ECOLOGY LETTERS

Ecology Letters, (2016) **19**: 992–1006 doi: 10.1111/ele.12624

REVIEW AND SYNTHESIS

Multidimensional biases, gaps and uncertainties in global plant occurrence information

Carsten Meyer,[1,2]* Patrick Weigelt[1] and Holger Kreft[1]*

Abstract

Plants are a hyperdiverse clade that plays a key role in maintaining ecological and evolutionary processes as well as human livelihoods. Biases, gaps and uncertainties in plant occurrence information remain a central problem in ecology and conservation, but these limitations remain largely unassessed globally. In this synthesis, we propose a conceptual framework for analysing gaps in information coverage, information uncertainties and biases in these metrics along taxonomic, geographical and temporal dimensions, and apply it to all *c.* 370 000 species of land plants. To this end, we integrated 120 million point-occurrence records with independent databases on plant taxonomy, distributions and conservation status. We find that different data limitations are prevalent in each dimension. Different metrics of information coverage and uncertainty are largely uncorrelated, and reducing taxonomic, spatial or temporal uncertainty by filtering out records would usually come at great costs to coverage. In light of these multidimensional data limitations, we discuss prospects for global plant ecological and biogeographical research, monitoring and conservation and outline critical next steps towards more effective information usage and mobilisation. Our study provides an empirical baseline for evaluating and improving global floristic knowledge, along with a conceptual framework that can be applied to study other hyperdiverse clades.

Keywords

Data bias, data deficiency, data uncertainty, Global Biodiversity Information Facility, Global Strategy for Plant Conservation, herbarium specimens, knowledge gaps, species distributions, survey effort, Wallacean shortfall.

Ecology Letters (2016) **19**: 992–1006

INTRODUCTION

Land plants (subkingdom Embryophyta, hereafter ‘plants’) are a hyperdiverse group of organisms and the principal providers of biochemical energy and habitat structure in most terrestrial ecosystems. Geographical distributions of plant species determine the spatio-temporal setting for evolutionary and ecological processes (Wright & Samways 1998; Kissling *et al.* 2008), and of the ecosystem functions and services upon which most other species, including humans, rely (Isbell *et al.* 2011; Gamfeldt *et al.* 2013). Advances in ecological theory and effective management of natural resources thus rest to a great extent on detailed information about spatio-temporal occurrences of plant species. For instance, improved occurrence information is presupposed by several international policy targets in the framework of the UN Convention on Biological Diversity’s Global Strategy for Plant Conservation (GSPC; www.cbd.int/gspc/targets.shtml; Paton 2009). To date, however, detailed distribution data sets typically required in ecological research and conservation only exist for a few plant groups and geographical regions (Riddle *et al.* 2011), a phenomenon termed the *Wallacean shortfall* (Lomolino 2004).

Most available data sets on plant distributions, including checklists, atlas data and range maps, are ultimately based on point-occurrence records. Such records represent the primary information on the three basic dimensions that characterise species distributions – taxonomy, space and time – as they provide direct evidence that a particular species occurred at a particular location at a particular point in time (Soberón & Peterson 2004). Over the last two decades, millions of digital plant records from herbarium specimens, field observations and other sources have been mobilised via international data-sharing networks, most notably that of the Global Biodiversity Information Facility (GBIF; Edwards 2000). In contrast to un-mobilised datasets or expert knowledge, these mobilised records represent the largest share of information that is both digital and easily accessible in a standard format (hereafter referred to as *digital accessible information* (DAI); originally referred to as *digital accessible knowledge*; Sousa-Baena *et al.*

[1]*Biodiversity, Macroecology and Conservation Biogeography Group, Faculty of Forest Sciences, University of Göttingen, Büsgenweg 1, 37077 Göttingen, Germany*

[2]*Synthesis Centre (sDiv), German Centre for Integrative Biodiversity Research (iDiv) Halle-Jena-Leipzig, Deutscher Platz 5e, 04103 Leipzig, Germany*

**Correspondence. E-mails*: *carsten.meyer@idiv.de, hkreft@uni-goettingen.de*

2014). Recent advances in unifying global plant taxonomic information (*The Plant List*, TPL 2014) now allow integrating thousands of floristic data sources under a common taxonomic framework.

Potential uses of DAI in ecology are manifold (Lavoie 2013), spanning from research on diversity patterns (Morueta-Holme *et al.* 2013), biological invasions (O'Donnell *et al.* 2012) or phenological changes (Calinger *et al.* 2013), to assessments and monitoring of threats (Brummitt *et al.* 2015), and conservation decision-making (Ferrier 2002; Guisan *et al.* 2013). However, broader application is limited by gaps, uncertainties and biases in each of the three basic dimensions taxonomy, space and time (Nelson *et al.* 1990; Soberón & Peterson 2004; Boakes *et al.* 2010; Schmidt-Lebuhn *et al.* 2013).

At least two major aspects of occurrence information directly influence opportunities for inference and application (Fig. 1). One aspect closely connected to the quantity of records is the *coverage* of the three dimensions with information. For instance, *taxonomic coverage,* i.e. how many of the existing species in different assemblages are documented, determines how reliably biodiversity can be compared across sites (Funk *et al.* 1999; Hortal *et al.* 2007). *Geographical coverage*, i.e. how well species' ranges are documented with records, affects the feasibility and reliability of species distribution modelling (Kadmon *et al.* 2003; Feeley & Silman 2011). Finally, high *temporal coverage*, i.e. continuous recording of species through time, is essential for monitoring species' responses to environmental change (Brummitt *et al.* 2015).

A second, more qualitative aspect of occurrence information is *uncertainty* regarding the taxonomic, geographical and temporal information that makes up occurrence records (Fig. 1). *Uncertainties* in DAI may have various sources related to precision, accuracy, ambiguity, credibility or age of information. For instance, ambiguous scientific names entail *uncertainty* regarding taxonomic identities (Jansen & Dengler 2010), imprecise sampling locations entail *uncertainty* regarding the environmental context in which species were found (Rocchini *et al.* 2011), and early sampling dates entail *uncertainty* regarding their continuing presence at those locations (Boitani *et al.* 2011).

Both gaps in information *coverage* and information *uncertainties* may be biased in the taxonomic, geographical and temporal dimensions (Fig. 1), potentially leading to biased ecological inferences (Prendergast *et al.* 1983; Hortal *et al.* 2008) and inefficient conservation (Grand *et al.* 2007). For instance, *taxonomic coverage* of plant assemblages may be geographically biased to certain regions (Yang *et al.* 2013; Sousa-Baena *et al.* 2014), and *geographical uncertainty* may be temporally biased towards older records (Murphey *et al.* 2004). Other types of ecologically relevant data bias are typically closely connected to the three basic dimensions, e.g. phylogenetic or functional biases (Schmidt-Lebuhn *et al.*

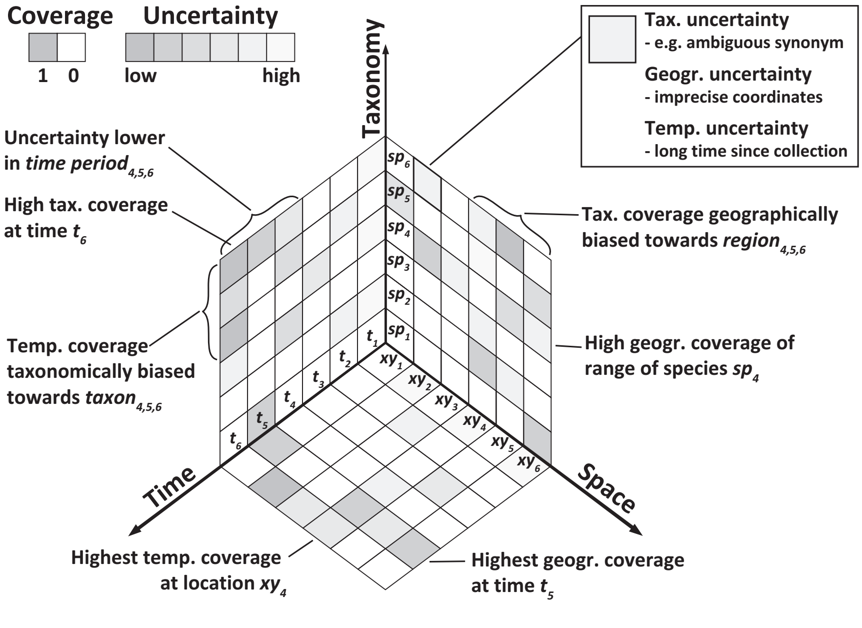
Coverage
Uncertainty
1 0
low high
Taxonomy
Tax. uncertainty
- e.g. ambiguous synonym
Geogr. uncertainty
- imprecise coordinates
Temp. uncertainty
- long time since collection
Uncertainty lower in *time period*$_{4,5,6}$
High tax. coverage at time t_6
Tax. coverage geographically biased towards *region*$_{4,5,6}$
sp_6
sp_5
sp_4
sp_3
sp_2
sp_1
Temp. coverage taxonomically biased towards *taxon*$_{4,5,6}$
High geogr. coverage of range of species sp_4
t_1
t_2
t_3
t_4
t_5
t_6
xy_1
xy_2
xy_3
xy_4
xy_5
xy_6
Time
Space
Highest temp. coverage at location xy_4
Highest geogr. coverage at time t_5

Figure 1 Framework for analysing limitations in occurrence information along taxonomic, geographical and temporal dimensions. Occurrence records cover different species ($sp_1,$ sp_2, ...), different locations (xy_1, xy_2, ...) and different points in time (t_1, t_2, ...). Planes of cells illustrate spread of information between pairs of dimensions, information from anywhere along the third dimension is vertically projected onto the plane. Applicability of occurrence information depends on: (1) *coverage* of the three dimensions with information (grey cells), and (2) *uncertainty* regarding the interpretation of information on the three dimensions (shade of grey cells). Thus, a cell may be *covered*, but *uncertainty* of the records may vary from low (dark grey) to high (light grey); white cells indicate no *coverage* with available information. Integrating across cells in one dimension summarises information per unit of the other dimension (e.g. bottom right: highest *geographical coverage* at time t_5 because four out of six locations covered). *Coverage* and *uncertainty* may be biased in each dimension (curly brackets; e.g. centre left: *temporal coverage* taxonomically biased because species of *taxon*$_{4,5,6}$ have systematically higher *coverage*, compared to *taxon*$_{1,2,3}$).

2013) to taxonomy, environmental bias (Funk *et al.* 2005) to space and seasonal bias (ter Steege & Persaud 1991) to time.

Understanding magnitude and biases in different metrics of *coverage* and *uncertainty* of DAI with regard to the three dimensions is crucial for evaluating prospects for research and other applications, and for prioritising and monitoring activities to improve DAI (Meyer *et al.* 2015; Peterson *et al.* 2015). Identifying botanical information gaps has a long history (Jäger 1976; Prance 1977; Kier *et al.* 2005), whereas most recent analyses emphasised effects on specific ecological applications of DAI (Feeley & Silman 2011; Yang *et al.* 2013; Maldonado *et al.* 2015). Despite the need to evaluate the multiple limitations in global DAI (Ladle & Hortal 2013), a quantitative assessment for the world's plants is lacking.

