA.

Brummitt, N., and Lughadha, E. N. 2003. Biodiversity: Where's Hot and Where's Not. Conservation Biology 17: 1442–1448.

Bühlmann P & Yu B. 2002. Analyzing Bagging. The Annals of Statistics 30: 927-961.

Busby, J. R. 1991a. BIOCLIM – a bioclimate analysis and prediction system. Nature conservation: cost effective biological surveys and data analysis (ed. by C.R. Margules and M.P. Austin), pp. 64-68. CSIRO, Melbourne.

Busby, J. R. 1991b. BIOCLIM – a bioclimate analysis and prediction system. Plant Protection Quarterly 6: 8–9.

Bush, M. B. 1994. Amazonian speciation: a necessarily complex model. Journal of Biogeography 21: 5–17.

Chamberlain, S. 2016. rgbif: Interface to the Global Biodiversity Information Facility 'API'. R package version 0.9.6. https://CRAN.R-project.org/package=rgbif

Chatrou, L. W., Pirie, M. D., Erkens, R. H. J., Couvreur, T. L. P., Neubig, K. M., Abbott, J. R., Mols, J. B., Maas, J. W., Saunders, R. M. K. & Chase, M. W. 2012. A new subfamilial and tribal classification of the pantropical flowering plant family Annonaceae informed by molecular phylogenetics. Botanical Journal of the Linnean Society 169: 5–40.

Cincotta, R. P., Wisnewski, J., and Engelman, R. 2000. Human population in the biodiversity hotspots. Nature 404, 990–992.

Clayton, W. D. & Hepper, F. N. 1974. Computer-aided chorology of West African grasses. Kew Bulletin 29: 213–234.

Colinvaux, P. A. 1989. Ice-age Amazon revisited. Nature 340: 188–189.

Colyn, M. M., Gautier-Hion, A. and Verheyen, W., 1991. A re-appraisal of the paleo-environmental history of Central Africa: evidence for a major fluvial refuge in the Zaire basin. Journal of Biogeography 18: 403-407

Conservatoire et Jardin botaniques de la Ville de Genève and South African National Biodiversity Institute, Pretoria. African Plants Database (version 3.4.0) 2016. http://www.ville-ge.ch/musinfo/bd/cjb/africa/ (accessed 2014-2016).

Couvreur, T. L. P. 2009. Monograph of the syncarpous African genera Isolona and Monodora (Annonaceae). Systematic Botany Monographs 87: 1-150.

Couvreur, T. L. P. 2014. Revision of the African genus Uvariastrum (Annonaceae). PhytoKeys 33: 1–40 doi: 10.3897/phytokeys.33.5907

Couvreur, T. L. P., Pirie, M. D., Chatrou, L. W., Saunders, R. M. K., Su, Y. C. F., Richardson, J. E. & Erkens, R. H. J. 2011. Early evolutionary history of the flowering plant family Annonaceae: steady diversification and boreotropical geodispersal. Journal of Biogeography 38: 664–680.

Couvreur, T. L. P., Chatrou, L. W., Sosef, M. S. M. & Richardson, J. E. (2008) Molecular phylogenetics reveal multiple tertiary vicariance origins of the African rain forest trees. BMC Biology 6: 54.

da Fonseca, G. A. B., Balmford, A., Bibby, C., Boitani, L., Corsi, F., Brooks, T., Gascon, C., Olivieri, S., Mittermeier, R. A., Burgess, N., et al. 2000. … following Africa's lead in setting priorities. Nature 405, 393–394.

Dauby, G., Zaiss, R., Blach-Overgaard, A., Catarino, L., Damen, T., Deblauwe, V. et al. 2016. RAINBIO: a mega-database of tropical African vascular plants distributions. PhytoKeys 74: 1–18.

Davis, C. C., Bell, C. D., Fritsch, P. W. & Mathews, S. 2002. Phylogeny of *Acridocarpus-Brachylophon* (Malpighiaceae): implications for Tertiary tropical floras and Afroasian biogeography. Evolution 56: 2395–2408.

Denys, E. 1980. A tentative phytogeographical division of tropical Africa based on a mathematical analysis of distribution maps. Bulletin du Jardin Botanique National de Belgique / Bulletin van de National Plantentuin van België, 50, 465–504.

Department of Plant Sciences, University of Oxford. 2013. Botanical Research and Herbarium Management System: Documentation - BRAHMS Online. http://herbaria.plants.ox.ac.uk/bol/brahms/Documentation (accessed 2014-2016).

Department of Plant Sciences, University of Oxford. The Oxford Plant Observatory TOPO 2016. http://herbaria.plants.ox.ac.uk/bol/topo/

Dietterich, T. G. 2000. An Experimental Comparison of Three Methods for Constructing Ensembles of Decision Trees: Bagging, Boosting, and Randomization. Machine Learning 40:139-157.

Dobignard, A. & Chatelain, C. 2010-2013. Synonymic and bibliographic index of North Africa plants. vol. 1-5.

Dowsett, H. J., M. M. Robinson, A. M. Haywood, D. J. Hill, A. M. Dolan, D. K. Stol, W .-L. Chan, A. Abe-Ouchi, M. A. Chandler, N. A.Rosenbloom, B. L. Otto-Bliesner, F. J. Bragg, D. J. Lunt, K. M. Foley and C. R. Riesselman. 2012. Assessing confidence in Pliocene sea surface temperatures to evaluate predictive models. Nature Climate Change 2: 365-371.

Droissart, V., Dauby, G., Couvreur, T. L. P. & the RAINBIO consortium. 2017. Half a million herbarium records to improve the florisitic regionalisation of tropical Africa. Unpublished poster presented at AETFAT 2017, publication *in prep.*

Drummond, A. J. & A. Rambaut. 2007. BEAST: Bayesian evolutionary analysis by sampling trees. BMC Evolutionary Biology 7:214.

Eig, A. 1931. Quelques faits de la phytogéographie palestinienne précédés par des remarques sur les notions phytogéographiques. Bulletin de la Société botanique de France 78: 297–305.

Endler, A. 1982. Pleistocene forest refugia – act or fancy? In Biological Diversification in the Tropics: Proceedings of the Fifth International Symposium of the Association for Tropical Biology, Held at Macuto Beach, Caracas, Venezuela, February 8-13, 1979. Ed G. Prance 1982.

Engler, A. 1964. Obersicht fiber die Florenreiche und Florengebiete der Erde. Syllabus der Pflanzenfamilien, ed. 12 (revised by H. Melchior), 2, 626-629.

FAO Africover Geomorphology/Landcover and Lithology Country Datasets 1998-2008. 2003, 1:350,000, for Burundi, Eritrea, Kenya, Rwanda, Somalia, and Tanzania South African Soil and Terrain (SOTERSAF) – 2003, 1:250,000 to 1:2,500,000 for Angola, Botswana, Mozambique, Namibia, South Africa, Swaziland, and Zimbabwe; Central African Soil and Terrain (SOTERCAF) – 2006, 1:1,000,000 to 1:2,000,000 for the Democratic Republic of the Congo; Northeastern African Soil and Terrain (SOTERNAF) – 1998, 1:1,000,000 for Djibouti, Egypt, Ethiopia, Sudan, and Uganda; Senegal Soil and Terrain (SOTWIS) – 2008, 1:1,000,000 for Senegal and Gambia; Madagascar Simplified Geology, Royal Botanic Gardens, KEW – 1998, 1:1,000,000, and Geologic Data Systems (GDS) Digital Geologic Map of the World: African Subset – 2008, 1:5,000,000.

Farjon, A. 2016. Conifers of the World: Resources for Conifer Research. http://herbaria.plants.ox.ac.uk/bol/conifers (accessed 2016).

Fayolle, A., Swaine, M. D., Bastin, J.-F., Bourland, N., Comiskey, J. A., Dauby, G., … Plumptre, A. J. (2014). Patterns of tree species composition across tropical African forests. Journal of Biogeography, 41, 2320–2331.

Fjeldså J. 1994. Geographical patterns for relict and young species of birds in Africa and South America and implications for conservation priorities. Biodiversity & Conservation 3: 207–226.

Fjeldså, J. & Lovett, J. 1997. Geographical patterns of old and young species in African forest biota: the significance of specific montane areas as evolutionary centres. Biodiversity and Conservation 6: 325-346.

Francis, A. P. & D. J. Currie. 2003. A globally consistent richness-climate relationship for angiosperms. American Naturalist 161: 523–536.

GADM (Global Administrative Areas database): http://www.gadm.org/

Gaston, K. 1996. Species-range-size distributions: patterns, mechanisms and implications. Trends in Ecology & Evolution 11: 197–201.

Gaston, K. 1998. Ecology: Rarity as double jeopardy. Nature 394: 229-230.

GBIF: The Global Biodiversity Information Facility: GBIF Backbone Taxonomy, 2013-07-01. Accessed via http://www.gbif.org/species/5407531 on 2014-06-16.

Geonames database: http://www.geonames.org/

Germishuizen, G. & Meyer, N. L. eds. 2003. Plants of Southern Africa: an annotated checklist Pretoria.

Gernhard, T. 2008. The Yule Process. Journal of Theoretical Biology 253: 769-778. Yule GU (1925) Philosophical Transactions of the Royal Society of London B: Biological Sciences 213: 21-87.

Global Land Cover Facility: www.landcover.org

Global Plants Initiative via JSTOR 2016. http://plants.jstor.org/.

Goevarts, R., M. Ruhsam, L. Andersson, E. Robbrecht, D. Bridson, A. Davis, I. Schanzer, B. Sonké. 2006. World checklist of Rubiaceae. Royal Botanic Gardens, Kew.

Good, R. 1974. The geography of the flowering plants. Ed. 4. London, UK: Longman.

Goodwin, Z. A., Harris, D. J., Filer, D., Wood, J. R. I., & Scotland, R. W. 2015. Widespread mistaken identity in tropical plant collections. Current Biology, 25, R1066–R1067.

Govaerts, R. et al. World Checklist of Selected Plant Families. Facilitated by the Royal Botanic Gardens, Kew. 2016. http://apps.kew.org/wcsp (accessed 2014-2016).

Govaerts, R. 2001. How many species of seed plants are there? Taxon 50: 1085–90.

Grenyer, R., Orme, C. D. L., Jackson, S. F., Thomas, G. H., Davies, R. G., Davies, T. J., Jones, K. E., Olson, V. A., Ridgely, R. S., Rasmussen, P. C., et al. 2006. Global distribution and conservation of rare and threatened vertebrates. Nature 444: 93–96.

Haffer, J. 1969. Speciation in Amazonian forest birds. Science 165: 131–137.

Hall, J. B., & Swaine, M. D. 1981. Distribution and ecology of vascular plants in a tropical rainforest. Forest vegetation in Ghana. The Hague, Nertherlands: Junk Publishers.

Hall, J.B. & Swaine, M. D. 1976. Classification and ecology of closed-canopy forest in Ghana. Journal of Ecology 64: 913-951

Hamilton, A. C. 1976. The significance of patterns of distribution shown by forest plants and animals in Tropical Africa for the reconstruction of upper Pleistocene palaeoenvironments: a review. Paleoecology of Africa 9: 63–97.

Hamilton, A. C. & Taylor, D. 1991. History of climate and forests in tropical Africa during the last 8 million years. In: N. Myers (ed.). Tropical forests and climate: 65-78. Kluwer Academic Publishers, Dordrecht (also Climatic Change 19: 65-78).

Hamilton, A. C. 1982. Environmental history of east Africa. A study of the Quaternary. Academic Press, London.

Hardy, O. J., Born, C., Budde, K., Daïnou, K., Dauby, G., Duminil, J., Ewédjé, E. B. K., Gomez, C., Heuertz, M., … Poncet, V. 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, D. J., Poulsen, A. D., Frimodt-Möller, C., Preston, J. & Cronk, Q. C. B. 2000 Rapid radiation in Aframomum (Zingiberaceae): evidence from nuclear ribosomal DNA internal transcribed spacer (ITS) sequences. Edinburgh Journal of Botany 57: 377

Harrison, S. P., Bartlein, P. J., Brewer, S., Prentice, I. C., Boyd, M., Hessler, I., Holmgren, K., Izumi, K., & Willis, K. 2014. Climate modelbenchmarking with glacial and mid-Holocene climates. Climate Dynamics 43: 671-688.

Harrison, S. & Noss, R. 2017. Endemism hotspots are linked to stable climatic refugia. Annals of Botany, 119, 207–214.

Hawthorne, W. D. 1995. Ecological profiles of Ghanaian forest trees. ODA Tropical Forestry Papers 29, Oxford Forestry Institute.

Hawthorne, W. D. 1996. Holes and the sums of parts in Ghanaian forest: regeneration, scale and sustainable use. Proceedings of the Royal Society of Edinburgh Section B: Biological Sciences 104: 75–176.

Hawthorne, W. D. 1998. Plants in Ankasa, Nini-Suhien, and Bia.Review of existing knowledge, results from a new survey and recommendations for management plans. Protected Areas Development Programme, Western Region, Ghana. ULG Consultants LTD in association with S. A. Agrer, N.V.

Hawthorne, W. D. 2002. Final report of the floral survey of the Biodiversity Component of the NRMP. Forestry Commission, Biodiversity Conservation Component. Ministry of Lands and Forestry, Ghana.

Hawthorne, W. D. & M. Abu-Juam. 1995. Forest Protection in Ghana. IUCN/ODA/Forest Department Republic of Ghana, Gland, Switzerland, and Cambridge, UK, xvii + 203 pp.

Hawthorne, W. D., & Jongkind, C. C. H. 2006. Woody plants of Western African forests. Kew, UK: Kew Publishing.

Hawthorne W. D. & Marshall, C. A. M. 2016. A Manual for Rapid Botanic Survey (RBS) and measurement of vegetation bioquality 2016. http://herbaria.plants.ox.ac.uk/bol/oxford/Survey.

Hély, C., Braconnot, P., Watrin, J., Zheng, W. 2009. Climate and vegetation: simulating the African humid period. Comptes Rendus Geoscience 341: 671–688.

Hijmans, R. J. 2016. raster: Geographic Data Analysis and Modeling. R package version 2.5-8. https://CRAN.R-project.org/package=raster

Hijmans, R. J., Cameron, S. E., Parra, J. L., Jones, P. G. & Jarvis, A. 2005. Very high resolution interpolated climate surfaces for global land areas. International Journal of Climatology 25: 1965-1978.

Hillebrand, H. 2004. On the generality of the latitudinal diversity gradient. American Naturalist 163: 192–211.

Hollingsworth, P. M. et al. 2009. A DNA barcode for land plants. PNAS 106: 12794-12797.

Holstein, N., & Renner, S. S. 2011. A dated phylogeny and collection records reveal repeated biome shifts in the African genus *Coccinia* (Cucurbitaceae). BMC evolutionary biology 11: 28.

Hothorn, T., Buehlmann, P., Dudoit, S., Molinaro, A. & Van Der Laan, M. 2006. Survival Ensembles. Biostatistics 7: 355--373.

Hughes, C. & Eastwood, R. 2006. Island radiation on a continental scale: Exceptional rates of plant diversification after uplift on the Andes. Proceedings of the National Academy of Sciences of the United States of America 103: 10334-10339.

Hunter, M. L., and Hutchinson, A. 1994. The Virtues and Shortcomings of Parochialism: Conserving Species That Are Locally Rare, but Globally Common. Conservation Biology 8: 1163–1165.

Hutchison, J., Dalziel, J. M. & Keay, R. W. J. 1954-1988. Flora of West Tropical Africa, Volumes 1-3. Royal Botanic Gardens, Kew, UK.

Ismail, I. M., & Elawad, A. A. 2015. Checklist of plants of Rashad and Alabassia localities (eastern Nuba Mountains), South Kordofan Sudan. Check List 11: 1805.

IUCN. The IUCN Red List of Threatened Species Version 2015-4 2016. http://www.iucnredlist.org/# (accessed May 2016).

Jongkind, C. C. H. 2004. Checklist of Upper Guinea forest species. In Poorter, L. et al. eds. Biodiversity of West African Forests. CAB International.

Jongkind, C. C. H. 2012. Checklist of Liberian vascular plants. Prepared for Fauna & Flora International. http://www.liberianfaunaflora.org/FFI/Document.aspx?p=30&ix=25&pid=3021&prcid=4&ppid=3021

Jongkind, C. C. H. 2015. Description of Gaertnera luteocarpa (Gentianales: Rubiaceae), with two subspecies, a new forest shrub species from Liberia, Ivory Coast and Ghana. European Journal of Taxonomy 126: 1–8.

Kaufman, D. M. & Willig, M. R. 1998. Latitudinal patterns of mammalian species richness in the New World: the effects of sampling method and faunal group. Journal of Biogeography 25: 795–805.

Kenfack, D., Goseline, G., Gereau, R. E., Schatz, G. E. 2003. The genus Uvariopsis (Annonaceae) in tropical Africa with a recombination and one new species from Cameroon. Novon 13: 443–449.

Kier, G. & Barthlott, W. 2001. Measuring and mapping endemism and species richness: A new methodological approach and its application on the flora of Africa. Biodiversity Conservation 10: 1513–1529.

Kier, G., Kreft, H., Lee, T., Jetz, W., Ibisch, P., Nowicki, C., Mutke, J. & Barthlott, W. 2009. A global assessment of endemism and species richness across island and mainland regions. Proceedings of the National Academy of Sciences of the United States of America 106: 9322–7.

Klopper, R. R., Gautier, L., Chatelain, C., Smith, G. F., 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.

Kreft, H., & Jetz, W. 2010. A framework for delineating biogeographical regions based on species distributions. Journal of Biogeography 37: 2029–2053.

Kriticos, D. J., Webber, B. L., Leriche, A., Ota, N., Macadam, I., Bathols, J. & Scott, J. K. 2012. CliMond: global high-resolution historical and future scenario climate surfaces for bioclimatic modelling. Methods Ecology and Evolution 3: 53-64.

Küper, W., Sommer, J. H., Lovett, J. C. & 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, J. H., Lovett, J. C., Mutke, J., Linder, H. P., Beentje, H. J., Rompaey, R. S. A. R. V., Chatelain, C., Sosef, M., and Barthlott, W. 2004. Africa's Hotspots of Biodiversity Redefined. Annals of the Missouri Botanic Garden 91: 525–535.

