Cell PRESS

Integrating biodiversity distribution knowledge: toward a global map of life

Walter Jetz[1], Jana M. McPherson[2,3] and Robert P. Guralnick[4,5]

[1] Department of Ecology and Evolutionary Biology, Yale University, 165 Prospect Street, New Haven, CT 06520, USA
[2] Centre for Conservation Research, Calgary Zoological Society, 1300 Zoo Road NE, Calgary, AB, T2E 7V6, Canada
[3] Biological Sciences, 8888 University Drive, Simon Fraser University, Burnaby, BC, V5A 1S6, Canada
[4] University of Colorado Museum of Natural History, Bruce Curtis Building, UCB 265, University of Colorado, Boulder, CO 80309, USA
[5] Department of Ecology and Evolutionary Biology, University of Colorado Boulder, Ramaley N122, Campus Box 334, University of Colorado, Boulder, CO 80309, USA

Global knowledge about the spatial distribution of species is orders of magnitude coarser in resolution than other geographically-structured environmental datasets such as topography or land cover. Yet such knowledge is crucial in deciphering ecological and evolutionary processes and in managing global change. In this review, we propose a conceptual and cyber-infrastructure framework for refining species distributional knowledge that is novel in its ability to mobilize and integrate diverse types of data such that their collective strengths overcome individual weaknesses. The ultimate aim is a public, online, quality-vetted 'Map of Life' that for every species integrates and visualizes available distributional knowledge, while also facilitating user feedback and dynamic biodiversity analyses. First milestones toward such an infrastructure have now been implemented.

Locating life on earth: the need and opportunity for improved knowledge

The geography of life on earth lies at the heart of ecology, evolution and the interaction of nature with human society. The spatiotemporal context of individual organisms, populations and species defines their environmental and biotic setting. This setting, in turn, drives ecological processes and provides the arena for micro- and macro-evolutionary mechanisms. Geographic data on the distribution of species are also vital for governments, agencies, and companies seeking to develop effective policies, and make sound decisions, regarding land management, health, climate change and biodiversity conservation. Such data thus provide a crucial intersection between the biological sciences and a diverse set of other disciplines. In the form of range map displays in zoos, aquaria and field guides, such data are also familiar to the broader public.

Given this pivotal role of species distribution information, it might be surprising to realize how poorly documented the geography of life on earth is, an impediment termed the Wallacean shortfall [1,2]. For even the best known species, information on their geographic distribution (i.e. where they are present or absent) is orders of magnitude coarser in spatial grain than almost all other important environmental information (Box 1). Here, we describe our vision of how this shortfall might be addressed through a global, collaborative infrastructure for storing, sharing, producing, serving, annotating and improving species distribution information. Our rationale is that the vast majority of existing information about species distributions is not readily available to scientists and the public, and has not been integrated. Undoubtedly, invaluable progress has recently been made with growing digital access to point records, expert range maps, and initiatives that compile and map information on species occurrence (Box 2), but these tend to focus on a single type of distributional knowledge. No service currently exists or is in sight that can host all the available types of information about species distributions, provide model-based integration, quantify uncertainty in individual and integrated data types, implement simple data upload and allow users spatially explicit, wiki-type feedback to advance the biodiversity knowledge base.

We here present our vision of such a service, a web-based infrastructure that provides a representation of the distribution of each species [3]. By drawing on and integrating diverse types of species distribution information, our vision aims to utilize the strengths of one data type to overcome weaknesses in another. The vision builds on national and international data compilation efforts, but goes beyond simple accumulation of data points by facilitating and incentivizing additional data and knowledge mobilization for scientists and amateurs, exposing existing data for feedback, and providing science-based integration of multiple types of information along with new tools for science, conservation monitoring and change projection. The intention is to enable integration of data across institutional, geopolitical and taxonomic boundaries, provide the synergies that arise from data pooling, and deliver spatial biodiversity products to all interested parties. Below, we outline a conceptual and informatics framework for realizing the vision for such a 'Map of Life', and highlight the exciting prospects for biodiversity monitoring, analysis and projection such a system provides. Thanks to seed funding from the US National Science Foundation, components of this framework are currently under implementation (see http://www.mappinglife.org).

Corresponding author: Jetz, W. (walter.jetz@yale.edu).

0169-5347/$ – see front matter doi:10.1016/j.tree.2011.09.007 Trends in Ecology and Evolution, March 2012, Vol. 27, No. 3

Box 1. Grain-size and knowledge gaps in global environmental data

Global environmental data vary both in granularity and reported accuracy. Thanks to advances in remote sensing and modelling efforts, terrestrial data on global topography, land cover and climate are now mostly provisioned at grain sizes of a few hundred meters or less (Figure I). Marine data remain coarser, but at least ocean surface properties are increasingly quantified at scales of a few kilometers (e.g. http://oceancolor.gsfc.nasa.gov/, http://www.myocean.eu/web/24-catalogue.php). Even future environmental change projections are increasingly highly-resolved and successfully down-scaled to a resolution of 50 kilometers and finer [59]. By contrast, from a global perspective, the grain of our knowledge on species distributions remains extremely unrefined for the vast majority of taxa. Among marine species, only mammals and commercially valuable taxa, especially those inhabiting waters above a continental shelf, are likely to be associated with distributional knowledge at scales finer than that of the entire exclusive economic zone (EEZ) of a country, a large marine ecosystem (LME), or FAO fishing area. This translates to a geographically highly variable grain (expressed as side length of an equivalent grid cell) of median ~450 km, ~1,000 km and ~4,000 km, respectively. For better-known (primarily non-marine) groups, reasonably reliable presence/absence information is globally available at the country or eco-region level (with respective medians of ~900 km and ~600 km grain). For a subset of these (e.g. birds, mammals, amphibians, select plant groups), expert opinion range maps exist (e.g. as compiled in secondary literature [60] or conservation assessments [61]) that increase the spatial resolution of our knowledge to ca. 100–200 km [12]. The spatial accuracy of such maps varies with species ecology [62] and tends to be higher in temperate than tropical regions [12]. Finer-grain distribution information exists for only few species in the form of gridded surveys. These are geographically restricted, however, and cover the full global range for only a tiny fraction of species. Thus, even for the best-studied set of 30 000–50 000 terrestrial species, a two to three order of magnitude knowledge gap exists relative to typical topography or land cover information. For the vast majority of species, the gap is four orders of magnitude and larger, because for many distributional knowledge is only available at continental or ocean-basin scales (Figure I). Given continuous improvements in sensor technology, this gap is widening daily. Further, not only is the spatial accuracy of species distribution data dramatically coarse, it also usually remains completely un-quantified (but see [12]).