Here, we provide such an assessment for all land plants, by integrating 120 million point-occurrence records facilitated via GBIF with comprehensive taxonomic databases, the World Checklist of Selected Plant Families, and the IUCN Global Red List. We examine DAI for gaps in *coverage*, data *uncertainties* and their biased variation along taxonomic, geographical and temporal dimensions, investigate pairwise and multivariate relationships between different metrics of *coverage* and *uncertainty*, and characterise geographical regions in terms of their multivariate data limitations. In light of these different limitations, we highlight how typical applications of DAI might be affected, and discuss prospects for using plant DAI in global ecological research, conservation and monitoring, with particular emphasis on GSPC targets. Finally, we outline next steps towards more effective information usage and mobilisation. Our work provides the first quantitative global synthesis of strengths and weaknesses in DAI for a hyperdiverse taxonomic group, and conceptual and empirical baselines for studying and addressing data limitations in future research and data mobilisation efforts.

METHODS

Data sources

We downloaded all records for land plants available via GBIF in January 2014 (*c.* 120 million). These records were contributed to the GBIF network by 238 data publishers in 48 countries (Table S1). The majority of these records (78%) came from field observations (e.g. from vegetation plot data) and from preserved herbarium specimens (17%). GBIF-facilitated records represent by far the largest source of DAI, and a substantial part of the digitised portion of the estimated 350 million records that exist in the World's herbaria (New York Botanical Garden 2014). Geographical gaps in global *coverage* of these records may represent genuinely undersampled regions, but also regions whose information is not yet digitised or integrated into international data-sharing networks (Meyer *et al.* 2015), such as Brazil or China (see Sousa-Baena *et al.* (2014) and Yang *et al.* (2013), respectively, for limitations in those regional databases).

We taxonomically standardised and validated verbatim scientific names, using comprehensive taxonomic information provided via *The Plant List* (TPL 2014) and iPlant's *Taxonomic Name Resolution Service* (TNRS 2014). We applied taxonomic and geographical filters (see section *Uncertainty* below) and excluded duplicate combinations of accepted species, sampling location and year-month combination (see Fig. S1 for an overview of our workflow, see *Supplementary Information (SI) 1* for details on standardisation and filtering). These steps led to a reduction of 119 058 280 raw records with 2 206 831 verbatim name strings to 55 929 317 unique point-occurrence records for 229 218 accepted species from 3 947 969 unique sampling locations and 3172 year-month combinations (*SI.1.1*).

DAI includes both erroneous and non-native species' records (Soberón & Peterson 2004), however, independent baseline information for validation (e.g. on species' native ranges) is lacking for most plants (Box 1). Therefore, we validated 16.8 million records for 105 031 species of seed plants (Spermatophyta; 34% of all plant species) against checklists for 'botanical countries' (level-3 regions of Biodiversity Information Standards, formerly Taxonomic Database Working Group – TDWG; www.tdwg.org/standards/109/), derived from the World Checklist of Selected Plant Families (WCSP, 2013). We determined the global conservation status of species using the International Union for Conservation of Nature's Global Red List (IUCN 2014).

A framework for assessing multidimensional data limitations

We developed a conceptual framework for assessing limitations in occurrence information along the three basic dimensions that define primary biodiversity data and characterise species distributions – taxonomy, space and time (Soberón & Peterson 2004). The framework consists of the quantification of the fundamental aspects of information *coverage* and *uncertainty* with regard to the three basic dimensions, and the assessment of variation and biases of each of the six resulting information metrics along each of these dimensions (Fig. 1). We followed this framework, by quantifying the different information metrics and assessing their variation across taxonomic, geographical and temporal units (see below). Note that while our framework can guide future assessments, our choice of indices to quantify the different information metrics is not prescriptive; depending on the goals of assessments, individual metrics may warrant more detailed analysis or quantification through other indices (see Box 1 and *SI 2* for challenges and limitations).

Coverage

We computed three metrics to estimate *taxonomic*, *geographical* and *temporal coverage*, i.e. the extent to which available records cover the three basic dimensions (Fig. 1). We estimated *taxonomic coverage* of 12 100 km² equal area grid cells (110 km × 110 km at the equator) as the ratio between recorded vascular plant richness and an estimate of actual richness (the co-kriging richness model of Kreft & Jetz (2007)). Similar metrics have been variously termed census-, inventory- or survey completeness (Colwell & Coddington 1994), but we use *taxonomic coverage* here for consistency with the general framework (Fig. 1). A general problem in assessing data *coverage* is that the very data limitations under study often preclude reliable baselines against which data gaps could be tested (see

Box 1 Inaccurate information, missing baselines and alien species

A number of fundamental issues in point-occurrence information compromise the study of data limitations as well as applications of DAI in research and conservation. A detailed assessment of these issues is beyond the scope of this study, but anyone using DAI should carefully consider these potential sources of error.

Information Inaccuracies: A potentially huge problem that is particularly difficult to address is information that is taxonomically or geographically inaccurate. It is largely unknown how often species were misidentified (Scott & Hallam 2002), how often direction and distance to known reference points were mismeasured or coordinates incorrectly recorded from GPS receivers (Murphey *et al.* 2004), and how often originally accurate information was subsequently rendered inaccurate during data curation, digitisation or mobilisation. Case studies found species misidentification rates between < 1 and 17% (Bisang & Urmi 1994; Scott & Hallam 2002; Ahrends *et al.* 2011). For large databases, taxonomic inaccuracies may be impossible to detect or reliably estimate without extremely labour-intensive reassessments of the original material and sufficiently rich metadata (e.g. on the experience of the identifying person), which, however, do not exist for most datasets. Similarly, only the most obvious geographical inaccuracies are likely to be detected. For example, apparent peaks in *taxonomic coverage* near country centroids likely reflect cases where indicated countries were later inaccurately geo-referenced to precise point localities (e.g. in Brazil; Fig 2b; compare Murphey *et al.* 2004; Maldonado *et al.* 2015). As another example, data-housing institutions or botanical gardens are sometimes inaccurately reported as 'natural' sampling locations, as seen in undated collections of hundreds of non-European species provided by the Bergius Herbarium (note the *taxonomic coverage* peak of 6.6 around Stockholm; Fig. 2b).

Missing Baselines: A further key problem for analysing data limitations is that the very subject of analysis often precludes reliable baselines against which limitations can be tested. For instance, our metric of *taxonomic coverage* will underestimate gaps for any localised plant diversity centre that may not be well-represented by the underlying plant-richness model. Species richness estimators (e.g. Chao & Jost 2012) are in turn highly sensitive to low record numbers and to non-natural relative abundances of species in natural history collections (ter Steege *et al.* 2011; *SI 2*, Fig. S2). Similarly, due to prevalent gaps and biases, it is rarely possible to reliably assess the *geographical coverage* of species' ranges, as meaningful distribution estimates for comparison are neither available nor feasible for the majority of species. Avoiding this problem by restricting assessments to well-known study systems naturally entails trade-offs for detail and comprehensiveness. Moreover, even the most authoritative baseline information is ultimately based on primary biodiversity records, and may thus change as new evidence becomes available (*SI 2*; Fig. S4a–b).

Alien Species: Finally, DAI includes an unknown quantity of valid occurrence records of species outside their native ranges. These records can play a crucial role in informing about the first occurrence and subsequent spatio-temporal spread of aliens in their invaded ranges, and thus facilitate the study and management of plant invasions (GSPC target 10; Broennimann *et al.* 2007; van Kleunen *et al.* 2015). However, if their alien status is undetected, these records can bias inferences concerning native biota, such as about the number of records usable for distribution estimations (Fig. 2d), or about the completeness of native plant inventories (note the higher-than-expected recorded richness in 3.6% of cells; Fig. 2b; also see Fig. S4c).

Box 1). In this respect, we acknowledge possible uncertainties in our index arising from the use of the plant-richness model, as well as continued debate among experts on the best ways for assessing *taxonomic coverage* (see *SI 2*). To address this problem, we compared our index against two alternative indices of *taxonomic coverage* that are estimated from the records themselves (*SI 2*, Fig S2) as well as a more robust, but less detailed and comprehensive, index derived from species checklists for selected plant families in botanical countries (WCSP 2013; Fig S4). As a complementary measure that emphasises the magnitude of total rather than proportional gaps in *taxonomic coverage*, we assessed the number of vascular plant species that were not recorded but expected to occur in a grid cell based on the richness model of Kreft & Jetz (2007).

To estimate *geographical coverage* of species' ranges and grid cells, respectively, we used the quantity of unique sampling locations per species and per grid cell land area (see Box 1 for limitations of this approach). To measure *temporal coverage* of species and of grid cells, we calculated the negative mean minimum time (in years) from all months between 1750 and 2010 to their respective temporally closest records available for that species or cell, respectively. This metric indicates the number of years that typically lie between a given point in time and the temporally nearest date when records were collected; it has large negative values if *temporal coverage* is low, i.e. if the entire time span contains large temporal gaps without any records. As a complementary measure that emphasises the magnitude of the most recent data gap, we assessed the time since the last records were collected for a given species or grid cell. We analysed temporal patterns of *taxonomic* and *geographical coverage* by comparing percentages of species and grid cells covered within, and cumulatively up to, 5-year periods.

Uncertainty

Uncertainty in DAI (Fig. 1) has multiple sources, including imprecise or inaccurate information (Murphey *et al.* 2004; Rocchini *et al.* 2011), uncertain status of species or ambiguous synonymy (Berendsohn 1995), and decay of information in space and over time (Ladle & Hortal 2013). Here, we focused on sources of *uncertainty* in reported information that can be readily assessed using available databases and tools (for data inaccuracies such as species misidentifications, which are not assessed here, see Box 1). To investigate *uncertainty* in DAI, we imposed increasingly stringent filters on records (basic, moderate and strict; described below) and investigated the impacts of these filters on numbers of retained records or species.

We defined three *taxonomic uncertainty* filters based on expert confidence, precision and ambiguity of scientific names, which we assessed during our taxonomic validation procedure (see *SI 1*):

(1) TaxStrict: Recorded name matches a species that TPL considers accepted with high expert confidence (three 'stars'; see www.theplantlist.org/about), with ≤ 5% orthographic distance (the number of changes that have to be applied to one string to match another; see *SI 1*), either directly or through an unambiguous synonym (one that only links to one accepted species);
(2) TaxModerate: Recorded name matches a species that TPL considers accepted with high or medium expert confidence (two or three 'stars') with ≤ 15% orthographic distance, either directly or through an unambiguous or ambiguous synonym;
(3) TaxBasic: Recorded name matches a species that TPL or TNRS considers accepted (no criteria for expert confidence in TPL) with ≤ 25% orthographic distance, either directly or through an unambiguous or ambiguous synonym. This basic filter was always applied before other analyses.