Lachenaud, O. 2013. Le genre Psychotria (Rubiaceae) en Afrique occidentale et centrale : taxonomie, phylogénie et biogéographie. Volume III: Planches, cartes de distribution et liste des specimens. Thèse présentée pour l'obtention du titre de Docteur en Sciences, Université Libre de Bruxelles.

Leal, M. E. 2004. The African rain forest during the Last Glacial Maximum, an archipelago of forests in a sea of grass. PhD thesis Wageningen University, Wageningen.

Lebrun, J. 1961. Les deux flores d'Afrique tropicale. Brussels, Belgium: Palais des Académies.

Lebrun, J. P. & Stork, A. L. 1991-2012. Enumération des plantes à fleurs d'Afrique tropicale et Tropical African Flowering Plants: Ecology and Distribution, vol. 1-7. Conservatoire et Jardin botaniques de la Ville de Genève.

Lemon, J. 2006. Plotrix: a package in the red light district of R. R-News 6(4): 8–12.

Léonard, J. 1965. Contribution à la subdivision phytogéographique de la Région Guinéo-Congolaise d'après la repartition géographique d'Euphorbiacées d'Afrique tropicale. Webbia 19: 627–649.

Liaw, A. & Wiener, M. 2002. Classification and Regression by randomForest. R News 2(3): 18-22.

Ligges, U., & Mächler, M. 2003. Scatterplot3d - an R Package for Visualizing Multivariate Data. Journal of Statistical Software 8(11): 1-20.

Lima-Ribeiro, M. S., Varela, S., González-Hernández, J., Oliveira, G., Diniz-Filho, J. A. F., Terribile, L. C. 2015. ecoClimate: a database of climate data from multiple models for past, present, and future for Macroecologists and Biogeographers. Biodiversity Informatics 10: 1-21.

Lima-Ribeiro, M. S., Varela, S., González-Hernández, J., Oliveira, G., Diniz-Filho, J. A. F., Peterson, A. T, Terribile, L. C. The ecoClimate Database, http://ecoclimate.org**.**

Linder, H. P. 2001. Plant diversity and endemism in sub-Saharan tropical Africa. Journal of Biogeography 28: 169–182.

Linder, H. P., Lovett, J. C., Mutke, J. M., Barthlott, W., Jürgens, N., Rebelo, T., & Küper, W. 2005. A numerical re-evaluation of the sub-Saharan phytochoria of mainland Africa. Biologiske Skrifter 55: 229–252.

Linder, H. P., de Klerk, H. M., Born, J., Burgess, N. D., Fjeldså, J., & Rahbek, C. 2012. The partitioning of Africa: statistically defined biogeographical regions in sub-Saharan Africa. Journal of Biogeography 39: 1189–1205.

Lomolino, M. V. 2004. Conservation biogeography. Frontiers of biogeography: new directions in the geography of nature (ed. by M.V. Lomolino and L.R. Heaney), pp. 293–296. Sinauer Associates, Sunderland, MA.

Lonnberg, E., 1929. The development and distribution of the African fauna in connection with and depending upon climatic changes. Arkiv for Zoologi 21 A, 4: 1-33.

Loreau, M. 2001. Biodiversity and Ecosystem Functioning: Current Knowledge and Future Challenges. Science 294: 804–808.

Lovett, J. C. 1998. Importance of the Eastern Arc Mountains for Vascular Plants. Journal of East African Natural History 87: 59–74.

Lovett, J. C. & Friis, I. 1996. Patterns of endemism in the woody flora of north-east and east Africa. The biodiversity of African plants (eds L.J.G. van der Maesen, X.M. van der Burgt and J.M. van Medenbach de Rooy), pp. 582–601. Kluwer Academic Publisher, Dordrecht.

Lovett, J. C., Rudd, S., Taplin, J., & Frimodt-Møller, C. 2000. Patterns of plant diversity in Africa south of the Sahara and their implications for conservation management. Biodiversity Conservation 9: 37–46.

Mace, G. M., Balmford, A., Boitani, L., Cowlishaw, G., Dobson, A. P., Faith, D. P., Gaston, K. J., Humphries, C. J., Vane-Wright, R. I., Williams, P. H., et al. 2000. It's time to work together and stop duplicating conservation efforts. Nature 405: 393–393.

Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M., Hornik, K. 2016. cluster: Cluster Analysis Basics and Extensions. R package version 2.0.5.

Mangenot, G. 1955. Etude sur les forêts des plaines et plateaux de la Côte-d'Ivoire. Etudes éburnéennes 4: 4-61.

Marshall, C. A. M., Wieringa, J. J., & Hawthorne, W. D. 2016. Bioquality hotspots in the tropical African flora. Current Biology 26: 3214–3219.

Maley, J., 1996. The African rain forest: main characteristics of changes in vegetation and climate from the Upper Cretaceous to the Quaternary. Proceedings of the Royal Society of Edinburgh B: Biological Sciences. 104B: 31–73.

Mbayngone, E., Schmidt, M., Hahn-Hadjali, K., Thiombiano, A., and Guinko, S. 2008. Magnoliophyta of the partial faunal reserve of Pama, Burkina Faso. Check List 4: 251–266.

Meerts, P. 2016. An annotated checklist to the trees and shrubs of the Upper Katanga (D.R. Congo). Phytotaxa 258 (3): 201–250.

Missouri Botanic Garden. Tropicos 2016. http://www.tropicos.org/. (accessed 2014-2016).

Mittermeier, R. A., Turner, W. R., Larsen, F. W., Brooks, T. M., and Gascon, C. 2011. Global Biodiversity Conservation: The Critical Role of Hotspots. In Biodiversity Hotspots, F.E. Zachos, and J.C. Habel, eds. (Berlin, Heidelberg: Springer Berlin Heidelberg), pp. 3–22.

Möller, M. & Cronk, Q. C. B. 2001. Phylogenetic studies in *Streptocarpus*: reconstruction of biogeographic history and distribution patterns in *Streptocarpus* (Gesneriaceae). Systematics & Geography of Plants 71: 545–555.

Moore, J. J. 1962. The Braun-Blanquet System: A Reassessment. Journal of Ecology 50: 761–769.

Moritz, C., Patton, J. L., Schneider C. J., & Smith, T. B. 2000. Diversification of faunas: An integrated molecular approach. Annual Review of Ecology and Systematics 31: 533-563.

Morley, R. J. 2000. Origin and evolution of tropical rain forests. John Wiley and Sons Ltd, New York.

Mutke, J., Kier, G., Braun, G., Schultz, C. & Barthlott, W. 2001. Patterns of African vascular plant diversity: A GIS based analysis. Systematics and Geography of Plants 71: 1125-1136.

Mutke, J. & Barthlott, W. 2005. Patterns of vascular plant diversity at continental to global scales. Biologiske Skrifter 55: 521–538.

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

Nakamura, N. 2013. Dissecting the Japanese hotspot. DPhil. University of Oxford.

NHN collection: Naturalis Biodiversity Center. Bioportal 2016. http://bioportal.naturalis.nl/.

Nelson, B. W., Ferreira, C. A. C., da Silva, M. F. & Kawasaki, M. L. 1990. Endemism centres, refugia and botanical collection density in Brazilian Amazonia. Nature 345: 714-716.

New, M., Lister, D., Hulme, M. & Makin, I. 2002. A high-resolution data set of surface climate over global land areas. Climate Research 21: 1-25.

NHN: BRAHMS Online 2016. http://herbarium.naturalis.nl/

O'Connor, C., Marvier, M., and Kareiva, P. 2003. Biological vs. social, economic and political priority-setting in conservation. Ecology Letters 6: 706–711.

Oksanen, F. Blanchet, G., Friendly, M., Kindt, R., Legendre, P., McGlinn, D., … Wagner, H. (2016). vegan: Community Ecology Package. R package version 2.4-0. https://CRAN.R-project.org/package=vegan

Oksanen, F., Kindt, R., & Simpson, G. L. 2016b. vegan3d: Static and Dynamic 3D Plots for the 'vegan' Package. R package version 1.0-1. https://CRAN.R-project.org/package=vegan3d

Olson, D. M. & Dinerstein, E. 1998. The Global 200: A Representation Approach to Conserving the Earth's Most Biologically Valuable Ecoregions. Conservation Biology 12: 502–515.

Olson, D. M., and Dinerstein, E. 2002. The Global 200: Priority Ecoregions for Global Conservation. Annals of the Missouri Botanic Garden 89: 199-224.

Ouédraogo, O., Schmidt, M., Thiombiano, A., Hahn, K., Guinko, S., and Zizka, G. (2011). Magnoliophyta, Arly National Park, Tapoa, Burkina Faso. Check List 7: 85–100.

Pebesma, E. J., Bivand, R. S. 2005. Classes and methods for spatial data in R. R News 5 (2) https://cran.r-project.org/doc/Rnews/.

Pirie, M. D. & Doyle, J. A. 2012. Dating clades with fossils and molecules: the case of Annonaceae. Botanical Journal of the Linnean Society 169: 84–116.

Poorter, L., Bongers, F. & Lemmens R. H. M. J. 2004. West African Forests: introduction. In: Poorter, L., F. Bongers, F. N'. Kouamé & W. D. Hawthorne. Biodiversity of West African forests: an ecological atlas of woody plant species. Wallingford, UK: CABI Publishing.

Poorter, L., Bongers, F. Kouamé, F. N'. & Hawthorne, W. D. 2004. Biodiversity of West African forests: an ecological atlas of woody plant species. Wallingford, UK: CABI Publishing.

Plana, V. 2004. Mechanisms and tempo of evolution in the African Guineo-Congolian rainforest. Philosophical transactions of the Royal Society of London. Series B, Biological Sciences 35: 1585–1594.

Prance, G. 1982. Biological Diversification in the Tropics: Proceedings of the Fifth International Symposium of the Association for Tropical Biology, Held at Macuto Beach, Caracas, Venezuela, February 8-13, 1979.

R Core Team. 2016. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/

Rabinowitz, D. 1981. Seven forms of rarity. In: Synge H (ed) The Biological Aspects of Rare Plant Conservation, pp 205–217. John Wiley & Sons, Cambridge,

Ratnasingham, S. & Hebert, P. D. N. 2007. BOLD: The Barcode of Life Data System (www.barcodinglife.org). Molecular Ecology Notes 7: 355-364.

Robbrecht, E. 1996. Geography of African Rubiaceae with reference to glacial rain forest refugia. The biodiversity of African plants (eds L.J.G. van der Maesen, X.M. van der Burgt and J.M. van Medenbach de Rooy), pp. 564–581. Kluwer Academic publishers, Dordrecht.

Rohde, K. 1992. Latitudinal gradients in species diversity: the search for the primary cause. Oikos 65: 514–527.

Rosenzweig, M. L. 1995. Species diversity in space and time. Cambridge University Press, Cambridge.

Roy, M. 1997. Recent diversification in African greenbuls (Pycnonotidae: Andropadus) supports a montane speciation model. Proceedings of the Royal Society London Series B 264: 1337–44.

Royal Botanic Gardens, Kew. The Flora of East Tropical Africa (FTEA) 1948-2012. https://plants.jstor.org/collection/FTEA (accessed 2014- 2016).

Salzmann, U. & Hoelzmann, P. 2005. The Dahomey Gap: an abrupt climatically induced rain forest fragmentation in West Africa during the late Holocene. The Holocene 15: 190–199.

Sanderson, M. J. 2002. Estimating absolute rates of molecular evolution and divergence times: A penalized likelihood approach. Molecular Biology & Evolution 19: 101-109.

Sayre, R., Comer, P., Warner, H. & Cress, J. 2009. A new map of standardized terrestrial ecosystems of the conterminous United States: U.S. Geological Survey Professional Paper 1768. 17 pages. (Also available online at http://pubs.usgs.gov/pp/1768.)

Sayre, R., P. Comer, J. Hak, C. Josse, J. Bow, H. Warner, M. Larwanou, E. Kelbessa, T. Bekele, H. Kehl, R. Amena, R. Andriamasimanana, T. Ba, L. Benson, T. Boucher, M. Brown, J. Cress, O. Dassering, B. Friesen, F. Gachathi, S. Houcine, M. Keita, E. Khamala, D. Marangu, F. Mokua, B. Morou, L. Mucina, S. Mugisha, E. Mwavu, M. Rutherford, P. Sanou, S. Syampungani, B. Tomor, A. Vall, J. Vande Weghe, E. Wangui, & L. Waruingi. 2013. A New Map of Standardized Terrestrial Ecosystems of Africa. Washington, DC: Association of American Geographers.

Simon, M. F., Grether, R., de Queiroz, L. P., Skema, C., Pennington, R. T. & Hughes, C. E. 2009. Recent assembly of the Cerrado, a neotropical plant diversity hotspot, by in situ evolution of adaptations to fire. Proceedings of the National Academy of Sciences USA 106: 20359–20364.

Smith, T. B., Wayne, R. K., Girman, D., Bruford, M. 1997. A role for ecotones in generating biodiversity. Science 276: 1855–57.

Smith, T. B., Schneider, C. J., Holder, K. 2001. Refugial isolation versus ecological gradients. Testing alternative mechanisms of evolutionary divergence in four vertebrates. Genetica 112–113: 383–398.

Sosef, M., 1994. Refuge begonias: taxonomy, phylogeny and historical biogeography of Begonia sect. Loasibegonia and sect. Scutobegonia in relation to glacial rain forest refugia in Africa. Agricultural University Wageningen Papers 94: 1–306.

Sosef, M. S. M. 1996. Begonias and African rain forest refuges: general aspects and recent progress. The biodiversity of African plants (eds L.J.G. van der Maesen, X.M. van der Burgt and J.M. van Medenbach de Rooy), pp. 602–611. Kluwer Academic Publishers, Dordrecht.

Sosef, M. S. M., Dauby, G., Blach-Overgaard, A., van der Burgt, X., Catarino, L., Damen, T., … Couvreur, T. L. P. 2017. Exploring the floristic diversity of tropical Africa. BMC Biology, 15:15.

Stattersfield, A. J., Crosby, M. J., Long, A. J. and Wege, D. C. 1988. Endemic Bird Areas of the World: Priorities for Biodiversity Conservation. BirdLife International (Cambridge, UK).

Stebbins, G. L. 1974. Flowering plants, evolution above the species level. Cambridge, MA: Harvard University Press.

Stevens, P. F. (2001 onwards). Angiosperm Phylogeny Website. Version 12, July 2012 [and more or less continuously updated since] http://www.mobot.org/MOBOT/research/APweb/.

Strobl, C., A.-L. Boulesteix, A. Zeileis & T. Hothorn. 2007. Bias in Random Forest Variable Importance Measures: Illustrations, Sources and a Solution. BMC Bioinformatics 8(25).

Strobl, C., Boulesteix, A.-L., Kneib, T., Augustin, T. & Zeileis, A. 2008. Conditional Variable Importance for Random Forests. BMC Bioinformatics 9(307).

Stropp, J., Ladle, R. J., Malhado, A. C. M., Hortal, J., Gaffuri, J., Temperley, W. H., … Mayaux, P. 2016. Mapping ignorance: 300 years of collecting flowering plants in Africa. Global Ecology and Biogeography 25: 1085–1096.

Strugnell, A. M. 2006. A Checklist of the Spermatophytes of Mount Mulanje, Malawi. Scripta Botanica Belgica 34 (National Botanic Garden of Belgium, Meise, Belgium).

Takhtajan, A. 1986. Floristic Regions of the World (translated by T. J. Crovello & A. Cronquist). Berkeley, CA: University of California Press.

Tamura, K., Stecher, G., Peterson, D., Filipski, A. & Kumar, S. 2013. MEGA6: Molecular Evolutionary Genetics Analysis version 6.0. Molecular Biology and Evolution: 30 2725-2729.

Taylor, K. E., Stouffer, R. J. & Meehl, G. A. 2012. An overview of CMIP5 and the experiment design. Bulletin of the American Meteorological Society 93: 485-498.

The Plant List Version 1.1 2016. http://www.theplantlist.org/. (accessed 2016)

Thulin, M. 1993. Flora of Somalia, Volume 1 (Royal Botanic Gardens, Kew).

True, D. 2002. Landforms of the Lower Mid-West. Missouri Resource Assessment Partnership. MoRAP Map Series MS-2003-001 (1:1,500,000). Available online at http://morap.missouri.edu/Posters. aspx.

URS. Volume 4, Part 1.1: Forest Botanical Baseline & Impact Assessment 2013. http://liberia.arcelormittal.com/corporate-responsibility/publications-and-reports.aspx (accessed May 27, 2016).

USGS for GMTED2010 tiles: https://lta.cr.usgs.gov/GMTED2010

Vane-Wright, R. I., Humphries, C. J., and Williams, P. H. 1991. What to protect?—Systematics and the agony of choice. Biological Conservation 55: 235–254.

Varela, S., Lima-Ribeiro, M. S., & Terribile, L. C. 2015. A Short Guide to the Climatic Variables of the Last Glacial Maximum for Biogeographers. PLOS ONE, 10, e0129037.

Versteegh, C. P. & Sosef, M. S. 2007. Revision of the African genus *Annickia* (Annonaceae). Systematics and Geography of Plants: 91-118.

Wenger, S. J. & Olden, J. D. 2012. Assessing transferability of ecological models: an underappreciated aspect of statistical validation. Methods in Ecology and Evolution 3: 260–267.

White, F. 1979. The Guineo-Congolian Region and its relationships to other phytochoria. Bulletin du Jardin botanique national de Belgique / Bulletin van de National Plantentuin van België 49: 11–55.

White, F. 1983. The vegetation of Africa - a descriptive memoir to accompany the Unesco/AETFAT/UNSO vegetation map of Africa. Natural Resources Research XX. Paris, France: UNESCO.

White, F. 1993. The AETFAT Chorological Classification of Africa: History, Methods and Applications. Bulletin du Jardin botanique national de Belgique / Bulletin van de National Plantentuin van België 62: 225-281.

