Multidimensional biases, gaps and uncertainties in global plant occurrence information

Plants are a hyperdiverse clade that plays a key role in maintaining ecological and evolutionary processes as well as human livelihoods. Glaring biases, gaps, and uncertainties in plant occurrence information remain a central problem in ecology and conservation, but these limitations have never been assessed globally. In this synthesis, we propose a conceptual framework for analyzing information biases, gaps and uncertainties 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 information metrics are largely uncorrelated, and filtering out specific limitations would usually lead to extreme trade-offs for other information metrics. In light of these multidimensional data limitations, we critically discuss prospects for global plant ecological and biogeographical research, monitoring and conservation, and outline critical next steps towards more effective information usage and mobilization. We provide an empirical baseline for evaluating and improving global floristic knowledge and our conceptual framework can be applied to the study of other hyperdiverse clades.

PeerJ PrePrints

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1326v2 | CC-BY 4.0 Open Access | rec: 26 Aug 2015, publ: 26 Aug 2015

Multidimensional biases, gaps and uncertainties in global plant occurrence information

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

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

[2] *Systemic Conservation Biology, University of Göttingen, Berliner Str. 28, 37073 Göttingen, Germany*.

PeerJ PrePrints

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

Correspondence to: Carsten Meyer (*cmeyer2@uni-goettingen.de*) or Holger Kreft (*hkreft@uni-goettingen.de*)

Abstract

Plants are a hyperdiverse clade that plays a key role in maintaining ecological and evolutionary processes as well as human livelihoods. Glaring biases, gaps, and uncertainties in plant occurrence information remain a central problem in ecology and conservation, but these limitations have never been assessed globally. In this synthesis, we propose a conceptual framework for analyzing information biases, gaps and uncertainties 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 information metrics are largely uncorrelated, and filtering out specific limitations would usually lead to extreme trade-offs for other information metrics. In light of these multidimensional data limitations, we critically discuss prospects for global plant ecological and biogeographical research, monitoring and conservation, and outline critical next steps towards more effective information usage and mobilization. We provide an empirical baseline for evaluating and improving global floristic knowledge and our conceptual framework can be applied to the study of other hyperdiverse clades.

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 datasets typically required in ecological research and conservation only exist for few plant groups and geographical regions (Riddle *et al.* 2011).

Most available datasets 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 characterize 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 mobilized via international data-sharing networks, most notably that of the Global Biodiversity Information Facility (GBIF; Edwards, 2000). Contrasting un-mobilized datasets or expert knowledge, these mobilized 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 DAK by Sousa-Baena *et al.* (2014a)). 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 are manifold (Lavoie 2013), spanning research on diversity patterns (Morueta-Holme *et al.* 2013), biological invasions (O'Donnell *et al.* 2012) or phenological changes (Calinger *et al.* 2013), 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 severe biases in each of the three basic dimensions (Nelson *et al.* 1990; Boakes *et al.* 2010; Schmidt-Lebuhn *et al.* 2013).

At least two aspects of DAI 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 in conservation prioritization (Funk *et al.* 1999). *Geographical coverage*, i.e., how well species' ranges are documented with records, affects the feasibility of species distribution modeling (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 interpretation of information on the three dimensions. For instance, ambiguous scientific names entail *uncertainty* regarding taxonomic identities (Jansen & Dengler 2010), imprecise sampling locations regarding the environmental context in which species were found (Rocchini *et al.*

2011), early sampling dates regarding their continuing presence at those locations (Boitani *et al.* 2011).