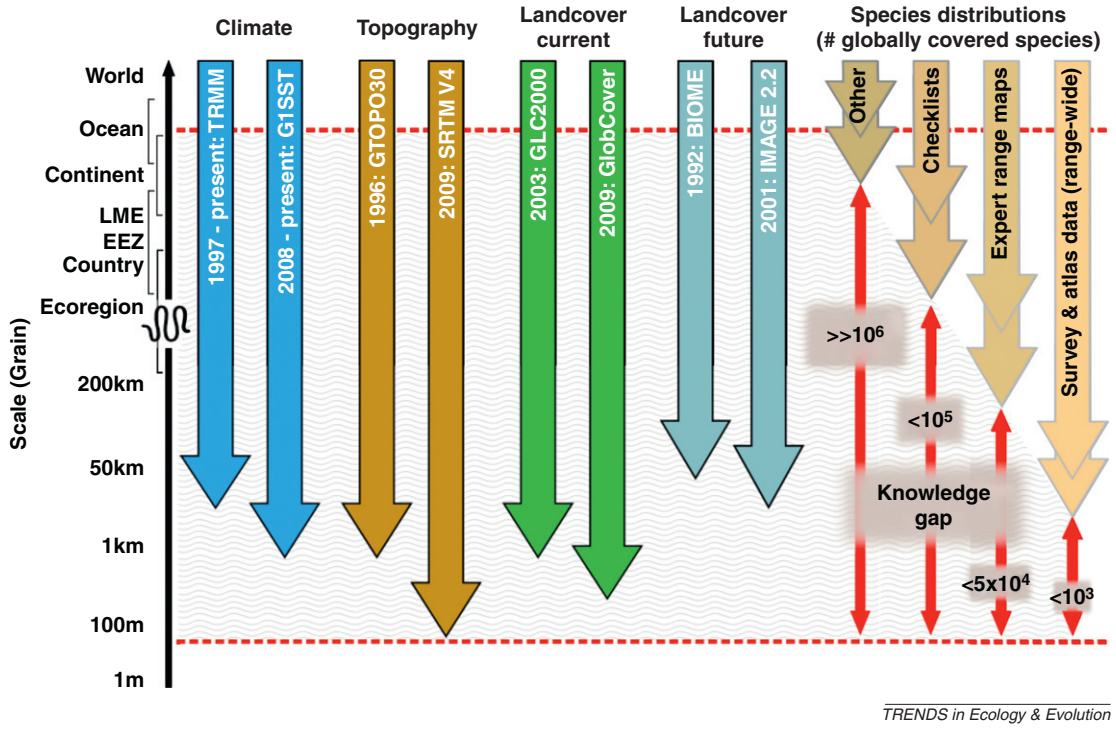

Figure I. Grain size of global-scale data on climate, topography, land cover and species distributions. Lighter shading for species distribution data indicates poorer taxonomic coverage; which is also quantified underneath arrows in terms of the number of species for which each data type is available. Abbreviations are as follows: TRMM = Tropical Rainfall Measurement Mission (http://eros.usgs.gov/#/Find_Data/Products_and_Data_Available/gtopo30_info); G1SST = Global 1 km Sea Surface Temperature (http://ourocean.jpl.nasa.gov/SST); GTOPO30 = global topography at 30 arc seconds (http://eros.usgs.gov); SRTM V4 = Shuttle Radar Topography Mission version 4 [21]; GLC2000 = Global Land Cover 2000 (http://bioval.jrc.ec.europa.eu/products/glc2000/glc2000.php); GlobCover = GlobCover project (http://ionia1.esrin.esa.int); BIOME = BIOME Model [63]; IMAGE 2.2 = Integrated Model to Assess the Global Environment (http://themasites.pbl.nl/en/themasites/image/index.html); LME = large marine ecosystems; EEZ = marine exclusive economic zones.

Conceptual framework for data integration

A key objective of the envisioned infrastructure is to derive the best-possible probabilistic estimate of the occurrence of each species at the finest possible scale over a given temporal range, using the maximum amount of available information. Select types of distribution data have previously been recognized and compared in their core attributes (e.g. spatial precision, autocorrelation structure, surrogacy value) [4–7], but there has been limited integration for distribution modelling (but see e.g. [8–11]). Beyond obvious data types, several other forms of distributional information exist, some more appreciated than others (see Box 3). Individually, each type suffers constraints given its spatial or temporal grain, false positive (absence wrongly declared as presence) or false negative rate (presence wrongly declared as absence), data availability, global uniformity or user input potential. Because the strengths of one data type often offset weaknesses in another, however, combining and integrating the different types dramatically improves the species distribution knowledgebase. Further, as explained

Box 2. A globally integrated infrastructure in the context of current web-based biodiversity initiatives

The past decade has seen several exciting taxon, region, project or data-type focused, web-based biodiversity informatics initiatives that now inspire the vision of an integrated spatial biodiversity infrastructure. The list below is illustrative rather than exhaustive, focusing on projects that innovatively address specific data type, informatics development, social, or taxonomic challenges.

- AmphibiaWeb (http://amphibiaweb.org) is an integrated, community-based web service that provides multiple sources of data about amphibians globally, including current names, photos and point occurrence data along with expert range maps.
- IUCN (http://www.iucnredlist.org), through its taxon assessments that draw on an extensive network of experts, has become a key aggregator and disseminator of expert range maps.
- For point data worldwide, especially museum-based records, GBIF (http://data.gbif.org) and OBIS (http://iobis.org) remain key clearing-houses and provide several online visualization tools and APIs to foster data exchange. Metadata provision and quality control remain problematic.
- Some citizen science initiatives are increasingly sophisticated in organizing and visualizing amateur surveys (e.g. http://ebird.org, http://www.reef.org)
- Movebank (http://www.movebank.org) has pioneered the integration and visualization of animal tracking data.
- Select spatio-temporal interpolation services generalize point occurrences via either interpolation or species distribution model-based predictions (http://www.lifemapper.org), sometimes using projected climate layers (e.g. 'Wallace Initiative'). Quality control and biases in source data remain problematic.
- The Aquamaps project (http://www.aquamaps.org) has taken simple occurrence-habit association modelling further by allowing for expert opinion input in the production of maps for approx. 9000 fish species.
- Regional projects have pushed several other frontiers. For example, the Atlas of Living Australia (http://spatial-dev.ala.org.au) now provides user-driven spatial biodiversity information together with innovative tools.
- Projects such as REBIOMA (http://www.rebioma.net) provide point data and distribution modelling applications for specific regions, and are pioneering web-collaborative approaches to validating and administering records and taxonomies.
- The US Gap Analysis (http://gapanalysis.usgs.gov/) stands out for a comprehensive habitat-model approach, using remote sensing, modelling and expert opinion.