Whittaker, R. J., Araujo, M.B., Jepson, P., Ladle, R.J., Watson, J.E.M. & Willis, K.J. 2005. Conservation biogeography: assessment and prospect. Diversity and Distributions 11: 3–23.

Wickens, G. E. 1976. The flora of Jebel Marra (Sudan Republic) and its geographical affinities. Kew, UK: Royal Botanic Gardens, Kew.

Wieringa, J. J. 1999. *Monopetalanthus* exit. A systematic study of *Aphanocalyx*, *Bikinia*, *Icuria*, *Michelsonia* and *Tetraberlinia* (Leguminosae, Caesalpinioideae). Wageningen Agricultural University Papers 99-4: 1-320.

Wieringa, J. J., & Poorter, L. 2004. Biodiversity hotspots in West Africa; patterns and causes. In Biodiversity of West African Forests: An Ecological Atlas of Woody Plant Species, L. Poorter, F. Bongers, F. N'. Kouamé & W. D. Hawthorne, eds. (CABI), pp. 61-72.

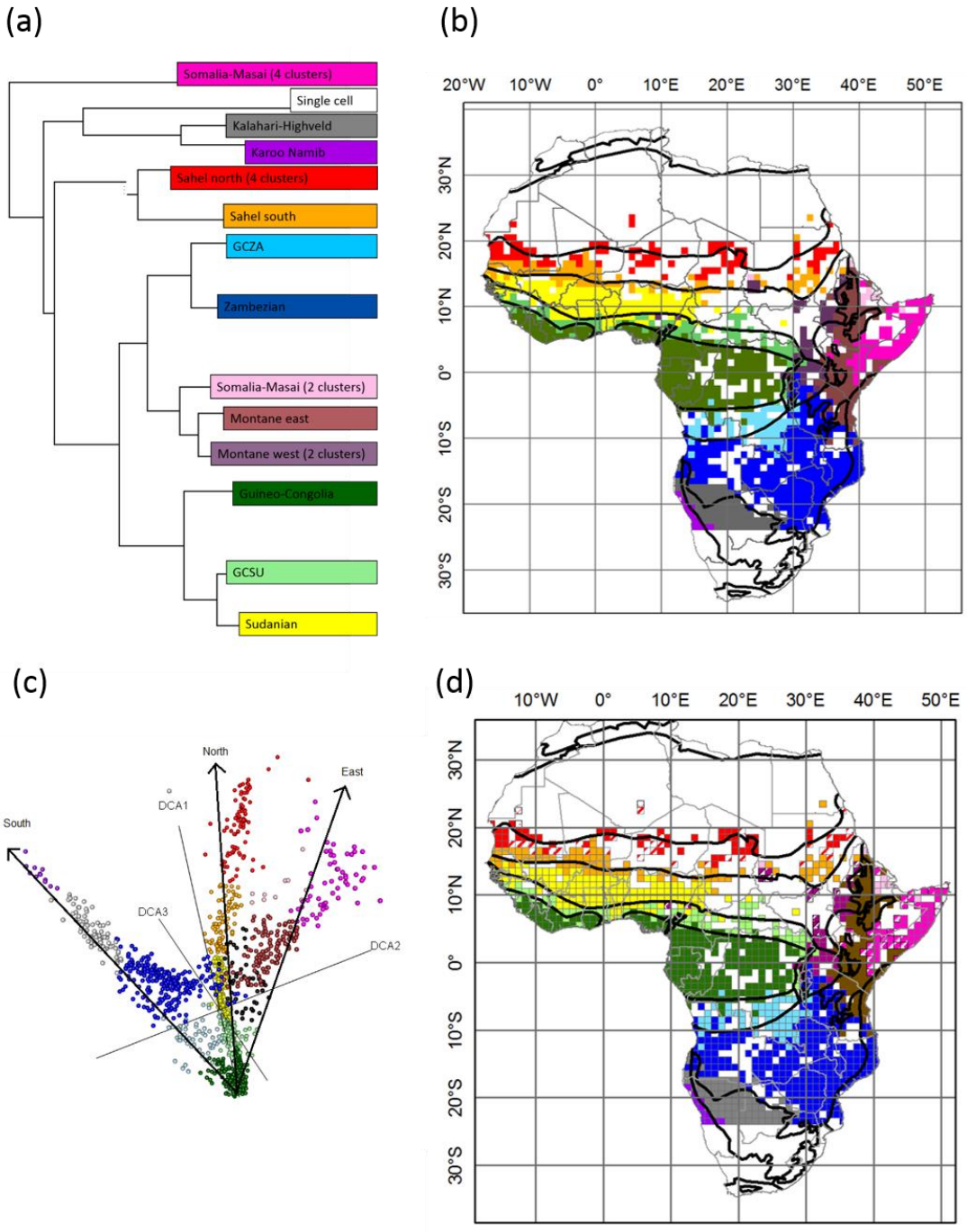
Williams, P., Gibbons, D., Margules, C., Rebelo, A., Humphries, C., & Pressey, R. 1996. A comparison of richness hotspots, rarity hotspots, and complementary areas for conserving diversity of British birds. Conservation Biology 10: 155–174.

Appendix 1

Defining regions in tropical Africa (Chapter 4)

To define our final regions as presented in Figure 3, we first ordinated and clustered all the tropical African cells. These results are shown in Fig A1.1. UPGMA 22-cluster solution of DCA ordination scores for all tropical African cells. (A) Dendrogram of regions; dotted branch line indicates where a node has been moved slightly for clearer presentation. (B) Mapped UPGMA regions. No cells have been reassigned. Some regions have been collapsed to a higher node as noted in regional names of panel A. No regions are nested, and no regions were combined from different branches of the dendrogram. (C) Cells in DCA ordination space, coloured by regions defined in panels a & b (montane west is shown in black). Arrows illustrate the floristic turnover from Guineo-Congolia to the south, north, and east. Total endemism in this framework is 43%. (d) 22-cluster solution, with no regions collapsed and no cells reassigned. Regions that were collapsed are indicated by the colour of the region they were collapsed into, but have different patterns. The single cells region is barely visible at 50E, 9N; it is shown in black. These regions have been collapsed as there seems to us no particular reason to maintain any of them – they are all either very small, or the result of dividing too far in the dendrogram to begin with.

Figure A1.1. UPGMA 22-cluster solution of DCA ordination scores for all tropical African cells.



To improve our regionalisation around eastern and southern Africa (for example, no Zanzibar Inhambane region was apparent), we took the cells of our GCZA, Zambezian, Somali-Masai 2 clusters, Montane west and montane east regions defined in Fig. A1.1 (all form a large clade together in the centre of the dendrogram) and reordinated and reclustered them. These results are shown in Figure A1.2.

Figure A1.2. (A & B) UPGMA 12-cluster solution of the cells shown in panel b, which together formed a large 'clade' in the whole tropical Africa analysis of Fig. A1.1. At 12 clusters, we can see the the Zanzibar Inhambane region, although it extends too far in land. We can also see that the 'GCZA' of Fig. A1.1, which was rather large in the SE relative to White's, has resolved into a central Zambezian region, and three western Zambezian regions, and there is no remaining 'signature' of a transition region. (C) In panel C, we have collapsed back all the Zambezian regions to their highest node. We have also collapsed the two, single cell, regions of panel a (shown in magenta and brown there) together, as they share a higher node. Additionally, we have visualised the four highest branches of the UPGMA dendrogram within the Zanzibar-Inhambane region: Two correspond to the inland cells which are implausibly ZI (shown in orange and purple), and two correspond to the northern and southern ZI (shown in dark grey and light grey). If we were not prepared to combine regions from different branches of the UPGMA dendrogram, we would have to accept this solution. This seems unreasonable, given the position of those cells in ordination space, and the large number of regions this would result in. (E & F) We have combined the regions shown in orange and purple in panel C into our main Zambezian region, as this does not make the ZI region (black) any less distinct from the Zambezian region than it was before. By the same logic, we have combined the light pink 'region' of panel C into the magenta region, and the three ZI regions into one. The regions of the GCZA, Zambezian, Somali-Masai 2 clusters, Montane west and

montane east regions defined in Fig. A1.1 are replaced by those shown in Figure 3 of the main paper.

Figure A1.2: Reordinating and reclustering eastern and southern Africa

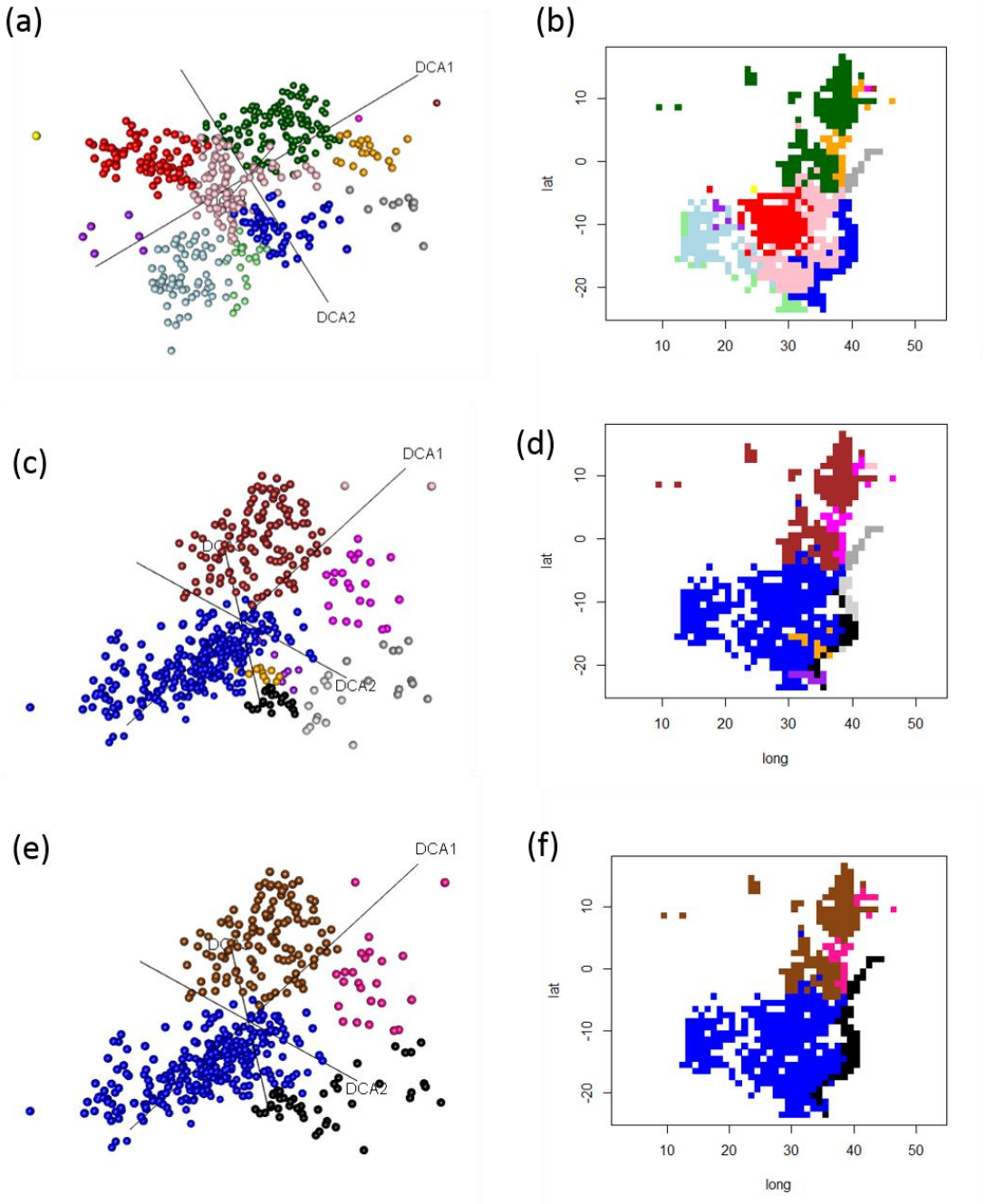


Figure A1.3: UPGMA clustering of beta-sim matrix. This analysis employs the same methods as Linder et al. 2012. UPGMA cluster analysis does not perform well with this large dataset of >30,000 species (cophenetic correlation coefficient = 0.44). Nonetheless, at only 6 clusters, Guineo-Congolia is recognisable as a region and includes Upper Guinea. By 12 clusters, the southern portion of Guineo-Congolia (White's Guineo-Congolian/Zambezian transition zone) is split off.

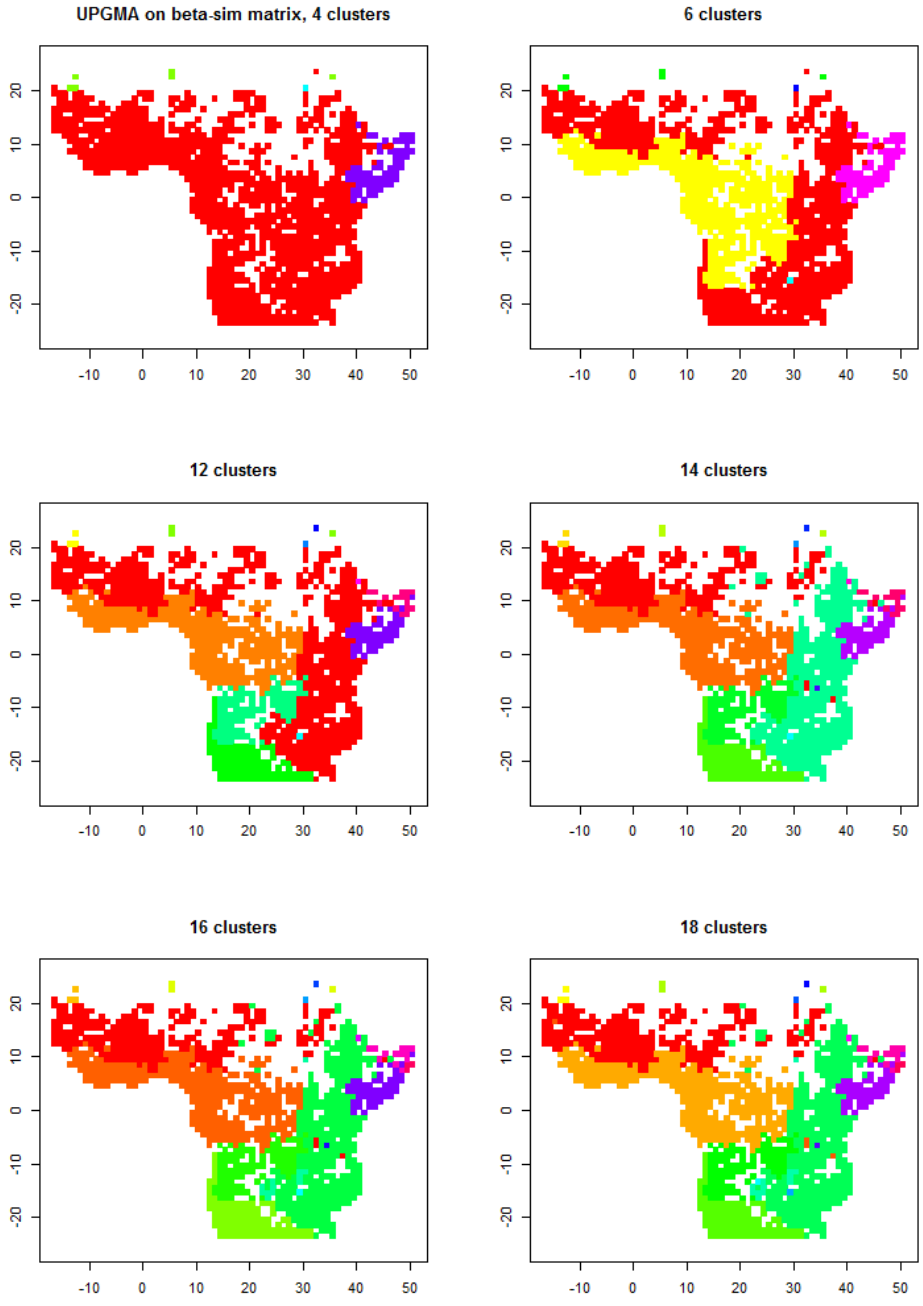


Figure A1.4: PAM/beta sim, the best way to lose Upper Guinea Diagnostic silhouette plot suggests 5 clusters as optimal, where Upper Guinea included in the Guineo-Congolian region. At 8 and 10 clusters, Upper Guinea has been subsumed into the GCSU. This is the closest we can get to a definition of Guineo-Congolia that excludes all of west Africa's forest zone. At 12 clusters and above, the forest zone of West Africa (Upper Guinea) is resolved again as distinct from the GCSU.

