Journal of Biogeography (J. Biogeogr.) (2011) **38**, 1164–1176

ORIGINAL ARTICLE

Predicting alpha diversity of African rain forests: models based on climate and satellite-derived data do not perform better than a purely spatial model

Ingrid Parmentier[1]*, Ryan J. Harrigan[2], Wolfgang Buermann[2], Edward T. A. Mitchard[3], Sassan Saatchi[2], Yadvinder Malhi[4], Frans Bongers[5], William D. Hawthorne[6], Miguel E. Leal[7], Simon L. Lewis[8], Louis Nusbaumer[9], Douglas Sheil[10,11], Marc S. M. Sosef[12], Kofi Affum-Baffoe[13], Adama Bakayoko[14], George B. Chuyong[15], Cyrille Chatelain[9], James A. Comiskey[16], Gilles Dauby[1], Jean-Louis Doucet[17], Sophie Fauset[8], Laurent Gautier[9], Jean-François Gillet[18], David Kenfack[19], François N. Kouamé[20], Edouard K. Kouassi[20], Lazare A. Kouka[21], Marc P. E. Parren[22], Kelvin S.-H. Peh[8], Jan M. Reitsma[23], Bruno Senterre[1], Bonaventure Sonké[24], Terry C. H. Sunderland[11], Mike D. Swaine[25], Mbatchou G. P. Tchouto[26], Duncan Thomas[27], Johan L. C. H. Van Valkenburg[28] and Olivier J. Hardy[1]

[1]Evolutionary Biology & Ecology, Faculté des Sciences, Université Libre de Bruxelles, Brussels, Belgium, [2]Center for Tropical Research, Institute of the Environment, University of California, Los Angeles, CA, USA, [3]School of Geosciences, University of Edinburgh, Edinburgh, UK, [4]Environmental Change Institute, School of Geography and the Environment, University of Oxford, Oxford, UK, [5]Centre for Ecosystem Studies, Wageningen University, Wageningen, The Netherlands, [6]Department of Plant Sciences, University of Oxford, Oxford, UK, [7]Missouri Botanical Garden, St. Louis, MO, USA, [8]Earth & Biosphere Institute, School of Geography, University of Leeds, Leeds, UK, [9]Conservatoire et Jardin botaniques de la Ville de Genève, Chambésy, Switzerland, [10]Institute of Tropical Forest Conservation (ITFC), Kabale, Uganda, [11]Centre for International Forestry Research (CIFOR), Bogor, Indonesia, [12]Netherlands Centre for Biodiversity Naturalis (section NHN), Biosystematics Group, Wageningen University, Wageningen, The Netherlands, [13]Resource Management Support Centre, Forestry Commission of Ghana, Kumasi, Ghana[14]Centre Suisse de Recherches Scientifiques en Côte d'Ivoire, Abidjan, Côte d'Ivoire, [15]Department of Plant and Animal Sciences, University of Buea, Buea, Cameroon, [16]Mid-Atlantic Network, Inventory and Monitoring Program, National Park Service, Fredericksburg, VA, USA, [17]Laboratory of

ABSTRACT

Aim Our aim was to evaluate the extent to which we can predict and map tree alpha diversity across broad spatial scales either by using climate and remote sensing data or by exploiting spatial autocorrelation patterns.

Location Tropical rain forest, West Africa and Atlantic Central Africa.

Methods Alpha diversity estimates were compiled for trees with diameter at breast height ≥ 10 cm in 573 inventory plots. Linear regression (ordinary least squares, OLS) and random forest (RF) statistical techniques were used to project alpha diversity estimates at unsampled locations using climate data and remote sensing data [Moderate Resolution Imaging Spectroradiometer (MODIS), normalized difference vegetation index (NDVI), Quick Scatterometer (QSCAT), tree cover, elevation]. The prediction reliabilities of OLS and RF models were evaluated using a novel approach and compared to that of a kriging model based on geographic location alone.

Results The predictive power of the kriging model was comparable to that of OLS and RF models based on climatic and remote sensing data. The three models provided congruent predictions of alpha diversity in well-sampled areas but not in poorly inventoried locations. The reliability of the predictions of all three models declined markedly with distance from points with inventory data, becoming very low at distances > 50 km. According to inventory data, Atlantic Central African forests display a higher mean alpha diversity than do West African forests.

Main conclusions The lower tree alpha diversity in West Africa than in Atlantic Central Africa may reflect a richer regional species pool in the latter. Our results emphasize and illustrate the need to test model predictions in a spatially explicit manner. Good OLS or RF model predictions from inventory data at short distance largely result from the strong spatial autocorrelation displayed by both the alpha diversity and the predictive variables rather than necessarily from causal relationships. Our results suggest that alpha diversity is driven by history rather

http://wileyonlinelibrary.com/journal/jbi
doi:10.1111/j.1365-2699.2010.02467.x

Journal of Biogeography

Tropical and Subtropical Forestry, Unit of Forest and Nature Management, Gembloux Agro-Bio Tech, University of Liege, Gembloux, Belgium, [18]*Nature +, Laboratory of Tropical and Subtropical Forestry, Unit of Forest and Nature Management, Gembloux Agro-Bio Tech, University of Liège, Gembloux, Belgium,* [19]*CTFS-SIGEO Africa Program Coordinator, Arnold Arboretum, Harvard University, Cambridge, MA, USA,* [20]*Laboratoire de Botanique, Université de Cocody, Abidjan, Côte d'Ivoire,*

*Correspondence: Ingrid Parmentier, Université Libre de Bruxelles, Evolutionary Biology & Ecology, CP 160/12, Av. F. Roosevelt 50, B-1050 Brussels, Belgium.
E-mail: inparmen@ulb.ac.be

than by the contemporary environment. Given the low predictive power of models, we call for a major effort to broaden the geographical extent and intensity of forest assessments to expand our knowledge of African rain forest diversity.

Keywords

African rain forests, biodiversity, climate, kriging, map, modelling, ordinary least squares, random forest, spatial autocorrelation, tree alpha diversity.

[21]*Institut Notre Dame, Brussels, Belgium,* [22]*Tropenbos International Congo-Basin Programme, Yaoundé, Cameroon,* [23]*Bureau Waardenburg bv, Consultants for Environment & Ecology, Culemborg, The Netherlands,* [24]*Ecole Normale Supérieure de Yaoundé, Université de Yaoundé I, Yaoundé, Cameroon,* [25]*Institute of Biological and Environmental Sciences, School of Biological Sciences, University of Aberdeen, Aberdeen, UK,* [26]*Ecole Nationale des Eaux et Forêts, Mbalmayo, Cameroon,* [27]*Department of Botany and Plant Pathology, Oregon State University, Corvallis, OR, USA,* [28]*Plant Protection Service, Wageningen, The Netherlands*

INTRODUCTION

Rain forests are species-rich ecosystems providing multiple services to humanity (Myers, 1997). In the last century, many tropical forest areas have been heavily exploited or cleared and converted to other land uses, leading to their fragmentation and alteration (Laurance *et al.*, 1999). The consequences of these changes include biodiversity loss and increasing atmospheric carbon dioxide concentrations, resulting in climate change, due to the conversion of high-carbon storage forest to low-carbon storage agriculture (Lewis, 2006). Despite the importance of rain forests for terrestrial biodiversity, the causes of the diversity gradients within and between the world's main rain forest areas remain poorly understood (Givnish, 1999; Parmentier *et al.*, 2007). In addition, large areas of rain forests are unexplored by scientists, giving a fragmentary view of spatial patterns of diversity, even for well-studied organisms such as plants (Burgess *et al.*, 2005).