Coverage and *uncertainty* may both be biased in the taxonomic, geographical and temporal dimensions (Fig. 1), potentially leading to biased ecological inferences (Prendergast *et al.* 1993; 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.* 2014a), and *geographical uncertainty* may be greater in older records (Murphey *et al.* 2004). Other types of ecologically relevant data bias are typically closely connected to these basic dimensions, e.g., phylogenetic or functional biases (Schmidt-Lebuhn *et al.* 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 *coverage* and *uncertainty* of DAI with regard to the three dimensions is crucial for evaluating prospects for research and other applications, and for prioritizing 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), while most recent analyses emphasized effects on specific applications (Feeley & Silman 2011; Yang *et al.* 2013; Sousa-Baena *et al.* 2014b). Despite the need to comprehensively evaluate global DAI, 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 biases, gaps and uncertainties in DAI along taxonomic, geographical and temporal dimensions, investigate pairwise and multi-variate relationships between alternative metrics of *coverage* and *uncertainty*, and characterize geographical regions in terms of their multivariate data limitations. In light of these limitations, we critically discuss prospects for using plant DAI in global ecological research, conservation and monitoring, with particular emphasis on GSPC targets. Finally, we outline critical next steps towards more effective information usage and mobilization. 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 mobilization efforts.

Methods

Point-occurrence information

We downloaded all data for land plants available via GBIF in January 2014 (c. 120 M). GBIF-facilitated records represent by far the largest source of DAI, and a substantial part of the digitized portion of the estimated 350 million records that exist in the World's herbaria (New York Botanical Garden 2014). Geographical gaps in global *coverage* in these records may represent genuinely under-sampled regions, but also regions whose information is not yet digitized or integrated into international data-sharing networks, such as Brazil (Sousa-Baena *et al.* 2014a) or China (Yang *et al.* 2013). We taxonomically standardized 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 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). These steps led to a reduction of 119,058,280 raw records with 2,206,831 verbatim name strings to 55,929,317 unique records for 229,218 accepted species from 3,947,969 unique sampling locations and 3,172 year-month combinations (*SI.1.1*). These records were contributed to the GBIF network by 238 data publishers in 48 countries. The majority of these records (78%) came from field observations (e.g., from vegetation plot data) and preserved specimens (17%).

Coverage

We computed three alternative metrics to estimate the extent to which available records *cover* the taxonomic, geographical and temporal dimensions (Fig. 1). We estimated *taxonomic coverage* of 12,100 km equal area grid cells (110 km x 110 km at the equator) as the ratio between recorded vascular plant richness and an environment-richness model for that group (Kreft & Jetz, 2007). Spatial patterns of *taxonomic coverage* may be affected by erroneous or non-native species' records. However, independent information on species' native ranges to geographically validate records does not exist for most plants. We thus additionally validated 16.8 M records for 105,031 species of seed plants (Spermatophyta; 34% of all) 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) and compared the ratio between DAI-recorded and checklist-based richness among botanical countries. 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. To measure *temporal coverage* of species and cells, we calculated the negative mean minimum time (in years) from all months between 1750 and 2010 to their respective temporally closest records. This metric has large negative values if *temporal coverage* is low, i.e., if the entire time span contains large temporal gaps without any records. We analyzed temporal patterns of *taxonomic* and *geographical coverage* by comparing percentages of species and grid cells covered within, and cumulatively up to, five-year periods.

Uncertainty

To investigate *uncertainty* in DAI (Fig. 1), we created three potential *uncertainty* filters ('basic', 'moderate', 'strict') that a user of DAI might consider to ensure data quality. We defined three *taxonomic uncertainty* filters based on criteria and decisions taken during taxonomic validation (see *SI 1*):

- TaxStrict: Recorded name matches an accepted species in TPL with high expert confidence (three 'stars'; www.theplantlist.org/about), with ≤5% orthographic distance (see *SI 1*), either directly or through an unambiguous synonym (i.e., one that only links to one accepted name);
- TaxModerate: Recorded name matches an accepted species in TPL with high or medium expert confidence (two or three 'stars') with ≤15% orthographic distance, either directly or through an unambiguous or ambiguous synonym;
- TaxBasic: Recorded name matches an accepted species in TPL or TNRS (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 indicated country:

- GeoStrict: Location reported with a precision of at least 1/1000 of a degree (~100 m at the equator);
- GeoModerate: Location reported with an precision of at least 1/100 of a degree;
- 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:

- TempStrict: Records collected after 1990;
- TempModerate: Records collected after 1970;
- 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 records that would be additionally excluded when applying moderate or strict *taxonomic* and *geographical uncertainty* filters, respectively, compared to the basic filter. We investigated patterns in *temporal uncertainty* by comparing percentages of species additionally excluded by moderate or strict *temporal uncertainty* filters. Similarly, we investigated patterns in combined *uncertainty* by comparing percentages of additionally excluded species if all three filters were applied.