We defined three *geographical uncertainty* filters, based on precision of coordinates and internal consistency with the indicated country (for inaccuracies such as measurement errors, see Box 1):

(1) GeoStrict: Location reported with a precision of at least 1/1000 of a degree (~ 100 m at the equator);
(2) GeoModerate: Location reported with a precision of at least 1/100 of a degree;
(3) GeoBasic: Location reported with a precision of at least 1/10 of a degree and falling within the indicated country. This filter was always applied before other analyses.

We defined three *temporal uncertainty* filters based on the principle that information quality decays with time (Boitani *et al.* 2011; Ladle & Hortal 2013):

(1) TempStrict: Records collected after 1990;
(2) TempModerate: Records collected after 1970;
(3) TempBasic: Records collected after 1950.

Unless stated otherwise, we hereafter refer to a dataset to which basic taxonomic and geographical filters, but no temporal filter, were applied. We investigated patterns in *taxonomic* and *geographical uncertainty* by comparing across species and grid cells the percentages of excluded records when additionally applying moderate or strict *taxonomic* and *geographical uncertainty* filters respectively. We investigated patterns in *temporal uncertainty* by comparing percentages of excluded species when additionally applying moderate or strict *temporal uncertainty* filters. Similarly, we investigated patterns in combined *uncertainty* by comparing percentages of additionally excluded species if all three moderate or strict filters (taxonomic, geographical and temporal) were applied.

Variation in occurrence information

To quantify and visualise taxonomic, geographical and temporal variation and biases in information *coverage* and *uncertainty*, we compared the respective metrics among major plant groups (bryophytes, pteridophytes, gymnosperms and angiosperms), geographical units (12 100 km² grid cells and TDWG level-3 'botanical countries') and 5-year periods. We defined bias as the non-random distribution of magnitude and/or prevalence of any data limitation (i.e. any metric of *coverage* or *uncertainty*) along a specified dimension. This definition corresponds to systematic error (Walther & Moore 2005), such as *unrecorded species* (i.e. low *taxonomic coverage*) or *imprecise coordinates* (i.e. high *geographical uncertainty*). For instance, we speak of 'geographical bias in *taxonomic coverage*' if gaps in species inventories are non-randomly distributed across spatial grid cells.

To test for taxonomic bias in information metrics, we compared mean species values of our information metrics among major plant groups using Tukey contrasts (Herberich *et al.* 2010). We quantified spatial bias using Moran's *I*, a commonly used spatial autocorrelation measure (Legendre & Legendre 2012), and temporal bias using lag-*k* correlations (ACF_{max}; Legendre & Legendre 2012). We assessed spatial and temporal autocorrelation over five spatial and temporal distance classes, respectively, and report the highest values. Spatial and temporal autocorrelation measures are not directly comparable, but in both cases, values approaching 0 mean that the information metric is even or random (i.e. unbiased) with regard to that dimension; values approaching 1 indicate highly biased information.

Relationships among information metrics

We investigated relationships between geographical patterns of nine different information metrics, including the three dimensions of *coverage* and *uncertainty,* combined *uncertainty* (see above; *uncertainty* measured here as information loss under moderate filtering), the number of missing vascular plant species, and the time since the last record was collected. We analysed pairwise and multivariate relationships between these nine metrics using pairwise Spearman rank correlations and principal component analysis (PCA) which reduces co-linear metrics to orthogonal principal components. We assigned red, green and blue components of the RGB colour space to the grid cells according to their positions in the three-dimensional space formed by the first three PCA axes (Weigelt *et al.* 2013). We then mapped these coloured grid cells to visualise which regions are characterised by the different aspects and dimensions of occurrence information. *P*-values for correlations between spatial patterns were adjusted to geographically effective degrees of freedom following Dutilleul (1993).

We assessed trade-offs between information *coverage* and *certainty* of DAI, and their implications for potential ecological and conservation applications, by counting species that would meet minimum data requirements of hypothetical distribution estimation methods (10–200 records; Kadmon *et al.* 2003; Rivers *et al.* 2011), if all three basic, moderate or strict *uncertainty* filters were applied. We performed these analyses globally and for TDWG level-1 continents, assigning species to continents if ≥ 80% of their records fell within their respective boundaries. All analyses were carried out in R versions 3.0.2–3.2.1 (R Core Team 2014).

RESULTS AND DISCUSSION

The large volume of plant records in global DAI (119 million; Fig. S1a) may misguide perceptions of the actual available information on plant occurrences. Our basic validation and filtering steps excluded 38.2 million records, including 12.5 million with non-validatable verbatim name strings (Fig. S1g, *SI 1*) and 27.9 million in the sea (Fig. S1c). Collecting duplicate specimens from the same plant individual is common practice in botany, and removing duplicated species-location-month combinations excluded a further 25 million records, leaving 56 million unique records for analyses (47% of all). The record number per species varied by five orders of magnitude, and by six orders of magnitude across grid cells (Fig. S1b). For instance, a single 12 100 km² cell in the Netherlands that is home to 38 data-contributing institutions and one of the world's largest vegetation plot datasets had 2.8 million records, whereas 21% of all cells had no records. All metrics assessed were severely biased in at least one of the three dimensions.

Coverage of the different dimensions

Taxonomic coverage

Globally, DAI on plant occurrences showed tremendous gaps in *taxonomic coverage*, with only about two-thirds of all plant species covered with at least one record that passed our basic filtering (229 218 out of 350 697 species accepted by TPL as of 2014). *Taxonomic coverage* was itself taxonomically biased, with 83% of pteridophytes but only 28% of bryophyte species represented (Fig. 2a). For most regions, *taxonomic coverage* of species assemblages was extremely low: 79% of cells had < 25% of species covered. Spatial autocorrelation furthermore demonstrated considerable geographical bias (Moran's $I = 0.60$, $P = 0$; Fig. 2b). These gaps and biases seriously impair important applications, from basic ecological studies (Yang *et al.* 2013) to site-based plant conservation prioritisation (GSPC target 5; Funk *et al.* 1999).

Recorded species richness was an almost perfect function of record number ($r_S = 0.94$, $P_{Dut} = 0$; Fig. S1b/f/k), demonstrating that centres of plant diversity perceived from occurrence records often reflect better documentation rather than true diversity patterns (Hortal *et al.* 2007; Yang *et al.* 2013; also compare *SI 2*, Fig S2c). Although absolute numbers of unrecorded species were highest in the tropics (e.g. Eastern

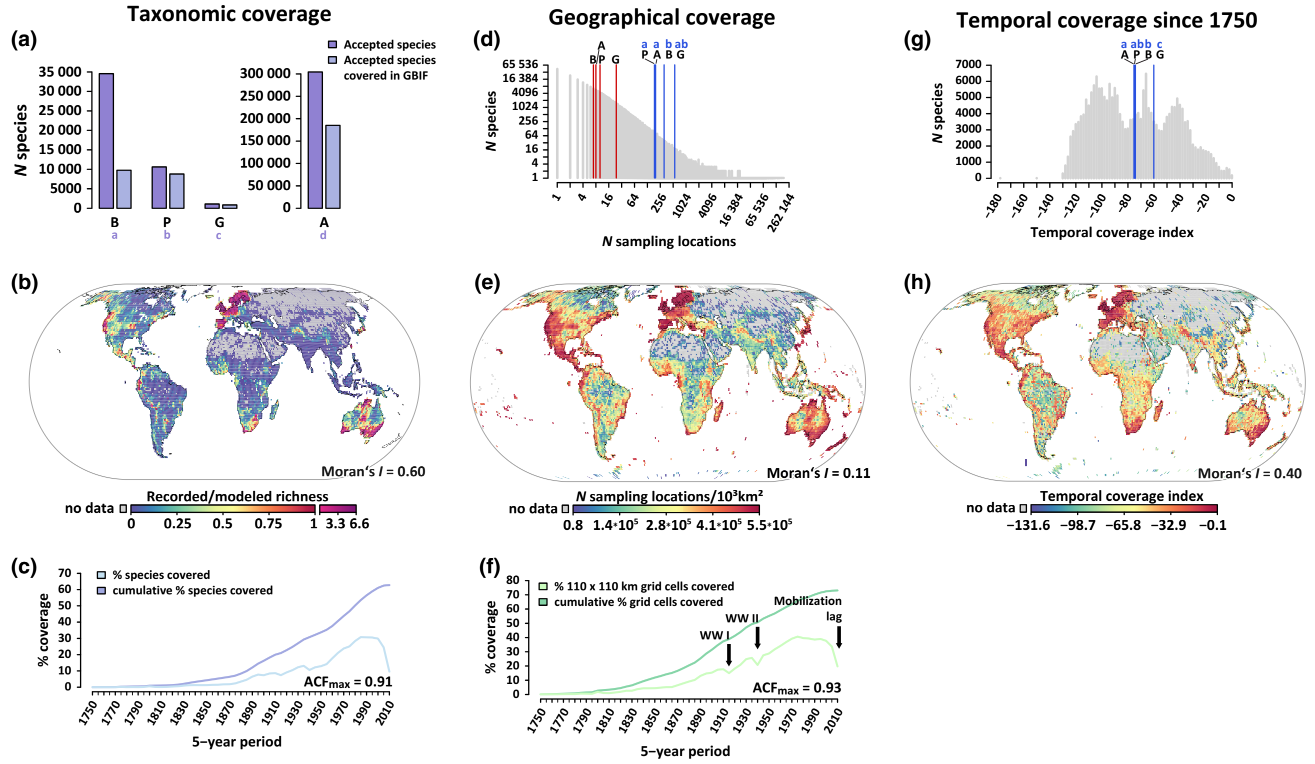

Figure 2 Global variation in occurrence information *coverage*. (a) *Taxonomic coverage* of major plant groups (accepted (TPL 2014) vs. recorded species; B – bryophytes, P – pteridophytes, G – gymnosperms, A – angiosperms); (b) Geographical variation across 12 100 km² grid cells of *taxonomic coverage* for vascular plants (recorded/modelled richness; Kreft & Jetz 2007); values > 1 indicate higher recorded than modelled richness; (c) Percentages of species *covered* within, and up to, 5-year periods since 1750; *Geographical coverage* of (d) species (*N* sampling locations) and (e) cells (*N* locations/10^4 km² land area); (f) Percentages of cells *covered* within, and up to, 5-year periods since 1750. *Temporal coverage* 1750–2010 of (g) species and (h) cells; small negative values denote high *coverage*. Blue/red bars in d/g: means/medians for major plant groups. In c/f, note dips during the World Wars and drop since the 1990s (possibly a time lag between record collection and mobilisation). Significant taxonomic/spatial/temporal biases indicated by lowercase letters/Moran's I/ACF_{max} (see Methods).