A potential way to compensate for the lack of inventory data is through modelling using georeferenced variables and determining their relationship with plant diversity (Ferrier *et al.*, 2004; Jarnevich *et al.*, 2006). Two main categories of models have been proposed: theoretical models and experimental models. Theoretical models are based on extrapolations of ecologically meaningful relationships between plants and environments, mainly water–energy dynamics (O'Brien, 2006), and provide coarse-scale diversity estimates. Experimental models are built upon the correlations of diversity datasets with a set of potentially predictive variables without a priori assuming the causal relationships. Experimental models can provide estimates at a local scale, in addition to the coarse scale of theoretical models. Two main types of variables appear to be correlated with plant diversity, and have been used in experimental modelling: (1) environmental variables that could potentially have a causal effect on plant diversity, and (2) variables that directly describe properties of the vegetation. Examples of environmental variables include those related to climate, topography and soil fertility (e.g. Currie & Paquin, 1987, in North America; ter Steege *et al.*, 2003, and Clinebell *et al.*, 1995, in Amazonia; Bongers *et al.*, 1999, and Field *et al.*, 2005, in Africa; Slik *et al.*, 2009, in Borneo). Examples of variables describing the vegetation itself include tree turnover (Phillips *et al.*, 1994, in Amazonia), stem density (ter Steege *et al.*, 2003, in Amazonia), remote sensing variables related to photosynthetic activity (normalized difference vegetation index, NDVI: Schmidt *et al.*, 2008, in Burkina Faso; leaf area index: Saatchi *et al.*, 2008, in Amazonia) and forest canopy roughness and moisture [Quick Scatterometer (QSCAT) microwave data: Saatchi *et al.*, 2008, in Amazonia]. Modelling a quantitative variable using explanatory variables has often been performed using linear models such as ordinary least squares (OLS) regressions. Other modelling approaches have also been developed that might better capture complex relationships among variables than models constrained by linear relationships. Among them, random forest models have been shown to perform particularly well for ecological predictions (Prasad *et al.*, 2006). Note that the word 'forest' relates to a collection of regression 'trees', whatever the variable studied. However, these new methods have not yet been evaluated for their ability to predict biodiversity patterns.

The reliability of various diversity modelling approaches has to be critically evaluated for different areas. Patterns and implied relationships can differ markedly within and among regions (e.g. see Parmentier *et al.*, 2007, for inconsistent correlations of tree alpha diversity with climatic variables between Africa and Amazonia). It is also unclear whether such correlations result from causal relationships (direct or indirect), as the mechanisms that could explain these relationships are still poorly understood and remain the subject of debate (Wright, 2002). Predictive models are generally based on contemporary factors, while historical factors could also explain at least part of the modern diversity gradients (McGlone, 1996). An additional problem arises due to the fact that both species richness and many of the variables used to predict richness are spatially autocorrelated (the similarity between samples for a given variable decreases with the spatial distance separating those samples; Legendre, 1993). When both the variable to be modelled and the predictor variables are spatially autocorrelated, classical methods to measure the association between these variables (i.e. regression, correlation, ANOVA) give false confidence because they tend to reject too often the null hypothesis that there is no association between these variables (Lennon, 2000; Diniz *et al.*, 2003; Bahn & McGill, 2007).

Here we focus on African tropical rain forests. The protection of African rain forests presents enormous challenges because most African forested countries have low incomes, weak conservation policies, inadequately developed institutions, growing populations and rising prices of food and energy (Balmford *et al.*, 2001; de Wasseige *et al.*, 2009). Moreover, the identification of areas most in need of protection is difficult because comparatively little is known about vast sections of these areas (Küper *et al.*, 2006). Burgess *et al.* (2005) presented a map of the plant species richness across sub-Saharan Africa at a 1° × 1° resolution for a dataset of 5958 plant species, using collection records from taxonomic revisions, distribution maps and herbarium specimen labels. However, as the collection intensity is very unequal across the domain, robust interpretation of the patterns documented is difficult. Using a similar dataset, Küper *et al.* (2006) modelled the individual species distributions at the same resolution and summed the individual species maps to obtain species diversity estimates, comparing them with field-measured diversity patterns. For several of the areas with very few collection records, such as the north-western Congolian lowland forests, their model predicted much higher species richness than currently documented. Another approach to estimate diversity measures is to focus on species inventories in standardized units of the rain forest. With such direct estimates of alpha diversity (local diversity) one can test the reliability of predictive models. In this study we compiled tree alpha diversity data using 573 inventory plots across West and Atlantic Central Africa. We also compiled data for 18 climatic and remote sensing variables for 1 km grid squares across the study domain. Our aim was to evaluate if we could map tree alpha diversity of African rain forests using inventory, climate and satellite-derived data combined with two distinct modelling approaches: a classical linear model using ordinary least squares (OLS) regression and a nonlinear multiple regression using random forest (RF). Specifically, we aimed to answer the following questions.

1. Using inventory data, what are the geographic patterns of tree alpha diversity?
2. What is the spatial dependency of tree alpha diversity in comparison to that of climatic and remote sensing data?
3. Using climatic and remote sensing data, do the OLS and RF models provide consistent geographic patterns of tree alpha diversity and congruent relationships with explanatory variables?
4. How do the OLS and RF models compare with a two-dimensional kriging based on the geographical coordinates only?
5. What is the reliability of model projections outside areas with inventory data?

MATERIALS AND METHODS

Data

Tree alpha diversity data

Data for trees with diameter at breast height (d.b.h.) ≥ 10 cm were compiled from the literature and unpublished data. The dataset comprises 573 plots and transects sections. We restricted our analysis to terra firme (upland, non-periodically flooded) plots in lowland (elevation < 900 m) old-growth forests. Alpha diversity is defined in this case as the local diversity in the community of trees with d.b.h. ≥ 10 cm. This stratum of the forest was chosen because it is a major component of the rain forest structure, and because data are available for a large number of plots. The plots and transects included in this study vary in shape, area and number of trees. Robust and meaningful comparisons of tree alpha diversity can be calculated from such a dataset provided that the diversity indices are correctly chosen (Condit *et al.*, 1998). To limit artificially increased alpha diversity due to species turnover between different forest types, a maximal plot or transect dimension of 500 m was used. The minimum number of trees with d.b.h. ≥ 10 cm was set to 50. The mean number of trees per plot was 145 (SD = 159, range = 50–684) and the mean plot size was 0.3 ha (SD = 0.35, range = 0.0375–1). As our dataset was unequally distributed within the African rain forest region (Fig. 1a), we had insufficient data points to include the extensive forests of the Democratic Republic of Congo in the analysis, and we restricted our study area to West Africa and Atlantic Central Africa (Upper Guinean and Lower Guinean phytogeographical regions according to White, 1979). Plot coordinates, plot size and alpha diversity values are provided in Appendix S1 in the Supporting Information.

Two measures of alpha diversity were used in this study: Fisher's alpha and $S(50)$. Fisher's alpha is a parametric index assuming that species abundances follow a log series distribution. It is defined implicitly by the formula $S = \alpha^*\ln(1 + n/\alpha)$, where S is number of taxa, n is number of individuals, and α is Fisher's alpha (Fisher *et al.*, 1943). $S(50)$ is a nonparametric sample-size unbiased estimator of alpha diversity: the expected number of species found in a subsample of size 50. It was computed following Hurlbert (1971) using the program BiodivR 1.1 (Hardy, 2009a). According to simulations (O.J. Hardy, unpublished data), Fisher's alpha is more sensitive to variations in plot shape, area and number of trees than *S*(50). Nevertheless, with geographic patterns of variation being very similar for both diversity measures and as Fisher's alpha has been more widely used in the literature, we present results for Fisher's alpha in the main text and results for *S*(50) in Appendices S2 and S3.

The quality of the botanical data was generally good, but not directly comparable between datasets because inventories were recorded by a large number of botanical teams. The number of individuals that will be identified or attributed to a species is dependent upon the experience of the botanist who made the inventory, as well as on the time spent on taxonomic analysis and the degree of specimen vouchering.

Climate data

A series of 1 km-scale bioclimatic metrics were obtained from WorldClim (WorldClim version 1.4; Hijmans *et al.*, 2005). These metrics are derived from monthly temperature and rainfall climatologies (1950–2000). They include 11 temperature and eight precipitation metrics, expressing spatial variations in annual means, seasonality (e.g. annual range in temperature and precipitation) and extreme or limiting climatic factors. The WorldClim monthly climatologies were developed using long-time series of a global network of weather stations. The station data were interpolated to monthly climate surfaces at 1 km spatial resolution by using a thin-plate smoothing spline algorithm with latitude, longitude and elevation as independent variables (Hijmans *et al.*, 2005). The accuracy of the data is mainly dependent upon the density of weather stations and on the quality of the interpolation.