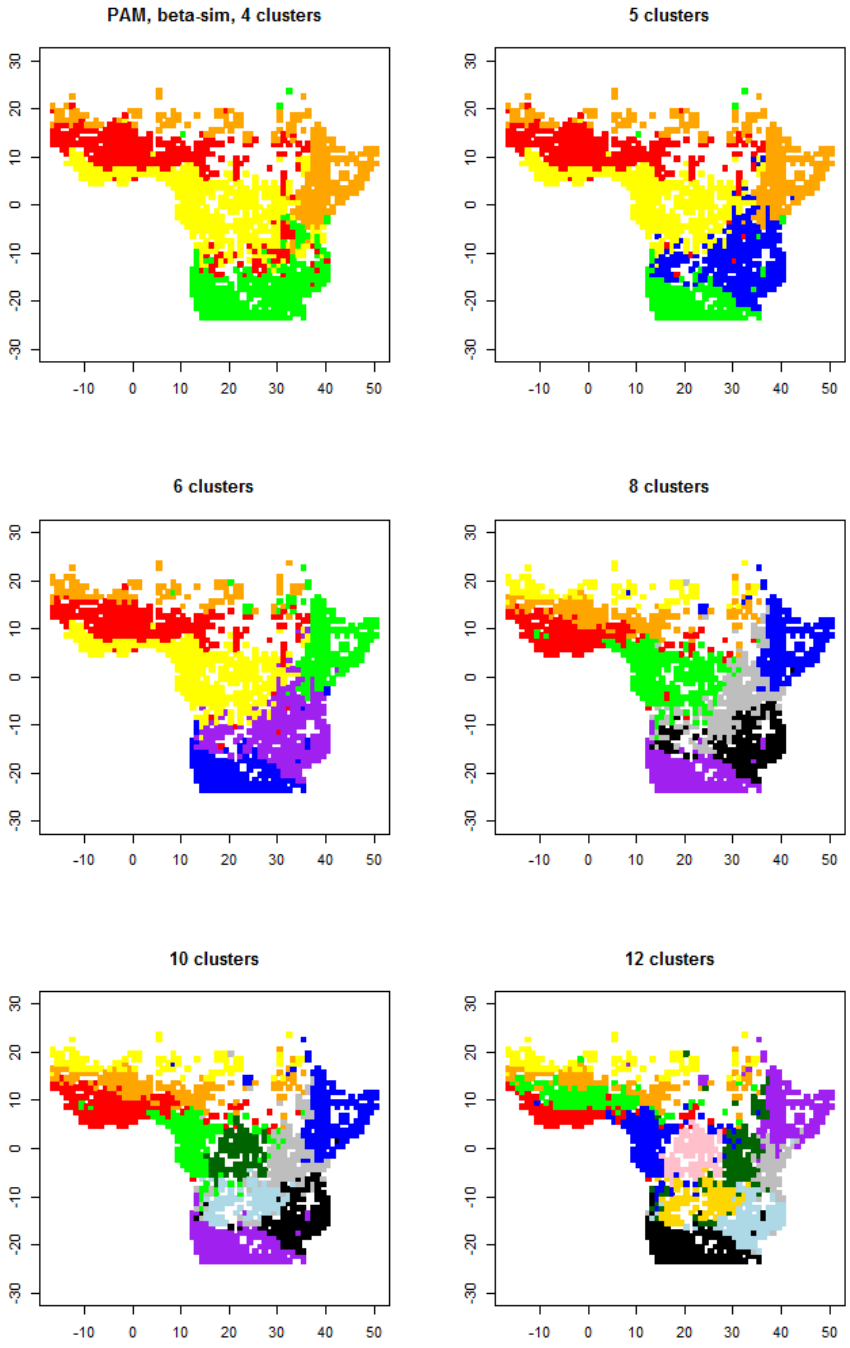


Figure A1.5. Divisive clustering of DCA ordination scores for tropical Africa (diana of the R package cluster). The cophenetic correlation coefficient here is 0.67, which is lower than for the UPGMA cluster results (0.75); we therefore presented UPGMA results in the main paper. Here again, it is not possible to recover a definition of Guineo-Congolia which excludes Upper Guinea.

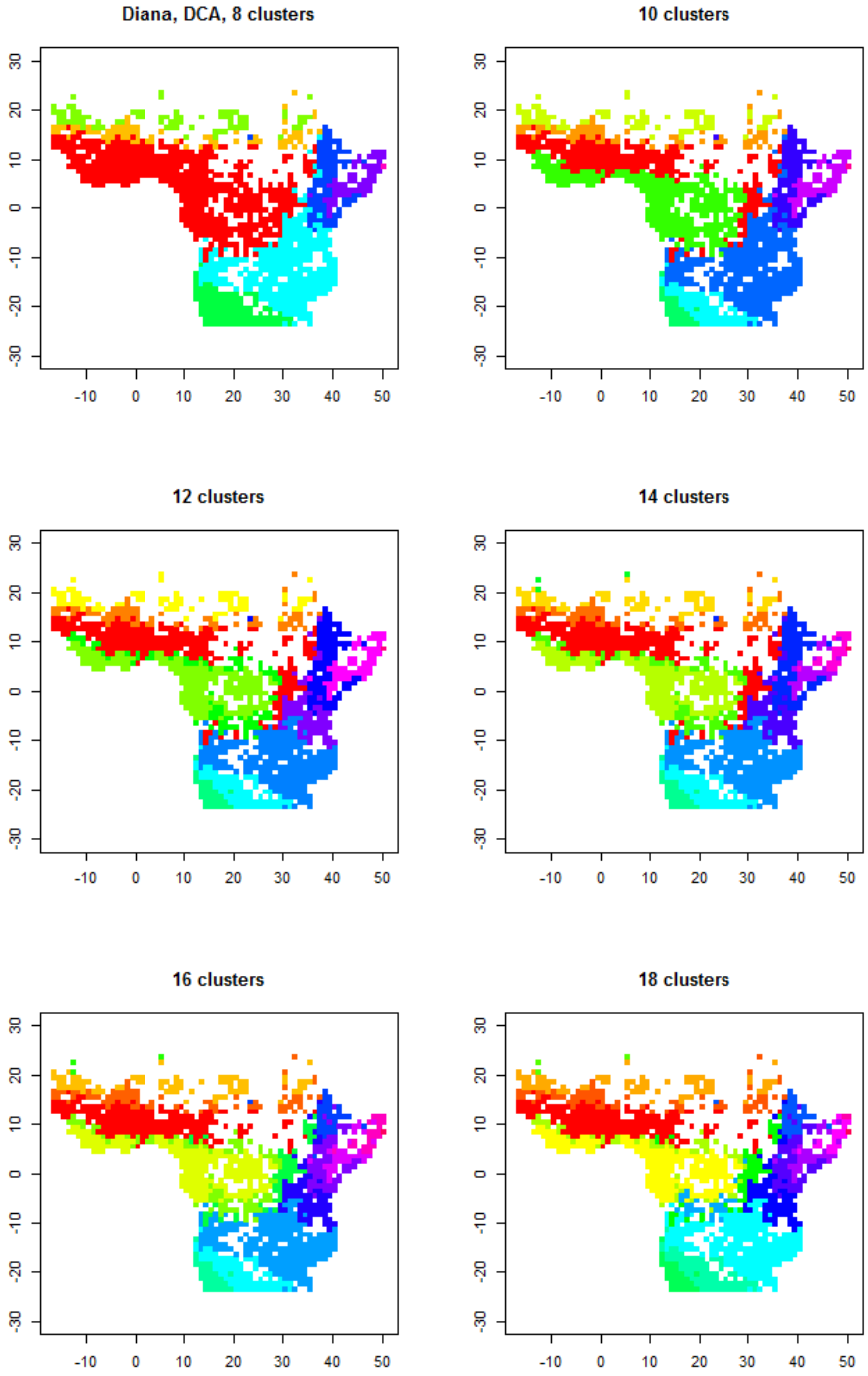


Figure A1.6. PAM 7-cluster solution of DCA axes scores, as suggested as optimal by the diagnostic silhouette plots. (A) Axes 1 and 2 of DCA ordination, coloured by 7 clusters defined with PAM. (B) 7 PAM clusters mapped in geographic space. Total endemism is 52% (52% of species are found in only one of these 7 regions).

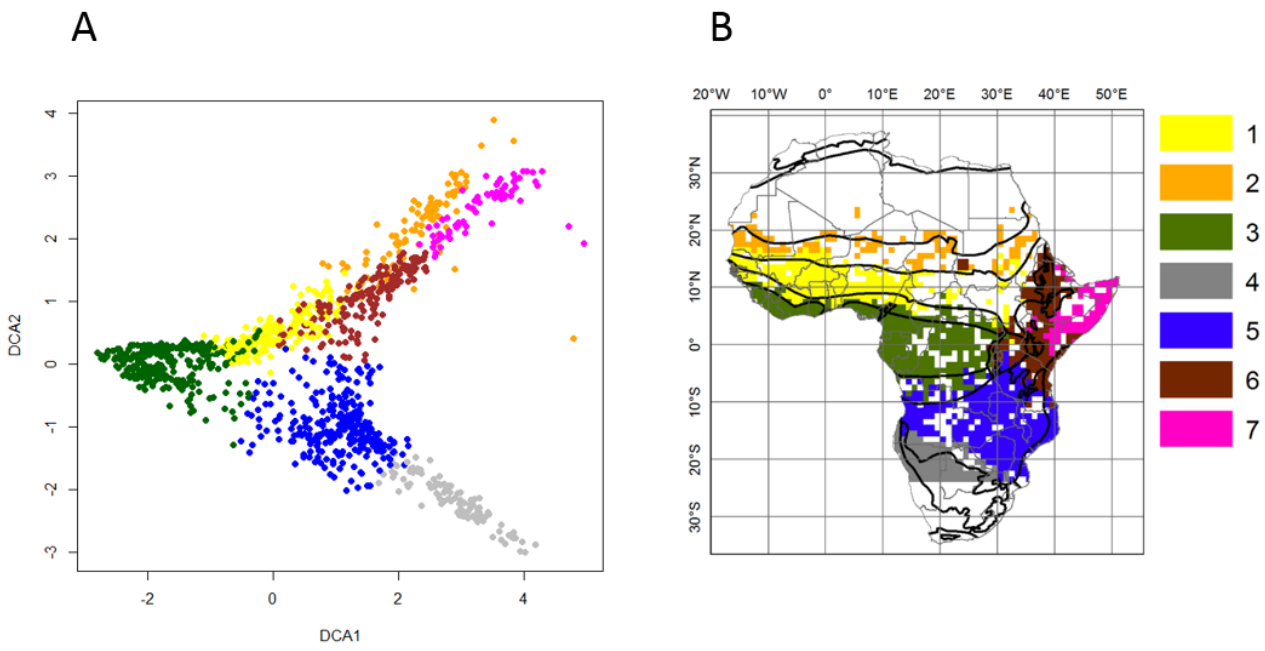


Table A1.1. Endemism rates for the PAM DCA 7 cluster solution.

	Region	No. taxa	No. endemic	% endemic
3	'Guineo-Congolia'	13,828	5681	41%
1	'Sudanian'	6391	400	6%
2	Sahel/Sahara	999	173	17%
5	Zambezian RCE	14,879	4165	28%
4	Kalahari-Highveld	5200	1163	22%
6	Afromontane RCE	12,790	3249	25%
7	Somalia-Masai RCE	2741	855	31%
	Total	**15,686**	**30,361**	**52%**

PAM analysis suggests that 7 clusters is the optimum number of clusters for the DCA ordination scores. This is the best we can do as regards dividing biological structure within this ordination, so we are more interested in how these 7 clusters fall in geographic space (Fig. A1.6b), than in how well the colours fit ordination space (Fig. A1.6a). Of course, 7 regions is fewer than White's 13. We have improved total endemism to 52%, but that is hardly surprising as the regions are larger. The Guineo-Congolian region is recognisable, and includes part of White's two transition zones. The Sudanian region is larger than in White's definition, and includes 1-2 degrees of both the Sahel and the GCSU transition zone. The Sahel (& Sahara) RTZ includes the grassland portion of White's Sahel, and White's Sahara. The Zambezian region has more or less the same definition as White's, except in the south east, where it is curtailed relative to the Kalahari-Highveld. The Kalahari-Highveld occupies around 25 degree squares more than in White's definition, and includes the Karoo-Namib. In the East, we have only 2 regions where White had 4. The Lake Victoria regional mosaic joins the Afromontane region. The Somalia-Masai region in our scheme is mainly restricted to Somalia, and the (mosaic of) Afromontane and Somalia-Masai cells from western Kenya and northern Tanzania are mostly assigned to our Afromontane region. White's Afromontane region occurs at high altitudes within a matrix of lower altitude woodlands, bushlands, thickets and secondary grasslands, assigned to the LVM or the Somalia-Masai regions. Our coarse resolution of 1 degree square means that cells with Afromontane vegetation also have these other types of vegetation, so cells with no Afromontane areas possessing those same 'other' species have joined them, resulting in a much larger 'Afromontane' region. We have not recovered the Zanzibar-Inhambane region.

The classification shown here is faithful to the DCA summary of the data, and maximises total endemism within the framework, but with only 7 regions is crude compared with White's regionalisation.

Figure A1.7. PAM 12-cluster solution of DCA axes scores. (A) Ordination scores clustered into 12 groups. (B) Geographic arrangement of these 12 clusters. Regions are named by the White chorion to which they most closely correspond (in inverted commas as they are defined differently), or are given provisional names.

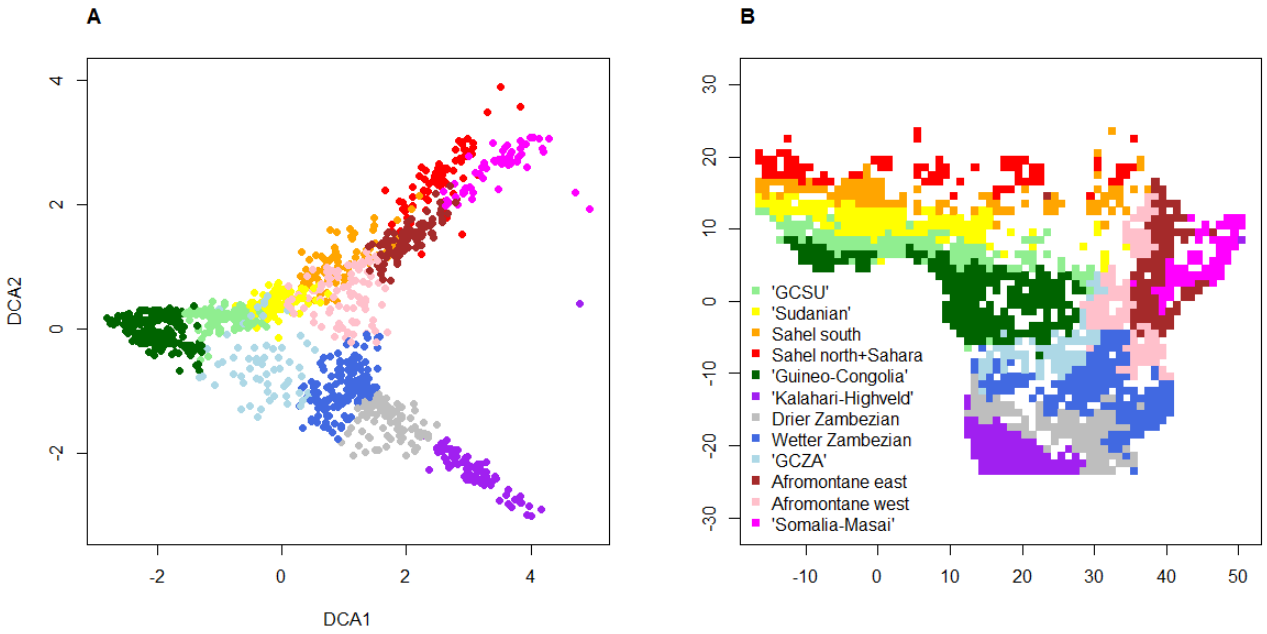


Table A1.2. Endemism rates for PAM DCA 12.

	Region	No. taxa	No. endemic	% endemic
1	GCSU	8023	388	5%
2	Sudanian	3997	133	3%
3	Sahel south	1791	63	4%
4	Sahel north	670	143	21%
5	Guineo-Congolia	11,660	3622	31%
6	Kalahari-Highveld	3324	724	22%
7	Drier Zambezian	6920	894	13%
8	Wetter Zambezian	10,899	1782	16%
9	GCZA	6935	448	6%
10	Eastern Afromontane	7936	1219	15%
11	Western Afromontane	11,757	1690	14%
12	Somalia-Masai	2408	746	31%
	Total	**11,852**	**30,361**	**39%**

Here we have 4 latitudinal regions to the north of Guineo-Congolia, where the 7 cluster solution had 2. The Afromontane region has divided into an eastern and a western portion, the Zambezian region has divided into a wetter portion and a drier portion; and the GCSU and GCZA have been recovered. We have not recovered White's Karoo-Namib, Zanzibar-Inhambane, or the LVM; instead we have gained an extra region to the north, an extra Zambezian region, an extra Afromontane region, and the two GC transition zones. The effect of the 'other' vegetation in our Afromontane samples is now clear, with the eastern Afromontane corresponding to the Somalia-Masai portion of the Afromontane matrix. The overall endemism rate is lower than for the 7 cluster solution as the regions are smaller. The total endemism rate is 39%, the same as in White's 12 region framework, but endemism within each region is higher, as these boundaries are drawn to fit this data set (except for Zambezian, which has divided into 2 and both are lower).

Figure A1.8. NMDS ordination of beta-sim matrix, with three axes, annotated with White's regions. The NMDS ordination does not have very strong structure relative to DCA. Here, stress=0.18. Even with 10 axes, and following several restarts, the stress was still more than 0.1. We discounted NMDS for the main paper for this reason.

