Ecological Applications, 25(5), 2015, pp. 1290–1302
© 2015 by the Ecological Society of America

Remote assessment of locally important ecological features across landscapes: how representative of reality?

Katherine J. Willis,[1,2,5,7] Alistair W. R. Seddon,[2] Peter R. Long,[1] Elizabeth S. Jeffers,[1] Neil Caithness,[3] Milo Thurston,[3] Mathijs G. D. Smit,[4] Randi Hagemann,[4] and Marc Macias-Fauria[1,6]

[1]*Biodiversity Institute, Oxford Martin School, Department of Zoology, South Parks Road, University of Oxford, Oxford OX1 3PS United Kingdom*
[2]*Department of Biology, University of Bergen, P.O. Box 7803, N-5020 Bergen, Norway*
[3]*Oxford e-Research Center, 7 Keble Road, University of Oxford, Oxford OX1 3QG United Kingdom*
[4]*Statoil ASA, Forusbeen 50, N-4035 Stavanger, Norway*
[5]*Royal Botanic Gardens, Kew, Richmond TW9 3AE United Kingdom*
[6]*School of Geography and the Environment, University of Oxford, Oxford OX1 3QY United Kingdom*

Abstract. The local ecological footprinting tool (LEFT) uses globally available databases, modeling, and algorithms to remotely assess locally important ecological features across landscapes based on five criteria: biodiversity (beta-diversity), vulnerability (threatened species), fragmentation, connectivity, and resilience. This approach can be applied to terrestrial landscapes at a 300-m resolution within a given target area. Input is minimal (latitude and longitude) and output is a computer-generated report and series of maps that both individually and synthetically depict the relative value of each ecological criteria. A key question for any such tool, however, is how representative is the remotely obtained output compared to what is on the ground. Here, we present the results from comparing remotely- vs. field-generated outputs from the LEFT tool on two distinct study areas for beta-diversity and distribution of threatened species (vulnerability), the two fields computed by LEFT for which such an approach is feasible. The comparison method consists of a multivariate measure of similarity between two fields based on discrete wavelet transforms, and reveals consistent agreement across a wide range of spatial scales. These results suggest that remote assessment tools such as LEFT hold great potential for determining key ecological features across landscapes and for being utilized in preplanning biodiversity assessment tools.

Key words: beta-diversity; biodiversity evaluation; ecological footprint; ecological risk; field forecast verification; local ecological footprinting tool; local field data; threatened species; vulnerability; web-based landscape planning tool.

Introduction

Throughout the world, there is widespread conversion of natural land for agriculture, urban growth, and industrial and extractive activities. Because most land for development is located outside of protected areas, the spatial arrangement of important ecological features across these landscapes is rarely considered in the preplanning stage; rather, this occurs once the footprint of the development has been decided through an environmental impact assessment (EIA). However, by this stage in the planning process, it is usually too late to relocate the facility and the EIA usually serves to minimize the overall impact through mitigation, restoration, and biodiversity offsetting (Kiesecker et al. 2009).

Usually, when a concession is granted for land conversion, the area of land available is far greater than the actual footprint of the development. It is in the preplanning stage, therefore, that trade-offs are considered in placement of the facility and its related environmental footprint. During the preplanning stage, consideration is usually given to physical features, such as bedrock, and location of existing infrastructures (e.g., road networks, existing pipelines), and there are a number of tools and maps available for providing such information. What has been lacking, however, is a tool to rapidly and remotely assess key ecological features across a landscape; thus, to determine those regions that carry a high ecological risk if damaged. At present, there are several tools that provide spatial mapping of landscapes to determine conservation priorities within protected landscapes (e.g., Possingham et al. 2000, Moilanen 2005, Ciarleglio et al. 2009, Pressey et al. 2009, Moilanen et al. 2011), and some of these have been used outside of protected areas for spatial planning purposes (Kiesecker et al. 2009). However, these are complex tools that require large amounts of technical expertise and time; they also rely upon a large amount of data input by the user. As a result, they are not suitable for quick preplanning assessment, and, as a conse-

Manuscript received 27 July 2014; accepted 17 September 2014. Corresponding Editor: J. Franklin
[7] E-mail: kathy.willis@zoo.ox.ac.uk

TABLE 1. Summary of the layers used in local ecological footprinting tool LEFT (Willis et al. 2012).

Layer	Definition	Data source	Calculation
Biodiversity (beta-diversity)	compositional turnover with respect to environmental covariates	Global Biodiversity Information Facility (GBIF), WorldClim (Hijmans et al. 2005), Harmonized World Soil Database, distance to water bodies (Lehner and Doll 2004)	Uses a generalized dissimilarity model (GDM; Ferrier et al. 2007) to map compositional turnover across the landscape.
Vulnerability	number of threatened species present	International Union for Conservation of Nature (IUCN) Red List, GBIF, WorldClim (Hijmans et al. 2005), Harmonized World Soil Database, distance to water bodies (Lehner and Doll 2004)	Species distribution models (MaxEnt; Philips et al. 2006) are run using GBIF occurrence data for those threatened IUCN species whose native range comprises the study area. Projected presences are summed across all modeled species, creating a map that represents the spatial arrangement of threatened IUCN species.
Fragmentation	size of the vegetation patch	GlobCover (Bontemps 2009)	Uses FRAGSTATS (McGarigal et al. 2002) to estimate patch sizes of GlobCover vegetation classes (Bontemps 2009).
Connectivity	river, lake, and wetland features; number of migratory species present	Global Lakes and Wetlands Database (GLWD), HydroSHEDS, and Global Registry of Migratory Species (GROMS)	Migratory species hotspots are mapped by overlaying GROMS species range polygons. Pixels situated within a riverine corridor or wetland region in the HydroSHEDS or GLWD layer are given a higher ecological score.
Resilience	defined as the ability to sustain high rates of net primary productivity in areas of low precipitation (as in Klein et al. 2009)	MOD17A3 annual net primary productivity (aNPP from MODIS, NASA's Moderate Resolution Imaging Spectroradiometer) and WorldClim (Hijmans et al. 2005)	Values of aNPP per vegetation class are mapped against typical precipitation of the driest quarter. Those pixels in which high productivity and low rainfall intersect are assigned a higher ecological value.

quence, many preliminary land-use planning decisions are made without consideration of ecological impact.

A research project was therefore initiated to develop a tool to provide fast and effective assessment of the spatial arrangement of key ecological features across landscapes. It set out with the objective of using existing web-based databases, models, and freely available satellite images. User requirements included minimum input by the user, output displayed as a series of maps, spatial arrangement of ecological features at a pixel scale relevant to most developments (<500 m), and an assessment that can be applied on any terrestrial landscape in the world. It was also determined that such a tool should provide three types of spatial data; first, location of the ecological properties of the current landscape (i.e., biodiversity, or beta-diversity, and threatened species, or vulnerability); second, the key features for supporting ecosystem functions (i.e., connectivity, or migration routes, wetlands, and habitat integrity); and third, a measure of the distribution of land resilient to current and future climatic perturbations (Willis et al. 2012).

In the first stage of the project, a methodology was devised that used globally available web-based databases, algorithms, and models to calculate maps depicting five key ecological features across landscapes: beta-diversity, vulnerability, fragmentation, connectivity and resilience (Table 1). The resultant methodology was then applied to three case-study regions in Canada, Russia, and Algeria (Willis et al. 2012). Generating output and maps using databases for these three regions, however, took approximately 12 months of skilled GIS technician time. In a second stage of the project, the aim was to therefore automate the process. The resulting local ecological footprinting tool (LEFT) is a simple web-based interface, in which the user submits a query by providing the coordinates of a site for analysis for a region up to $40° \times 40°$ ($\sim 4000 \times 4000$ km; tool *available online*).[8] The output is a series of maps, a report highlighting the spatial arrangement of the different features across the landscape, and a summary ecological risk map at a 300-m resolution within the given area. This is provided in usually less than 1 hour for typical sizes of concession areas.