The integrated infrastructure envisioned here as 'Map of Life' aims to build on and complement these and other efforts. By addressing key storage, query, visualization and modelling challenges common to all, and providing mapping and data integration services, the platform is intended to empower region and taxon-specific efforts, freeing resources for investment in core competencies.

below, the value of some data types only materializes when, as envisioned here, the entire (global) ranges of a whole suite of taxa are modelled simultaneously.

Species presence

A variety of data types (see Box 3) can provide a first rough delineation of the maximum area of potential presence over a recent timeframe. Boundaries in the original source(s) can be widened to reflect the spatial uncertainty of underlying data (e.g. 100–200 km for many terrestrial expert range maps [12]), dispersal capacity, or environmental contiguity [13]. Beyond, the species is assumed absent because conditions are either unsuitable or the locations are out of reach. The area inside identifies the maximum area of possible presence during the given time period (i.e. the geographic space and full suite of environmental conditions available to the species), whether suitable or not.

Crucial information on distribution within this potential range is provided by confirmed sites of presence at a known time, based on point records, area inventories, survey and atlas data. These distribution data types vary in spatial grain (finer for points, coarser for inventories), level and spatial evenness of presence probability (high but concentrically declining for point records, low but even for large-area inventories), shape (circular, quadratic, uneven), and temporal extent (day or month for many point data, potentially years or decades for area inventories). What unites all types, however, is their ability to estimate the presence of a species probabilistically in a given bounded area encompassed by them, which in turn reveals associations with the environmental conditions of that area during the observation period.

Species absence

High-quality (i.e. search-effort intensive) area inventories and survey or atlas data might be available to identify confirmed sites of absence within the potential range and time period. Additional probable absences might be derived for locations where allied species (those subject to similar sampling regimes) have been recorded, but not the focal species [14]. If strong dependencies between the focal species and other taxa are known, confirmed or modelled occurrences for those other taxa, too, can be used to deduce sites of probable absence. Moreover, for terrestrial species, opportunities to infer absences have recently arisen from new remote sensing-based vegetation assessments of the surface of the earth (e.g. [15–17], Box 1).

For tens of thousands of terrestrial species, sufficient information on habitat preferences exists to link to high-resolution global vegetation classifications and quantify (un-) suitability of land cover [18–20]. Pixel-level suitability scores might be binary (e.g. absent in 'water' for a purely terrestrial species), ordinal or continuous, reflecting the relative suitability of habitat types (e.g. woodland type, tree height or percent tree cover layers for 'interior forest specialist'). Opportunities for expert-based quantification of suitability extend to all other environmental parameters that have been mapped globally in digital form such as topography, hydrography [21], soil [22], bioclimate, or water-chemistry conditions. Whereas inherent limitations exist [23–26], such expert suitability models have been successfully applied at broad scale for select taxa, terrestrial and marine [20,27,28]. The full potential of such deductive modelling remains underexploited, however, perhaps because spatial environmental layers are often inaccessible to species experts: a shortcoming our vision would address.

Model-based data fusion

The presence, absence and suitability data compiled for a species might subsequently be integrated in some form of species distribution model (SDM). Such models exploit correlations between known occurrence and environmental

Box 3. Data types pertinent to refining species distribution knowledge

We outline twelve major types of information on species distributions and their respective strengths and weaknesses for fine-scale inference (Figure I).

- **Regional checklists** indicate species occurrence within broad geopolitical, geographic, or bioclimatic regions. If thorough, they help infer (temporally crude) absence over large areas.
- **Expert range maps** are expert-drawn outlines of species distributions. Coarse-grained, they suffer high false positive rates that vary with species ecology [12,62,64], but reliably indicate absence outside their boundaries.
- **Modelled distributions** are rule-based predictions of species occurrence based on usually statistically derived environmental suitability. Their utility for inference of absence and probable presence depends on quality, documentation, spatial and temporal scale.
- **Focal species point records** refer to geographically localised specimen, field observation or tracking data. Collection biases impede their use in determining absence [65–68], but they contain spatially and temporally highly resolved information on presence [but see 69].
- **Allied species point records** comprise geographically localized records for species with similar sampling regimes as the focal species. As a proxy measure of survey effort, they can signal probable focal species absence [14].
- **Area inventories** consist of long-term checklists for defined areas (e.g. nature reserves, islands). They can indicate fine-scale presence if the area is reasonably small and, assuming sufficient cumulative effort, also help infer absence.
- **Survey and atlas data** comprise whole-taxon inventories of standardized search units (e.g. transects, field plots, atlas grid cells). Depending on scale, quality and intensity, they reliably indicate presence or both presence and absence, often with high temporal precision.
- **Habitat preferences**, elevation and physiological tolerance limits as documented in the literature or through expert assessment can be coupled with fine-scale land cover, topography and climate data to delineate absence [19,28].
- **Species dependencies** describe tight interactions (beneficial or antagonistic) between the focal and one or more other species (e.g. via crucial resource provision or competitive exclusion [70,71]). If the distribution of involved species is known, such data helps pinpoint areas of presence or absence.
- **Dispersal capacity and related traits** (e.g. body size [72,73]) might inform about the accessibility of unsampled and isolated but environmentally suitable locations, and about patterns of occupancy, space use and abundance, and spatial autocorrelation.
- **Phylogenetic relatedness** often confines species to similar geographic or environmental space [74], so that information about the distributions of related species can assist predictions for phylogenetically affiliated species that lack data.
- **Detectability** differs between species, observers, survey/collection methods and period, and ideally is quantified via repeated sampling or expert knowledge and included in range modelling [75–77].