Assessing variation in occurrence information

To quantify and visualize 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 km grid cells and TDWG level-3 'botanical countries'), and five-year periods.

We investigated relationships between geographical patterns of nine different information metrics, including the three dimensions of *coverage* and *uncertainty* and combined *uncertainty* (see above; *uncertainty* measured here as information loss under moderate filtering). We also included two further aspects of limitations in DAI: the number of vascular plant species that were not recorded but expected to occur in an area based on an environment-richness model (Kreft & Jetz 2007), and the time (in years) since the last record in a grid cell was recorded. We analyzed 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 color 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 colored grid cells to visualize which regions are characterized 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). All analyses were carried out in R versions 3.0.2-3.2.1 (R Core Team 2014).

We assessed opportunities for selected research and conservation applications of DAI globally and for TDWG level-1 regions, by counting species that would meet minimum data requirements of hypothetical distribution estimation methods if all three basic, moderate or strict *uncertainty* filters were applied. We assingned species to regions if >80% of their records fell within their respective boundaries.

Results and Discussion

The high number of globally mobilized plant records (119 M; Fig. S1A) may misguide perceptions of the actual available information on plant occurrences. Our basic validation and filtering steps excluded 38.2 M records, including 12.5 M with non-validatable verbatim name strings (Fig. S1G, *SI 1*) and 27.9 M in the sea (Fig. S1C). Note that the latter included records with imprecise or erroneous coordinates (Yesson *et al.* 2007) but also potentially valid records of marine species (e.g., sea grasses). Collecting duplicate specimens from the same plant individual is common practice in botany, and removing duplicated species-location-month combinations excluded a further 25 M records, leaving 56 M unique records for analyses (47% of all). Record numbers varied by five orders of magnitude across species, and by six orders of magnitude across 12,100 km grid cells (Fig. S1B). For instance, a single cell in the Netherlands had 2.8 M records, whereas 21.2% of all cells had no records.

Coverage of the different dimensions

Taxonomic coverage

Globally, 229,218 plant species (65% of all 350,697 accepted by TPL as of 2014) were represented with ≥1 record that passed our basic filtering. *Taxonomic coverage* was itself taxonomically biased, with only 28.3% of bryophytes but 82.9% of pteridophyte species represented (Fig. 2A).

Recorded species richness in grid cells was an almost perfect function of record number (r_S=0.94, P_{Dut}=0; Fig. S1B/F/K), demonstrating that centers of plant diversity perceived from occurrence records often reflect better documentation rather than true diversity patterns (Nelson *et al.* 1990; Yang *et al.* 2013; Engemann *et al.* 2015). Accordingly, *taxonomic coverage* of plant assemblages was extremely heterogeneous in space (Fig. 2B). Only 5.4% of cells had a ratio between recorded and modeled species richness >0.8 and could thus be considered taxonomically well-*covered*. Regions which high *taxonomic coverage* were Europe, parts of Australia, North America, South Africa, Ecuador, Costa Rica, and scattered parts in the rest of the World (Fig. 2B). Conversely, 78.6% of the world was severely under-inventoried with ratios below 0.25 (Fig. 2B). Large numbers of 'missing' species, i.e., that portion of modeled richness that was not confirmed by records, were typical for Eastern Amazonia and Borneo (Fig. S2A). Surprisingly, our results did not confirm previous observations that data gaps are higher in the tropics than in the non-tropics (Collen *et al.*, 2008; P_{Dut}=0.37), nor that they are higher in Neo- than in Palaeotropical areas (Prance, 1977;

P_{Dut}=0.64). The overall low *taxonomic coverage* over vast extents seriously impairs estimations of plant diversity (Yang *et al.* 2013) and site-based plant conservation prioritization (GSPC target 5; Funk *et al.* (1999)).