Amazonia, Borneo; Fig. S3a), patterns of proportional *taxonomic coverage* did not confirm previous observations of a 'tropical data gap' (Collen *et al.* 2008; $P_{\mathrm{Dut}} = 0.37$), nor of higher data gaps in Neo- than in Palaeotropical areas (Prance 1977; $P_{\mathrm{Dut}} = 0.64$). Instead, severe gaps emerged across most of Asia, Northern and Central Africa, Amazonia and Arctic Canada (Fig. 2a). Despite limitations in our index of *taxonomic coverage* (and in indices of other information metrics; see Box 1), these broad-scale patterns are largely robust against alternative indices (*SI 2*, Fig S2), and provide an important first step in identifying priority regions for improving botanical baseline information (GSPC target 3; Sousa-Baena *et al.* 2014).

Geographical coverage

One of the most prominent applications of point-occurrence records in ecology is the estimation of species distributions. However, available records typically *covered* individual plant species at only seven unique locations (median across species with ≥ 1 record; Fig. 2d), too few to construct meaningful species distribution models (Guisan *et al.* 2007; Feeley & Silman 2011) or extent-of-occurrence range maps (Gaston & Fuller 2009; Rivers *et al.* 2011). Our *geographical coverage* index showcased a significant taxonomic bias (Fig 2d), as well as a geographical bias, mainly towards well-studied North America, Western Europe and Australia (Fig. 2e). Outside those regions, high *geographical coverage* often appeared associated with specific botanical interest and major research and data mobilisation programs. For instance, Madagascar has exceptional plant diversity and endemism (> 11 000 species, 82% endemic; Callmander 2011). Missouri Botanical Garden has long focused on the botanical exploration of Madagascar (Raven & Axelrod 1974), was one of the first institutions to engage in data mobilisation (Crosby & Magill 1988), and as a consequence now contributes 66% of Madagascan records.

Temporal coverage

Continuous *temporal coverage* of species and regions is necessary for monitoring changes in biodiversity (Boakes *et al.* 2010) and to provide historical baselines (Willis *et al.* 2007). Given the general paucity of long-term datasets in ecology, identifying continuities in existing DAI may uncover vantage points for future monitoring activities (Johnson *et al.* 2011). Most species had extremely low *temporal coverage* since 1750, with a given point in time typically decades away from the nearest record (median: 77.3 years; Fig. 2g). In addition, *temporal coverage* was geographically highly biased (Moran's $I = 0.40$, $P = 0$), with less than 2 months typically lying between a given point in time and the closest sampling date in the best-*covered* cell in eastern England, in contrast to 73 years for Amazonia and Asia (medians; Fig. 2h). For many global change questions, such as monitoring of pole-ward range expansions or land-use driven range contractions (Feeley 2012), *temporal coverage* specifically of recent decades may be most relevant and *coverage* since 1950 was indeed higher (Fig. S3b–c). Worryingly, however, several tropical and high arctic regions undergoing very rapid land-cover or climate change (Burrows *et al.* 2011; Hansen *et al.* 2013) were characterised both by poor *temporal coverage* and ageing records, notably in Canada, central Africa and Asia (Fig. S3c–d). For instance the last record in a given Angolan grid cell was typically collected 36 years ago (median, measured from 2010).

Temporal variation in coverage

Globally, *coverage* of species and grid cells mostly increased through time, apart from dips during the World Wars (Fig. 2c/f). *Geographical coverage* appears to have levelled off since the 1970s and *taxonomic coverage* since the 1980s, whereas cumulative *coverage* continued to increase at lower rates (Fig. 3e–f). The steep drops in global *coverage* since the mid-1990s may partly reflect time lags between field collection and mobilisation of records (Gaiji *et al.* 2013), but also decreasing survey effort (Prather *et al.* 2004). The latter would be alarming, as new, up-to-date records are crucial both for studying recent environmental change and for securing the data foundations of botanical research in coming decades (Johnson *et al.* 2011). These general trends in global *coverage* hide strong spatio-temporal variation in certain regions (Fig. S5).

Uncertainty regarding the interpretation of information

Uncertainties in point-occurrence information increase the likelihood that available records are misinterpreted, such as when ambiguous synonyms link records to false accepted species, or when imprecise coordinates link them to the false environmental conditions. Compared to mere gaps in information *coverage*, misinterpretations are even doubly harmful, as false records are added and true records are ignored. Unlike aspects of *coverage* (e.g. Yang *et al.* 2013; Sousa-Baena *et al.* 2014; Meyer *et al.* 2015), *uncertainties* in DAI have received relatively little attention (e.g. Feeley & Silman 2010; Ahrends *et al.* 2011).

Taxonomic uncertainty

Taxonomic uncertainty regarding interpretations of scientific names can arise from missing clarity on whether names are accepted or synonyms, from ambiguous synonyms linked to several accepted names, or from orthographic variations and spelling mistakes (Berendsohn 1995; Jansen & Dengler 2010; see Box 1 for additional *uncertainties* due to species misidentifications). We found extremely high rates of *taxonomically uncertain* information in global plant DAI, as showcased by the 67% of information that was lost when applying our strict taxonomic filter.

The proportion of records with taxonomically *uncertain* information was taxonomically and geographically highly biased. For instance, pteridophytes disproportionately lost records under moderate filtering (Fig. 3i), possibly reflecting continuing major changes in fern taxonomy (Christenhusz & Chase 2014). Furthermore, different degrees of *uncertainty* showed very different geographical biases: Depending on the strictness of taxonomic filtering, geographical peaks in lost information appeared either in insular South-East Asia (moderate filter, Fig. 3a) or in Europe and North America (strict filter, Fig. 3b). High *taxonomic uncertainty* for the latter regions might appear counterintuitive, given their long history

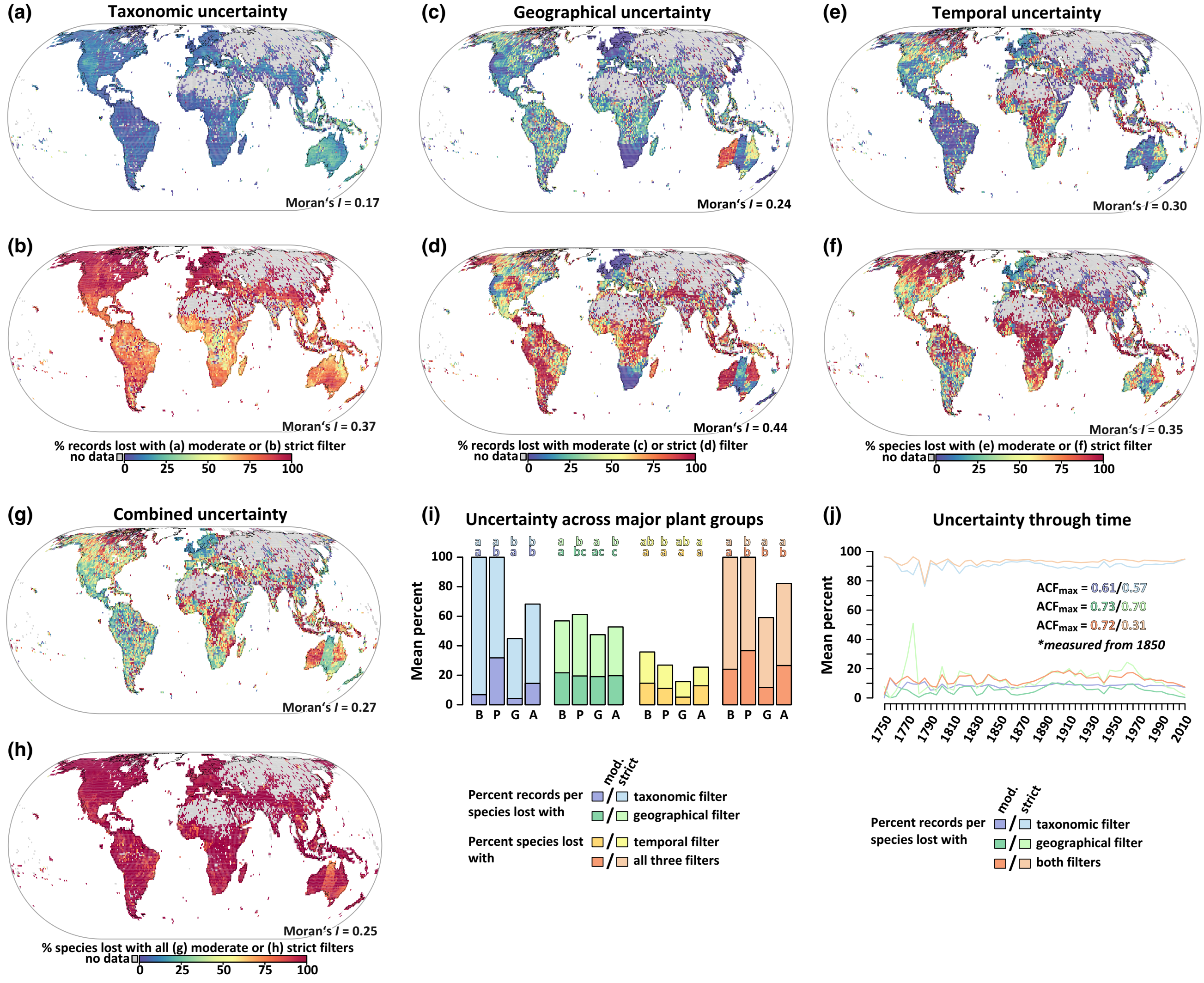

Figure 3 Global variation in occurrence information *uncertainty*. Geographical patterns across 12 100 km² grid cells of percentages of records excluded by (a) moderate and (b) strict *taxonomic uncertainty* filtering, by (c) moderate and (d) strict *geographical uncertainty* filtering; geographical patterns of percentages of species excluded by (e) moderate and (f) strict *temporal uncertainty* filtering; by applying all three (g) moderate and (h) strict filters. (i) Taxonomic patterns across major plant groups (B – bryophytes, P – pteridophytes, G – gymnosperms, A – angiosperms) of mean percentages of records per species excluded by taxonomic and geographical filters, and of species entirely excluded by temporal and combined filters; (j) Temporal patterns across 5-year periods between 1750 and 2010 of percentages of records excluded by taxonomic, geographical and the two combined filters. Significant taxonomic/spatial/temporal biases indicated by lowercase letters/Moran's I/ACF_{max} (see Methods).

of intensive taxonomic work. However, precisely this work has resulted in many taxonomic revisions, which, despite their obvious benefits, increased *taxonomic uncertainties*, as intended taxonomic delimitations of ambiguous species names usually cannot be inferred for most of the disparate data sources that make up DAI. Potentially, any DAI-based application for species with ambiguous names (e.g. that were historically used for broader species concepts) could thus be biased, as inferences may be partly based on records for related species that were historically lumped together (Berendsohn 1995).

Geographical uncertainty

Imprecisely geo-referenced sampling locations lead to *uncertainty* regarding the geographical and environmental context of species' occurrences. Such *geographical uncertainty* is widespread in global DAI: Applying our basic geographical filter already lead to a 38% loss in accepted species from our data set, confirming a strong trade-off between geographical precision and *taxonomic coverage* of information (Feeley & Silman 2010).