We tested for covariance among the original 19 bioclimatic metrics by using 1000 random points across West Africa and Atlantic Central Africa. For metric pairs showing high correlation (Pearson's correlations in the order of 0.9 or larger), we retained the metric most commonly used in distribution modelling. The final bioclimatic subset used for this study included nine bioclimatic variables; annual mean temperature (Bio1), mean diurnal temperature range (Bio2), temperature seasonality (Bio4), maximum temperature of the warmest month (Bio5), minimum temperature of the coldest month (Bio6), annual mean rainfall (Bio12), rainfall seasonality (Bio15), rainfall of the wettest quarter (Bio16) and rainfall of the driest quarter (Bio17).

Remote sensing data

We used both optical passive and microwave active sensors layers. The optical data used in this study stem from the Moderate Resolution Imaging Spectroradiometer (MODIS) sensors and include the normalized difference vegetation index (NDVI), a measure of vegetation greenness, and the vegetation continuous field product of the Global Land Cover Facility

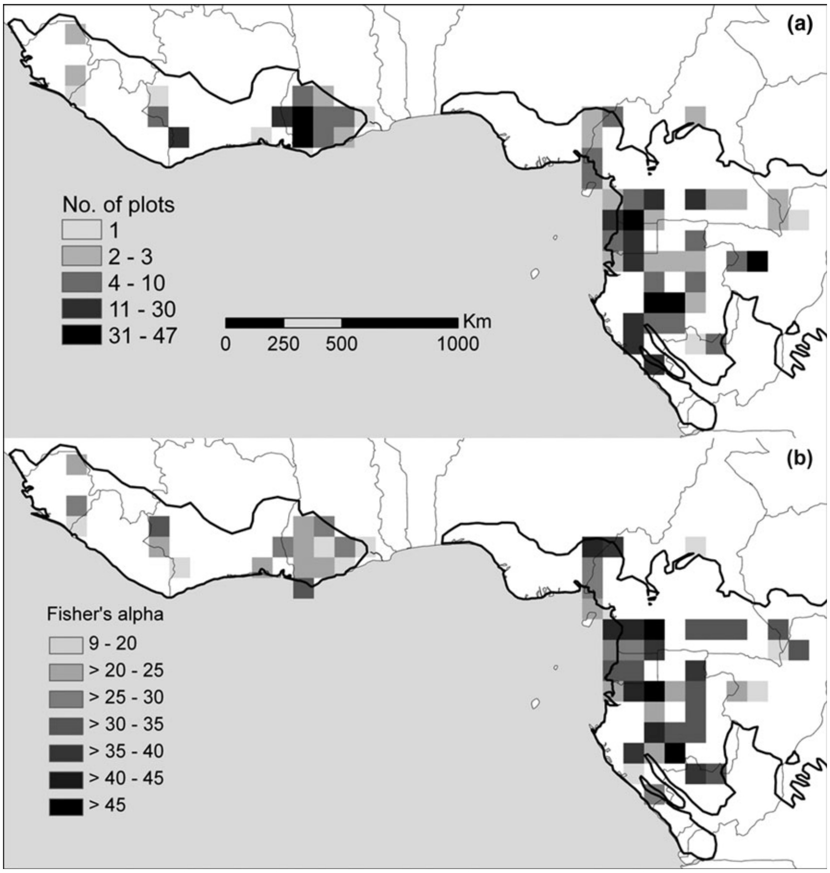

Figure 1 Inventory data maps for the alpha diversity of trees in the rain forest of West Africa and Atlantic Central Africa for 0.8° grid cells. (a) The number of plots included in each grid cell. (b) Mean of the Fisher's alpha values of the plots included in each grid cell.

(GLCF) as a measure of the percentage of tree cover (TREE) (Hansen *et al.*, 2003). For this study, monthly NDVI as well as annual tree cover data from the year 2001 were compiled over our study region at the original 1 km MODIS resolution. Based on the monthly NDVI data, several metrics were generated (avoiding contaminated data, most often due to cloud cover) to capture temporal and spatial characteristics of vegetation: (1) NDVI max: maximum based on 12 months, and (2) NDVI green: greenest (maximum NDVI) quarter based on four seasons (DJF, MAM, JJA and SON). Radar backscatter measurements are sensitive to surface canopy roughness, surface canopy moisture, and other seasonal attributes, such as the deciduousness of vegetation (Imhoff, 1995). For this study, we included the microwave QuickSCAT (QSCAT), at wavelengths of *c.* 2 cm (Ku-band), available in 3-day composites at 2.25 km resolution. The 3-day data of the year 2001 with horizontal polarization and complete data coverage were used to create average monthly composites and then further processed to produce two metrics that included annual mean (QSCAT mean) and standard deviation of radar backscatter over the 12 months (QSCAT std). In a final step, the QSCAT metrics were re-aggregated to the 1 km spatial resolution of the optical data (each 1 km pixel was given the value of the 2.25 km pixel covering it). We also included data from the Japanese Earth Resource Satellite (JERS-1), a radar sensor with a wavelength of 23 cm. These longer wavelength data are more strongly related to biomass than the QSCAT layers, but are also influenced by aspects of vegetation structure. We used the mosaic produced from 1996 data at a 100 m resolution by the Global Rain Forest Mapping Project (GRFM; Rosenqvist *et al.*, 2000), which we resampled to the MODIS 1 km grid. We obtained digital elevation data from the Shuttle Radar Topography Mission (SRTM). These data were aggregated from the native SRTM 90 m resolution to 1 km to match the target resolution of our study. In addition to mean elevation (SRTM mean), we also included the standard deviation within each 1 km^2 pixel (SRTM std) based on the 90 m data as an indicator of surface ruggedness.

Spatial dependency of the variables

The spatial autocorrelation of quantitative variables is described using Moran's I statistic. The latter expresses the correlation of the values of a given variable, defined for a set of locations, between pairs of locations situated at given physical distances apart. It is computed for each distance class d as (Sokal & Oden, 1978):

$$I(d) = \sum_{i}^{n}\sum_{j}^{n} w_{ij}(d).(x_i - \bar{x}).(x_j - \bar{x})/\mathrm{Var}(x)\sum_{i}^{n}\sum_{j}^{n} w_{ij}(d)$$

where x_i is the value of variable x for sample i; $\bar{x}$ and $\mathrm{Var}(x)$ are, respectively, the mean and variance of variable x estimated from the whole data set; n is the total sample size and $w_{ij}(d)$ are weights equalling one if the distance between samples i and j is included in class d, and zero otherwise. Values of Moran's I plotted against d produce a correlogram, that is the function $I(d)$. When a variable is spatially autocorrelated, $I(d)$ is expected to be positive at short distances, decreasing with increasing distances and eventually reaching negative values. Without spatial autocorrelation, the correlogram is horizontal, except for local fluctuations due to limited sample size. The spatial autocorrelation was tested by randomizing the values of the variable among all samples. Analyses were run with Torocor 1.0 (Hardy, 2009b).