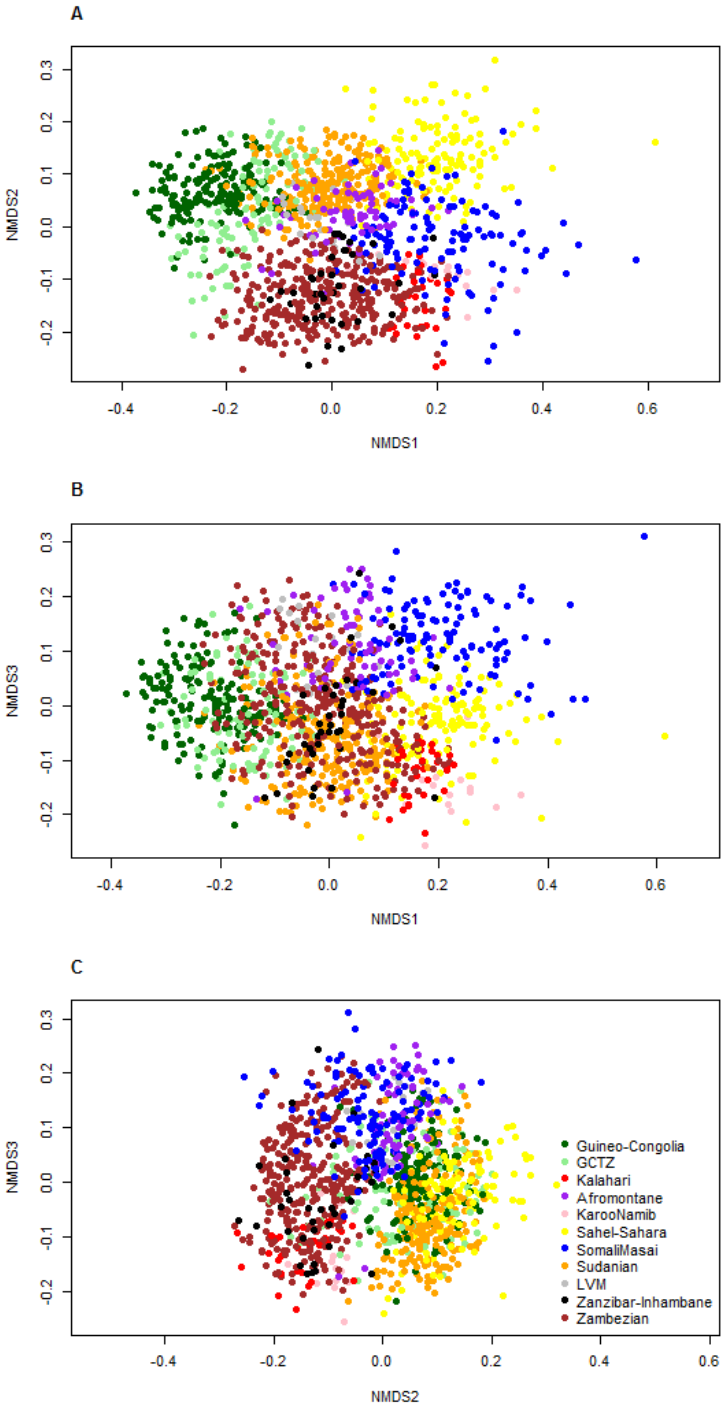
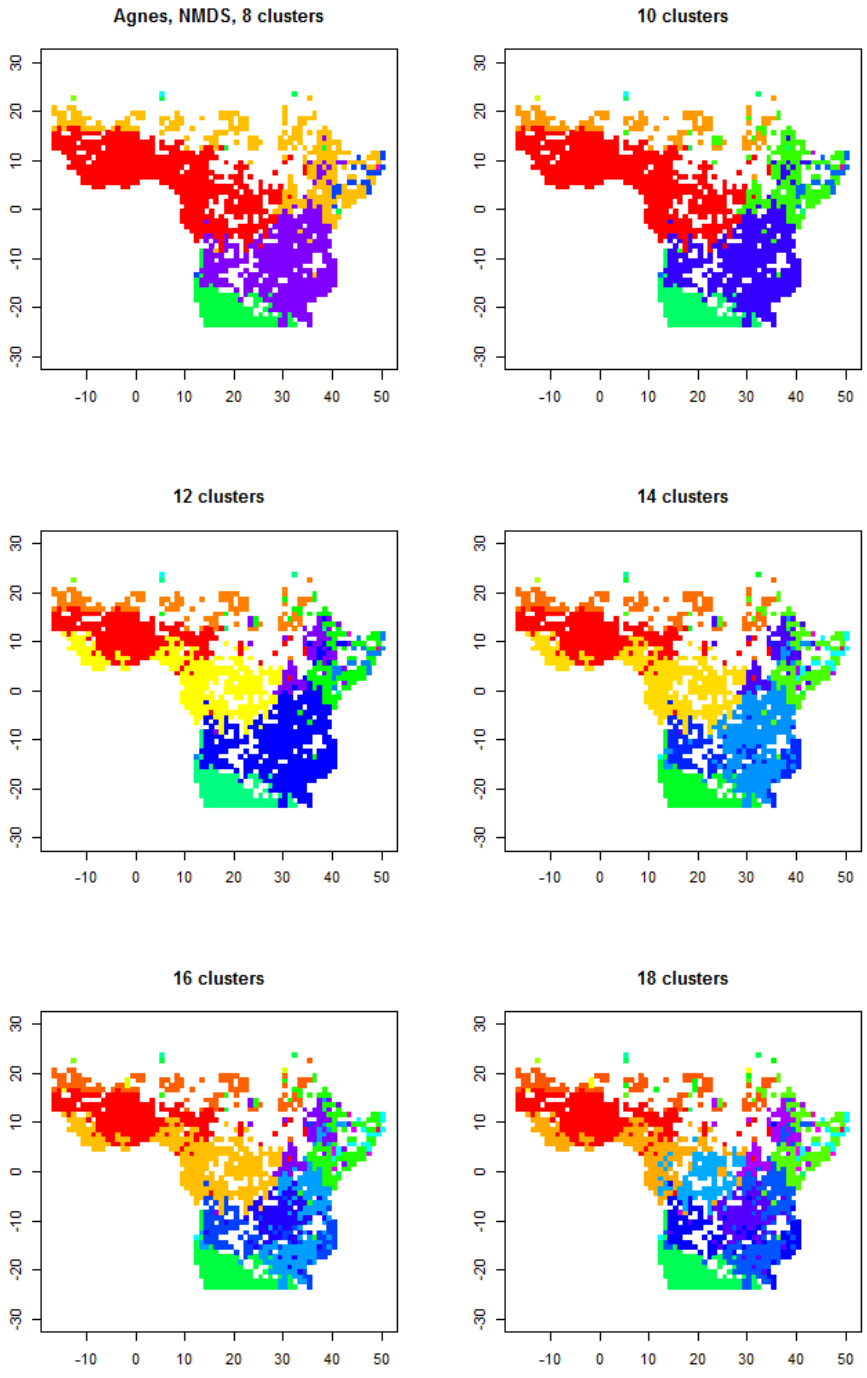


Figure A1.9. UPGMA clustering of NMDS axes 1-3 scores. Western Upper Guinea is partially subsumed into the Sudanian region, although eastern Upper Guinea is not. The geographic structure is generally poor and hard to interpret, relative to the clustering analyses of DCA axes, as we would predict from the NMDS ordination results themselves.



Appendix 2

MRI results for Chapter 6

Predicting bioquality across tropical Africa

Models presented here use MRI GCM.

set.seed=42

cf.mri <- cforest(GHI~.,data=alldata,control=cforest_unbiased(mtry=23,ntree=1500,

trace=TRUE))

> cf.mri

Random Forest using Conditional Inference Trees

Number of trees: 1500

Response: GHI

Inputs: LAT, LONG, AREA, ALTMEAN, ALTSTD, SLMEAN, SLSTD, LSVARIETY, LSMAJORITY, LTVARIETY, LTMAJORITY, MGVARIETY, TOVARIETY, TOMAJORITY, TMVARIETY, TMMAJORITY, CHORIONMAJ, CHORVAR, CLASS, SUBCLASS, FORMATION, DIVISION, MG1, MG2, MG3, MRI_Modern.bio.1, MRI_Modern.bio.2, MRI_Modern.bio.3, MRI_Modern.bio.4, MRI_Modern.bio.5, MRI_Modern.bio.6, MRI_Modern.bio.7, MRI_Modern.bio.8, MRI_Modern.bio.9, MRI_Modern.bio.10, MRI_Modern.bio.11, MRI_Modern.bio.12, MRI_Modern.bio.13, MRI_Modern.bio.14, MRI_Modern.bio.15, MRI_Modern.bio.16, MRI_Modern.bio.17, MRI_Modern.bio.18, MRI_Modern.bio.19, stability01, stability02, stability03, stability04, stability05, stability06, stability07, stability08, stability09, stability10, stability11, stability12, stability13, stability14, stability15, stability16, stability17, stability18, stability19

Number of observations: 3426

Mean of squared residuals: 13089.99

root-mean-square error (RMSE): 114.4115

Mean absolute error: 75.35231

St. dev. Absolute error: 86.10565

Figure A2.1. Plot of OOB predictions against input (observed) GHIs. Grey dotted line shows the 1:1 line, solid grey line shows the linear regression of the two datasets. Predicted values are more average than the input values.

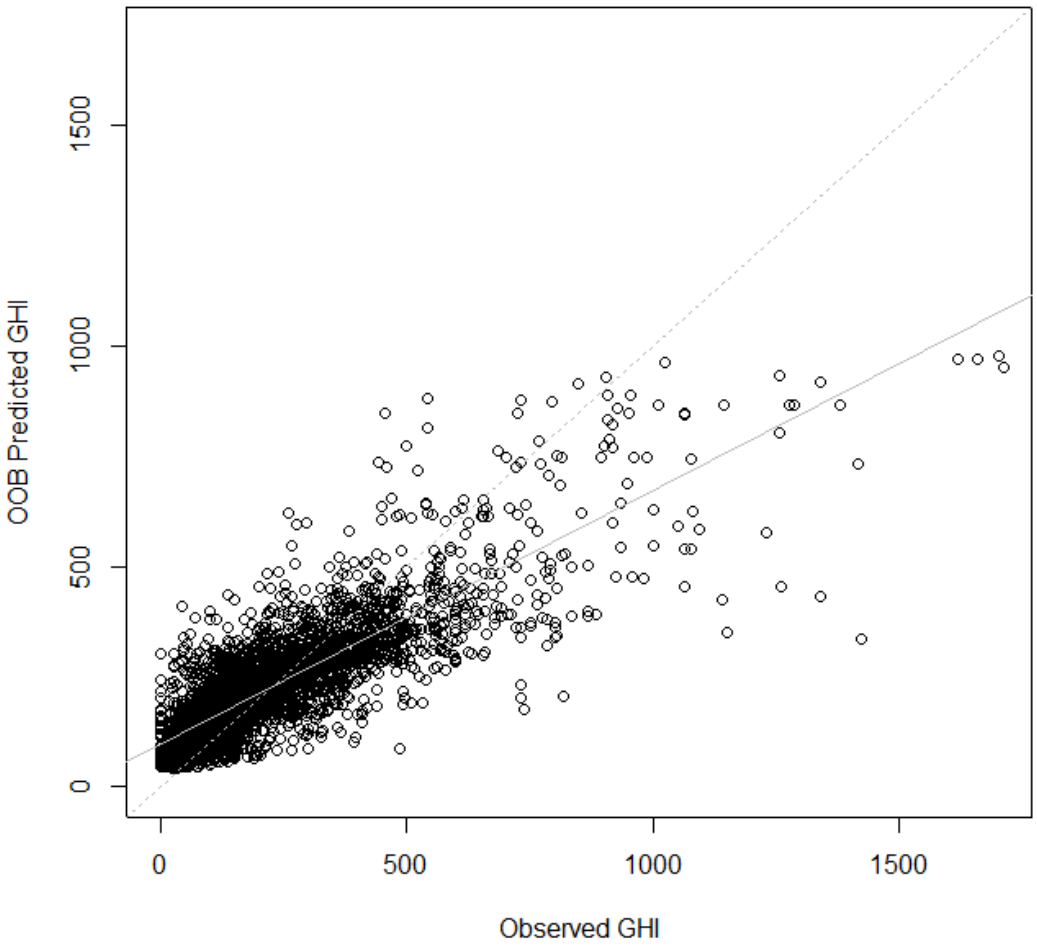
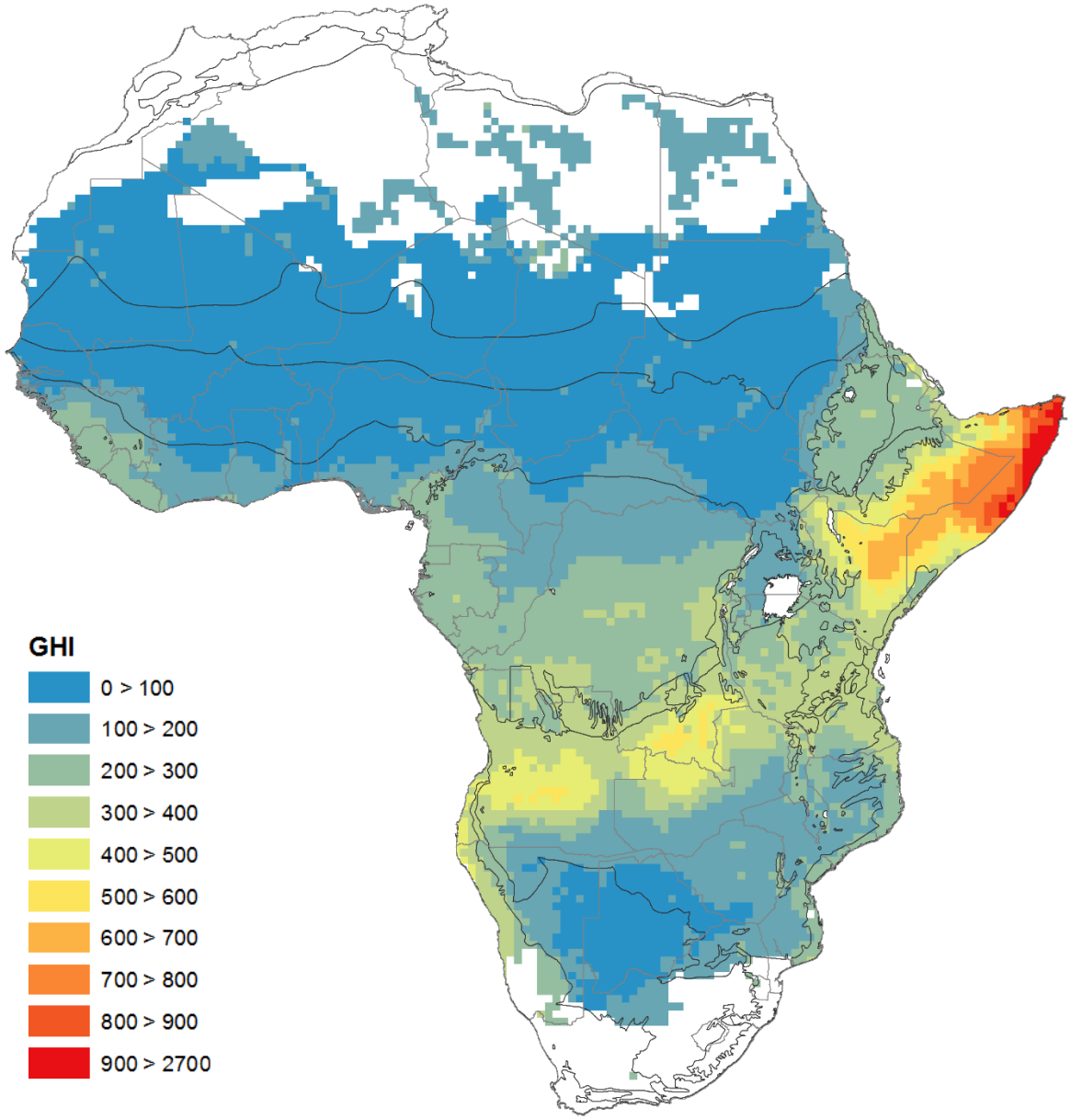


Figure A2.2. Conditional forest modelled output GHI for tropical Africa.



Variable importance

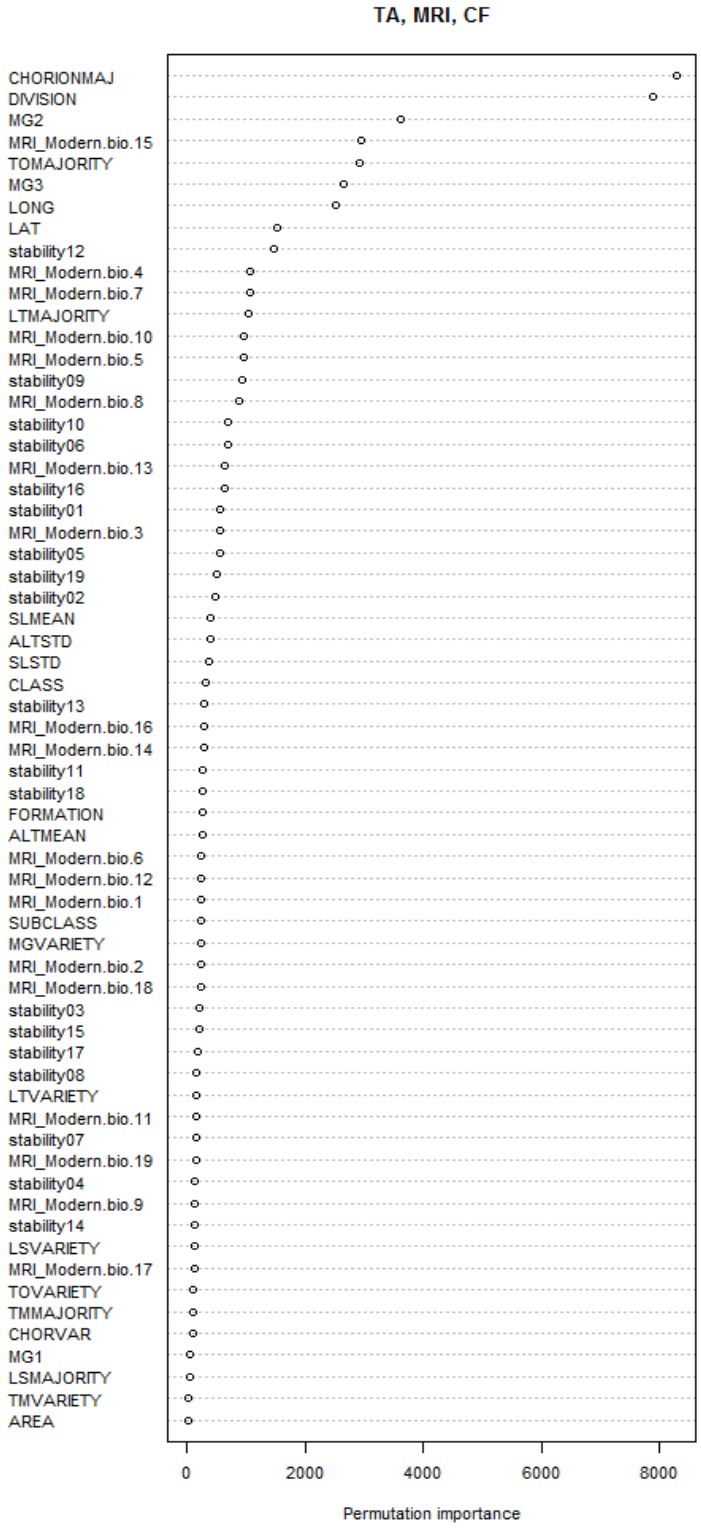
Importance values from 3 runs, data file: C/R/RF/ecoClimate/MRI/cf_mri_impruns.csv

Friedman test on rank orders, Friedman chi-squared = 1.5556, df = 2, p-value = 0.4594; Accept null (rank importance is the same on each run).