	Spatial grain	Temporal precision	Availability (# species)	Geographic bias	False positive rate	False negative rate	User input potential	Inference	Examples
Regional checklists	<1000km	10-50 yrs.	Most	Even	High	Low	High	A	http://gis.wwfus.org/wildfinder http://www.kew.org/wcsp/
Expert range maps	<200km*	10-50 yrs.	Many	Even	High	Low	High	A	http://www.iucnredlist.org [19, 61, 78-80]
Modelled distributions	Variable	10-50 yrs.	Some	Biased	Variable	Variable	High	A/P	statistically-derived: [30] expert-derived [20,28]
Focal species point records	<20km	Days	Some	Biased	Low	High	Medium	P	http://www.movebank.org http://www.gbif.org
Allied species point records	<20km	Days	Some	Biased	Low	Medium	Medium	A	http://www.iobis.org http://www.gbif.org
Area inventories	<100km	10-50 yrs.	Some	Biased	Medium	Low	Medium	A/P	http://www.ice.ucdavis.edu/project/bioinventory
Survey & atlas data	<1km	1-10 yrs.	Some	Biased	Low	Variable	Medium	A/P	http://sabap2.adu.org.za http://www.pwrc.usgs.gov/bba
Habitat suitability	<1km	10-50yrs.	Some	Even	Medium	Low	High	A	http://www.iucnredlist.org/initiatives amphibians/analysis/habitat
Species dependencies	Variable	100+ yrs.	Select	Biased	Variable	Variable	High	A/P	[70-71]
Dispersal capacity & related traits	Variable	Days-yrs.	Select	Even	Variable	Variable	Medium	A	[72-73]
Detectability	<1km	Days-yrs.	Select	Biased	Low	Low	Medium	A	[76-77]
Phylogenetic relationships	<1000km	100+ yrs.	Select	Even	High	Low	Medium	A	http://www.treebase.org [74]

* Likely coarser for marine taxa

TRENDS in Ecology & Evolution

Figure I. Twelve major types of data that contribute information on species distributions. Cell entries indicate the nature (text labels) and quality (colour, from white=poor to green=good) of the available spatial, temporal and taxonomic coverage, error rates and user input potential, and their consequence for inference of fine-grained presence (**P**) or absence (**A**). Different data types have different strengths and weaknesses, as evident when examining individual columns (e.g. false positive rates are low for point records and survey data, high for expert range maps and regional checklists).

variables that describe specific dimensions of the ecological niche of species [29–32]. The type and scale of environmental variables that best capture niche dimensions remain under investigation, and vary with species motility, primary habitat and physiology [33–35]. Moreover, the integration of disparate data requires extraction of environmental attributes specific to observation periods. To date, the temporal and spatial grain of analysis in SDM studies is often driven by the resolution of available environmental rather than species data, and only a single distribution data type is considered [36]. The integration of disparate species distribution data types with varying spatial extents, shapes, and spatial uncertainties presents new challenges for which current SDM methods are not well suited. One potential approach is to perform analyses at the coarsest grain of available data (e.g. 100 km) by aggregating finer-grained information. Alternatively, greater ecological relevance and detail might be retained via randomized, probabilistic sampling of, for example, 1 km pixels within areas of potential presence or absence. Such probabilistic sub-sampling should permit diverse data with varying spatial precision to then be jointly analyzed with established SDM techniques (e.g. Generalized Linear Models, Maximum Entropy models, Boosted regression trees, etc.) to produce multi-model-based predictions of species presence probabilities over a given period [37]. Community (e.g. dissimilarity-) based models offer a promising complementary modelling path [38,39]. Ultimately, however, scale-dependence of different environmental constraints on species occurrences [33,40] might favour approaches that are explicitly hierarchical or otherwise integrate across grains [41–43]. Bayesian approaches, for example, seem well suited for combining different types of data, including expert opinions, across grains [8,32,44,45]. These and related methods might also facilitate the inclusion of expert- or survey-derived estimates of detectability, of species dispersal or occupancy correlates, or of relative phylogenetic position; all hitherto underused but potentially powerful forms of information if many species are assessed in a single framework (Box 3).

Resulting *integrative models* (i.e. models that, unlike ordinary SDMs, manage to combine information form a broader range of data types and uncertainties) can subsequently be used to produce probabilistic and binary output maps that depict the best possible estimate of the distribution of a species over a specific time-frame. These maps might be displayed in conjunction with confirmed sites of presence and absence and measures of uncertainty, quantified via internal cross-validation and visualized as, for example, maps depicting confidence intervals [46,47], or consensus among alternative models [48].

Benefits

Although significant challenges remain, the potential benefits of the outlined data integration are probably tremendous. These include improved quality control and cross-validation among distributional data types. For example, points falling inside regions of firmly expected absence could be flagged for exclusion from integrative modelling or potential correction in original data. Data integration should also alleviate the geographic and environmental biases in single distribution data types that plague the performance of ordinary SDMs [14,40,49]. Whereas modelled distributions might remain coarse-grained and uncertain for many species, we submit that even the documentation and quantification of our collective lack of knowledge of species distributions is important progress. Most crucially, the approach proposed will integrate data sets and types that individually might have limited use. Its implementation necessitates infrastructure and services that facilitate the envisioned data provision and integration.

Informatics framework for data integration, mapping and crowd-sourced quality improvement

Besides conceptual challenges, fundamental technical issues for the realization of our vision exist: how to consolidate all the necessary data sources, automate the steps to create integrative models, store their outputs, and provide users with tools to explore and use all the data from this growing repository. Doing so requires an informatics framework with four essential components (Figure 1): (i) an upload and storage mechanism for all data types listed in Box 3; (ii) a workbench for user-defined or semi-automated data integration and modelling to produce estimates of species distributions; (iii) a user interface for uploading, searching and visualizing data, and for editing, commenting and voting on map outputs; and (iv) application programming interfaces (APIs) and unique identifiers that allow other biodiversity data services to seamlessly access individual and integrated data products and accompanying analytical services. Underlying this infrastructure is a common architecture to store all data types and associated metadata, and to track linkages between different data components.

Data mobilization and vetting

The envisioned platform for uploading, storing and publishing distributional information must be designed to encourage data mobilization from a diverse set of data producers, including individual researchers, government or non-governmental organizations, and other applications that aggregate species distribution information (see Box 2). The platform will allow datasets to be uploaded one at a time, in batch-mode or via web-services that remotely harvest existing data repositories. Producers will be prompted to furnish their product(s) with appropriate metadata, some of which (e.g. observer effort) might be used in integrative modelling. They can choose to make their products fully or partially available, or instead store them in private workbenches only accessible to themselves or collaborators via simple user authentication services.