Ecological modelling

The relationships between alpha diversity and the climatic and remote sensing variables described above were modelled using two different methods: an ordinary least squares (OLS) multiple regression and random forest (RF) developed by Breiman (2001). Modelling was performed in R 2.1.8 (http://www.r-project.org/). The OLS model was run with the log-likelihood maximized option. To apply the RF model we used the package randomForest v. 4.5–30 in the R statistical framework (Liaw & Wiener, 2002). The RF model is an ensemble classifier, a statistical learning procedure based on multiple decision trees used to predict a response variable (here local diversity) according to explanatory variables (here climate and remote sensing variables). A random subset of both the predictor variables and the data records themselves (here the 573 sample plots) are used for training of each decision tree. That tree statement is then tested on the remaining records (in RF procedures, each record is excluded from the training set in *c.* 36% of the runs; Liaw & Wiener, 2002), and the collection of these tests results in an overall statement of random forest performance. This algorithm modelling technique has proved to be very powerful in previous applications, for instance in predictive vegetation mapping under current and future climate, using climate, soil, land use, landscape and topography as explanatory variables (e.g. Prasad *et al.*, 2006). It has the advantage over OLS regression in that there is no assumption on the type of relationships (e.g. linear relationship) between response and explanatory variables, so it can handle very complex relationships involving interaction and nonlinearity across a response variable. Two estimations of variable importance provided by the RF model were used in this study (Kuhn *et al.*, 2008): (1) mean decrease accuracy (MDA) based on mean square error, and (2) mean decrease Gini (MDG) based on decision tree node purity. We limited the extent of the model predictions to roughly the rain forest limits defined by Mayaux *et al.* (2004) in order to mask out savannas and to avoid predicting tree diversity where we estimate there is no forest. We also predicted tree alpha diversity with a two dimensional kriging based on the latitude and longitude using the Krig function in the R package 'field' 6.3 (Fields Development Team, 2006) with default parameters. Kriging is an interpolation technique that fits a surface to irregularly spaced data and accounts for spatial autocorrelation. Here it estimates the alpha diversity at a location as a weighted average over data points where weights decrease with the distance to the focal location. The kriging model assumes that the unknown function is a realization of a Gaussian random spatial process. It optimizes the parameters of the general equation $Y = P(x) + Z(X) + e$, where Y is the vector of tree alpha diversity, $P(x)$ is a first order trend surface, $Z(X)$ is a mean zero, Gaussian stochastic process with a covariance structure across space modelled as an exponential function of distance, and e is a residual error term. Hence, the kriging model takes into account a possible linear gradient of diversity through space [$P(x)$ term] and the spatial autocorrelation of diversity [$Z(X)$ term].

Validation of the model prediction according to the distance to inventory data

We designed and implemented a novel approach to assess the quality of the predictions according to their distance to available inventory data. The principle is based on evaluating model predictions by defining the training dataset in a spatially explicit way. For each of the 573 data points considered in sequence (focal point), we excluded this focal point from the training dataset, as well as all other surrounding data points within a defined radius (Fig. 2). We ran the model and recorded the value estimated for that focal point. Once estimations were obtained for all 573 data points, the squared Pearson's correlation coefficient (pseudo-R^2) between estimated and real values was calculated. Computations were performed in R using OLS, RF or kriging models for the exclusion radii 0, 0.01°, 0.02°, 0.05°, 0.1°, 0.2°, 0.3°, 0.4°, 0.5°, 1°, 2°, 5° and 10°, corresponding to *c.* 0, 1.1, 2.2, 5.6, 11, 22, 33, 45, 56, 111, 222, 557 and 1113 km, respectively.

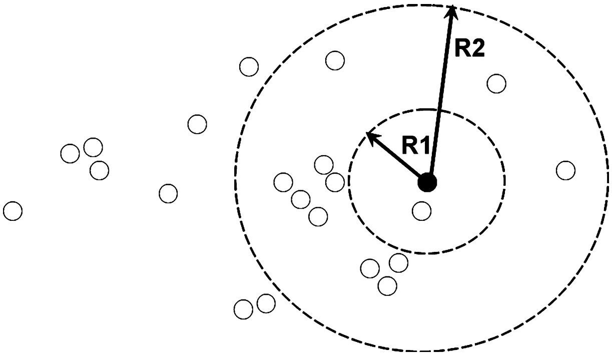

Figure 2 Method for evaluating the power of a model in a spatially explicit way. The goal is to assess the predictive power of a model trained on available inventory data points according to the distance to data points. A focal data point (black dot) is suppressed from the training dataset as well as all other data points (open circles) that are within a defined radius (R1 or R2) of that focal point. In the example shown, when radius = 0, only the focal point is suppressed while one or 12 additional points are also suppressed when the radius = R1 or R2, respectively. The tested model is then used to estimate the value of the focal point using the remaining data points. The procedure is repeated, with each data point sequentially becoming the focal point and excluding surrounding points. The pseudo-R^2 of the regression of estimated values on real values characterizes the predictive power of the model at a given minimal distance from existing data points. If pseudo-R^2 quickly decreases as the exclusion radius increases, it means that the model can provide reliable predictions in the vicinity of existing data points but not at large distances.

RESULTS

Diversity patterns: inventory data

The highest alpha diversity values are located in Atlantic Central Africa [see Fig. 1b and Fig. S1 in Appendix S2, which present

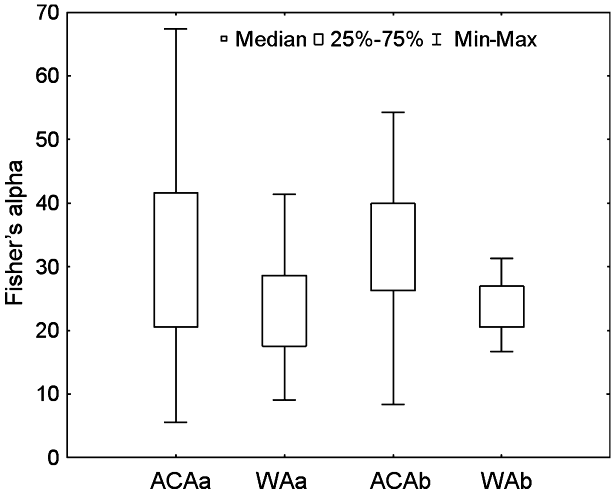

Figure 3 Boxplots of Fisher's alpha values for trees with diameter at breast height ≥ 10 cm in 573 inventory plots in the rain forests of West Africa (WA) and Atlantic Central Africa (ACA), in all plots (a) and for 0.4° grid cells (b).

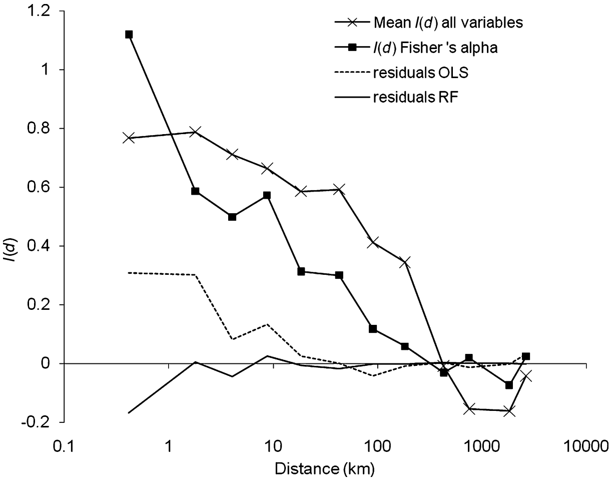

Figure 4 Spatial autocorrelograms [$I(d)$ = Moran's I as a function of distance] for Fisher's alpha values of trees with diameter at breast height ≥ 10 cm in 573 inventory plots in the rain forests of West Africa and Atlantic Central Africa, for the climatic and the remote sensing variables (mean values) corresponding to the plot locations, and for the residuals of ordinary least squares (OLS) and random forest (RF) models used for predicting Fisher's alpha from this dataset.