Taxonomic coverage scores exceeded 1 in 3.6% of cells (Fig. 2B). Scores >1 may stem from an underestimation of richness by the environment-richness model, records of non-native species, or inaccurate information on sampling locations. For instance, the score of 6.6 around Stockholm was mainly due to undated records for non-native species provided by the Bergius Herbarium, possibly from collections assembled in the 19th century by the East India Company. As another example, peaks in *taxonomic coverage* often emerged in cells around country centroids, likely reflecting erroneous geo-referencing of records lacking precise information on sampling locations (e.g., in Bazil; compare Maldonado *et al.* (2015)). Such factors could influence recorded/modeled richness ratios anywhere in the world, therefore *taxonomic coverage* cannot directly be interpreted as completeness of native plant inventories. Anyone using DAI to study native biodiversity should carefully consider these potential sources of error.

A more robust measure of *taxonomic coverage* could be attained for 'botanical countries' and 105,031 species of native seed plants (spermatophytes), based on records that were geographically validated against botanical-country checklists (WCSP (2013); Fig. S3A). However, these coarser patterns only moderately correlated with mean grid-level *coverage* (r_P=0.68, P_{Dut}=0), and underestimated local data gaps in botanical countries where *coverage* was achieved by combining scattered species records over vast areas, such as in Argentina or the Democratic Republic of the Congo (Fig. 2B, Fig. S3A). Due to their higher spatial resolution, grid-level metrics therefore better indicate global data gaps, and provide an important first step in identifying priority regions for improving botanical baseline information (GSPC target 3; Sousa-Baena *et al.* (2014a)).

10.1% (1.7 M) of records for WCSP-listed species were collected outside the species' currently recognized native ranges, but even these records could play an important role for progress towards GSPC targets. 45% of these were collected immediately adjacent to recognized native ranges, and potentially represent valid additions to those regions' native floras, notably in the Neotropics (Fig. S3B). This highlights the importance of DAI for target 1, the completion of an online flora of all plants (Paton 2013). The 0.9 M records collected well beyond native ranges possibly represent non-native plants (Fig. S3C) and this information could support target 10 by facilitating study and effective management of plant invasions (Broennimann *et al.* 2007; van Kleunen *et al.* 2015).

Geographical coverage

If a species has been recorded at a sufficient number of sampling locations, records may be used to estimate the extent of occurrence (Gaston & Fuller 2009; Rivers *et al.* 2011) or to model probabilities of occurrence at finer scales (Guisan *et al.* 2007; Feeley & Silman 2011). However, mobilized records for a given species were typically collected from only 7 unique sampling locations (median across species with ≥1 record; Fig. 2D), hampering meaningful estimations for the majority of plant species.

Estimates of *geographical coverage* of regions may aid in pinpointing under-collected areas where new species might be found (Bebber *et al.* 2010), and in controlling for uneven survey effort in biodiversity analyses (Schulman *et al.* 2007; Lobo 2008). As expected, *geographical*

coverage was generally high in traditionally 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 mobilization 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 mobilization (Crosby & Magill 1988), and as a consequence now contributes 66% of Madagascan records.

Temporal variation in taxonomic and geographical coverage

Globally, percentages of *covered* species and grid cells mostly increased through time, apart from dips during the World Wars (Fig. 2C/F). *Geographical coverage* appears to have leveled off since the 1970s and *taxonomic coverage* since the 1980s, while 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 mobilization 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).