Table A2.1. T-tests for (logged) variable importance, means not logged. Results from seed 42 are shown here (arbitrarily).

Variable	Mean	p
Bio 1-11	998.7621	p-value = 0.5058
Bio12-19	605.6802	
Stability 1-11	433.2434	p-value = 0.8722
Stability 12-19	452.9001	
Bio 1-19	593.2246	p-value = 0.643
Stability 1-19	441.5199	

Figure A2.3. Importance scores (seed 42) for explanatory variables within the CF GHI model.



Major relationships

Figure A2.4. GHI (y) plotted against stability measures for the temperature related bio1-11.

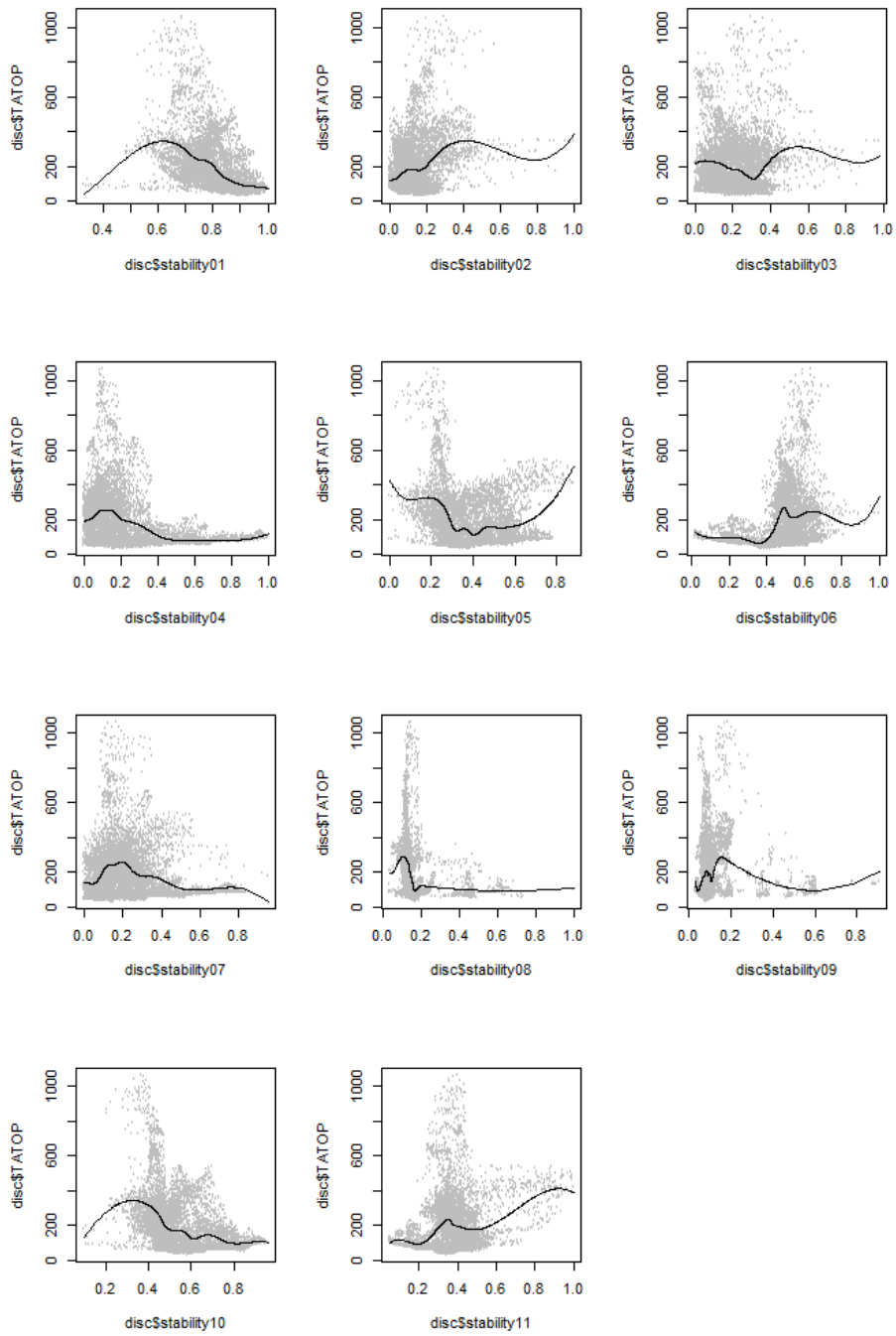


Figure A2.5. GHI (y) plotted against stability measures for the rain related bio12-19.

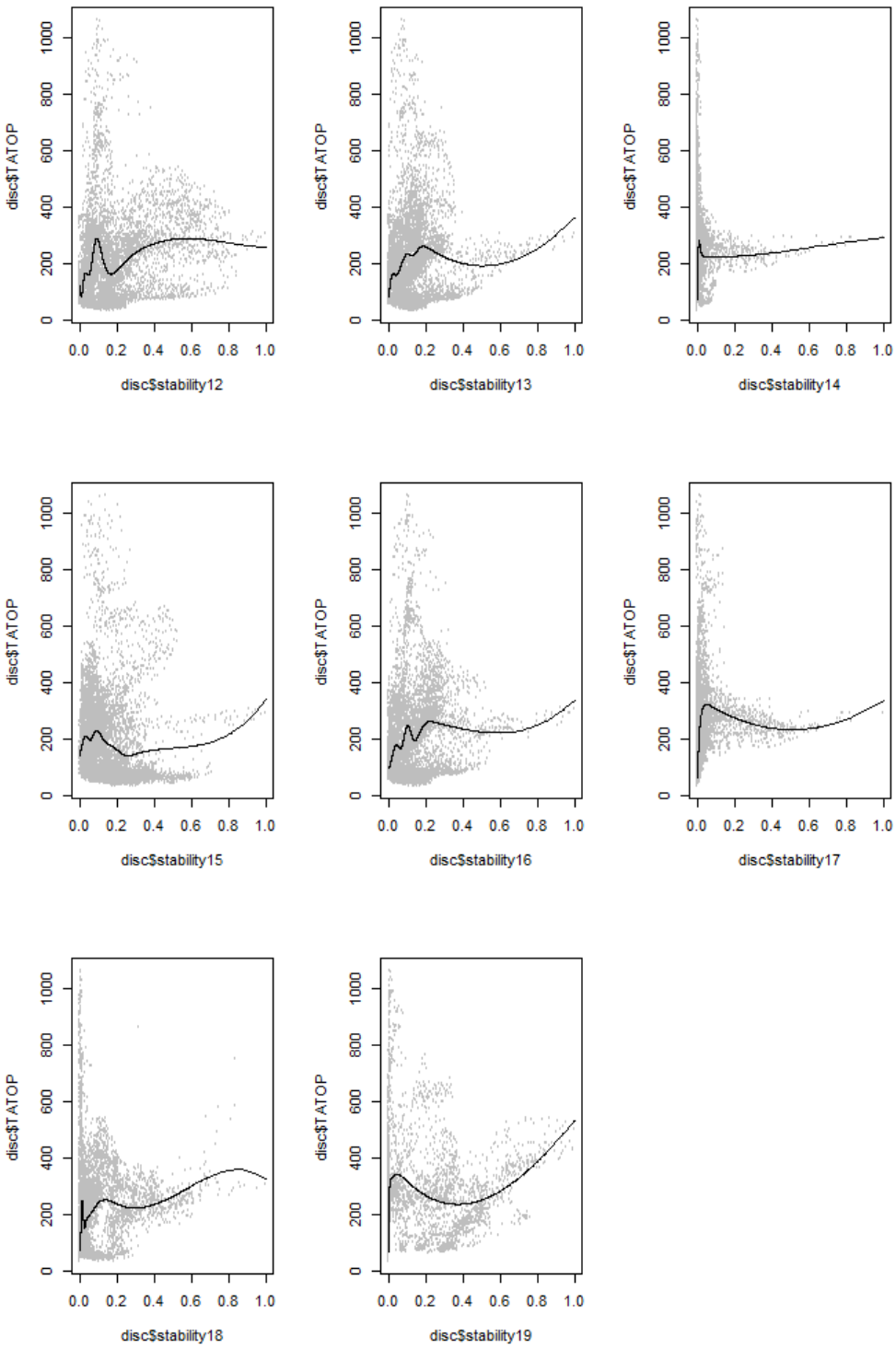


Figure A2.6. GHI (y) plotted against mean modern measures for the temperature related bio1-11.

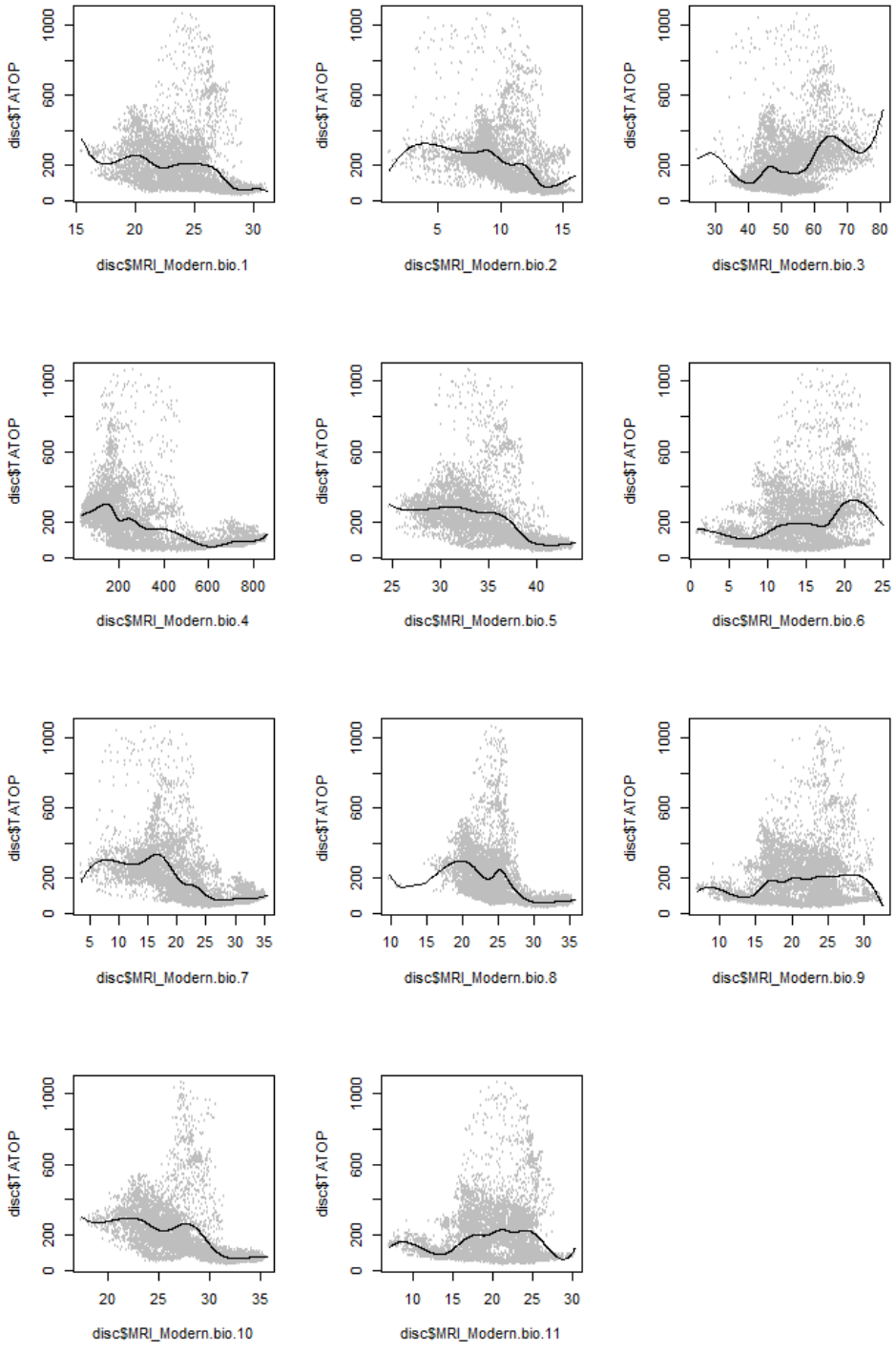


Figure A2.7. GHI (y) plotted against mean modern measures for the rainfall related bio12-19.

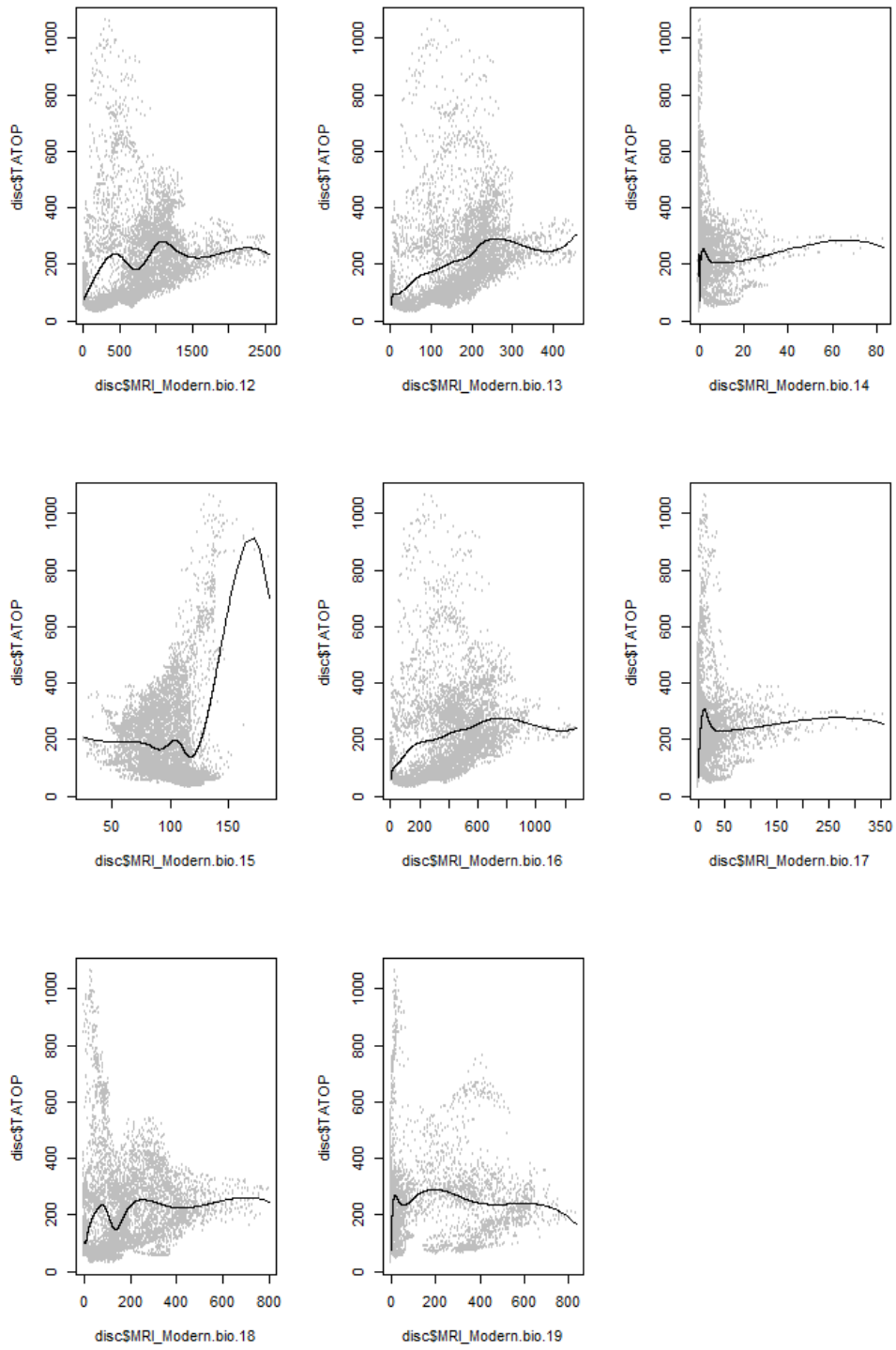


Figure A2.8. GHI (y) plotted against mean altitude, mean slope, latitude, longitude and area.