A key question with any such tool, however, is the accuracy of the output; how does this compare to what is on the ground? The third stage of this project, therefore, was to evaluate the performance of this tool and the confidence that can be placed in the output. We thus undertook analyses to compare outputs from LEFT with high-quality field-based data sets. A methodology was devised to determine the scale of error associated with the output from LEFT through comparison of outputs obtained for the beta-diversity

[8] http://www.biodiversity.ox.ac.uk/left

TABLE 2. Species occurrence data sets used to produce FIELD- and LEFT-based beta-diversity layers for the two study areas.

Data source	Spatial scale of sample collection	Years of data collection	Amphibians	Birds	Mammals	Plants	Reptiles
Cusuco National Park, Honduras							
FIELD (Operation Wallacea)	35 × 3 km sample routes used to collect continuous observations	2004–2010					
No. species			29	191	83	146	49
No. samples			376	3707	1401	997	501
LEFT (GBIF)	300-km rectangular buffer around each study area (limited to ecoregions in study area)	mostly 20th and 21st centuries					
No. species			156	607	179	8461	270
No. samples			915	2764	1397	46483	1286
Mahamavo Forest, Madagascar							
FIELD (Operation Wallacea)	14 × 3 km sample routes used to collect continuous observations	2009–2011					
No. species			10	127	18	176	53
No. samples			146	5714	434	245	1768
LEFT (GBIF)	300-km rectangular buffer around each study area (within the same ecoregions as the study area)	Mostly 20th and 21st centuries					
No. species			10	50	15	4278	23
No. samples			16	101	18	19848	32

Notes: Cusuco National Park, Honduras; 15.46°–15.58° N, 88.36°–88.15° W; area, 310 km^2; biome is tropical moist forest (Schultz 1995). Mahamavo Forest, Madagascar; 16.00°–15.25° S, 46.28°–47.10° E; area, 5215 km^2; biomes are tropical dry forest, wetlands, and savannas. Sample per species ratios between FIELD and GBIF data sets for each study area were tested for difference in the mean (paired two-tailed Student's *t* test); Cusuco $P = 0.03$; Mahamavo $P = 0.053$.

and the vulnerability layers (the two layers in LEFT that can be fully validated with biological field data) with two independent, high-quality field data sets in Honduras and Madagascar. This paper describes the results of this evaluation and discusses the potential of web-based tools such as LEFT to remotely assess these two important ecological features across global landscapes.

Study Areas

Validation sites were selected using the following criteria: spatial extent, availability of data to assess vulnerability and beta-diversity, and representation of different biome types. In addition, it was essential that none of these data sets had been uploaded into the Global Biodiversity Information Facility (GBIF), which is accessed by LEFT in its remote assessment of ecological properties; they therefore needed to provide a set of field data independent of the global databases used in LEFT (Table 2; Appendix A; *available online*)[9]. This exhaustive set of criteria largely limited the amount of available validation data sets. Field sites from two different vegetation biomes that fit these criteria were obtained for comparison, namely the Mahamavo watershed, Madagascar (tropical dry forest, wetlands, and savanna) and a site within Cusuco National Park, Honduras (montane tropical moist forest; Fig. 1).

Maps of beta-diversity and vulnerability (i.e., distribution of vulnerable species) were automatically generated for each site using the web-based tool, and then maps were created for the same sites manually, using the same algorithms and GIS projections, but, in the construction of these, the data obtained from the global databases was substituted with independent, local field data.

Methods

Mapping landscape beta-diversity

Following the methods of Ferrier et al. (2007), LEFT implements a series of functions designed to map beta-diversity across the landscape using a generalized dissimilarity model (GDM). In order to calculate beta-diversity, data for vascular plants and terrestrial vertebrates are automatically downloaded from GBIF and assigned to a 300-m resolution grid. Biodiversity data are obtained for a larger area than the one selected by the user (a 300-km rectangular buffer is used around

[9] http://oxlel.zoo.ox.ac.uk/wp-content/uploads/2013/11/Willis_et_al_2015_Ecol_Appl_biodiversity_records.zip

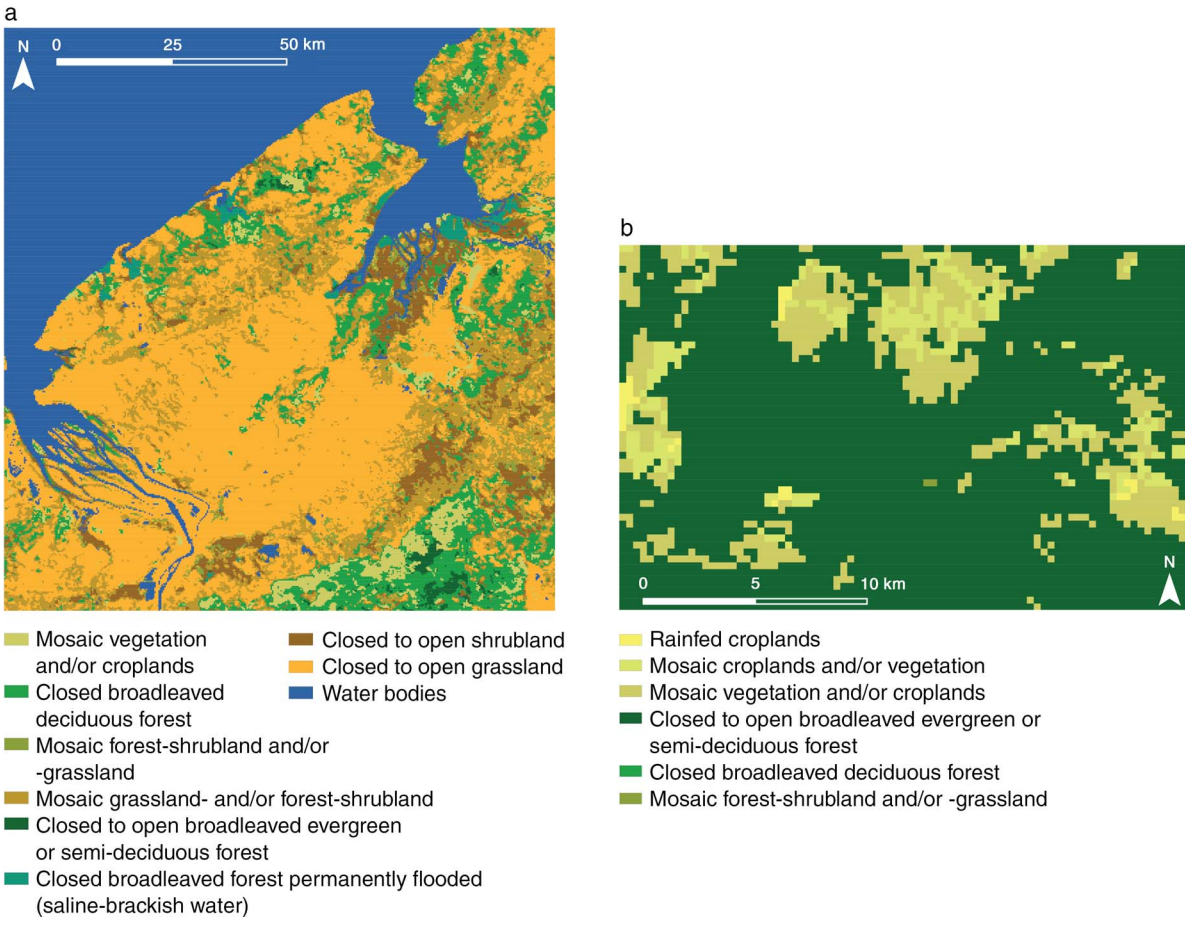


FIG. 1. Land cover classification of the two study sites as depicted in global land cover map 2009 (300-m resolution; Bontemps et al. 2009): (a) Mahamavo, Madagascar; (b) Cusuco, Honduras. Land cover type is included in the legend under each of the two maps.

each study area) for constructing the GDM. These data are filtered to retain only those occurrences recorded in locations within the same set of eco-regions (Olson et al. 2001) as contained in the study area (Willis et al. 2012). Any duplicate records of species occurrence in one location are removed. The result is a matrix of species occurrence within 300-m pixels.