Before contributed records are visualized and included in models, the harmonizing of taxonomic nomenclature, quality control (e.g. regarding misidentifications) and precision assessment should occur. Up-to-date and automated taxon name services will strongly facilitate content provision and reconciliation for the envisioned platform [50]. The planned combination of different data sources and types and the cross-validation it enables is expected to offer substantial additional opportunities for assessing, flagging, and filtering the quality of species distribution

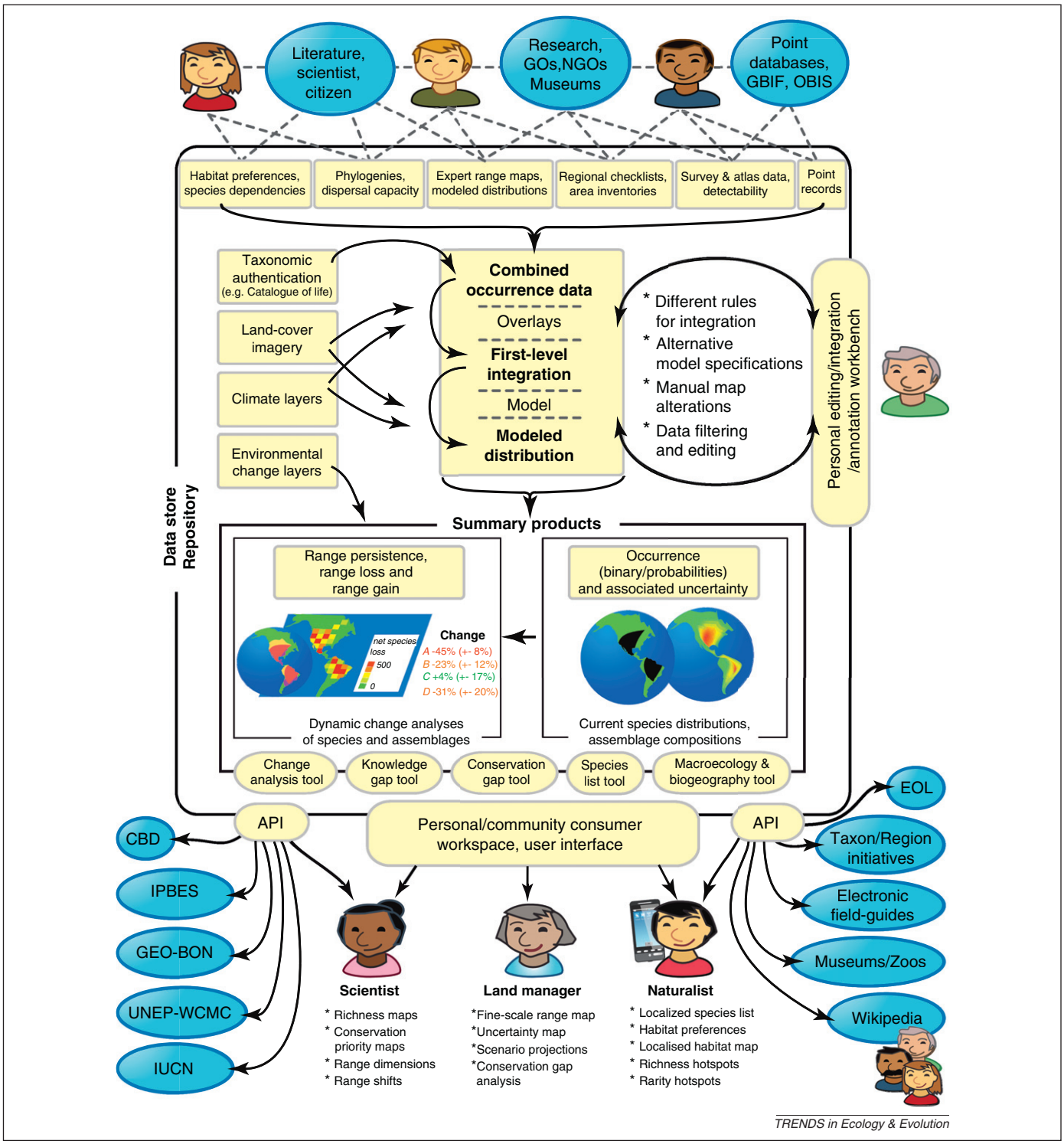

Figure 1. Schematic diagram showing how producers and consumers of species distribution information interact with the envisioned infrastructure, currently under implementation as 'Map of Life'. The planned web platform facilitates the uploading of species distribution information from many different organizations and sources, including data on habitat preferences, point occurrences and expert range maps. The infrastructure stores these data and provides a workbench for integrating them for one or many species. The data compiled, resulting summary information such as binary and probabilistic occurrence maps, and products from analysis tools can be provided to individual consumers, or served via Application Programming Interfaces (APIs) to other services or institutions such as Encyclopedia of Life (EOL, http://www.eol.org), GEO Biodiversity Observation Network (GEO BON, http://www.earthobservations.org/geobon.shtml), initiatives connected to the Convention on Biodiversity (CBD, http://www.cbd.int) or the Intergovernmental Platform on Biodiversity and Ecosystem Services (IPBES, http://www.ipbes.net).

records (e.g. via predicted presence probabilities for outlying points).

Data integration

The user interface of the envisioned platform will then permit all stakeholders to select, combine, query, view, annotate and edit distributional data sources from a workbench (Figure 1). Users will be able to retrieve and download several distributional data products for a single species, higher-level taxon, species sets, or regions of interest. First level integration of data sources will allow users to overlay multiple sources of geographical

distribution data on base maps and retrieve appropriate metadata for each data source. More sophisticated integration will use a system of lookup tables to identify linkages between species habitat preferences and global land cover maps or to flag potential sites of species absence based on other data.

To produce an overall 'best estimate' of the distribution of every species of interest, the platform is expected to contain a customized species distribution modelling application that permits different types of distributional and environmental data to feed into the construction of integrative models. Users will be able to select a set of modelling parameters or accept defaults to run models for individual or sets of species. Model outputs will consist of metadata about the model run adhering to the Ecological Metadata Language, a predicted species distribution map providing presence probabilities and confidence bounds, and uncertainty of modelled predictions (Figure 1). Pixels in the predicted map will initially indicate the relative likelihood of species occurrence, convertible to binary presence-absence maps via user-defined or default thresholding [51].