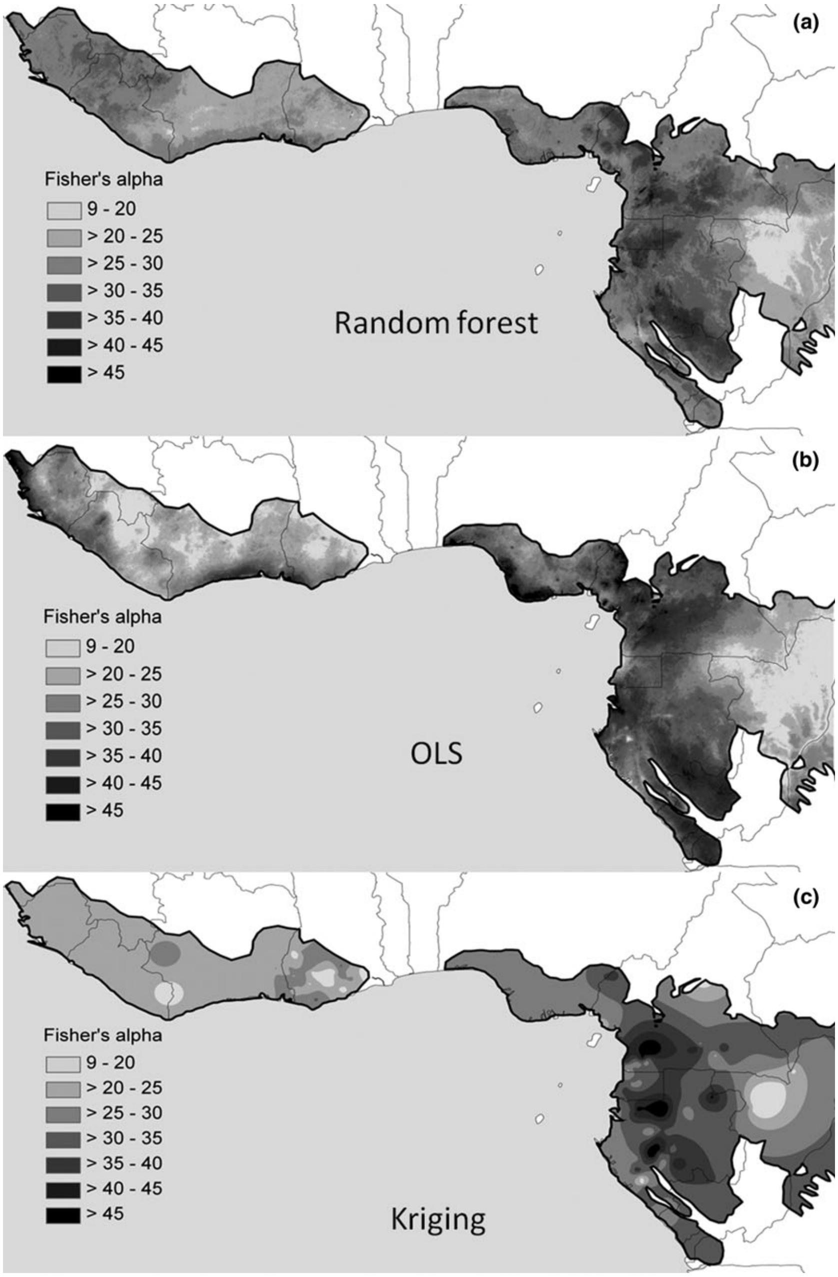

Figure 5 Model predictions of Fisher's alpha values for trees with diameter at breast height ≥ 10 cm in 573 inventory plots in the rain forests of West Africa and Atlantic Central Africa with three different models: (a) random forest (RF) model, (b) ordinary least squares (OLS) model, (c) kriging model.

mean values for 0.8° grid cells, and Fig. S2b in Appendix S2 for higher resolution maps with 0.4° grid cells, as well as Fig. 3 for a comparison of the median of the Fisher's alpha values for all plots (a) and for 0.4° grid cells (b)]. Mean Fisher's alpha values are significantly higher in Atlantic Central Africa (ACA; mean = 32.1 ± 0.7 SD, $n = 428$) than in West Africa (WA; mean = 23.4 ± 1.2 SD, $n = 145$) according to the Mann–Whitney U-test ($P < 0.001$) and remain significant using mean estimates per grid cell at 0.4° resolution (ACA: mean = 33.4 ± 1.4 SD, $n = 40$; WA: mean = 23.3 ± 1.9 SD, $n = 21$, only grid cells containing at least three plots or transects were considered). Similar results were obtained for the $S(50)$ diversity measure (see Figs S1b, S2c and S3 in Appendix S2).

The lack of inventory data is illustrated by the fact that no estimate was available for 65% and 84% of rain forest grid cells at 0.8° and 0.4° resolution, respectively. There are no obvious east–west or north–south geographic gradients within either forest block. Highest diversity values (for 0.4° pixels with mean Fisher's alpha > 40 and including at least three plots) are observed in Cameroon in the Campo Ma'an National Park and in Gabon in the Bélinga Mountain range, the Massif du Chaillu, and a region stretching from the Monts de Cristal to the Waka National Park.

Spatial dependency of alpha diversity and predictive variables

Fisher's alpha is strongly spatially autocorrelated (Fig. 4). Moran's I statistic computed for different distance intervals, $I(d)$, is very high (close to unity) at short distances and

Table 1 Importance of climatic and remote sensing variables for predicting tree alpha diversity in the rain forests of Atlantic Central Africa and West Africa using a linear model (ordinary least squares, OLS) and the random forest (RF) model. A higher percentage of mean decrease accuracy (% MDA) or mean decrease Gini (MDG) represents higher variable importance in the RF model. A higher absolute *t*-value represents a higher variable importance in the OLS model.

		OLS	RF	RF	OLS	RF rank	RF rank
Variable	OLS coefficient	*t*-value	% MDA	MDG	rank	% MDA	MDG
elevation	0.046	5.594	31	8803	1	3	3
bio1	1.715	4.817	20	4387	2	13	12
QSCAT mean	0.059	3.542	31	7979	3	2	4
bio5	−0.755	−3.013	26	6059	4	7	9
NDVI green	0.002	2.098	20	4218	5	12	14
NDVI mean	−0.002	−1.897	17	6569	6	15	6
bio6	−0.359	−1.477	24	4001	7	9	16
bio15	0.341	1.401	33	12281	8	1	1
QSCAT std	0.114	1.018	21	4136	9	11	15
NDVI max	−0.001	−0.895	16	4462	10	17	11
JERS	0.007	0.709	16	3808	11	18	17
bio2	−0.150	−0.580	24	4238	12	8	13
bio12	0.006	0.572	22	6382	13	10	8
bio4	0.003	0.482	28	6529	14	5	7
bio17	−0.004	−0.267	30	6579	15	4	5
bio16	−0.006	−0.254	26	9496	16	6	2
TREE	−0.005	−0.156	16	3781	17	16	18
elevation std	−0.001	−0.015	17	4879	18	14	10

bio1, annual mean temperature; bio2, mean diurnal temperature range; bio4, temperature seasonality; bio5, maximum temperature of the warmest month; bio6, minimum temperature of the coldest month; bio12, annual mean rainfall; bio15, rainfall seasonality; bio16, rainfall of the wettest quarter; bio17, rainfall of the driest quarter; QSCAT, QuickSCAT microwavedata related to canopy roughness and humidity; NDVI, normalized difference vegetation index; JERS, radar sensor related to biomass; TREE, percentage of tree cover.

Table 2 Predictive power of the models [ordinary least squares (OLS) linear model, random forest (RF) model and spatial kriging (KR)] assessed from the pseudo-R^2 (%) of real versus estimated Fisher's alpha values according to the minimal distance from points with inventory data. The number of plots represents the mean size (± SD) of the dataset used for prediction, which includes the whole dataset (573 plots in the rain forests of West Africa and Atlantic Central Africa, trees with diameter at breast height ≥ 10 cm) except for the focal plot as well as all plots within the threshold distance (Fig. 2).

Distance (°)	Distance (km)	No. plots	R^2 (OLS)	R^2 (RF)	R^2 (KR)
0	0	572	26	40	42
0.01	1.1	570 ± 4	19	25	29
0.02	2.2	569 ± 4	18	19	25
0.05	5.6	566 ± 6	17	17	23
0.1	11	564 ± 7	16	12	21
0.2	22	559 ± 12	15	8	15
0.3	33	555 ± 12	14	6	13
0.4	45	551 ± 14	12	7	12
0.5	56	546 ± 16	9	5	4
1	111	517 ± 30	7	6	2
2	222	466 ± 47	4	3	3
5	557	294 ± 118	1	0	1
10	1113	216 ± 123	3	0	5

decreases fairly linearly with the log of the distance up to *c.* 400 km, beyond which it levels out. Similar spatial dependency is observed for the 18 climatic and remote sensing variables included in the models [$I(d)$ values between 0.5 and 1 at short distance, see Fig. 4 and Table S1 in Appendix S3].