Dissimilarity matrices based on the presence–absence version of the Bray-Curtis index are constructed for those pixels containing biodiversity data for each of five groups: plants, amphibians, birds, mammals, and reptiles. The current version of the GDM R package (Ferrier et al. 2007) is limited to 32-bit processors (*available online*).[10] This causes memory limitations when computing the GDM algorithm, which will be largely overcome when a new, 64-bit version of the GDM package will be available. Meanwhile, the pixel-by-species matrix for each group is currently limited to 2000 pixels. Any group consisting of more than 2000 pixels is randomly reduced to this maximum size (in our study, this procedure was only necessary in one case, plants for Mahamavo in the LEFT run). These matrices are then regressed against dissimilarity matrices computed for four WorldClim (Hijmans et al. 2005) bioclimatic variables (annual mean temperature, total annual precipitation, temperature seasonality, and precipitation seasonality), plus distance to water bodies (Lehner and Doll 2004) in the animal groups, and percentage of nitrogen in the soil and soil water holding capacity (Harmonized World Soil Database) in plants (Willis et al. 2012; *available online*).[11] The aim is to build a model that explains the relationship between environmental variable dissimilarities and beta-diversity that can then be used to interpolate values of beta-diversity in 300-m grid cells for which there is no biodiversity information.

GDM is then projected on 300 randomly selected pixels within the study site and an inverse distance weighted (IDW) interpolation implemented to obtain the compositional dissimilarity value for each 300-m

[10] https://sites.google.com/site/gdmsoftware/home

[11] http://daac.ornl.gov/SOILS/guides/igbp-surfaces.html

pixel. This process is iterated 10 times without replacement, and the median dissimilarity value among the 10 runs is taken. The calculation is performed individually for each taxonomic group and a single value of beta-diversity for each pixel is assigned to the final beta-diversity map by finding the maximum value calculated from all the groups (Willis et al. 2012).

For validation, we followed the same procedure, but using the high quality geo-referenced field data instead of the GBIF data.

Mapping distribution of vulnerable species across the landscape

In LEFT, the distribution of vulnerable species across the landscape is calculated using the International Union for Conservation of Nature Red List (IUCN; IUCN 2012). This list has assessments for >70 000 species globally, of which about 40 000 have spatial data. The list of countries making up the geographic native range of each threatened species is compiled from this list. For each species, a list of locality observations is obtained from GBIF and filtered by boundary polygons of the countries of its native geographic range. Next, the potential distribution of each species is modeled using the maximum entropy algorithm implemented in the software MaxEnt (Phillips et al. 2006, 2008) and six environmental and geophysical variables (mean annual temperature, temperature seasonality, total annual precipitation, precipitation seasonality, slope, and aspect). The bioclimatic data are obtained from WorldClim (Hijmans et al. 2005).

MaxEnt uses presence-only data and has been implemented in a wide range of applications, including mapping species distributions to design conservation priorities (e.g., Pawar et al. 2007, Thorn et al. 2009). Its advantage relies on its ability to use low numbers of species presence data to map distributions over a large area. Since threatened species are, by their nature, difficult to find and commonly occur in low abundances, MaxEnt is an effective method to build species distribution models (SDMs) in this context. Nevertheless, we limited the implementation of MaxEnt to species containing five or more observations, and only well-performing models were retained (defined as having an area under the curve of the receiver operating characteristic [AUC] $\geq$ 0.7; Araújo et al. 2011).

For each LEFT analysis, the models for those species whose native range is included in the study area are projected on the area under study, and the projected probability values hardened to [0, 1] at a cut-off value of 0.5. The hardened (binary, presence–absence) values are then summed across all modeled species. The resulting map represents the relative numbers of threatened species potentially present across the study area. Pixels with potentially more globally threatened species are treated as having a higher ecological value. The resultant map indicates areas across the landscape with higher concentrations of vulnerable species. LEFT output also provides an estimate of the total number of vulnerable species present in the region.

For validation, we modeled the distribution of any species listed at least as threatened in the IUCN Red List that also occurred in our field data sets using MaxEnt (Phillips et al. 2006) with the same criteria as implemented in LEFT.

Map comparison

Encapsulating the similarity between two maps (or fields) in a single score of a given statistic (e.g., correlation) is limited by the complexity of the information contained in a map. A field is not a collection of independent values, and this essential information is lost in most comparison cases in which each matrix is treated in practice as a vector (Briggs and Levine 1997). This has been of paramount concern in the field of meteorology, where comparisons between weather forecasts and observed weather outcomes are essential. Multivariate measures of similarity have been proposed in these cases (e.g., Livezey et al. 1995), which have the advantage of offering a richer picture of the similarity between fields and offer information at several spatial scales. In particular, the implementation of discrete wavelet transforms to orthogonally decompose two fields into dyadic (2^h) scales has been successfully utilized over other multivariate techniques when aiming to compare two single maps (as opposed to having many replicates of each map; Briggs and Levine 1997). This method (field forecast verification, or FFV) decomposes a given original field into a number h of maps, each at a different spatial scale and orthogonal (i.e., independent) from the rest. The outcomes from this partitioning of the original field can be used to compute standard similarity scores at different spatial scales, resulting in a multivariate (multi-scale) similarity score that enables a much richer comparison between two maps and avoids both the limitations of one-dimensional scores and the need for replication of other multivariate comparison techniques.

In our case, our aim was not to compare a forecast with an observed map (such as in meteorology), but rather two outcomes from a given algorithm produced using independent sources of data (the web-based, globally sourced LEFT and the field-based local data sets). The field-based results can be seen as a ground-truthed (validation) data set, although they are not free of noise and potential bias (i.e., not all the features stemming from the field-based data set will necessarily be closer to reality than the ones stemming from LEFT). In any case, similarity between both maps will strongly suggest that the patterns depicted by each are robust and, hence, reliable. We, thus, applied FFV twice, to compare the beta-diversity fields and to compare the vulnerability fields obtained from LEFT and from field-based data sets. For each dyadic scale, Pearson correlation (r) and mean squared error (mse) were computed between both fields. Correlation is a good measure of linear relationship between two fields but