The prevalence of *geographical uncertainties* was highly biased in space (Fig 3c–d, Moran's $I = 0.24–0.44$, $P = 0$) and time (Fig 3j, $ACF_{max} = 0.70–0.73$). Imprecisely geo-referenced locations were most prevalent in tropical and remote non-tropical regions (e.g. Alaska, temperate Asia, Western Australia; Fig. 3d), likely due to lower quality maps and more sparsely distributed settlements, which frequently serve as geographical references during surveys. Analogously, *geographical uncertainty* increased during two major periods that saw intensive explorations of tropical and remote regions, first during the second wave of European colonial expansion, 1860–1910, and again between 1940 and 1965 (Fig. 3j; Fig. S5b–d; Fig. 3c–d). The subsequent decrease in *geographical uncertainty* may reflect the increasing availability of better maps and, later, GPS technology. However, *geographical uncertainty* can also be generated during data mobilisation. For instance, patterns in Australia closely mirrored administrative boundaries, reflecting different mobilisation policies of Australian state departments, which contributed 54% of Australian records (Fig. 3c). At the time of downloading our records, several Australian datasets were mobilised into the GBIF network via intermediaries that deliberately generalised location coordinates of any potentially sensitive information. Mobilisation pathways since changed and generalisations are now restricted to much lower percentages of Australian records (e.g. of species threatened by illegal collecting; Klazenga & Vaughan 2014).

The overall high levels of *geographical uncertainty* severely compromise applications like distribution modelling, that rely on linking species occurrences with fine-scale environmental data for extrapolation (Feeley & Silman 2010; Rocchini *et al.* 2011). This problem is aggravated in environmentally heterogeneous regions, where even slight errors would substantially alter the perceived environmental associations of species (Feeley & Silman 2010; e.g. note the high *uncertainty* in the tropical Andes; Fig. 3c–d). In this context, we shall stress that even precisely reported information cannot assure that locations are in fact accurate (see discussion in Box 1). Nevertheless, such precision-based assessments can offer a starting point for focusing additional geo-referencing activities.

Temporal uncertainty

Despite the importance of early collected records for informing about past biota, old records also inherit greater *temporal uncertainty* regarding the continuing presence of species at or near sampling locations (Soberón & Peterson 2004; Boitani *et al.* 2011), as distributions may have responded to land-use and climatic changes (Thuiller *et al.* 2008) or biological processes (Schurr *et al.* 2012). Therefore, important applications like conservation planning or distribution modelling, that link DAI with modern habitat data, usually require similarly modern occurrence information (Boitani *et al.* 2011). Sufficiently modern information for such applications was, however, extremely scarce for most taxa and regions. On average, 62% of species in a given grid cell had no record collected after 1990, 32% even had no record from after 1970 (Fig. 3e–f). Particularly high *uncertainty* levels emerged for much of Arctic Canada, central Africa, Iraq, eastern India, Myanmar and Java (Fig. 3e). Apart from continuously well-sampled areas like north-western Europe, only regions that *only* saw intensive surveying during recent decades appeared as having generally low *temporal uncertainty* (e.g. Benin, Indochina, the circum-Tibetan mountains; Fig. 3f, Fig. S5f).

Combined uncertainty

Nearly all plant records are subject to some form of data *uncertainty* (Fig. 3g). Thus, minimising *uncertainty* in all three dimensions by combining taxonomic, geographical and temporal filters would lead to substantial trade-offs for *coverage* (compare Feeley & Silman 2010; Boitani *et al.* 2011). Of all species in our dataset, 79% had no record that passed all strict filters; 52% even had no record passing all moderate filters. North-western Europe was the only larger region where typically ≥ 80% of species in a given grid cell had at least one record that passed moderate combined filters (Fig. 3g). No region retained much of available information under strict combined filtering; even regions where 20% of recorded species would withstand such filters were confined to Benin, Indochina and central and south-eastern Australia (Fig. 3h).

Given these pervasive levels of data *uncertainty*, it is highly likely that species identities and their environmental associations are frequently misinterpreted in ecological studies (Feeley & Silman 2010; Jansen & Dengler 2010). Furthermore, our documented patterns of *uncertainty* demonstrate that the likelihood of such misinterpretations is biased to particular taxonomic groups, geographical regions and time periods. Overall, these issues seriously hamper opportunities for ecological inference and application, and need to be carefully accounted for whenever records of variable or unknown quality are used in biodiversity analyses (Rocchini *et al.* 2011). Furthermore, results from past studies should be rigorously scrutinised and ecological insights critically re-evaluated, as uncertainties around estimates may have been grossly underestimated and many conclusions may ultimately not be supported by available data.

Relationships between different aspects of occurrence information

In addition to some obviously diverging patterns of individual information metrics (Fig. 2 and 3), we found clear quantitative evidence for a multidimensionality of limitations in DAI. Different metrics showed clearly distinct patterns and predominantly characterise different parts of the world.

Pairwise Spearman rank correlations across nine metrics of occurrence information varied strongly but mostly yielded weak to moderate spatial associations ($|r_S| = 0.00$–0.86, median = 0.23; Fig. S6). Some *coverage* metrics were moderately to strongly correlated ($r_S = 0.63$–0.86), mainly because *coverage* of any dimension is constrained by the number of available records (correlations with record number: $r_S = 0.65$–0.92; compare Yang *et al.* 2013). *Taxonomic* and *geographical coverage* were also moderately and negatively correlated with time since the last recording activities (r_S: -0.67 to -0.70). In contrast, most *uncertainty* metrics showed no or only weak correlations, with only *temporal* and combined *uncertainty* being highly correlated ($r_S = 0.75$). Notably, most metrics correlated poorly with quantities of mobilised raw records (Fig. S6), providing evidence that such simplistic indicators cannot reliably inform about different quantitative and qualitative aspects of occurrence information.

The first three axes of the PCA of the nine metrics accounted for 69.8% of their variation (Fig. 4). Plotting ordination site scores on a world map characterised regions in terms of their multidimensional data limitations (Fig. 4d). The most important axis (38%) mainly separated regions of high *taxonomic* and *geographical coverage*, e.g. in Europe ($r_S = 0.86/0.85$; bright green cells in Fig. 4a–b/d), from regions where a long time has passed since the last recording

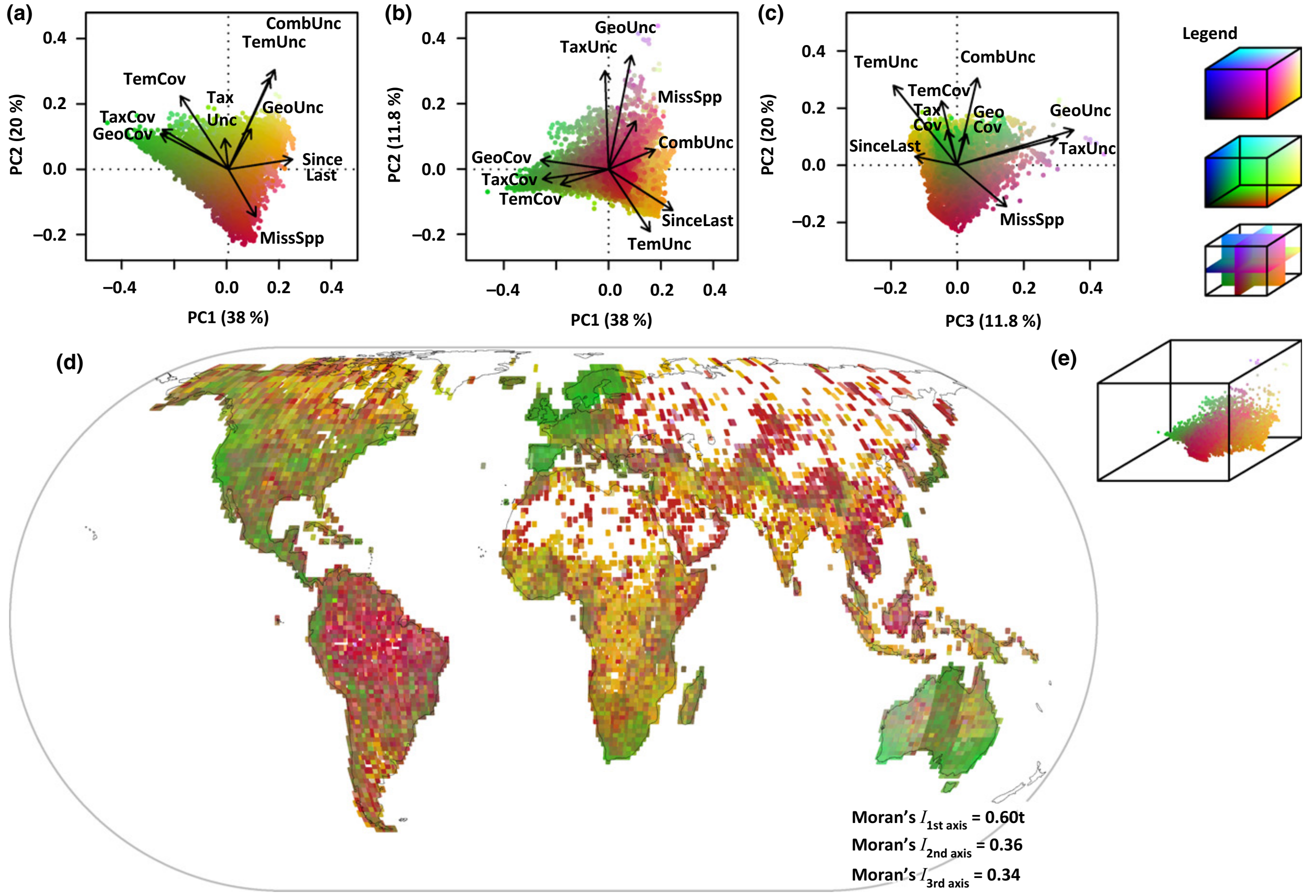

Figure 4 Principal component analysis (PCA) of 9 metrics of plant occurrence information across 12 100 km² grid cells with ≥ 1 record. (a–c) Biplots of the first three PCA axes. (d) Global map of ordination site scores; similar colours denote regions characterised by similar information metrics. Colours refer to a red–green–blue (RGB) colour space (legend) projected onto the (e) 3D PCA space (Weigelt *et al.* 2013). **TaxCov**: *taxonomic coverage* (recorded/modelled richness; Kreft & Jetz 2007); **GeoCov**: *geographical coverage* (*N* sampling locations/10^4 km² land area); **TempCov**: *temporal coverage* 1750–2010, estimated as mean minimum time between all months since 1750 and their respective closest recording date; **TaxUnc**: % records lost under moderate taxonomic filtering; **GeoUnc**: % records lost under moderate geographical filtering; **TempUnc**: % species lost under moderate temporal filtering; **CombUnc**: % species lost under combined filtering. **MissSpp**: *N* species modelled, but not recorded; **SinceLast**: Years since last recording activity.

activities, e.g. in Central Africa and South Asia ($r_S = -0.85$; yellow cells in Fig. 4a–b/d). The second axis (20% of variance) mainly correlated with combined and *temporal uncertainty* ($r_S = 0.74/0.75$; Fig. 4a/c/d), highlighting, e.g. Arctic Canada. Combined *uncertainty* also characterised much of Asia, such as the Altai or the mountain ranges between Eastern Tibet and Sichuan (Fig. 4d). *Taxonomic* and *geographical uncertainty* varied mainly along the third axis (11.8% of variance; r_S: 0.69/0.47; Fig. 4b–c), characterising, e.g. Borneo.