Dissemination and feedback tools

Maps that are produced will be available for download in various formats and will also be added to the repository for further use and editing. Users logged-in to authenticated accounts will have access to familiar, interactive graphics tools, particularly 'pencil' and 'erase,' enabling them to spatially edit range polygons by re-drawing polygon boundaries. Results, accompanied by author-provided annotation and updated metadata, might be submitted back to the application as a new representation of the species distribution. Maps will be dynamically updated based on the latest data and input, and community voting tools will enable user communities to rate and select those most accurate at the scale in question. These 'best' maps will then be passed to downstream services. All consumers, therefore, will consistently have the most extensive, integrated, vetted, well-documented and updated knowledge products possible. The platform outlined above is currently under development as 'Map of Life' (see http://www.mappinglife.org for more information and project status)

Dynamic global biodiversity analyses and change assessments

A repository of global, quality-assessed and temporally explicit geographic distributions for thousands of species will allow the provision of several key basic and applied analyses and syntheses, either internally or externally via API (see Figure 1). Implemented as tools, these will enable push-of-a-button analyses of the spatial dimensions of biodiversity and facilitate analyses of past or future biodiversity change. All analyses can be dynamically updated as new environmental or species data comes online, or automated altogether. Such tools dramatically simplify the process of scientific data exploration and analysis, leading to potential novel research findings and a much wider use of distributional data by a broad set of stakeholders. We list some intriguing examples.

Revealing biodiversity

A Species List Tool will allow users to draw or select regions of interest to determine which taxa occur within. Users of location-aware mobile devices, for example, will thus be able to download a list of species and associated probabilities of occurrence in their immediate vicinity. In the form of a Conservation Gap Tool**,** probabilistic species lists might be provided simultaneously for a large set of nature reserves to support conservation discovery and inform dynamic analyses of the representativeness of the reserve network. Observed or modelled species lists could easily link to ancillary species data (such as threat status, ecological traits or phylogenetic information). Summary information that builds on such lists (e.g. assemblage species richness, average attributes of species in assemblage) might also be queried over regional or global grids and thus allow on-the-fly basic scientific analyses as part of a Macroecology and Biogeography Tool.

Assessing change

Directly linked to automatically updated environmental (e.g. satellite-based) layers (e.g. [52], http://www.google.org/earthengine), occurrence predictions can be revised or applied to specific time periods, ultimately allowing an ongoing analysis of change in species distributions (or assemblage compositions) given observed environmental change. Key assessments of changes in species- or area-focused (e.g. region, reserve) indicators proffered through a Change Analysis Tool may be implemented and easily updated. Assessments might also be conducted automatically in the background, thereby supporting chief mandates of organizations such as IUCN, UNEP-WCMC, IPBES, etc. [53]. Change analyses would need to account for sampling effort in original sources [54] and might be extended to include projected climate and land-cover layers that allow forecasts of distributional change. By quantifying range gain and loss as well as compositional change for different models and alternative scenarios, the tool, or other global change assessment services that connect to the envisioned platform, will facilitate ongoing, dynamic bioclimate and habitat loss analyses. Of direct use in identifying potentially impacted species and regions (including reserves), these analyses will radically increase the speed with which scientists and decision-makers can use and respond to new information entering the system.

Highlighting information gaps

Another key use of the envisioned infrastructure will be to quantify the magnitude of current taxonomic and geographic knowledge gaps and to track the speed of their closure, for example via a Knowledge Gap Tool. Such a tool will provide crucial documentation and dynamic indicators of our spatial understanding of biodiversity. It should also provide incentives and quantitative support for new data collection by museums, (non-) governmental organizations, researchers and citizen scientists.

The Map of Life vision: challenges, opportunities

Exciting scientific uses for the envisioned 'Map of Life' infrastructure abound and will continue to grow, through additional data linkages for example to abundance,

genotypic, palaeontological, and trait information and the assessment of distributions and niche evolution over time. 'Static' global change assessments in research publications might thereby become dynamic as the underlying methods and algorithms are implemented in, or connected with, this growing architecture. The unprecedented access to spatio-temporal species information might even enable numerous additional applications far beyond our current vision.

Although challenges remain in developing statistically robust models to integrate heterogeneous distribution data types, they are unlikely to represent a permanent obstacle. All computational tools to implement the envisioned cyberinfrastructure already exist. Thus, the greatest challenge for fully realizing our vision might be more sociological than technological. Ever-advancing digital sources and targeted mobilization efforts will allow many datasets to be readily provided, but other crucial information in both analog and digital form is much less accessible.

We hope that Map of Life will both empower and entice the community to actively participate in creating the best possible species distributional knowledge and to recognize the major knowledge gaps that still constrain science and society. Organizations, consortia and scientists conducting atlas and survey work or species distribution modelling might chose Map of Life or affiliated sites as an outlet for making certain primary or modelled species distribution information available to the public. For scientists in particular, who are under increasing pressure to sustainably and accessibly archive their datasets, Map of Life opens up a low-effort opportunity to provide data access while facilitating immediate integration with other information and ultimately, through partner services, continued data citation [55–58].

A project such as Map of Life will not magically close the staggering biodiversity data knowledge gaps that are constraining science and management. However, to date even much of what we as society do know remains un-mobilized, non-integrated, unquantified and underused. Given community participation in building and contributing to a global endeavour such as Map of Life, the concepts, methods, and technologies clearly exist to take a large step forward in geographic understanding and appreciation of biodiversity.

Acknowledgements

We are extremely thankful to Encyclopedia of Life for funding workshops that explored the desirability, feasibility and possible implementation of a 'Map of Life'. The following participants and colleagues contributed invaluable support, ideas and feedback: James Edwards, Jane Elith, Simon Ferrier, Nick King, Brian McGill, Craig Moritz, Cynthia Parr, Steven Phillips, Thiago Rangel, Dan Rosauer, Tim Robinson, Florencia Sangermano, Aaron Steele, Aimee Stewart, Javier de ala Torre, Woody Turner, John Wieczorek. Additionally, we thank Nick Dulvy, Tien Ming Lee, Axel Moehrenschlager, Paul Craze and three anonymous reviewers for comments on the manuscript, and Leigh Anne McConnaughey for crucial help preparing the figures. Andrew Hill has been especially instrumental in development of the Map of Life vision and implementation. The project was also supported by NSF grant DBI 0960550 (Map of Life: An infrastructure for integrating global species distribution knowledge) to WJ and RPG; NSF grant DEB 1026764 to WJ; an NCEAS working group grant to Brian McGill, JMM, WJ and RPG; support by the African Conservation Centre (Nairobi) to WJ, and an NSERC postdoctoral fellowship to JMM.