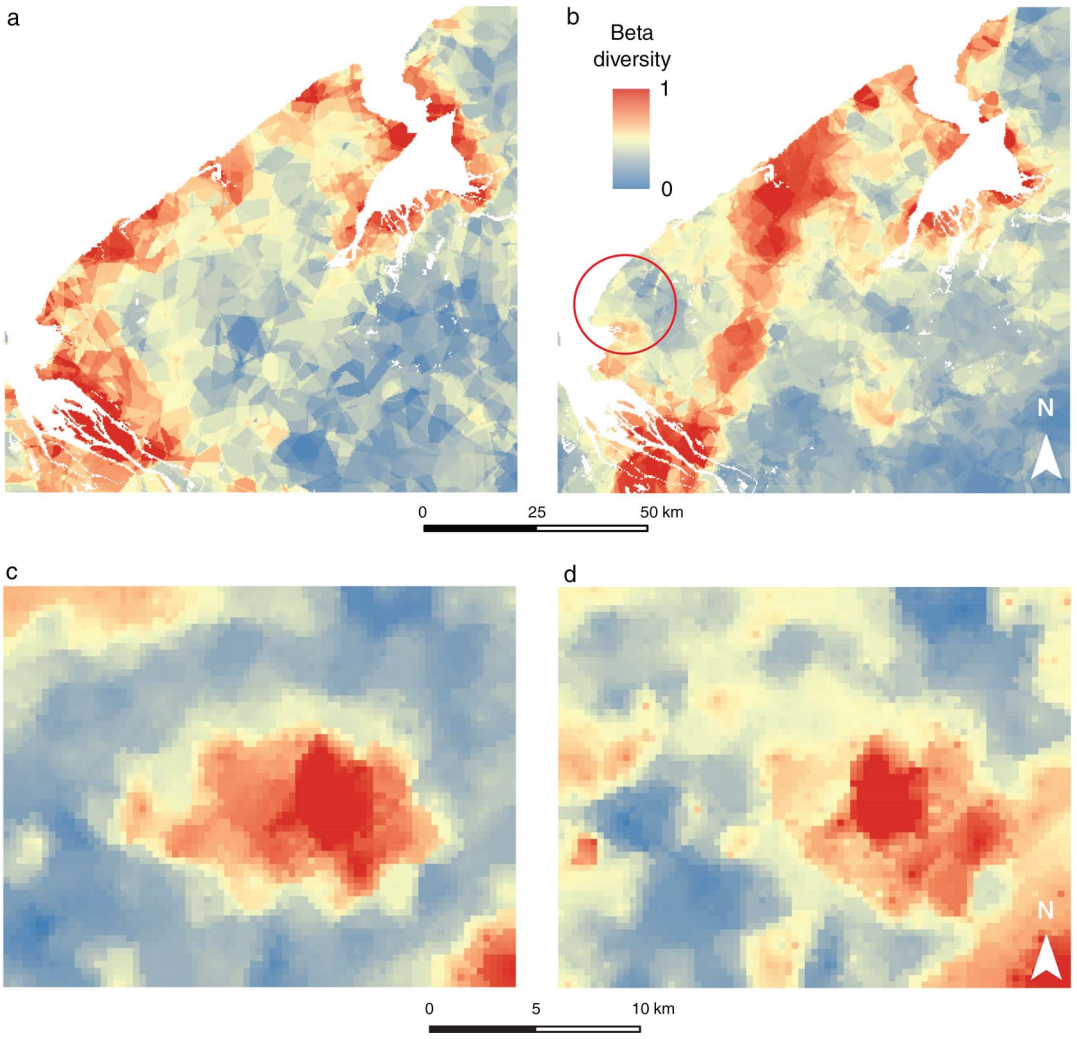


FIG. 2. Beta-diversity maps for (a and b) Mahamavo, Madagascar and (c and d) Cusuco, Honduras computed using generalized dissimilarity modeling (GDM). Maps computed using the local ecological footprinting tool (LEFT) sourced with global online databases (a and c); and computed using local field data (b and d). The red circle in (b) indicates the location of the city of Mahajanga, which causes a disagreement between the LEFT and local field data beta-diversity outputs.

ignores bias; whereas, mse ignores the linear relationship but is a good measure of accuracy (Murphy 1995). Hence, r and mse give different and complementary information important to understanding aspects of field (map) closeness (Briggs and Levine 1997).

Selection of the best mother wavelet (i.e., the type of wavelet used to decompose the initial map into dyadic components of increasing coarseness) was done according to a measure of minimum entropy (i.e., the wavelet type producing a transform with minimum entropy, as measured by the Shannon index, was selected). This, in practice, meant the selection of the wavelet type producing maps with the greatest ratio of large to small coefficients (i.e., the wavelet type that best compressed the information at each spatial scale). Haar wavelet (Daubechies 1992, Press et al. 1992) showed the best performance in all cases. Further, noise reduction was implemented through soft thresholding (i.e., shrinking towards zero small coefficients from each dyadic map in an effort to remove insignificant information; Briggs and Levine 1997). All input maps were standardized and ranged from 0 to 1. Whereas the beta-diversity fields were analyzed without transformation, the vulnerability fields were log-transformed prior to the FFV, since they were highly left-skewed. FFV was implemented using the software map comparison kit (Visser and de Nijs 2006).

RESULTS

Beta-diversity mapping

Results for the validation of the beta-diversity layer were highly encouraging (Fig. 2). We found consistently high r (>0.8 in Cusuco, >0.7 in Mahamavo) and low mse scores (<0.005) across all scales between both fields for Mahamavo and Cusuco (Figs. 3–4), which indicate a

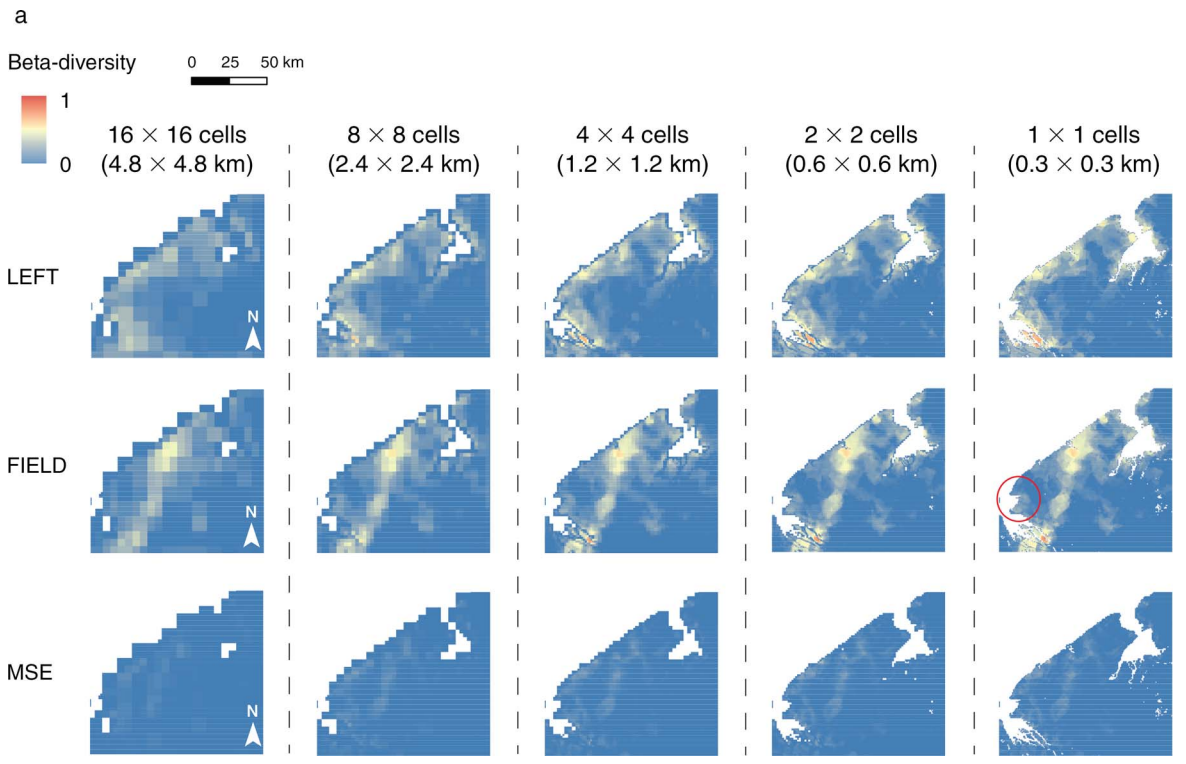


FIG. 3. Maps of (a) Mahamavo, Madagascar and (b) Cusuco, Honduras comparing local ecological footprinting tool (LEFT) and field-based (FIELD) beta-diversity outputs at each of the five successively finer orthogonal wavelet scales (see *Methods*). For each successive dyadic spatial scale (column labels, with spatial resolutions ranging from ~25 km^2 to 0.09 km^2), a map is produced that is statistically independent from maps at other scales, allowing for a multivariate assessment of the agreement between the LEFT and FIELD outputs. The bottom row depicts the mean square error (mse) between LEFT- and FIELD-based maps. The red circle indicates the location of the city of Mahajanga, which causes a disagreement between the LEFT and FIELD beta-diversity outputs.

robust similarity over a full range of geographical dimensions.

In Mahamavo, the main patterns depicted by both maps included lower beta-diversity values through the central and southern region of the site (which is currently covered in savanna), and higher values along the northwest coast and in the western regions of the study region (which currently contain mangroves and biodiversity-rich coastal tropical dry forest; Figs. 1–2). In Cusuco, the algorithm consistently picked the highest beta-diversity in the central part of the region, which is a mountainous region with higher species diversity, abundance, and turnover.

The overall higher similarity statistics in Cusuco might be due to two factors; the physical structure of the study area in Cusuco consists of a large elevational gradient in its center, which might be easier to replicate than the more complex gradients in Mahamavo; and the lack of large populated areas in Cusuco (the Mahamavo study area includes the city of Mahajanga in its western coast; Fig. 3) means that the bioclimatic variables implemented in the analyses can better mimic the observed patterns (i.e., LEFT is blind to cities, whereas local field data is sensitive to them).