Overall, the results of our study highlight the differences, rather than the similarities, between geographical patterns of different aspects and dimensions of occurrence information. Different limitations predominate in different regions. Similar differences can be expected among taxonomic and temporal patterns of the different information metrics. For instance, pteridophytes stood out for their high *taxonomic coverage* but also showed the highest levels of *taxonomic uncertainty*. This multidimensionality of limitations in DAI deserves careful attention in research and conservation applications, as well as in future efforts to assess and improve information.

Prospects for using DAI in global plant research, conservation and monitoring

Despite the showcased limitations in DAI, there is an urgent need to use this information in plant research and conservation. For instance, DAI-based estimates of distributions (extent of occurrence or area of occupancy) will play a vital role in conservation assessments (GSPC target 2; Schatz 2009; Rivers *et al.* 2011), threatened species management (GSPC target 7; McLane & Aitken 2012) and monitoring (Brummitt *et al.* 2015). As shown below, the potential for such applications largely depends on the ability of distribution estimation methods to deal with low record numbers and high data *uncertainty*.

If useful species distribution estimates could be made based on 10 sampling locations (Rivers *et al.* 2011) and estimation methods were robust towards relatively high data *uncertainty*, DAI could currently facilitate distribution estimates and thus preliminary conservation assessments for 85 787 non-red-listed or 'Data-Deficient' species globally (*c.* 25% of all plants; Fig. 5). This represents a potential seven-fold increase

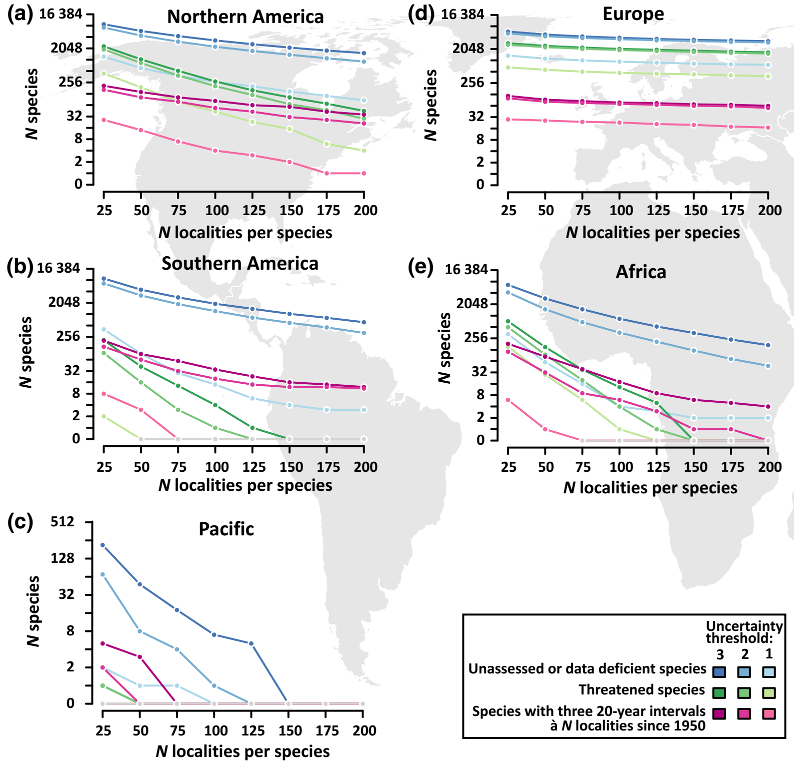

Figure 5 Global trade-offs between plant occurrence information *coverage* and *uncertainty*. Shown are numbers of species whose distributions could be estimated with hypothetical methods, depending on those methods' minimum requirements (10–200 sampling locations; Kadmon *et al.* 2003; Rivers *et al.* 2011) and robustness towards different levels of data *uncertainty*. (a) Northern America, (b) Southern America, (c) Pacific, (d) Europe, (e) Africa, (f) Temperate Asia, (g) Tropical Asia, (h) Australasia. Blue colours: species that are either un-assessed or 'Data-Deficient' on the IUCN Red List (2014). Violet colours: species with Red List categories 'Vulnerable', 'Endangered' or 'Critically Endangered'. Green colours: species for which the indicated number of sampling locations exists in each of three twenty-year periods since 1950. Colour shadings indicate filters (basic, moderate, strict) used to reduce taxonomic, geographical and temporal *uncertainty*. World regions are level-1 regions of Biodiversity Information Standards (TDWG).

compared to the IUCN Red List (i.e. ignoring national red lists; as of Aug 2014). However, this number would drop to only 865 or 0.2% for *uncertainty*-sensitive methods requiring ≥ 200 locations (Feeley & Silman 2011). Similarly, depending on methods' data requirements, distribution estimates might be feasible for 0.07–15.7% of 'Threatened' plants, and for 0.03–6.6% of all plants for each of three twenty-year periods since 1950. While these figures do demonstrate considerable potential for DAI applications, this potential is geographically highly uneven (Fig. 5). For instance, DAI-based monitoring of distributional changes since 1950 might be feasible for 48-3682 European but only 0-26 Pacific plant species (Fig. 5).

Most distribution modelling methods are highly sensitive to both number and quality of records (Guisan *et al.* 2007), yet few and uncertain records are the reality for the vast majority of plant species. While restricting analyses to highest-quality records is often recommended (Feeley & Silman 2010), cut-offs are usually arbitrary, and strict filters wipe out most available information (Fig. 4h, Fig. 5). Moreover, different filters may introduce different biases to already-biased datasets (Fig. 4). More effective usage of DAI would be to explicitly incorporate biases and *uncertainties* into analyses. Methods for doing so are increasingly available (McInerny & Purves 2011; Beale & Lennon 2012; Dorazio 2014; Velásquez-Tibatá *et al.* 2015), and further developing such methods holds great potential for advancing global plant research and conservation. Hierarchical Bayesian methods might be particularly well-suited (Beale & Lennon 2012; Iknayan *et al.* 2014). Theoretically, *uncertainty* of each record could be accounted for individually, e.g. by sampling possible interpretations of ambiguous synonyms from distributions of candidate accepted species, and by sampling possible interpretations of imprecise coordinates from distributions of potentially true locations around the indicated coordinates.

Taxonomic standardisation and basic geographical plausibility checks, as carried out in this study, are an essential part of any analysis using DAI (Chapman 2005). However, even thorough post-processing cannot fully eliminate information inaccuracies such as taxonomic misidentifications or incorrectly recorded sampling locations (Soberón & Peterson 2004), as these usually cannot be detected in DAI (Box 1). Sampled taxonomic re-assessments of original material (Scott & Hallam 2002; Ahrends *et al.* 2011) and sampled ground-truthing of occurrences (Miller *et al.* 2007) could provide vital information on typical rates of such errors for different taxa, regions and data sources. If additionally, the reporting and curating of appropriate metadata could be improved, the combined information could be used to explicitly model the likelihood of data inaccuracies, which could additionally be accounted for in biodiversity models.

Our analyses demonstrate that after two decades of intensive data mobilisation, options for using plant DAI in global research and conservation are still severely compromised by different data limitations. Even under our most optimistic scenario regarding methods' data requirements and robustness to *uncertainty*, DAI-based distribution estimations would be unfeasible for three quarters of all plants. Better integration of regional data sources into global DAI could provide some remedy, but these sources exhibit similar limitations (Yang *et al.* 2013; Sousa-Baena *et al.* 2014). The multidimensionality of data limitations also implies flaws in the accuracy of distribution datasets that are ultimately derived from primary biodiversity records, such as checklists, range maps and atlas data. This is exemplified by the many WCSP-listed species that are recorded in regions immediately adjacent to their supposedly correct native ranges, which may very well represent valid additions to those regions' native floras (Fig. S4b). Botanical inventorying will never be complete and severe data gaps will likely persist for decades to come, as evident in slow progress towards regional and global floras (GSPC target 1; Paton 2013). Meeting GSPC targets on plant conservation seems unlikely without substantial increases in funding and personnel allocated to data collection, curation and mobilisation. Given difficulties in securing adequate and sustained financing for such activities (Vollmar *et al.* 2010; Bradley *et al.* 2014; Costello *et al.* 2014), efforts to improve DAI should be globally coordinated and prioritised (Meyer *et al.* 2015).

Towards more effective improvement of DAI

Our analyses provide an important first step towards prioritising efforts to enhance global DAI on plant occurrences. Distinguishing between metrics of information *coverage* and *uncertainty* in taxonomic, geographical and temporal dimensions allows narrowing down critical improvements. For instance high *taxonomic uncertainty* in South-East Asian and pteridophyte floras may be addressed by targeted taxonomic revisions and better integration of taxonomic resources into *The Plant List*. New surveys to update information seem most urgently needed for Central Africa, Mozambique, tropical Asia and Arctic Canada. In general, Asian and bryophyte floras are woefully under-represented in DAI, and mobilising respective occurrence datasets seems like an obvious priority. To maximise leverage for applicability in research and conservation, such preliminary priorities could be further refined, by considering, e.g. current or projected threats (Pyke & Ehrlich 2010), geographical and environmental distance to well-sampled regions (Funk *et al.* 2005; Sousa-Baena *et al.* 2014), and opportunities for continuing or closing gaps in long time series (Johnson *et al.* 2011). Relevant collections for such targeted data mobilisation may be identified through metadata digitisation (Berendsohn & Seltmann 2010), while identifying socio-economic drivers of information gaps can help prioritise key activities likely to have a large impact (Yang *et al.* 2014; Meyer *et al.* 2015). Specialised biodiversity informatics infrastructures (e.g. Jetz *et al.* 2012; Atlas of Living Australia 2015) could play an important role in highlighting and tracking the various data limitations. Our conceptual framework for analysing quantitative and qualitative data limitations along different dimensions may serve as a model for future assessments for plants as well as for other hyperdiverse clades.

The multidimensional and largely un-correlated limitations in DAI also raise the question of how to effectively monitor progress towards international targets on improving and sharing biodiversity knowledge (GSPC target 3, Aichi target 19). Simplistic indicators like global or per-country record quantities (e.g. Tittensor *et al.* 2014) cannot inform about data *uncertainties* or fine-scale biases in *coverage*. To monitor improvements in the usefulness of DAI, rather than mere increases in data volume, we recommend evaluating a suite of indicators that inform about both quantitative and qualitative aspects of DAI at relevant scales.