Model predictions

The Spearman correlations between Fisher's alpha, $S(50)$, the remote sensing variables and the climatic variables in the inventory data are presented in Table S2 in Appendix S3. Fisher's alpha values correlate best with rainfall seasonality (Spearman's $r_s = 0.36$), mean NDVI ($r_s = -0.33$) and rainfall in the wettest quarter ($r_s = 0.31$). Diversity maps obtained from the predictions of the three models (OLS, RF and kriging) are presented in Fig. 5. These patterns are generally consistent between the three models in areas where inventory data are available, but they differ considerably in areas lacking inventory data (e.g. in Sierra Leone, south-eastern Liberia, southern Nigeria, southern Congo; see Fig. S5 in Appendix S2). The OLS and RF models agree on two of the three most important variables (Table 1). According to the *t*-statistics of the OLS model, the most important variables are elevation ($t = 5.6$), mean annual temperature ($t = 4.8$), and QSCAT mean ($t = 3.5$). In the RF model, according to the mean decrease accuracy values, the most important variables are rainfall seasonality (33%), QSCAT mean (31%) and elevation (31%).

The majority of the variance remains unaccounted for in both the OLS and RF models: only 31% of variance is explained in the OLS model and 40% for the RF model. Note that these percentages may not be directly comparable, because whereas they are based on the complete dataset for the OLS model, they are based on iterations of sub-sampling used for testing and training in the RF model. Residuals of the OLS model display short-distance spatial autocorrelation (Fig. 4). Hence, the climatic and remote sensing variables introduced in the OLS model cannot completely explain the similarity in diversity values of closely located points. In this respect, the RF model seems to perform better as there is no positive (actually a slightly negative) spatial autocorrelation at short distance (Fig. 4).

Validation of model predictions outside of well-sampled areas

We assessed the quality of the predictions according to the distance to available inventory data. The kriging spatial model, based on geographical coordinates only, performs better in these validations than the RF and OLS models constructed with climate and remote sensing data (Table 2). The performances of the three models decay very quickly when the exclusion radius increases, even when only a few points are excluded. For the RF and OLS models, the squared Pearson's correlation coefficient (pseudo-R^2) between estimated and real values drops to 18–19% with an exclusion radius of 2.2 km (excluding, on average, three additional plots). Pseudo-R^2 becomes < 10% as soon as all plots at < 22 km from the focal plot were excluded for the RF model, a pseudo-R^2 value > 10% is retained until the exclusion radius exceeds 50 km. Note that for an exclusion radius of 50 km, < 5% of the plots from the training dataset are excluded. These results can explain why the three models show congruent patterns in areas well covered in the training dataset, but fail to do so in regions with no or few data.

DISCUSSION

Diversity patterns

Our confidence in the predictive maps derived from the modelling approaches decreases rapidly with distance from sampled areas. Therefore, little can be inferred for large areas across West and Atlantic Central Africa. Although alpha diversity is highly variable even at a local scale, forest plots tend to have decreasing similarity in tree alpha diversity values when the spatial distance between these plots increases. At the continental scale, tree alpha diversity patterns in the reliable regions of our maps (< 50 km from points with inventory data, approximately the areas covered by the pixels in Fig. 1a) are similar to those for vascular plant species richness (including herbs and shrubs) mapped by Linder (2001) and Barthlott *et al.* (2007).

Tree alpha diversity in West Africa is, on average, lower than in Atlantic Central Africa. This may be explained by a smaller regional species pool adapted to wet and humid conditions in West Africa (Fox & Srivastava, 2006). Most high diversity areas identified in Atlantic Central Africa correspond to hilly regions at medium elevation, which could explain the positive correlation observed in our dataset between alpha diversity and both the elevation and the standard deviation of elevation. It must be noted that the inventory plots used are not a random sample of the domain. Most inventory plots are located in protected areas, which biases the maps towards the diversity of mature and relatively undisturbed forest vegetation. Moreover, many areas within the African rain forest have now been deforested or degraded (particularly in West Africa) and the maps for those areas are mostly predicting potential or historical tree diversity rather than actual diversity.

Correlations and causality

A predictive model of tree alpha diversity from contemporary climate and remote sensing data is likely to be accurate under the following conditions: (a) the model is able to predict diversity in a subset of the dataset using the rest of the dataset, (b) the variables used to build the model have similar distributions in areas with no training data as they do in sampled areas, (c) there is a causal relationship between the variables in the model and tree alpha diversity, or these variables have spatial patterns similar to those of other variables that do have a causal effect on tree alpha diversity. The performances of the RF and OLS models decay very quickly with increasing distance to the inventory data points. RF performs better than the OLS regression in areas where ample inventory data are available, while the linear model is more effective when inventory data are scarce, possibly because the linear model better captures the diversity gradient between West Africa and Central Africa. The kriging model based on geographical coordinates performs better than the OLS and RF models. Hence, condition (a) is partially met for all three models; they are able to predict actual diversity in the training dataset, with the restriction that nearby data points are available in the training dataset. But, unfortunately, they predict very poorly outside well-sampled areas. Not all environmental conditions and forest types are equally represented in our training dataset because many areas remain unexplored, or no compatible dataset was available for inclusion into our models. Some of the variables used to build the models have values outside the range of those present in our training dataset in part of the study area (see Fig. 3 in Parmentier *et al.*, 2007). Condition (b) is thus not totally met, and this may affect the predictive power of the OLS and RF models. However, this cannot explain the low predictive power at distances beyond 50 km because environmental variables are still highly autocorrelated at 50 km (Fig. 4), such that the training dataset must include representative environmental conditions.

The average plot size (0.3 ha) is much smaller than the grid cell size of each of the explanatory variables (100 ha), potentially obscuring the relationship between diversity values and these variables. However, the high spatial autocorrelation of diversity for nearby plots (< 1 km, see Fig. 4) suggests that the intra-pixel heterogeneity or the lack of intra-pixel resolution cannot explain the low predictive power of the models based on climatic and remote sensing data.

Is there a causal relationship between the important variables in the models and tree alpha diversity? The RF and OLS models agree on the importance of elevation and of the QSCAT mean. However, the positive relationship between elevation and alpha diversity observed here is in contradiction with what has been traditionally reported in the literature (see Givnish, 1999; for a review). Parmentier *et al.* (2007) have previously shown that, for similar elevational ranges, the relationships of alpha diversity with elevation were reversed in Amazonia and Africa. QSCAT was the best predictor of tree alpha diversity in the Maxent model for the Amazonian rain forest (Saatchi *et al.*, 2008) and was interpreted as describing properties of the forest canopy: roughness and moisture. This mix of consistent and inconsistent results between studies does not help in answering the question of the causality, so that condition (c) might not be met.

If there are weak direct or indirect causal relationships between local diversity and the climate or remote sensing variables considered, how could a predictive model show good performances in the vicinity of existing data points? An explanation may lie in the strong spatial dependency displayed by all the variables and the heterogeneous distribution of plot data. The very high Moran's *I* values at short distances imply that nearby plots share similar diversity values as well as similar values for explanatory variables. Because the regression models fit a dataset where a majority of data points are concentrated in a limited number of areas, there is a high level of spatial pseudoreplication. Actually, if available plots were concentrated in only two narrow areas displaying contrasting diversity values, the apparent prediction power would approach pseudo-$R^2 = 1$, typical for a regression based on two points. A nonlinear model such as RF could perform better than a model constrained by linear relationships (OLS) because higher degrees of freedom permit a larger number of particular local associations among variables to be accounted for. Spatial pseudoreplication thus overestimates the intrinsic predictive power of modelling approaches based on georeferenced variables. Within the RF model, performance is evaluated by training the algorithm on random subsets of the dataset and testing the performance on the remaining subsets. Because subsets taken at random from the inventory data are used, there is no control of the impact of spatial dependency. Kriging probably performs better than the two other models because it avoids local components of non-spatially structured environmental variation driving richness. Similarly, in a species distribution modelling study, Bahn & McGill (2007) showed that spatial interpolation led to better predictive models than habitat-based models (environmental variables).