Vulnerability

Results for the validation of the vulnerability layer were also promising (Figs. 4–6). We found moderately high r (>0.7 in Cusuco, >0.6 in Mahamavo) and low mse scores (≤0.012) across all scales between both fields for Mahamavo and Cusuco, indicating robust similarities over a wide range of geographical dimensions. The fact that no major dips in the similarity statistics were observed across all scales is remarkable and highlights a very consistent signal. The overall spatial patterns in both regions followed the same trends as for the beta-diversity layer, with more endangered species in the central mountainous part of Cusuco and in the coastal lands of the Mahamavo region. Again, the overall lower similarity statistics for Mahamavo relative to Cusuco can be attributed to the abovementioned factors (i.e., lack of large elevational gradients and the presence of a large populated area in the Mahamavo study area).

DISCUSSION

In an ideal world, landscape planners would use high-resolution field data sets to assess the spatial distribution of key ecological features across landscapes. However, with the exception of a few regions globally (e.g., parts

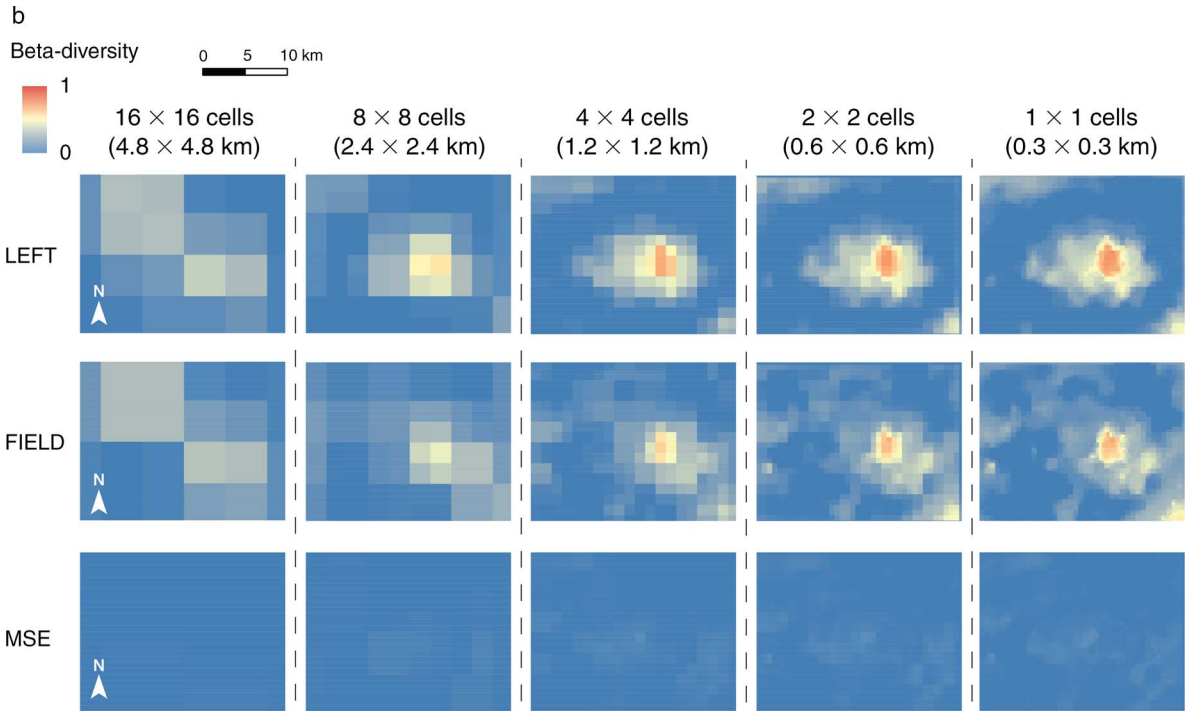


FIG. 3. Continued.

of North America, Finland, New Zealand, and the United Kingdom), the paucity of existing field data and the time and costs involved in collecting new data makes this task near to impossible. Increasingly, therefore, many individuals and organizations are starting to utilize the information contained within web-based global databases to obtain a preliminary ecological assessment. With these databases, however, there is always the question of whether the combination of databases and modeling is a good enough representation of what is on the ground to enable remote assessment to be a viable option. In *Field forecast verification*, we critically examine the results from our study to address this question and, then, go on to discuss the potential sources of uncertainty stemming from the use of web-based global databases and tools associated with them.

Field forecast verification

The comparison of the results of the beta-diversity and vulnerability layers calculated in LEFT with those obtained using local field data suggests that similar results can be obtained for the two study regions (Figs. 2–6). This occurs despite large variation in the total numbers of observations and species between the different data sets (Table 2). FFV successfully produced a multivariate assessment of the agreement between LEFT- and field-based maps, enabling a richer and less biased analysis than one-dimensional statistics. The consistent low inter-map mse and moderately high r scores (Fig. 4) exhibited across all spatial scales suggest that a coherent common signal was captured by both methods and both data sets. In contrast with the field of meteorology where FFV was developed, we do not start from a prediction and an observation map to be compared, in which one (the observation) is taken as correct, and the other (the prediction) is taken as imperfect. Instead, we have two maps created with independent data, both of which can be seen as predictions. The observed agreement between both can, thus, be interpreted as evidence of robust and realistic modeling of patterns of beta-diversity and vulnerability. This is the first time that FFV has been applied to evaluate regional biodiversity assessments, and this study demonstrates its excellent potential to estimate the agreement between two maps.

GBIF-based vs. intensive field-based biodiversity data

The GBIF data sets used by LEFT contain, in general, a greater number of species for each taxon group compared to the field data, but have generally fewer discrete observations per group (Table 2). This is probably due to the fact that the GBIF data has been collected by a larger number of contributors for a longer period of time, and, therefore, represents a less intensive sampling effort compared to field data used in this study. As a result, in terms of the total number of observations, the data used by LEFT may highlight the potential presence of rare, threatened species better than field data sets. However, the resultant over-estimation in LEFT is not necessarily bad, since it provides a precautionary approach towards threatened species and flags up those that that may occur in the region.

Beta-diversity

Cell size (km²)	Mahamavo Corr	Mahamavo MSE	Cusuco Corr	Cusuco MSE
16 × 16	0.74	0.003	0.95	0.001
8 × 8	0.69	0.004	0.84	0.002
4 × 4	0.67	0.004	0.81	0.003
2 × 2	0.68	0.005	0.81	0.003
1 × 1	0.68	0.005	0.81	0.003

Vulnerability

Cell size (km²)	Mahamavo Corr	Mahamavo MSE	Cusuco Corr	Cusuco MSE
16 × 16	0.62 (0.55)	0.012 (0.013)	0.82	0.008
8 × 8	0.66 (0.55)	0.010 (0.013)	0.75	0.008
4 × 4	0.65 (0.56)	0.010 (0.012)	0.77	0.008
2 × 2	0.62 (0.56)	0.012 (0.012)	0.75	0.009
1 × 1	0.61 (0.55)	0.012 (0.012)	0.72	0.010

FIG. 4. Field forecast verification statistics (FFV, see *Methods*) for beta-diversity and vulnerability layers. The statistics depict the agreement between outputs computed using the local ecological footprinting tool (LEFT; sourced with global online databases) and those computed using local field data. Statistics are computed for each successive dyadic spatial scale (cell size from ~25 km^2, or 16 × 16 300-m cells, to 0.09 km^2, or 1 × 1 300-m cell). At each scale, the data compared is statistically independent from other scales, allowing for a multivariate assessment of the agreement between the LEFT and FIELD outputs. Corr is Pearson correlation; MSE is mean squared error. Note the overall high agreement in all cases, which is sustained at all spatial scales. Red lines show the statistics in a graphical format. Note that FFV for vulnerability in Mahamavo depicts two values; the value in parenthesis (and its correspondent dashed red line in the plot) corresponds to an absolute lower bound for each statistic and spatial scale when the effect of a discretization artefact in the analysis is accounted for (see Appendices B–D).