References

1 Brown, J. and Lomolino, M.V. (1998) *Biogeography*, Sinauer Associates, Inc.
2 Whittaker, R.J. *et al.* (2005) Conservation biogeography: assessment and prospect. *Divers. Distrib.* 11, 3–23
3 Wilson, E.O. (2003) The encyclopedia of life. *Trends Ecol. Evol.* 18, 77–80
4 Graham, C.H. and Hijmans, R.J. (2006) A comparison of methods for mapping species ranges and species richness. *Global Ecol. Biogeogr.* 15, 578–587
5 McPherson, J.M. and Jetz, W. (2007) Type and spatial structure of distribution data and the perceived determinants of geographical gradients in ecology: the species richness of African birds. *Global Ecol. Biogeogr.* 16, 657–667
6 Rondinini, C. *et al.* (2006) Tradeoffs of different types of species occurrence data for use in systematic conservation planning. *Ecol. Lett.* 9, 1136–1145
7 Ferrier, S. (2002) Mapping spatial pattern in biodiversity for regional conservation planning: where to from here? *Syst. Biol.* 51, 331–363
8 Murray, J.V. *et al.* (2009) How useful is expert opinion for predicting the distribution of a species within and beyond the region of expertise? A case study using brush-tailed rock-wallabies *Petrogale penicillata*. *J. Appl. Ecol.* 46, 842–851
9 Smith, C.S. *et al.* (2007) Using a Bayesian belief network to predict suitable habitat of an endangered mammal - The Julia Creek dunnart (*Sminthopsis douglasi*). *Biol. Conserv.* 139, 333–347
10 Williams, N.S.G. *et al.* (2008) A dispersal-constrained habitat suitability model for predicting invasion of alpine vegetation. *Ecol. Appl.* 18, 347–359
11 Kaschner, K. *et al.* (2008) *AquaMaps: predicted range maps for aquatic species.* World Wide Web Electronic Publication, Version 08/2010, (www.aquamaps.org)
12 Hurlbert, A.H. and Jetz, W. (2007) Species richness, hotspots, and the scale dependence of range maps in ecology and conservation. *Proc. Natl. Acad. Sci. U.S.A.* 104, 13384–13389
13 Eastman, J.R. and Sangermano, F. (2007) An automated procedure for the bulk re-processing of species range polygons. In *Proceedings, 7th IALE World Congress*
14 Phillips, S.J. *et al.* (2009) Sample selection bias and presence-only distribution models: implications for background and pseudo-absence data. *Ecol. Appl.* 19, 181–197
15 ESA (2008) *GlobCover Land Cover v2 2008 database*, European Space Agency GlobCover Project, led by MEDIAS-France
16 Hansen, M.C. *et al.* (2003) Global percent tree cover at a spatial resolution of 500 meters: first results of the MODIS vegetation continuous fields algorithm. *Earth Interact.* 7, 1–15
17 Turner, W. *et al.* (2003) Remote sensing for biodiversity science and conservation. *Trends Ecol. Evol.* 18, 306–314
18 Corsi, F. *et al.* (2000) Modeling species distribution with GIS. In *Research Techniques in Animal Ecology: Controversies And Consequences* (Boitani, L. and Fuller, T.K., eds), pp. 389–434, Columbia University Press
19 Jetz, W. *et al.* (2007) Projected impacts of climate and land-use change on the global diversity of birds. *PLoS Biol.* 5, 1211–1219
20 Scott, J.M. *et al.* (1993) Gap analysis: a geographic approach to protection of biological diversity. *Wildl. Monogr.* 3–41
21 CIAT (2004) *CGIAR-CSI SRTM 90m Database: Void-filled seamless SRTM data V2, 2004.* International Centre for Tropical Agriculture (CIAT) (http://srtm.csi.cgiar.org)
22 Nachtergaele, F. *et al.* (2010) *Harmonized World Soil Database*
23 Dettmers, R. *et al.* (2002) Testing habitat-relationship models for forest birds of the southeastern United States. *J. Wildl. Manag.* 66, 417–424
24 Doswald, N. *et al.* (2007) Testing expert groups for a habitat suitability model for the lynx *Lynx lynx* in the Swiss Alps. *Wildl. Biol.* 13, 430–446
25 Johnson, C.J. and Gillingham, M.P. (2004) Mapping uncertainty: sensitivity of wildlife habitat ratings to expert opinion. *J. Appl. Ecol.* 41, 1032–1041
26 Yamada, K. *et al.* (2003) Eliciting and integrating expert knowledge for wildlife habitat modelling. *Ecol. Model.* 165, 251–264
27 La Sorte, F.A. and Jetz, W. (2010) Projected range contractions of montane biodiversity under global warming. *Proc. R. Soc. B* 277, 3401–3410