We argue that the simple approach proposed here, whereby the quality of model prediction was evaluated in a spatially explicit way (removal of all nearby data points from the training dataset for increasing distances, Fig. 2), should be applied when predicting spatial patterns from experimental models to assess the spatial extent of model reliability.

Other studies attempting to map diversity using climate and remote sensing data may provide more reliable predictions, for example, when the study area encompasses several diversity-contrasted vegetation types displaying distinctive spectral signatures and/or responding to climatic gradients. We did not use theoretical models based on water–energy dynamics (e.g. O'Brien, 2006) because we limited our study to one biome, the rain forest, and to an area with limited variations in latitude and elevation (< 900 m). It is possible that these models would provide good first-order predictions, for instance a lower diversity in West Africa than in Atlantic Central Africa, but would not provide detailed patterns of tree alpha diversity within these two regions.

We do, however, observe diversity patterns in the inventory data. These diversity patterns might well reflect the history of past climate and vegetation changes rather than adaptation to contemporary environmental conditions (Hawthorne, 1996; Wieringa & Poorter, 2004; Parmentier *et al.*, 2007). Fossil pollen data document strong modifications of the floristic composition of the African rain forest through time: the most important occurring during the Last Glacial Maximum (Maley, 1991). According to Bonnefille (2007), the modern composition of the different types of rain forest known today in Atlantic Central Africa only dates from between a few centuries to 1000 years. Considering the limited dispersal capacities of most rain forest trees (Muller-Landau *et al.*, 2008), it is likely that part of the rain forest species pool has not had enough time to reach all the contemporary suitable environments. The idea that diversity might be in equilibrium with local deterministic environmental factors is thus questioned. In Amazonia, Stropp *et al.* (2009) partitioned total tree alpha diversity into regional and local components, which are controlled by evolutionary and ecological processes, respectively. Regional diversity was correlated with palaeoclimatic stability and long-term large-scale ecosystem dynamics, both mechanisms contributing to high diversity in the central to western Amazon. It is likely that in African rain forests as well, regional scale differences in tree alpha diversity originate at least partly from historical factors.

At first sight, the absence of a clear relationship between environmental variables and diversity patterns might be interpreted as supporting Hubbell's neutral community theory, where all individuals are assumed to have the same behaviour irrespective of their species (Hubbell, 2001). However, the distribution of each species could be affected by environmental variables (implying a non-neutral behaviour) while the superposition of all distributions could result in diversity gradients uncorrelated with environmental gradients.

CONCLUSIONS

Predicting tree alpha diversity within unsampled rain forests from climatic and remote sensing data is an attractive approach to compensate for the lack of inventory data. Yet, its effectiveness has to be carefully tested (Araújo, 2003). Our models failed to produce reliable predictions in areas > 50 km from the nearest sampled data points. The predictive powers of the models at short distances are likely to be due to the strong spatial autocorrelation displayed by tree alpha diversity and climatic or remote sensing explanatory variables. This explains why models based on these explanatory variables were less powerful than a kriging model based on spatial data alone. Similar studies may suffer the same kind of problems, and it is strongly recommended that model validation is tested while controlling for spatial dependency (Currie, 2007), such as performed in this study (Fig. 2). This is particularly important when producing predictive maps, as projections from fine resolution variables (e.g. Fig. 5a,b) produce maps with high levels of detail. In fact, high resolution predictions do not guarantee reliability (the less detailed Fig. 5c was shown to be more reliable than Fig. 5a,b) and all the maps in Fig. 5 were shown to be potentially misleading in non-sampled areas. Predictions could possibly be improved by using other contemporary variables and/or considering historical factors in the models. Due to the difficulties in modelling tree alpha diversity in the African rain forests and given the large amount of unsampled area on our raw data map, more fieldwork in unexplored areas would certainly be a key to progress.

ACKNOWLEDGEMENTS

We thank the editor, Michael Patten, and two anonymous referees who gave useful comments and suggestions on an earlier draft of this manuscript. We wish to thank Lourens Poorter, Renaud Cortay, Hannsjörg Wöll, John Weingart and all the people who contributed to the inventory data collection and taxonomic identifications. I.P. (postdoctoral researcher) and O.J.H. (research associate) are financed by the Belgian Fund for Scientific Research (FRS-FNRS).

REFERENCES

Araújo, M.B. (2003) Predicting species diversity with ED: the quest for evidence. *Ecography*, **26,** 380–383.

Bahn, V. & McGill, B.J. (2007) Can niche-based distribution models outperform spatial interpolation? *Global Ecology and Biogeography*, **16,** 733–742.

Balmford, A., Moore, J.L., Brooks, T., Burgess, N., Hansen, L.A., Williams, P. & Rahbek, C. (2001) Conservation conflicts across Africa. *Science*, **291,** 2616–2619.

Barthlott, W., Hostert, A., Kier, G., Koper, 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.

Bongers, F., Poorter, L., Van Rompaey, R. & Parren, M.P.E. (1999) Distribution of twelve moist forest canopy tree species in Liberia and Cote d'Ivoire: response curves to a climatic gradient. *Journal of Vegetation Science*, **10,** 371–382.

Bonnefille, R. (2007) Rainforest responses to past climate changes in tropical Africa. *Tropical rainforest responses to climatic change* (ed. by M.B. Bush and J.R. Flenley), pp. 117–170. Springer, Berlin.

Breiman, L. (2001) Random forests. *Machine Learning*, **45,** 5–32.

Burgess, N., Kuper, W., Mutke, J., Brown, J., Westaway, S., Turpie, S., Meshack, C., Taplin, J., McClean, C. & Lovett, J.C. (2005) Major gaps in the distribution of protected areas for threatened and narrow range Afrotropical plants. *Biodiversity and Conservation*, **14,** 1877–1894.

Clinebell, R.R., Phillips, O.L., Gentry, A.H., Stark, N. & Zuuring, H. (1995) Prediction of Neotropical tree and liana species richness from soil and climatic data. *Biodiversity and Conservation*, **4,** 56–90.

Condit, R., Foster, R.B., Hubbell, S.P., Sukumar, R., Leigh, E.G., Manokaran, N., Loo de Lao, S., LaFranckie, J.V. & Ashton, P.S. (1998) Assessing forest diversity on small plots: calibration using species-individual curves from 50 ha plots. *Forest biodiversity research, monitoring and modelling* (ed. by F. Dallmeier and J.A. Comiskey), pp. 247–266. UNESCO & The Parthenon Publishing Group, Paris.

Currie, D.J. (2007) Disentangling the roles of environment and space in ecology. *Journal of Biogeography*, **34,** 2009–2011.

Currie, D.J. & Paquin, V. (1987) Large-scale biogeographical patterns of species richness of trees. *Nature*, **329,** 326–327.

Diniz, J.A.F., Bini, L.M. & Hawkins, B.A. (2003) Spatial autocorrelation and red herrings in geographical ecology. *Global Ecology and Biogeography*, **12,** 53–64.

Ferrier, S., Powell, G.V.N., Richardson, K.S., Manion, G., Overton, J.M., Allnutt, T.F., Cameron, S.E., Mantle, K., Burgess, N.D., Faith, D.P., Lamoreux, J.F., Kier, G., Hijmans, R.J., Funk, V.A., Cassis, G.A., Fisher, B.L., Flemons, P., Lees, D., Lovett, J.C. & Van Rompaey, R. (2004) Mapping more of terrestrial biodiversity for global conservation assessment. *BioScience*, **54,** 1101–1109.

Field, R., O'Brien, E.M. & Whittaker, R.J. (2005) Global models for predicting woody plant richness from climate: development and evaluation. *Ecology*, **86,** 2263–2277.

Fields Development Team (2006) *Fields: tools for spatial data*. National Center for Atmospheric Research, Boulder, CO. Available at: http://www.image.ucar.edu/GSP/Software/Fields/ (accessed 20 February 2010).

Fisher, R.A., Corbet, A.S. & Williams, C.B. (1943) The relation between the number of species and the number of individuals in a random sample of an animal population. *Journal of Animal Ecology*, **12,** 42–58.