Sources of uncertainty

Selected study regions.—Obtaining high-quality, GBIF-independent biodiversity data sets that would enable the validation of LEFT proved to be a very difficult task, which highlights the momentum that the GBIF portal is taking as a hub for biodiversity data. Our study sites contained a series of characteristics that made them demanding and challenging, and, therefore, optimal validation material. Although Cusuco consisted of a region with a strong climatic gradient, which facilitates SDM's performance, Honduras is not a member of the Organization for Economic Co-operation and Development (OECD) and, thus, belongs in the regions poorly represented by GBIF (84.45% of GBIF records in 2012 originated in OECD countries [Gaiji et al. 2013]). Mahamavo contains a large urban area, and is located in a highly diverse and fragmented terrain comprising transitions from dry savannah to closed tropical forest and mangroves, as well as in a hotspot of biodiversity (and, hence, traditionally under-sampled by GBIF [Gaiji et al. 2013]). The fact that both regions resulted in satisfactory LEFT- vs. field-based comparisons is, thus, especially positive.

Despite the ideal characteristics of the test sites as validation material, the promising results reported in this study are based on two study regions, and, thus, a need exists to find or produce other independent biodiversity records in order to further validate the algorithm with a larger number of test areas in a wide range of ecosystem types and land use regimes.

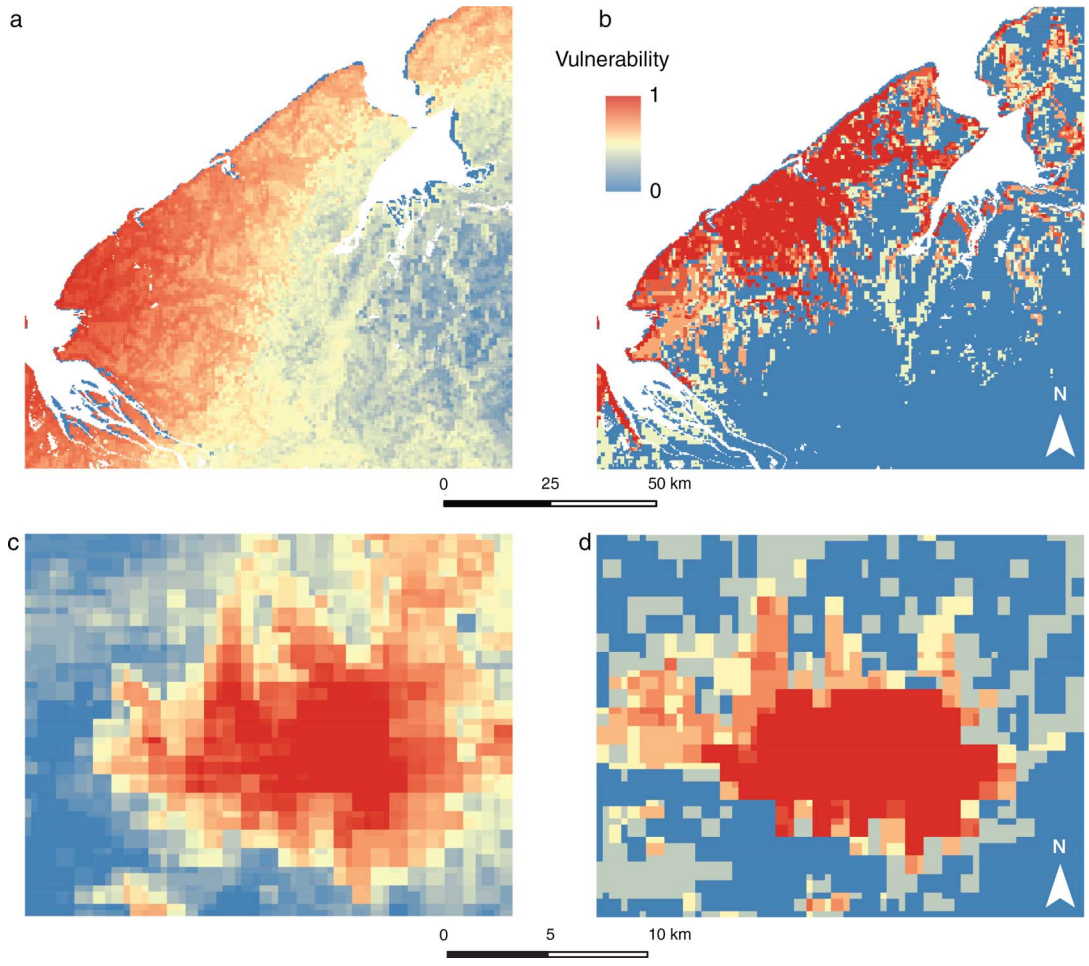


FIG. 5. Vulnerability maps for (a and b) Mahamavo, Madagascar and (c and d) Cusuco, Honduras depicting the relative number of endangered species in each region (vulnerability). Maps computed using the local ecological footprinting tool (LEFT; a and c), sourced with global online databases; maps computed using local field data (c and d).

Urban areas.—In this study, climatic variables and species point observations were employed to estimate probabilities of species occurrence through SDM. This bears consequences when important urban areas are included in the study regions. If cities, or other intensely human-modified landscapes, occur within the climatic space of high-biodiversity assemblages, these will be predicted to contain high biodiversity when, in fact, they might be biodiversity poor, due to human activities. This is clearly seen in Figs. 1–3 for the Mahamavo study area, which contains a large city (Mahajanga, population estimate >250 000, and the third largest seaport in Madagascar) occurring within potentially high-biodiversity conditions. The local nature of the field-based species occurrence data allowed the SDM to predict lower relative diversity over Mahajanga's area of influence, whereas the more regional (stemming from a 300-km buffer zone around the study area) nature of the LEFT GBIF-based data predicted large biodiversity and presence of vulnerable species over the same area (Figs. 2 and 3). The overestimation of biodiversity over human-modified areas will be addressed in future developments of LEFT by adding a land-cover layer as a modeling covariate besides climatic variables. Nevertheless, the usefulness of depicting potentially high biodiversity values over urban and human-modified landscapes where this resource has been depleted needs to be considered.

Global biodiversity databases.—It is necessary to discuss whether the amount of data provided by GBIF is sufficient to be representative of the biodiversity on the ground. A recent report on the quality of the GBIF database assessed three main sources of uncertainty, namely spatial, temporal, and taxonomic (Gaiji et al. 2013). Temporal uncertainty does not directly apply to our study, since we did not infer trends in biodiversity through time. Taxonomy presents two potentially relevant sources of uncertainty. Whereas taxonomical identification in GBIF has substantially improved since 2011 and is not considered a significant problem, there is still a strong bias in the number of records toward taxonomic groups belonging in the kingdoms Animalia and Plants, which