CONCLUSIONS

As demonstrated, severe multidimensional biases, gaps and uncertainties are prevalent in global DAI on plant occurrences, hampering opportunities for using this information in global biodiversity research and for achieving international targets on plant conservation. Either goal would require both substantial up-scaling and prioritisation of efforts to collect and mobilise additional, and enhance the quality of available, occurrence information. Progress in improving DAI should be monitored using meaningful indicators. However, it should be stressed that severe data limitations will remain the norm for most species and regions. Greater effort should therefore be made to make best-possible use of limited information. This includes developing easy-to-use routines for explicitly incorporating data limitations into analyses, more widely adopting such methods, and clearly articulating remaining uncertainties.

ACKNOWLEDGEMENTS

We thank the many botanists collecting, curating and sharing plant distribution data. We thank Kenneth Feeley, Joaquín Hortal, two anonymous reviewers and the editor Janneke Hille Ris Lambers for constructive comments that greatly improved this manuscript. We thank Tim Robertson, Janet Scott and Martin Turjak for assistance with GBIF data retrieval, IUCN red list data and WCSP checklist data respectively. CM acknowledges funding from the Deutsche Bundesstiftung Umwelt (DBU). HK acknowledges funding by the German Research Foundation (DFG) in the framework of the German Excellence Initiative within the Free Floater Program at the University of Göttingen. PW acknowledges funding in the scope of the BEFmate project from the Ministry of Science and Culture of Lower Saxony.

AUTHOR CONTRIBUTIONS

All authors designed this study, C.M. compiled data, C.M. and P.W. performed taxonomic harmonisation, C.M. performed the analyses and wrote the first draft of the manuscript and all authors contributed substantially to revisions.

REFERENCES

Ahrends, A., Rahbek, C., Bulling, M.T., Burgess, N.D., Platts, P.J., Lovett, J.C. *et al.* (2011). Conservation and the botanist effect. *Biol. Conserv*., 144, 131–140.

Atlas of Living Australia. (2015). Spatial Portal. Available at: http://spatial.ala.org.au/. (Last accessed 8 January 2015).

Beale, C.M. & Lennon, J.J. (2012). Incorporating uncertainty in predictive species distribution modelling. *Philos. Trans. R. Soc. Lond. B Biol. Sci*., 367, 247–58.

Berendsohn, W.G. (1995). The concept of 'potential taxa' in databases. *Taxon*, 44, 207–212.

Berendsohn, W.G. & Seltmann, P.S. (2010). Using geographic and taxonomic metadata to set priorities in specimen digitization. *Biodivers. Informatics*, 7, 120–129.

Bisang, I. & Urmi, E. (1994). Studies on the status of rare and endangered bryophytes in Switzerland. *Biol. Conserv*., 70, 109–116.

Boakes, E.H., McGowan, P.J.K., Fuller, R.A., Chang-qing, D., Clark, N.E., O'Connor, K. *et al.* (2010). Distorted views of biodiversity: spatial and temporal bias in species occurrence data. *PLoS Biol*., 8, e1000385.

Boitani, L., Maiorano, L., Baisero, D., Falcucci, A., Visconti, P. & Rondinini, C. (2011). What spatial data do we need to develop global mammal conservation strategies? *Philos. Trans. R. Soc. Lond. B Biol. Sci*., 366, 2623–32.

Bradley, R.D., Bradley, L.C., Garner, H.J. & Baker, R.J. (2014). Assessing the value of natural history collections and addressing issues regarding long-term growth and care. *Bioscience*, 64, 1150–1158.

Broennimann, O., Treier, U.A., Müller-Schärer, H., Thuiller, W., Peterson, A.T. & Guisan, A. (2007). Evidence of climatic niche shift during biological invasion. *Ecol. Lett*., 10, 701–709.

Brummitt, N., Bachman, S.P., Aletrari, E., Chadburn, H., Griffiths-Lee, J., Lutz, M. *et al.* (2015). The sampled red list index for plants, phase II: ground-truthing specimen-based conservation assessments. *Philos. Trans. R. Soc. Lond. B Biol. Sci*., 370, 20140015.

Burrows, M.T., Schoeman, D.S., Buckley, L.B., Moore, P., Poloczanska, E.S., Brander, K.M. *et al.* (2011). The pace of shifting climate in marine and terrestrial ecosystems. *Science*, 334, 652–656.

Calinger, K.M., Queenborough, S. & Curtis, P.S. (2013). Herbarium specimens reveal the footprint of climate change on flowering trends across north-central North America. *Ecol. Lett*., 16, 1037–44.

Callmander, M. (2011). The endemic and non-endemic vascular flora of Madagascar updated. *Plant Ecol. Evol*., 144, 121–125.

Chao, A. & Jost, L. (2012). Coverage-based rarefaction and extrapolation: standardizing samples by completeness rather than size. *Ecology*, 93, 2533–2547.

Chapman, A.D. (2005). Principles and Methods of Data Cleaning – Primary Species and Species-Occurrence Data, version 1.0. Report for the Global Biodiversity Information Facility, Copenhagen.

Christenhusz, M.J.M. & Chase, M.W. (2014). Trends and concepts in fern classification. *Ann. Bot*., 113, 571–94.

Collen, B., Ram, M., Zamin, T. & McRae, L. (2008). The tropical biodiversity data gap: addressing disparity in global monitoring. *Trop. Conserv. Sci*., 1, 75–88.

Colwell, R.K. & Coddington, J.A. (1994). Estimating terrestrial biodiversity through extrapolation. *Philos. Trans. R. Soc. Lond. B Biol. Sci*., 345, 101–18.

Costello, M.J., Appeltans, W., Bailly, N., Berendsohn, W.G., de Jong, Y., Edwards, M. *et al.* (2014). Strategies for the sustainability of online open-access biodiversity databases. *Biol. Conserv*., 173, 155–165.

Crosby, M.R. & Magill, R.E. (1988). *TROPICOS. A Botanical Database System at the Missouri Botanical Garden*. Missouri Botanical Garden, St. Louis, MI.

Dorazio, R.M. (2014). Accounting for imperfect detection and survey bias in statistical analysis of presence-only data. *Glob. Ecol. Biogeogr*., 23, 1472–1484.

Dutilleul, P. (1993). Modifying the t test for assessing the correlation between two spatial processes. *Biometrics*, 49, 305–314.

Edwards, J.L. (2000). Interoperability of biodiversity databases: biodiversity information on every desktop. *Science*, 289, 2312–2314.

Feeley, K.J. (2012). Distributional migrations, expansions, and contractions of tropical plant species as revealed in dated herbarium records. *Glob. Chang. Biol*., 18, 1335–1341.

Feeley, K.J. & Silman, M.R. (2010). Modelling the responses of Andean and Amazonian plant species to climate change: the effects of georeferencing errors and the importance of data filtering. *J. Biogeogr*., 37, 733–740.

Feeley, K.J. & Silman, M.R. (2011). Keep collecting: accurate species distribution modelling requires more collections than previously thought. *Divers. Distrib*., 17, 1132–1140.

Ferrier, S. (2002). Mapping spatial pattern in biodiversity for regional conservation planning: where to from here? *Syst. Biol*., 51, 331–363.

Funk, V.A., Zermoglio, M.F. & Nasir, N. (1999). Testing the use of specimen collection data and GIS in biodiversity exploration and conservation decision making in Guyana. *Biodivers. Conserv*., 8, 727–751.

Funk, V.A., Richardson, K.S. & Ferrier, S. (2005). Survey-gap analysis in expeditionary research: where do we go from here? *Biol. J. Linn. Soc*., 85, 549–567.

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

Gamfeldt, L., Snäll, T., Bagchi, R., Jonsson, M., Gustafsson, L., Kjellander, P. *et al.* (2013). Higher levels of multiple ecosystem services are found in forests with more tree species. *Nat. Commun*., 4, 1340.

Gaston, K.J. & Fuller, R.A. (2009). The sizes of species' geographic ranges. *J. Appl. Ecol*., 46, 1–9.

Grand, J., Cummings, M.P., Rebelo, T.G., Ricketts, T.H. & Neel, M.C. (2007). Biased data reduce efficiency and effectiveness of conservation reserve networks. *Ecol. Lett*., 10, 364–74.

Guisan, A., Zimmermann, N.E., Elith, J., Graham, C.H., Phillips, S. & Peterson, A.T. (2007). What matters for predicting the occurrences of trees: techniques, data, or species' characteristics? *Ecol. Monogr*., 77, 615–630.

Guisan, A., Tingley, R., Baumgartner, J.B., Naujokaitis-Lewis, I., Sutcliffe, P.R., Tulloch, A.I.T. *et al.* (2013). Predicting species distributions for conservation decisions. *Ecol. Lett*., 16, 1424–1435.

Hansen, M.C., Potapov, P.V., Moore, R., Hancher, M., Turubanova, S.A., Tyukavina, A. *et al.* (2013). High-resolution global maps of 21st-century forest cover change. *Science*, 342, 850–3.

Herberich, E., Sikorski, J. & Hothorn, T. (2010). A robust procedure for comparing multiple means under heteroscedasticity in unbalanced designs. *PLoS ONE*, 5, e9788.

Hortal, J., Lobo, J.M. & Jiménez-Valverde, A. (2007). Limitations of biodiversity databases: case study on seed-plant diversity in tenerife, canary islands. *Conserv. Biol*., 21, 853–863.

Hortal, J., Jiménez-Valverde, A., Gómez, J.F., Lobo, J.M. & Baselga, A. (2008). Historical bias in biodiversity inventories affects the observed environmental niche of the species. *Oikos*, 117, 847–858.

Iknayan, K.J., Tingley, M.W., Furnas, B.J. & Beissinger, S.R. (2014). Detecting diversity: emerging methods to estimate species diversity. *Trends Ecol. Evol*., 29, 97–106.

Isbell, F., Calcagno, V., Hector, A., Connolly, J., Harpole, W.S., Reich, P.B. *et al.* (2011). High plant diversity is needed to maintain ecosystem services. *Nature*, 477, 199–202.

IUCN. (2014). IUCN Red List of Threatened Speciess. Version 2014.4. Available at: http://www.iucnredlist.org. (Last accessed 29 August 2014).

Jäger, E.J. (1976). Areal- und Florenkunde (Floristische Geobotanik). *Prog. Bot*., 38, 314–330.

Jansen, F. & Dengler, J. (2010). Plant names in vegetation databases – a neglected source of bias. *J. Veg. Sci*., 21, 1179–1186.

Jetz, W., McPherson, J.M. & Guralnick, R.P. (2012). Integrating biodiversity distribution knowledge: toward a global map of life. *Trends Ecol. Evol*., 27, 151–159.

Johnson, K.G., Brooks, S.J., Fenberg, P.B., Glover, A.G., James, K.E., Lister, A.M. *et al.* (2011). Climate change and biosphere response: unlocking the collections vault. *Bioscience*, 61, 147–153.

Kadmon, R., Farber, O. & Danin, A. (2003). A systematic analysis of factors affecting the performance of climatic envelope models. *Ecol. Appl.*, 13, 853–867.