Fox, J.W. & Srivastava, D. (2006) Predicting local–regional richness relationships using island biogeography models. *Oikos*, **113,** 376–382.

Givnish, T.J. (1999) On the causes of gradients in tropical tree diversity. *Journal of Ecology*, **87,** 193–210.

Hansen, M., DeFries, R., Townshend, J.R., Carroll, M., Dimiceli, C. & Sohlberg, R. (2003) *Vegetation continuous fields MOD44B, 2001 percent tree cover, collection 3*. University of Maryland, Maryland, MD.

Hardy, O.J. (2009a) *BiodivR 1.1. A program to compute statistically unbiased indices of species diversity within sample and species similarity between samples using rarefaction principles*. Available at: http://ebe.ulb.ac.be/ebe/Software.html (accessed 1 July 2009).

Hardy, O.J. (2009b) *TOROCOR: a program to assess the association between spatially autocorrelated variables using a torus-translation test on multiple grids*. Available at: http://ebe.ulb.ac.be/ebe/Software.html (accessed 1 July 2009).

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.

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.

Hubbell, S.P. (2001) *The unified neutral theory of biodiversity and biogeography*. Princeton University Press, Princeton, NJ.

Hurlbert, S.H. (1971) The nonconcept of species diversity: a critique and alternative parameters. *Ecology*, **52,** 577–586.

Imhoff, M.L. (1995) A theoretical-analysis of the effect of forest structure on synthetic-aperture radar backscatter and the remote-sensing of biomass. *IEEE Transactions on Geoscience and Remote Sensing*, **33,** 341–352.

Jarnevich, C.S., Stohlgren, T.J., Barnett, D. & Kartesz, J. (2006) Filling in the gaps: modelling native species richness and invasions using spatially incomplete data. *Diversity and Distributions*, **12,** 511–520.

Kuhn, S., Egert, B., Neumann, S. & Steinbeck, C. (2008) Building blocks for automated elucidation of metabolites: machine learning methods for NMR prediction. *BMC Bioinformatics*, **9,** 400.

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.

Laurance, W.F., Fearnside, P.M., Laurance, S.G., Delamonica, P., Lovejoy, T.E., Rankin-de Merona, J., Chambers, J.Q. & Gascon, C. (1999) Relationship between soils and Amazon forest biomass: a landscape-scale study. *Forest Ecology and Management*, **118,** 127–138.

Legendre, P. (1993) Spatial autocorrelation: trouble or new paradigm. *Ecology*, **74,** 1659–1673.

Lennon, J.J. (2000) Red-shifts and red herrings in geographical ecology. *Ecography*, **23,** 101–113.

Lewis, S.L. (2006) Tropical forests and the changing earth system. *Philosophical Transactions of the Royal Society B: Biological Sciences*, **361,** 195–210.

Liaw, A. & Wiener, M. (2002) Classification and regression by randomForest. *R News: the Newsletter of the R Project*, **2,** 18–22.

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

Maley, J. (1991) The African rain forest vegetation and palaeoenvironments during the late Quaternary. *Climatic Change*, **19,** 79–98.

Mayaux, P., Bartholome, E., Fritz, S. & Belward, A. (2004) A new land-cover map of Africa for the year 2000. *Journal of Biogeography*, **31,** 861–877.

McGlone, M.S. (1996) When history matters: scale, time, climate and tree diversity. *Global Ecology and Biogeography Letters*, **5,** 309–314.

Muller-Landau, H.C., Wright, S.J., Calderon, O., Condit, R. & Hubbell, S.P. (2008) Interspecific variation in primary seed dispersal in a tropical forest. *Journal of Ecology*, **96,** 653–667.

Myers, N. (1997) The world's forests and their ecosystem services. *Nature's services: societal dependence on natural ecosystems* (ed. by G.C. Daily), pp. 215–235. Island Press, Washington, DC.

O'Brien, E.M. (2006) Biological relativity to water–energy dynamics. *Journal of Biogeography*, **33,** 1868–1888.

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

Phillips, O.L., Hall, P., Gentry, A.H., Sawyer, S.A. & Vásquez, R. (1994) Dynamics and species richness of tropical rain forests. *Proceedings of the National Academy of Sciences USA*, **91,** 2805–2809.

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

Rosenqvist, A., Shimada, M., Chapman, B., Freeman, A., De Grandi, G., Saatchi, S. & Rauste, Y. (2000) The Global Rain Forest Mapping project – a review. *International Journal of Remote Sensing*, **21,** 1375–1387.

Saatchi, S., Buermann, W., ter Steege, H., Mori, S. & Smith, T.B. (2008) Modeling distribution of Amazonian tree species and diversity using remote sensing measurements. *Remote Sensing of Environment*, **112,** 2000–2017.

Schmidt, M., Konig, K. & Muller, J.V. (2008) Modelling species richness and life form composition in Sahelian Burkina Faso with remote sensing data. *Journal of Arid Environments*, **72,** 1506–1517.

Slik, J.W.F., Raes, N., Aiba, S.I., Brearley, F.Q., Cannon, C.H., Meijaard, E., Nagamasu, H., Nilus, R., Paoli, G., Poulsen, A.D., Sheil, D., Suzuki, E., van Valkenburg, J., Webb, C.O., Wilkie, P. & Wulffraat, S. (2009) Environmental correlates for tropical tree diversity and distribution patterns in Borneo. *Diversity and Distributions*, **15,** 523–532.

Sokal, R.R. & Oden, N.L. (1978) Spatial autocorrelation in biology. 1. Methodology. *Biological Journal of the Linnean Society*, **10,** 199–228.

ter Steege, H., Pitman, N., Sabatier, D. *et al.* (2003) A spatial model of tree alpha-diversity and tree density for the Amazon. *Biodiversity and Conservation*, **12,** 2255–2277.

Stropp, J., ter Steege, H. & Malhi, Y. (2009) Disentangling regional and local tree diversity in the Amazon. *Ecography*, **32,** 46–54.

de Wasseige, C., Devers, D., de Marcken, P., Eba'a Atyi, R., Nasi, R. & Mayaux, P. (eds) (2009) *Les forêts du Bassin du Congo – etat des forêts 2008*. Office des publications de l'Union européenne, Luxembourg.

White, F. (1979) The Guineo-Congolian Region and its relationships to other phytochoria. *Bulletin du Jardin Botanique National de Belgique*, **49,** 11–55.

Wieringa, J.J. & Poorter, L. (2004) Biodiversity hotspots in West Africa; patterns and causes. *Biodiversity of West African forests: an ecological atlas of woody plant species* (ed. by L. Poorter, F. Bongers, F.N. Kouamé and W.D. Hawthorne), pp. 61–72. CABI Publishing, Wallingford, UK.

Wright, S.J. (2002) Plant diversity in tropical forests: a review of mechanisms of species coexistence. *Oecologia*, **130,** 1–14.

SUPPORTING INFORMATION

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

Appendix S1 Details and diversity values for 573 plots and transect sections in the rain forests of West Africa and Atlantic Central Africa (trees with diameter at breast height ≥ 10 cm).

Appendix S2 Predicting alpha diversity of African rain forests: Figures S1–S5.

Appendix S3 Predicting alpha diversity of African rain forests: Tables S1–S3.

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

BIOSKETCH

Ingrid Parmentier has been assembling tree inventory data in African rain forests since 2006. With **Olivier Hardy**, she is studying the diversity of African rain forest trees at different spatial scales (local to inter-continental).

Author contributions: I.P., S.S. and Y.M. initiated this research project. I.P. and O.J.H. performed most analyses with the help of R.H., S.S. and W.B.; O.J.H. developed the R routines to test the map predictions. I.P. and O.J.H. wrote the manuscript with substantial contributions by F.B., E.M., M.E.L., R.H., S.D., S.L., M.S.M.S., W.B., W.H. and Y.M. The remaining authors as well as F.B., M.E.L., M.S.M.S., S.D., S.L. and W.H. contributed with data and commented on the analyses and the manuscript.

Editor: Michael Patten