28 Kaschner, K. *et al.* (2006) Mapping world-wide distributions of marine mammal species using a relative environmental suitability (RES) model. *Mar. Ecol. Prog. Ser.* 316, 285–310
29 Hirzel, A.H. and Le Lay, G. (2008) Habitat suitability modelling and niche theory. *J Appl. Ecol.* 45, 1372–1381
30 Guisan, A. and Thuiller, W. (2005) Predicting species distribution: offering more than simple habitat models. *Ecol. Lett.* 8, 993–1009
31 Austin, M. (2007) Species distribution models and ecological theory: a critical assessment and some possible new approaches. *Ecol. Model.* 200, 1–19
32 Latimer, A.M. *et al.* (2006) Building statistical models to analyze species distributions. *Ecol. Appl.* 16, 33–50
33 Guisan, A. *et al.* (2007) Sensitivity of predictive species distribution models to change in grain size. *Divers. Distrib.* 13, 332–340
34 Belmaker, J. and Jetz, W. (2011) Cross-scale variation in species richness–environment associations. *Global Ecol. Biogeogr.* 20, 464–474
35 Robinson, L.M. *et al.* (2011) Pushing the limits in marine species distribution modelling: lessons from the land present challenges and opportunities. *Global Ecol. Biogeogr.* 20, 789–802
36 Barry, S. and Elith, J. (2006) Error and uncertainty in habitat models. *J. Appl. Ecol.* 43, 413–423
37 Thuiller, W. (2003) BIOMOD – optimizing predictions of species distributions and projecting potential future shifts under global change. *Global Change Biol.* 9, 1353–1363
38 Ferrier, S. *et al.* (2007) Using generalized dissimilarity modelling to analyse and predict patterns of beta diversity in regional biodiversity assessment. *Divers. Distrib.* 13, 252–264
39 Ferrier, S. *et al.* (2004) Mapping more of terrestrial biodiversity for global conservation assessment. *Bioscience* 54, 1101–1109
40 Menke, S.B. *et al.* (2009) Characterizing and predicting species distributions across environments and scales: Argentine ant occurrences in the eye of the beholder. *Global Ecol. Biogeogr.* 18, 50–63
41 Osborne, P.E. *et al.* (2007) Non-stationarity and local approaches to modelling the distributions of wildlife. *Divers. Distrib.* 13, 313–323
42 Hooten, M. *et al.* (2003) Predicting the spatial distribution of ground flora on large domains using a hierarchical Bayesian model. *Landsc. Ecol.* 18, 487–502
43 Gelfand, A.E. *et al.* (2005) Modelling species diversity through species level hierarchical modelling. *J. R. Stat. Soc. C: Appl. Stat.* 54, 1–20
44 Kéry, M. and Royle, J.A. (2008) Hierarchical Bayes estimation of species richness and occupancy in spatially replicated surveys. *J. Appl. Ecol.* 45, 589–598
45 Royle, J.A. and Dorazio, R.M. (2008) *Hierarchical Modeling and Inference in Ecology: The Analysis of Data from Populations, Metapopulations and Communities*, Academic Press
46 Elith, J. *et al.* (2002) Mapping epistemic uncertainties and vague concepts in predictions of species distribution. *Ecol. Model.* 157, 313–329
47 Hamilton, G. *et al.* (2009) Bayesian model averaging for harmful algal bloom prediction. *Ecol. Appl.* 19, 1805–1814
48 Marmion, M. *et al.* (2009) Evaluation of consensus methods in predictive species distribution modelling. *Divers. Distrib.* 15, 59–69
49 Randin, C.F. *et al.* (2006) Are niche-based species distribution models transferable in space? *J. Biogeogr.* 33, 1689–1703
50 Patterson, D.J. *et al.* (2010) Names are key to the big new biology. *Trends Ecol. Evol.* 25, 686–691
51 Liu, C.R. *et al.* (2005) Selecting thresholds of occurrence in the prediction of species distributions. *Ecography* 28, 385–393
52 Nemani, R. *et al.* (2009) Monitoring and forecasting ecosystem dynamics using the Terrestrial Observation and Prediction System (TOPS). *Remote Sens. Environ.* 113, 1497–1509
53 Ferrier, S. (2011) Extracting more value from biodiversityc observations through integrated modeling. *Bioscience* 61, 96–97
54 Tingley, M.W. and Beissinger, S.R. (2009) Detecting range shifts from historical species occurrences: new perspectives on old data. *Trends Ecol. Evol.* 24, 625–633
55 Staff, S. (2011) Challenges and opportunities. *Science* 331, 692–693
56 Reichman, O.J. *et al.* (2011) Challenges and opportunities of open data in ecology. *Science* 331, 703–705
57 Whitlock, M.C. (2011) Data archiving in ecology and evolution: best practices. *Trends Ecol. Evol.* 26, 61–65
58 McDade, L.A. *et al.* (2011) A challenge to biologists to create and embrace a new assessment system for modern professional productivity. *Bioscience* 61, 619–625
59 van Vuuren, D.P. *et al.* (2010) Downscaling socioeconomic and emissions scenarios for global environmental change research: a review. *Wiley Interdiscip. Rev.: Climate Change* 1, 393–404
60 del Hoyo, J. *et al.*, eds (1992–2010) *Handbook of the Birds of the World*, Lynx Editions
61 IUCN (2005) *Global Mammal Assessment.* IUCN, CABS
62 Jetz, W. *et al.* (2008) Ecological correlates and conservation implications of overestimating species geographic ranges. *Conserv. Biol.* 22, 110–119
63 Prentice, I.C. *et al.* (1992) A global biome model based on plant physiology and dominance, soil properties and climate. *J. Biogeogr.* 19, 117–134
64 Hurlbert, A.H. and White, E.P. (2005) Disparity between range map- and survey-based analyses of species richness: patterns, processes and implications. *Ecol. Lett.* 8, 319–327
65 Schulman, L. *et al.* (2007) Analysing botanical collecting effort in Amazonia and correcting for it in species range estimation. *J. Biogeogr.* 34, 1388–1399
66 Reddy, S. and Davalos, L.M. (2003) Geographical sampling bias and its implications for conservation priorities in Africa. *J. Biogeogr.* 30, 1719–1727
67 Soberon, J. (2005) Completeness of databases at different spatial scales, In *DIVERSITAS International Conference Integrating Biodiversity Science for Human Well-being 2005: Biodiversity informatics: Acquisition, analysis, archiving and applications*,
68 Boakes, E.H. *et al.* (2010) Distorted views of biodiversity: spatial and temporal bias in species occurrence data. *PLoS Biol.* 8, e1000385
69 Wieczorek, J. *et al.* (2004) The point-radius method for georeferencing locality descriptions and calculating associated uncertainty. *Int. J. Geogr. Inf. Sci.* 18, 745–767
70 Guisan, A. *et al.* (2006) Making better biogeographical predictions of species' distributions. *Ecology* 43, 386–392
71 Heikkinen, R.K. *et al.* (2007) Biotic interactions improve prediction of boreal bird distributions at macro-scales. *Global Ecol. Biogeogr.* 16, 754–763
72 Jenkins, D.G. *et al.* (2007) Does size matter for dispersal distance? *Global Ecol. Biogeogr.* 16, 415–425
73 Jetz, W. *et al.* (2004) The scaling of animal space use. *Science* 306, 266–268
74 Cooper, N. *et al.* (2011) Phylogenetic conservatism of environmental niches in mammals. *Proc. R. Soc. B* 278, 2384–2391
75 Gu, W. and Swihart, R.K. (2004) Absent or undetected? Effects of non-detection of species occurrence on wildlife-habitat models. *Biol. Conserv.* 116, 195–203
76 MacKenzie, D.I. *et al.* (2003) Estimating site occupancy, colonization, and local extinction when a species is detected imperfectly. *Ecology* 84, 2200–2207
77 Dorazio, R.M. *et al.* (2006) Estimating species richness and accumulation by modeling species occurrence and detectability. *Ecology* 87, 842–854
78 IUCN *et al.* (2004) *Global Amphibian Assessment.* (www.globalamphibians.org)
79 Tittensor, D.P. *et al.* (2010) Global patterns and predictors of marine biodiversity across taxa. *Nature* 466, 1098–1101
80 Ridgely, R.S. *et al.* (2003) *Digital Distribution Maps of the Birds of the Western Hemisphere. Version 1.0.* NatureServe