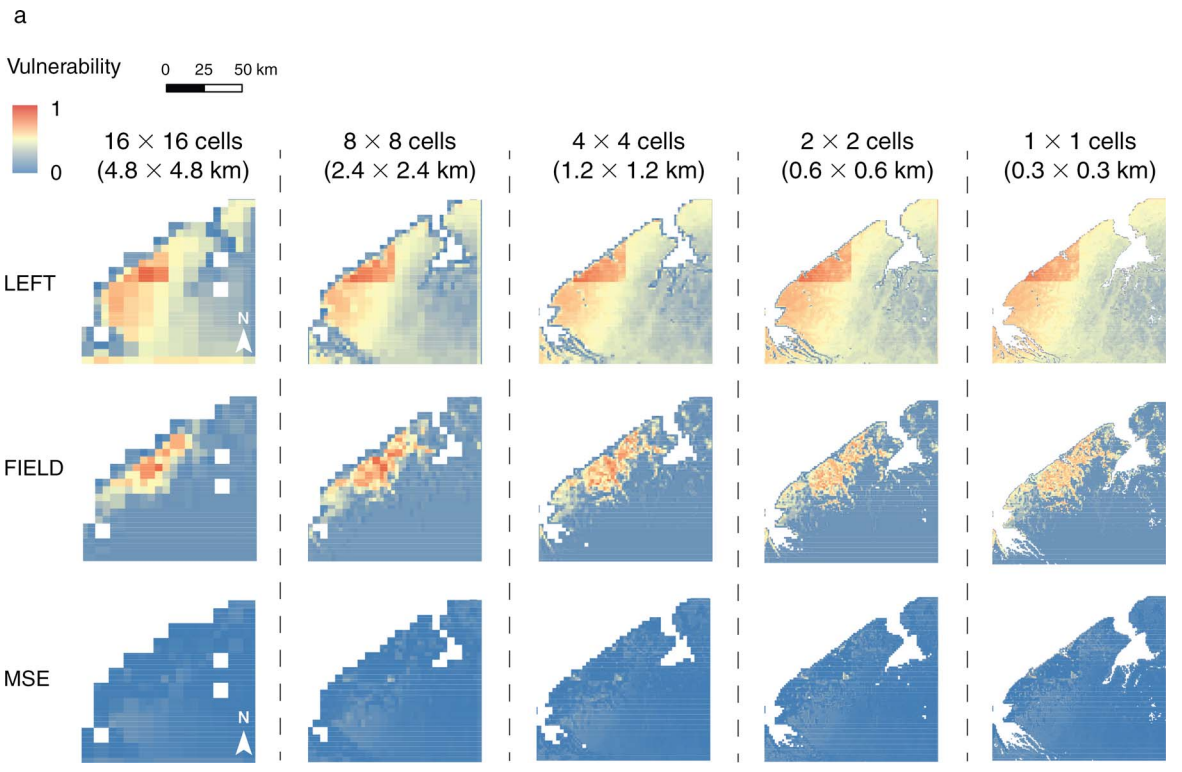


FIG. 6. Maps of (a) Mahamavo, Madagascar and (b) Cusuco, Honduras comparing local ecological footprinting tool (LEFT) and field-based (FIELD) vulnerability outputs at each of the five successively finer orthogonal wavelet scales (see *Methods*). For each successive dyadic spatial scale (shown in the upper labels, from ~25 km^2 to 0.09 km^2), a map is produced that is statistically independent from maps at other scales, allowing for a multivariate assessment of the agreement between the LEFT and FIELD outputs. The lower row depicts the mean square error (MSE) between LEFT- and FIELD-based maps. The discretization artefact seen in the upper images (a sharp block of higher values than its neighbor cells) in (a) is due to the blocky nature of the discrete wavelet transform. A sensitivity analysis showed it to have a limited effect on the computed agreement statistics (see Fig. 4; Appendices B–D).

may alter estimates of overall biodiversity (Gaiji et al. 2013). Finally, GBIF's relative abundances of records do not match global patterns of biodiversity, showing a peak in the northern hemisphere mid-latitudes, instead of the tropics and biodiversity hotspots, such as the Cape region (spatial bias; Yesson et al. 2007, Gaiji et al. 2013). However, these numbers become much less biased when accounting for the number of species recorded instead of number of observations (Gaiji et al. 2013).

LEFT focuses on especially well-represented taxonomic groups in GBIF. Although there is no doubt that accounting for fungal, protozoan, or bacterial biodiversity would inform of fundamental ecological properties of the landscape, this is, at the moment, an unattainable goal, given the scarcity of data in these taxa, that we hope will become feasible in the future. The adaptive and flexible structure of LEFT would surely allow such implementation. Within Plants and Animalia, large differences occur within different families, but species records are generally abundant; even in the case of large families (>1000 species), 5–10% of the species within each family were found to contain optimal data for SDM (Gaiji et al. 2013).

In the assessment of the impact of GBIF biases in the presented approach, it is also important to clearly take into consideration how GBIF data was used in our methodology. LEFT uses GBIF records as observations. In the computation of beta-diversity, all possible records are taken (with the exception of non-natives and duplicates) and their compositional dissimilarities modeled using spatially explicit environmental data. In the computation of vulnerability, only species listed as threatened in the IUCN are used to feed SDMs, employing a set of spatially explicit climatic variables to statistically define each species' climate envelope and project it onto the selected study area, and only the species with successfully validated models (i.e., AUC > 0.7; Araújo et al. 2011) are included. The use of robust modelling with spatially explicit environmental data sets minimizes the issue of the spatial representativeness of the obtained biodiversity patterns.

The numbers discussed in this paper suggest that there is presently an already very large amount of biodiversity information in GBIF, and that, despite its limitations in terms of spatial and taxonomic completeness, this information proves useful in providing robust insights regarding biodiversity patterns over the study areas and

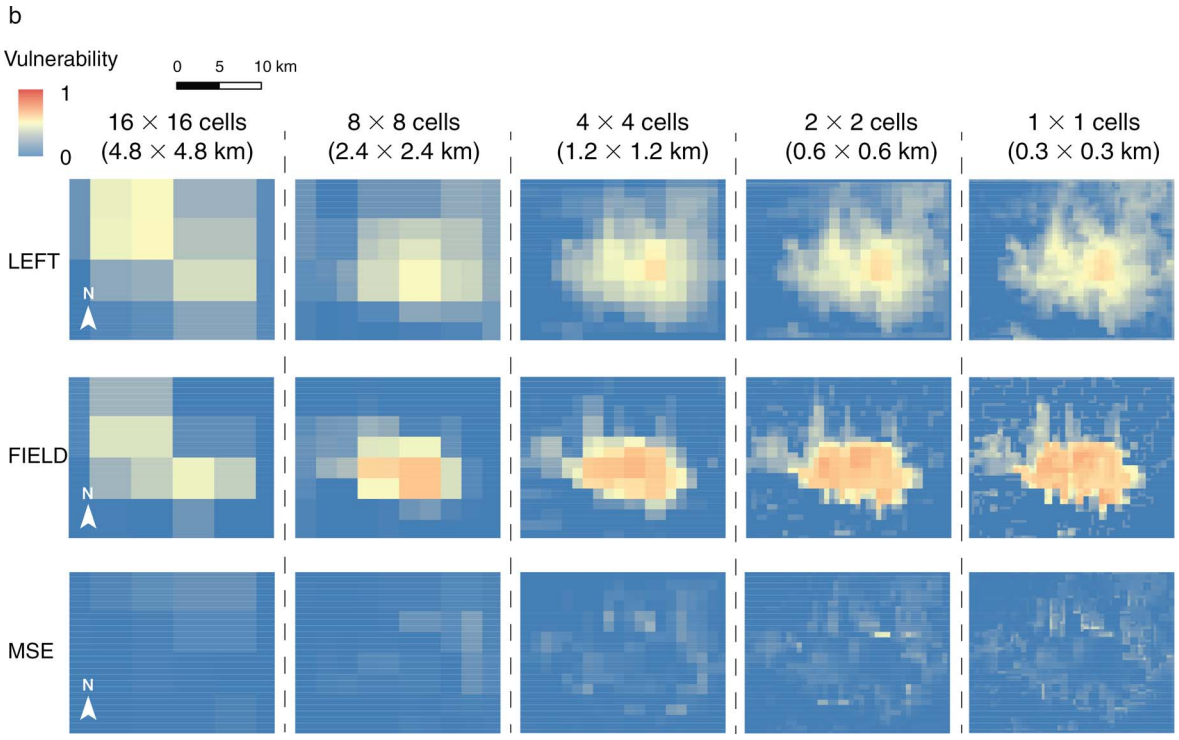


FIG. 6. Continued.

the target taxa. Given the characteristics of the data used, LEFT's strengths lie in estimating relative measures of biodiversity through space (as suggested by Boakes et al. 2010) and not absolute values of biodiversity (e.g., species lists), which, although also provided, are less correlated with on-the-ground situations. Absolute biodiversity values may, in fact, be obtained, not in the preplanning phase for which LEFT has been devised, but in later phases with on-the-ground assessments. Finally, it is important to note that global biodiversity databases, such as GBIF and the IUCN Red List, are set to only improve with time, and with them the potential of methods such as those outlined in this study.