Kier, G., Mutke, J., Dinerstein, E., Ricketts, T.H., Küper, W., Kreft, H. *et al.* (2005). Global patterns of plant diversity and floristic knowledge. *J. Biogeogr.*, 32, 1107–1116.

Kissling, W.D., Field, R. & Böhning-Gaese, K. (2008). Spatial patterns of woody plant and bird diversity: functional relationships or environmental effects? *Glob. Ecol. Biogeogr.*, 17, 327–339.

Klazenga, N. & Vaughan, A. (2014). Australia's virtual herbarium hits 5 million records. *Australas. Syst. Bot. Soc. Newsl.*, 159, 7–10.

van Kleunen, M., Dawson, W., Essl, F., Pergl, J., Winter, M., Weber, E. *et al.* (2015). Global exchange and accumulation of non-native plants. *Nature*, 525, 100–103.

Kreft, H. & Jetz, W. (2007). Global patterns and determinants of vascular plant diversity. *Proc. Natl Acad. Sci. USA*, 104, 5925–30.

Ladle, R. & Hortal, J. (2013). Mapping species distributions: living with uncertainty. *Front. Biogeogr.*, 5, 4–6.

Lavoie, C. (2013). Biological collections in an ever changing world: Herbaria as tools for biogeographical and environmental studies. *Perspect. Plant Ecol. Evol. Syst.*, 15, 68–76.

Legendre, P. & Legendre, L.F. (2012). *Numerical Ecology*, 24th edn. Elsevier, Amsterdam, Netherlands.

Lomolino, M. (2004). Frontiers of biogeography: new directions in the geography of. In: *Frontiers of Biogeography: New Directions in the Geography of Nature*. (eds Lomolino, M.V. & Heaney, L.R.). Sinauer Associates, Sunderland, MA, USA, pp. 293–296.

Maldonado, C., Molina, C.I., Zizka, A., Persson, C., Taylor, C.M., Albán, J. *et al.* (2015). Estimating species diversity and distribution in the era of big data: to what extent can we trust public databases? *Glob. Ecol. Biogeogr.*, 24, 973–984.

McInerny, G.J. & Purves, D.W. (2011). Fine-scale environmental variation in species distribution modelling: regression dilution, latent variables and neighbourly advice. *Methods Ecol. Evol.*, 2, 248–257.

McLane, S.C. & Aitken, S.N. (2012). Whitebark pine (*Pinus albicaulis*) assisted migration potential: testing establishment north of the species range. *Ecol. Appl.*, 22, 142–153.

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

Miller, B.P., Enright, N.J. & Lamont, B.B. (2007). Record error and range contraction, real and imagined, in the restricted shrub Banksia hookeriana in south-western Australia. *Divers. Distrib.*, 13, 406–417.

Morueta-Holme, N., Enquist, B.J., McGill, B.J., Boyle, B., Jørgensen, P.M., Ott, J.E. *et al.* (2013). Habitat area and climate stability determine geographical variation in plant species range sizes. *Ecol. Lett.*, 16, 1446–1454.

Murphey, P.C., Guralnick, R.P., Glaubitz, R., Neufeld, D. & Ryan, J.A. (2004). Georeferencing of museum collections: a review of problems and automated tools, and the methodology developed by the Mountain Informatics Initiative (Mapstedi). *PhyloInformatics*, 21, 1–29.

Nelson, B.W., Ferreira, C.A.C., da Silva, M.F. & Kawasaki, M.L. (1990). Endemism centres, refugia and botanical collection density in Brazilian Amazon. *Nature*, 345, 714–716.

New York Botanical Garden. (2014). Index Herbariorum: A Global Directory of Public Herbaria and Associated Staff. Available at: http://sciweb.nybg.org/science2/IndexHerbariorum.asp. (Last accessed 24 March 2015).

O'Donnell, J., Gallagher, R.V., Wilson, P.D., Downey, P.O., Hughes, L. & Leishman, M.R. (2012). Invasion hotspots for non-native plants in Australia under current and future climates. *Glob. Chang. Biol.*, 18, 617–629.

Paton, A. (2009). Biodiversity informatics and the plant conservation baseline. *Trends Plant Sci.*, 14, 629–37.

Paton, A. (2013). From working list to online flora of all known plants – looking forward with hindsight. *Ann. Missouri Bot. Gard.*, 99, 206–213.

Peterson, A.T., Soberón, J. & Krishtalka, L. (2015). A global perspective on decadal challenges and priorities in biodiversity informatics. *BMC Ecol.*, 15, 15.

Prance, G.T. (1977). Floristic inventory of the tropics: where do we stand? *Ann. Missouri Bot. Gard.*, 64, 659–684.

Prather, L.A., Alvarez-Fuentes, O., Mayfield, M.H. & Ferguson, C.J. (2004). The decline of plant collecting in the United States: a threat to the infrastructure of biodiversity studies. *Syst. Bot.*, 29, 15–28.

Prendergast, J.R., Wood, S.N., Lawton, J.H. & Eversham, B.C. (1993). Correcting for variation in recording effort in analyses of diversity hotspots. *Biodivers. Lett.*, 1, 39–53.

Pyke, G.H. & Ehrlich, P.R. (2010). Biological collections and ecological/environmental research: a review, some observations and a look to the future. *Biol. Rev. Camb. Philos. Soc.*, 85, 247–66.

R Core Team. (2014). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Available at: http://www.R-project.org/. (Last accessed 15 January 2016).

Raven, P.H. & Axelrod, D.I. (1974). Angiosperm biogeography and past continental movements. *Ann. Missouri Bot. Gard.*, 61, 539–673.

Riddle, B.R., Ladle, R.J., Lourie, S.A. & Whittaker, R.J. (2011). Basic biogeography: estimating biodiversity and mapping nature. In: *Conservation Biogeography* (eds Ladle, R.J. & Whittaker, R.J.). John Wiley & Sons, Oxford, UK, pp. 47–92.

Rivers, M.C., Taylor, L., Brummitt, N.A., Meagher, T.R., Roberts, D.L. & Lughadha, E.N. (2011). How many herbarium specimens are needed to detect threatened species? *Biol. Conserv.*, 144, 2541–2547.

Rocchini, D., Hortal, J., Lengyel, S., Lobo, J.M., Jimenez-Valverde, A., Ricotta, C. *et al.* (2011). Accounting for uncertainty when mapping species distributions: the need for maps of ignorance. *Prog. Phys. Geogr.*, 35, 211–226.

Schatz, G.E. (2009). Plants on the IUCN Red List: setting priorities to inform conservation. *Trends Plant Sci.*, 14, 638–42.

Schmidt-Lebuhn, A.N., Knerr, N.J. & Kessler, M. (2013). Non-geographic collecting biases in herbarium specimens of Australian daisies (Asteraceae). *Biodivers. Conserv.*, 22, 905–919.

Schurr, F.M., Pagel, J., Cabral, J.S., Groeneveld, J., Bykova, O., O'Hara, R.B. *et al.* (2012). How to understand species' niches and range dynamics: a demographic research agenda for biogeography. *J. Biogeogr.*, 39, 2146–2162.

Scott, W.A. & Hallam, C.J. (2002). Assessing species misidentification rates through quality assurance of vegetation monitoring. *Plant Ecol.*, 165, 101–115.

Soberón, J.M. & Peterson, A.T. (2004). Biodiversity informatics: managing and applying primary biodiversity data. *Philos. Trans. R. Soc. Lond. B Biol. Sci.*, 359, 689–98.

Sousa-Baena, M.S., Garcia, L.C. & Peterson, A.T. (2014). Completeness of digital accessible knowledge of the plants of Brazil and priorities for survey and inventory. *Divers. Distrib.*, 20, 369–381.

ter Steege, H. & Persaud, C.A. (1991). The phenology of Guyanese timber species: a compilation of a century of observations. *Vegetatio*, 95, 177–198.

ter Steege, H., Haripersaud, P.P., Bánki, O.S. & Schieving, F. (2011). A model of botanical collectors' behavior in the field: never the same species twice. *Am. J. Bot.*, 98, 31–7.

Thuiller, W., Albert, C., Araújo, M.B., Berry, P.M., Cabeza, M., Guisan, A. *et al.* (2008). Predicting global change impacts on plant species' distributions: future challenges. *Perspect. Plant Ecol. Evol. Syst.*, 9, 137–152.

Tittensor, D.P., Walpole, M., Hill, S.L.L., Boyce, D.G., Britten, G.L., Burgess, N.D. *et al.* (2014). A mid-term analysis of progress toward international biodiversity targets. *Science*, 346, 241–244.

TNRS (2014). The Taxonomic Name Resolution Service. iPlant Collaborative. Version 3.2. Available at: http://tnrs.iplantcollaborative.org. (Last accessed 7 April 2014).

TPL (2014). The Plant List. Version 1.1; Published on the Internet. Available at: http://www.theplantlist.org/. (Last accessed 23 March 2014).

Velásquez-Tibatá, J., Graham, C.H. & Munch, S.B. (2015). Using measurement error models to account for georeferencing error in species distribution models. *Ecography*, 38, 001–012.

Vollmar, A., Macklin, J.A. & Ford, L.S. (2010). Natural history specimen digitization: challenges and concerns. *Biodivers. Informatics*, 1, 93–112.

Walther, B.A. & Moore, J.L. (2005). The concepts of bias, precision and accuracy, and their use in testing the performance of species richness estimators, with a literature review of estimator performance. *Ecography*, 28, 815–829.

WCSP. (2013). World Checklist of Selected Plant Families. Facilitated by the Royal Botanic Gardens, Kew. Available at: http://apps.kew.org/wcsp/. (Last accessed 20 March 2014).

Weigelt, P., Jetz, W. & Kreft, H. (2013). Bioclimatic and physical characterization of the world's islands. *Proc. Natl Acad. Sci. USA*, 110, 15307–12.

Willis, K.J., Araújo, M.B., Bennett, K.D., Figueroa-Rangel, B., Froyd, C.A. & Myers, N. (2007). How can a knowledge of the past help to conserve the future? Biodiversity conservation and the relevance of long-term ecological studies. *Philos. Trans. R. Soc. Lond. B Biol. Sci.*, 362, 175–86.

Wright, M.G. & Samways, M.J. (1998). Insect species richness tracking plant species richness in a diverse flora: in the Cape Floristic South Africa Region, South Africa. *Oecologia*, 115, 427–433.

Yang, W., Ma, K. & Kreft, H. (2013). Geographical sampling bias in a large distributional database and its effects on species richness-environment models. *J. Biogeogr.*, 40, 1415–1426.

Yang, W., Ma, K. & Kreft, H. (2014). Environmental and socio-economic factors shaping the geography of floristic collections in China. *Glob. Ecol. Biogeogr.*, 23, 1284–1292.

SUPPORTING INFORMATION

Additional Supporting Information may be found online in the supporting information tab for this article.

Editor, Janneke Hille Ris Lambers
Manuscript received 8 February 2016
First decision made 17 March 2016
Manuscript accepted 14 April 2016