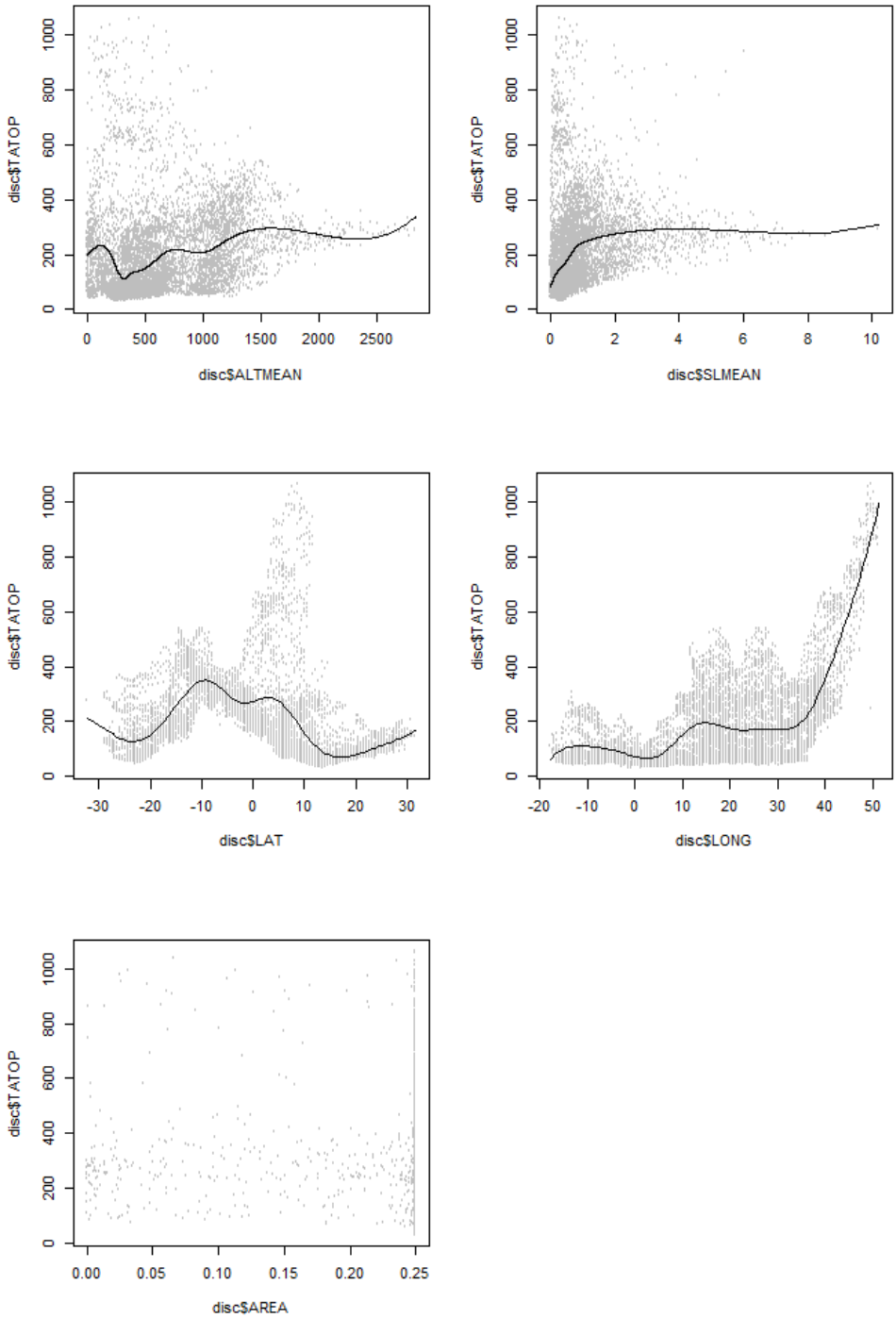


Figure A2.9. Boxplot showing median GHI for each chorion. IX Afroalpine was majority chorion in no cell; XI GC/Sud transition zone I summarised with GC; XV (TongaPondo), V (Cape), XVIII (Sahara-Med) and VII (Med) have no training data; (XIX and XX are Madagascan); 0 denotes missing; MA mangrove, WA water.

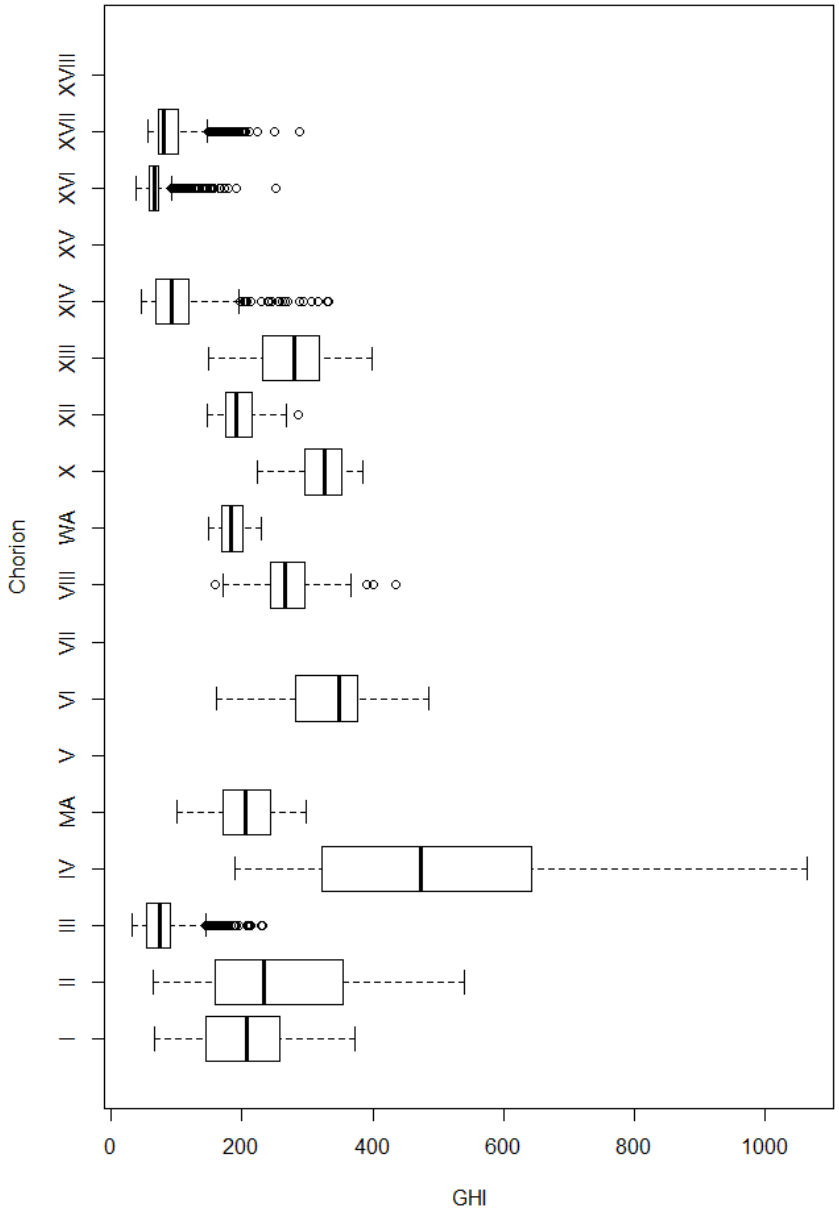


Figure A2.10. Boxplot showing median GHI for each lithology class. 17=allvuim-Gypsum, found only in Somalia.

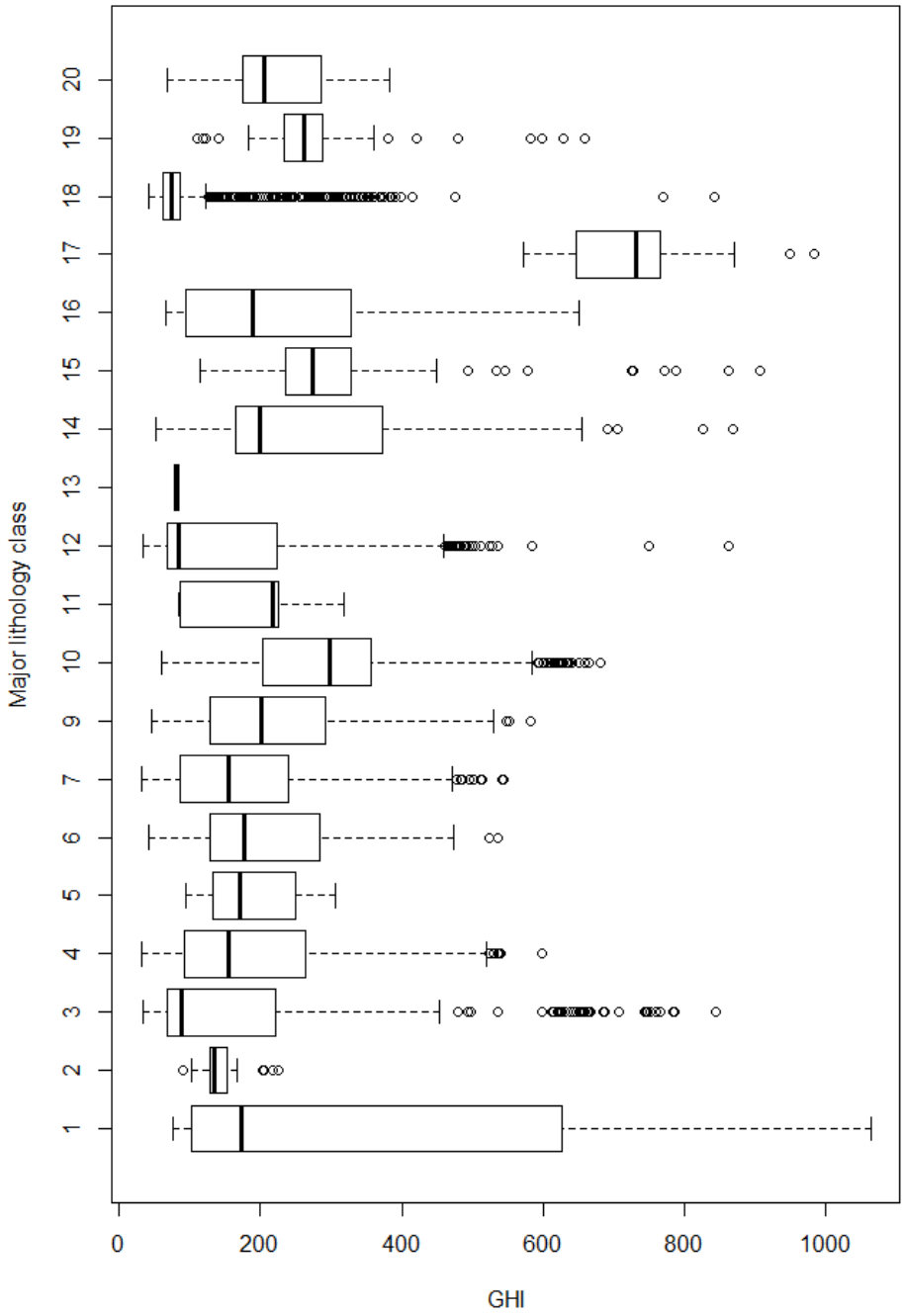
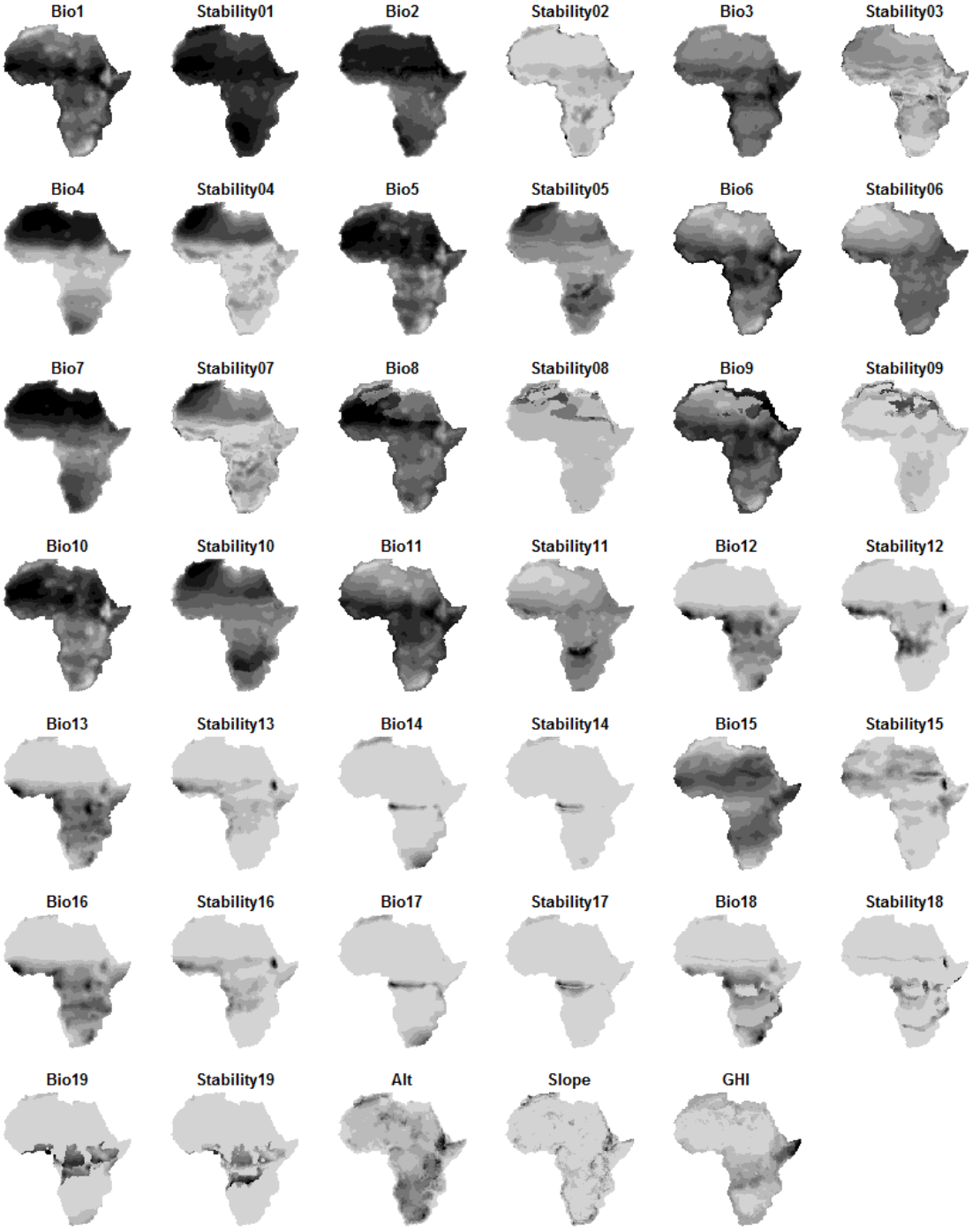


Figure A2.11. MRI Modern Bioclim variables and stability measures. Dark=high values; 10 colours from light grey to black; equal intervals (not quantiles), e.g. for stability, 0-0.1, 0.1-0.2 etc. High values represent max change, low values represent max stability.

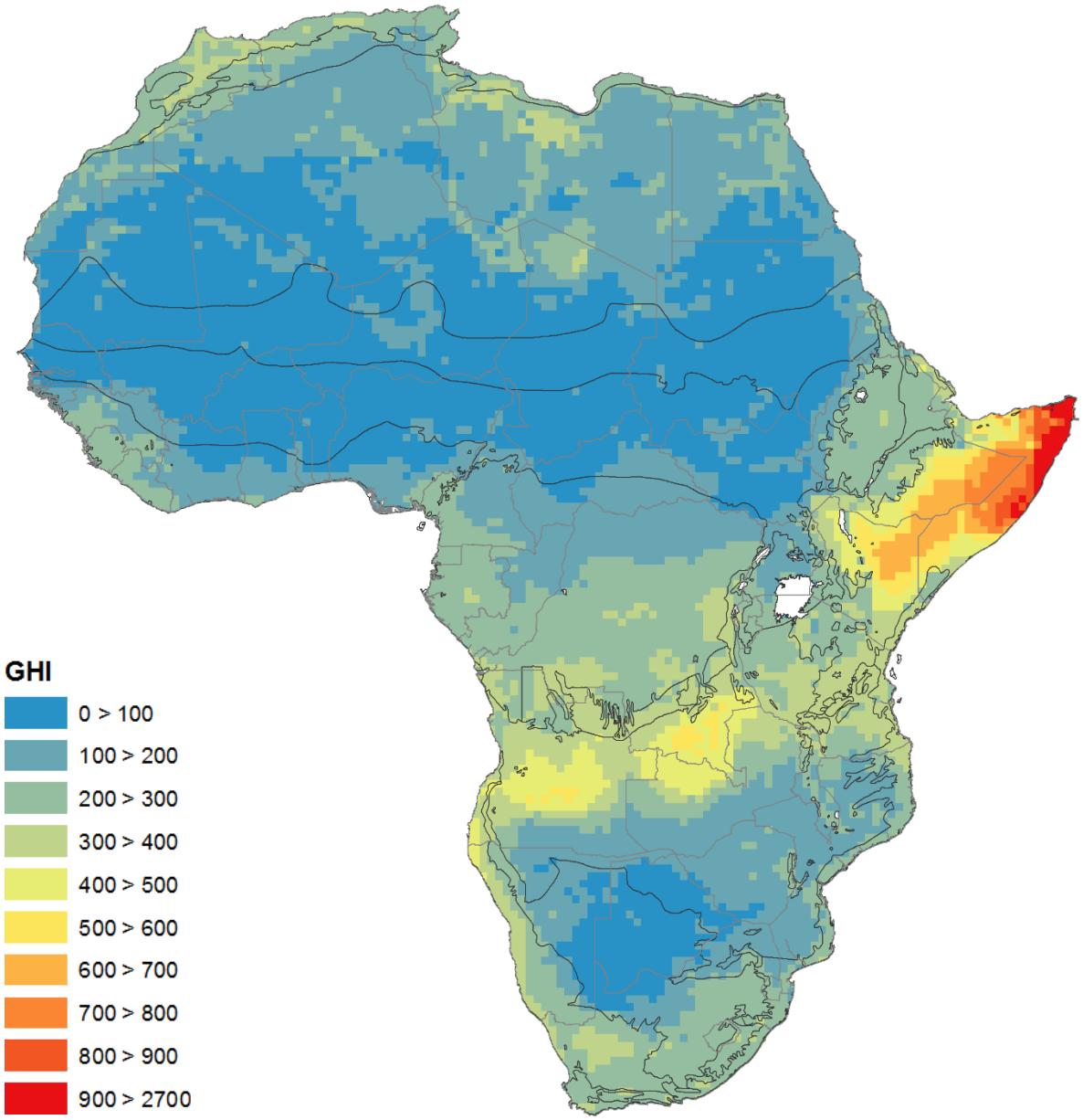


Extrapolating to mainland Africa

All Africa model:

```
set.seed=42

cf.mri.AA <- cforest(GHI~.,data=alldata,control=cforest_unbiased(mtry=23,ntree=1500,

        trace=TRUE))

> cf.mri.AA

      Random Forest using Conditional Inference Trees

Number of trees:  1500

Response:  GHI

Inputs:  LAT, LONG, AREA, ALTMEAN, ALTSTD, SLMEAN, SLSTD, LSVARIETY, LSMAJORITY,
LTVARIETY, LTMAJORITY, MGVARIETY, TOVARIETY, TMVARIETY, TMMAJORITY, CHORVAR, CLASS,
SUBCLASS, FORMATION, DIVISION, MRI_Modern.bio.1, MRI_Modern.bio.2, MRI_Modern.bio.3,
MRI_Modern.bio.4, MRI_Modern.bio.5, MRI_Modern.bio.6, MRI_Modern.bio.7,
MRI_Modern.bio.8, MRI_Modern.bio.9, MRI_Modern.bio.10, MRI_Modern.bio.11,
MRI_Modern.bio.12, MRI_Modern.bio.13, MRI_Modern.bio.14, MRI_Modern.bio.15,
MRI_Modern.bio.16, MRI_Modern.bio.17, MRI_Modern.bio.18, MRI_Modern.bio.19, stability01,
stability02, stability03, stability04, stability05, stability06, stability07, stability08, stability09,
stability10, stability11, stability12, stability13, stability14, stability15, stability16, stability17,
stability18, stability19

Number of observations:  3426

Mean of squared residuals: 12939.31

root-mean-square error (RMSE): 113.7511

Mean absolute error: 75.39054

St. dev. Absolute error: 85.19212
```

Figure A2.12. Conditional forest modelled GHI map of mainland Africa.