Similar taxonomic biases as discussed for GBIF can be expected from the IUCN Red List of vulnerable species. Since only the list of endangered species is used in LEFT and the spatial extent of these is inferred from GBIF occurrences, no other sources of uncertainty need to be discussed.

CONCLUSION

This study tested the potential of using large databases, models, and automated tools to determine important ecological features across global landscapes. The combination of modeling techniques and large biodiversity databases, such as the Global Biodiversity Information Facility (GBIF) and the International Union for Conservation of Nature Red List (IUCN), constitutes a powerful resource able to provide remote species occurrence data. This has enabled assessments of what is on the ground that capture relative spatial patterns of beta-diversity and vulnerability robustly. Potential sources of uncertainty inherent in this approach involve spatial biases in GBIF, which are minimized due to the way in which LEFT employs GBIF data and are set to decrease as the databases grow and become more complete; the inability of LEFT to take into account major urban and human-modified landscapes that modify biodiversity patterns on the ground, which can be surmounted by using land use maps as covariates to model species distributions; and the limited number of regions over which LEFT has been validated, which is a consequence of the lack of GBIF-independent biodiversity data sets.

Despite these potential sources of uncertainty, the results from this study suggest that the information provided by LEFT is robustly similar to field data-generated biodiversity and vulnerability patterns in the two case study regions. The similarity of the spatial patterns obtained through LEFT and field-based approaches can only be explained by the fact that they are reflecting fundamental and realistic properties of the landscape. Validating the output of LEFT represents an important step in the development of a tool that enables the assimilation of biodiversity data into a format that is easily accessible and interpretable by non-specialists, citizen scientists, and landscape planners alike.

ACKNOWLEDGMENTS

The authors acknowledge funding for this project from Statoil, Operation Wallacea and the Oxford Martin School

(Biodiversity Institute and the Oxford Martin Program on Resource Stewardship), and interesting and helpful discussions and collaboration with GBIF and members of the Biodiversity Institute, Oxford. We also thank Alex Hagen-Zanker (University of Surrey, UK) and Ton de Nijs (National Institute for Public Health and the Environment at Bilthoven, The Netherlands) for their help in dealing with the field forecast verification methodology.

Literature Cited

Araújo, M. B., D. Alagador, M. Cabeza, D. Nogués-Bravo, and W. Thuiller. 2011. Climate change threatens European conservation areas. Ecology Letters 14(5):484–492.

Boakes, E. H., P. J. K. McGowan, R. A. Fuller, D. Chang-Qing, N. E. Clark, K. O'Connor, and G. M. Mace. 2010. Distorted views of biodiversity: spatial and temporal bias in species occurrence data. PLoS Biology 8(6):e1000385.

Bontemps, S., P. Defourny, E. Van Bogaert, O. Arino, V. Kalogirou, and J. R. Perez. 2011. Globcover 2009: product description and validation report. Université Catholique de Louvain and European Space Agency. http://due.esrin.esa.int/globcover/LandCover2009/GLOBCOVER2009_Validation_Report_2.2.pdf

Briggs, W. M., and R. A. Levine. 1997. Wavelets and Field Forecast Verification. Monthly Weather Review 125:1329–1341.

Ciarleglio, M., J. W. Barnes, and S. Sarkar. 2009. ConsNet: new software for the selection of conservation area networks with spatial and multi-criteria analyses. Ecography 32:205–209.

Daubechies, I. 1992. Ten lectures on wavelets. SIAM, Philadelphia, Pennsylvania, USA.

Ferrier, S., G. Manion, J. Elith, and K. Richardson. 2007. Using generalized dissimilarity modelling to analyse and predict patterns of beta diversity in regional biodiversity assessment. Diversity and Distributions 13:252–264.

Gaiji, S., V. Chavan, A. H. Ariño, J. Otegui, D. Hobern, R. Sood, and E. Robles. 2013. Content assessment of the primary biodiversity data published through GBIF network: status, challenges and potentials. Biodiversity Informatics 8:94–172.

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

IUCN (International Union for Conservation of Nature). 2012. Red list of threatened species. Version 2012.1. http://www.iucnredlist.org

Kiesecker, J. M., H. Copeland, A. Pocewicz, N. Nibbelink, B. McKenney, J. Dahlke, M. Holloran, and D. Stroud. 2009. A framework for implementing biodiversity offsets: selecting sites and determining scale. BioScience 59:77–84.

Klein, C., K. Wilson, M. Watts, J. Stein, S. Berry, J. Carwardine, M. S. Smith, B. Mackey, and H. Possingham. 2009. Incorporating ecological and evolutionary processes into continental-scale conservation planning. Ecological Applications 19:206–217.

Lehner, B., and P. Doll. 2004. Development and validation of a global database of lakes, reservoirs and wetlands. Journal of Hydrology 296:1–22.

Livezey, R. E., J. D. Hoopingarner, and J. Huang. 1995. Verification of official monthly mean 700-hPa height forecasts: an update. Weather Forecasting 10:512–527.

McGarigal, K., S. A. Cushman, M. C. Neel, and E. Ene. 2002. FRAGSTATS: spatial pattern analysis program for categorical maps. University of Massachusetts, Amherst, Massachusetts, USA.

Moilanen, A. 2005. Reserve selection using nonlinear species distribution models. American Naturalist 165:695–706.

Moilanen, A., B. J. Anderson, F. Eigenbrod, A. Heinemeyer, D. B. Roy, S. Gillings, P. R. Armsworth, K. J. Gaston, and C. D. Thomas. 2011. Balancing alternative land uses in conservation prioritization. Ecological Applications 21:1419–1426.

Murphy, A. H. 1995. The coefficients of correlation and determination as measures of performance in forecast verification. Weather Forecasting 10:681–688.

Olson, D. M., et al. 2001. Terrestrial ecoregions of the worlds: a new map of life on Earth. BioScience 51:933–938.

Pawar, S., M. S. Koob, C. Kelley, M. Firoz Ahmed, S. Chaudhuri, and S. Sarkar. 2007. Conservation assessment and prioritization of areas in Northeast India: priorities for amphibians and reptiles. Biological Conservation 136(3):346–361.

Phillips, S. J., R. P. Anderson, and R. E. Schapire. 2006. Maximum entropy modeling of species geographic distributions. Ecological Modelling 190:231–259.

Phillips, S. J., and M. Dudik. 2008. Modeling of species distributions with MaxEnt: new extensions and a comprehensive evaluation. Ecography 31:161–175.

Possingham, H., I. Ball, and S. Andelman. 2000. Mathematical methods for identifying representative reserve networks. Pages 291–305 *in* S. Ferson and M. Burgman, editors. Quantitative methods for conservation biology. Springer-Verlag, New York, New York, USA.

Press, W. H., S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery. 1992. Numerical recipes in C. Second edition. Cambridge University Press, Cambridge, Massachusetts, USA.

Pressey, R. L., M. E. Watts, T. W. Barrett, and M. J. Ridges. 2009. The C-plan conservation planning system: origins, applications, and possible futures. Pages 211–234 *in* A. Moilanen, K. A. Wilson, and H. P. Possingham, editors. Spatial conservation prioritization: quantitative methods and computational tools. Oxford University Press, Oxford, UK.

Thorn, J. S., V. Nijman, D. Smith, and K. A. I. Nekaris. 2009. Ecological niche modelling as a technique for assessing threats and setting conservation priorities for Asian slow lorises (Primates: *Nycticebus*). Diversity and Distributions 15(2):289–298.

Visser, H., and T. de Nijs. 2006. The map comparison kit. Environmental Modelling and Software 21:346–358.

Willis, K. J., E. S. Jeffers, C. Tovar, P. R. Long, N. Caithness, M. G. D. Smit, R. Hagemann, C. Collin-Hansen, and J. Weissenberger. 2012. Determining the ecological value of landscapes beyond protected areas. Biological Conservation 147:3–12.

Yesson, C., et al. 2007. How global is the global biodiversity information facility? PLoS ONE 2(11):e1124.

Supplemental Material

Ecological Archives

Appendices A–D are available online: http://dx.doi.org/10.1890/14-1431.1.sm