Appendix 3

Climate appendix for Chapter 6

Figure A3.1 shows that variation between the different GCMs in their estimation of key climate parameters, as represented in the ecoClimate database, is big. Panel A shows the estimate of annual mean temperature for every half degree cell in Africa as estimated by IPSL, against the same variable estimated by CCSM, their linear regression (grey line), and the 1:1 line. Although the correlation is fairly strong (0.83), the residuals are still large, and reveal that discrepancies in annual temperature estimates of 3°C for a cell are not uncommon. Panel C shows annual modern rainfall for each cell in Africa, as estimated by two different GCMs in the ecoClimate database, and shows even larger residuals than in panel A. The largest discrepancy for a cell is around 1500 mm rainfall a year, which in Ghana would equate to the difference between the Wet Evergreen forest zone (2000 mm/year) and the savanna zone (500 mm/year). Panel B shows IPSL's estimate of *modern* annual mean temperature for each half degree cell in Africa against its estimate for *LGM* annual mean temperature. Not only is the correlation stronger (1.06) but the residual values are smaller than in panel A. To find out what IPSL estimates the current temperature to be for a cell, it is more reliable to find out what it estimates for that cell 20,000 years ago and add 6.5°C, than it is to ask a different climate model what it estimates the modern temperature of that cell to be. The residuals in Panel D, LGM rainfall against modern rainfall, are also smaller than in Panel C. This wide variation between the GCMs, combined with less variation amongst a single model over modern to paleo timescales, makes it unsurprising that we found no trend between age and importance when all the GCM data were included together in a single randomForest model.

Figure A3.1. Correlation between modern and LGM estimates of mean temperature and rainfall from the IPSL and GISS general circulation models.

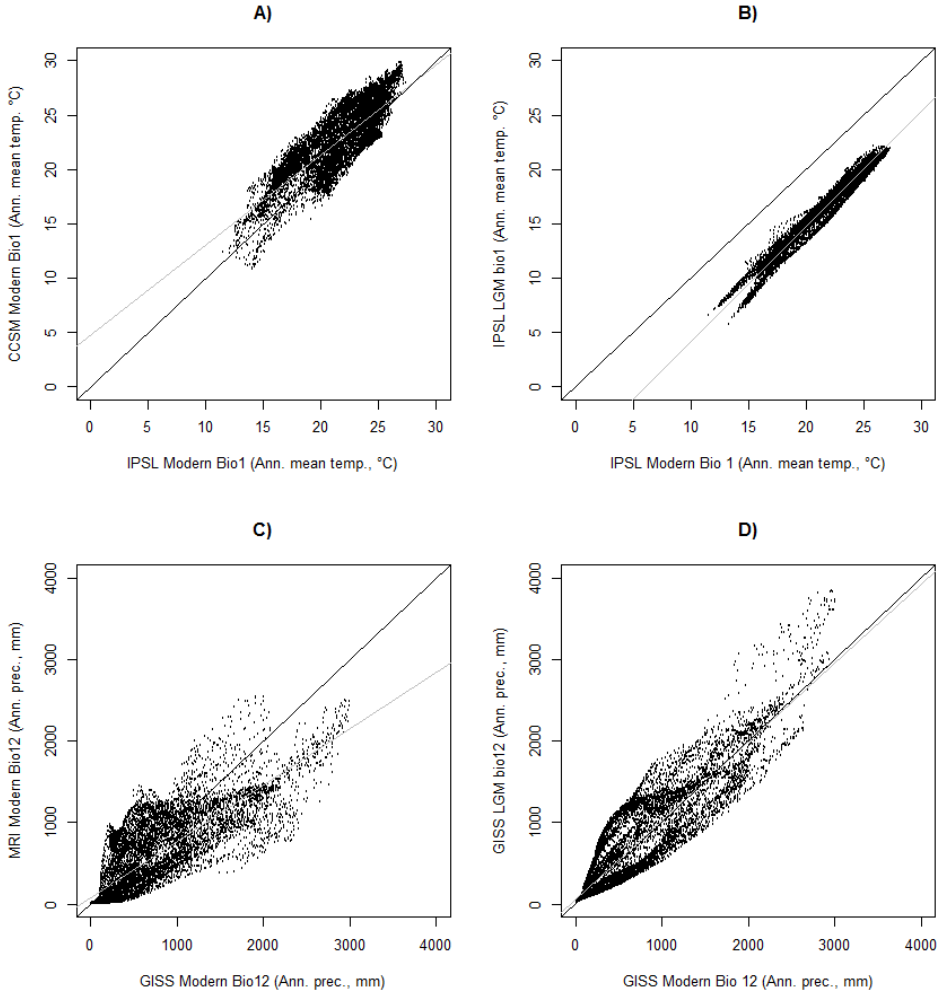


Figure A3.2. Plots of modern annual mean temperature against LGM annual mean temperature, as estimated by 8 of the GCMs, as represented in the ecoClimate database. With the possible exception of the MIROC model, the correlations are very linear (collinear), such that LGM annual mean temperature is unlikely to hold much additional predictive power to modern mean temperature (and vice versa).

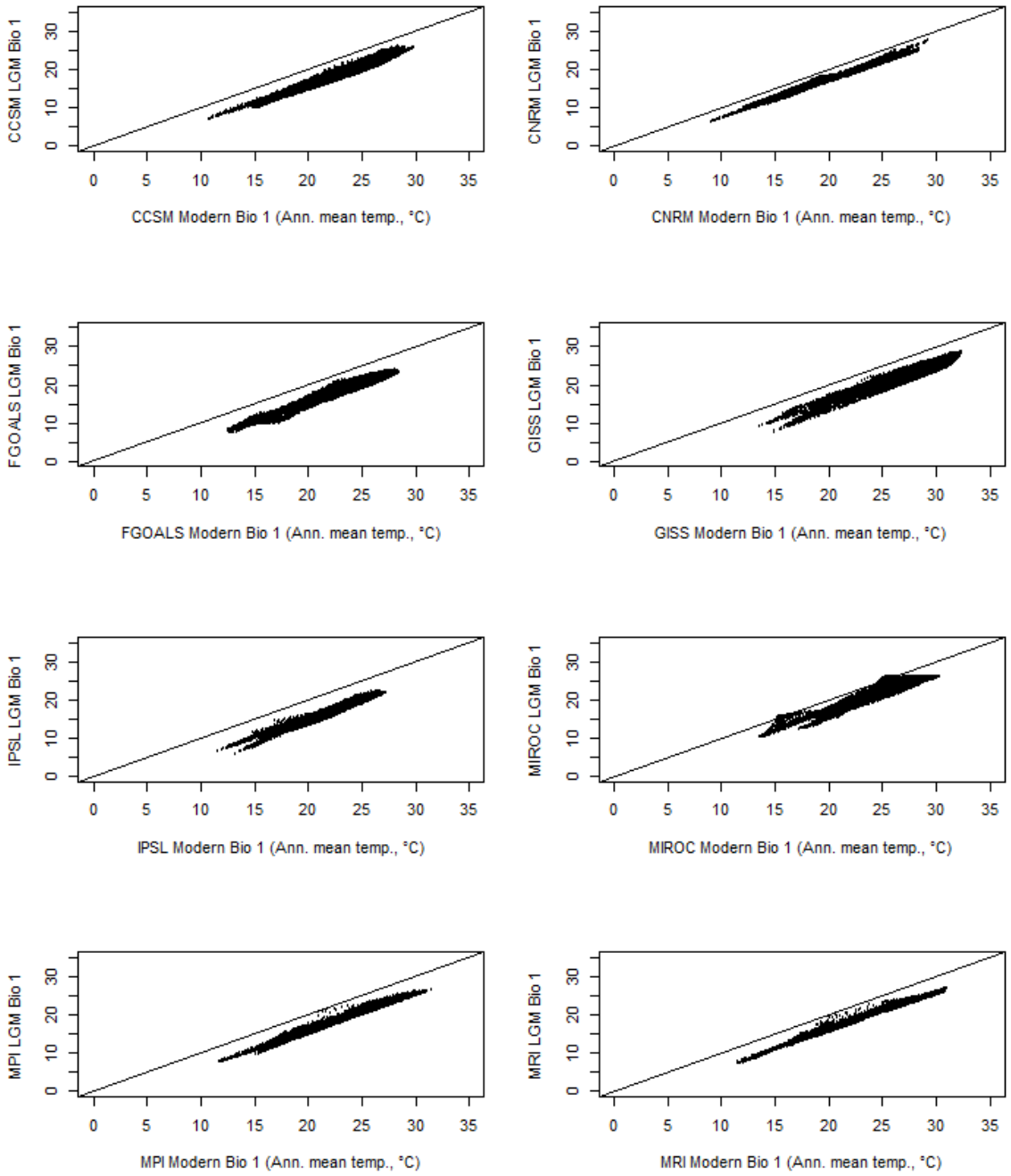


Figure A3.3. Plots of modern annual mean temperature against mid Holocene (6000 years ago) annual mean temperature, as estimated by 7 of the GCMs, as represented in the ecoClimate database. Correlation is strong. GISS does not provide an estimate for the mid Holocene.

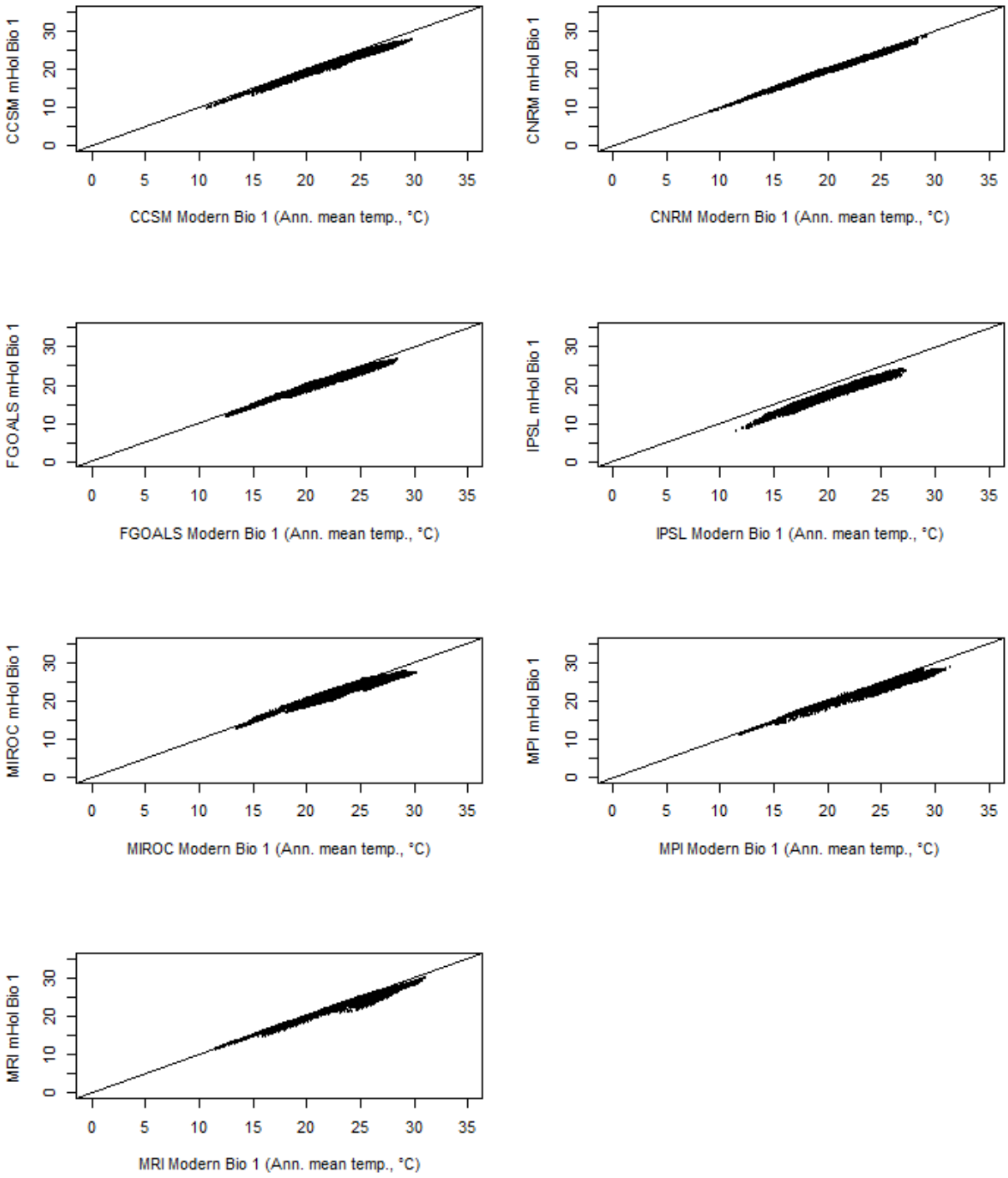


Figure A3.4. Plots of modern annual precipitation against LGM annual precipitation, as estimated by 8 of the GCMs, as represented in the ecoClimate database. There is less agreement between the models than for temperature (Figure 8), suggesting in general that rainfall parameters are subject to greater uncertainty than temperature estimates (Varela et al. 2015). However, the correlations are less strong, meaning that LGM rainfall values hold some information that is additional to modern rainfall estimates (whether or not this information is reliable is a separate question).

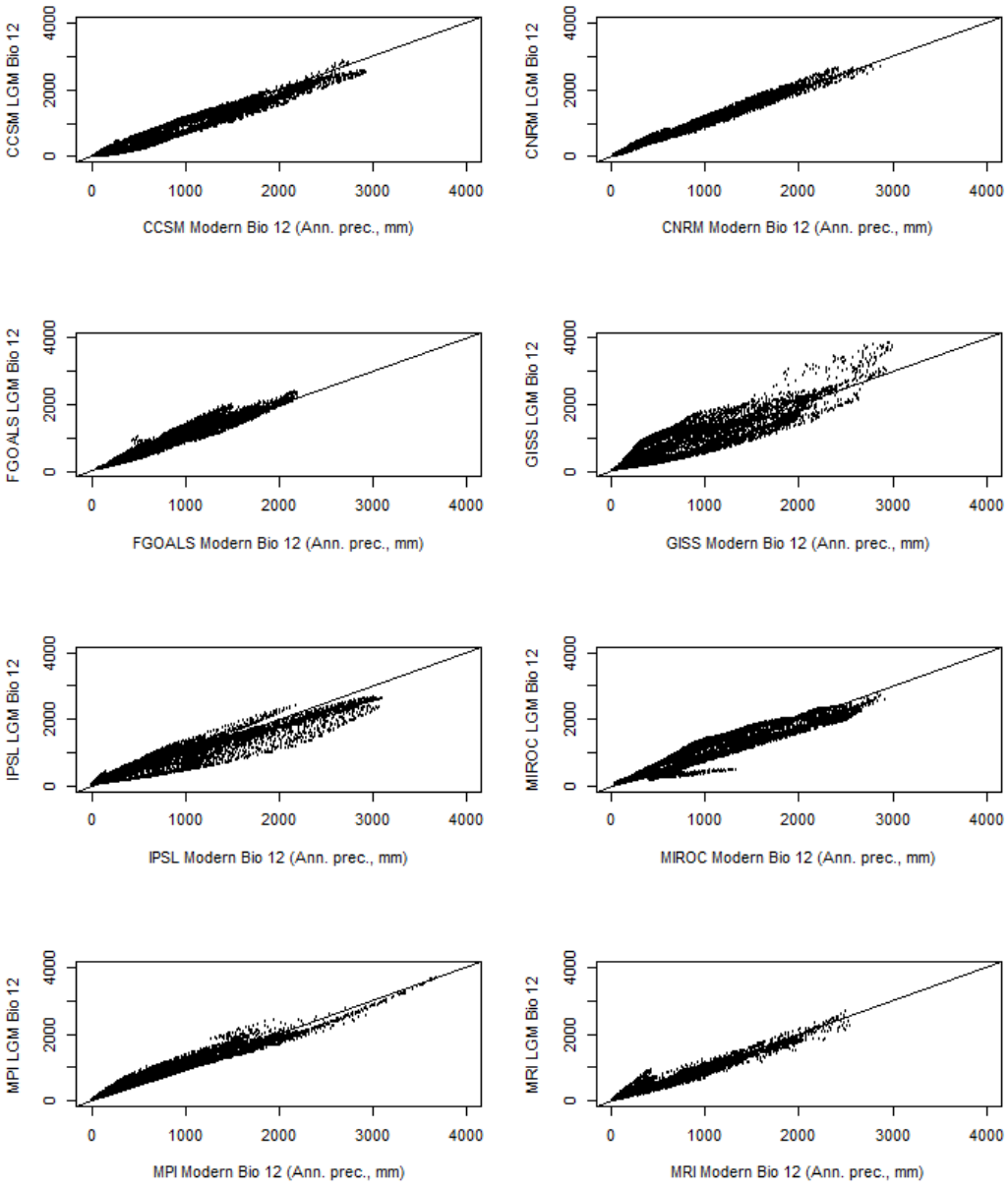


Figure A3.4. Plots of modern annual precipitation against mid Holocene annual precipitation, as estimated by 7 of the GCMs. There is less agreement between the models than for temperature (Figure 9), and the relationships are also different compared with Figure 10, suggesting that mHol rainfall values hold some information that is additional to both modern and LGM rainfall estimates.