While *covered* species and areas mostly increased through time globally, there was strong spatio-temporal variation in certain regions (Fig. S4). For instance, since the 1950s, sampling activity decreased in the Afrotropics and Middle East, while it increased in the Neotropics and circum-Tibetan mountain ranges (Fig. S4D-F). Accordingly, regional percentages of *covered* species also changed over recent decades (Fig. S4K-M). In many parts of the world, *taxonomic coverage* during a given time period was always well below cumulative *coverage*, demonstrating that regionally high *coverage* is often reached only by aggregating information over very long time spans.

Temporal coverage

Continuous *temporal coverage* of species and regions is important to monitor 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 continuouities 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). *Temporal coverage* of grid cells was very high across non-eastern Europe. For instance, less than two months typically lay between a given point in time and the closest sampling date in the best-*covered* cell in eastern England. Large temporally well-*covered* areas also spanned North America, Central America, the Caribbean, the northern Andes, south-eastern Brazil, South Africa, Madagascar, the Kashmir region, south-western Australia, and New Zealand (Fig. 2H). In contrast, most of Amazonia and Asia showed extremely poor *temporal coverage* (median: 73.1 years; Fig. 2H).

For many global change questions such as the monitoring of poleward range expansions or land-use driven range contractions, *temporal coverage* specifically of recent decades may be more relevant, and *coverage* since 1950 was indeed higher (Fig. S2B-C). Worryingly however, several tropical and high arctic regions undergoing rapid land cover or climate

change (Burrows *et al.* 2011; Hansen *et al.* 2013) were characterized both by poor *temporal coverage* and aging records, notably in Canada, central Africa and Asia (Fig. S2C-D). For instance, the last record in a given Angolan grid cell was typically collected 36 years ago (median, measured from 2010).

Uncertainty regarding the interpretation of information

Compared to *coverage*–related aspects of species occurrence information (Yang *et al.* 2013; Sousa-Baena *et al.* 2014a; Meyer *et al.* 2015), patterns in more qualitative aspects like information *uncertainty* have received little attention.

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 (Jansen & Dengler, 2010; see *SI 1*). We found that applying our moderate filter to reduce *taxonomic uncertainty* lead to a minor loss of only 8.4% of DAI (see *Methods*). However, applying a very strict taxonomic filter lead to a loss of the majority of available information (66.5% of records; 62.7% of species). Pteridophytes disproportionately lost records under moderate filtering, compared to other groups (Fig. 3I), possibly due to the continuing major changes in fern taxonomy (Christenhusz & Chase 2014). Bryophytes and pteridophytes would altogether be excluded by our strict taxonomic filter (Fig. 3I), because *The Plant List* only assigns highest confidence levels to names sourced from taxonomically comprehensive and peer-reviewed databases, which do not exist for these groups (www.theplantlist.org/1.1/about). Depending on the rigor of taxonomic filtering, geographical peaks in lost information appeared in insular South-East Asia (moderate filter, Fig. 3A) or in the North American Midwest and the Caribbean (strict filter, Fig. 3B). Contrasting these strong taxonomic and geographical patterns, *taxonomic uncertainty* varied very little through time, with usually around 10% and 70% of records in a given five-year period falling above our moderate and strict *taxonomic uncertainty* threshold, respectively (Fig. 3J).

Geographical uncertainty

Imprecisely geo-referenced sampling locations lead to *uncertainty* regarding the geographical and environmental context of species' occurrences. This *uncertainty* hampers applications built on linking occurrences with high-resolution environmental data in species distribution models (Feeley & Silman 2010; Rocchini *et al.* 2011). Applying our basic geographical filter already lead to a 38% loss in accepted species (from 367,703 to 229,218), confirming a strong trade-off between geographical precision and *taxonomic coverage* of occurrence information (Feeley & Silman 2010). Compared to our pre-filtered dataset, further applying moderate and strict geographical filters would lead to an additional reduction of, respectively, 1.9% and 13.9% in records and 5.5% and 25.3% in species. The relatively low percentages of globally excluded records were mainly due to high numbers of precisely geo-referenced records in North-Western Europe (Fig. S1J). However, such global statistics of data *uncertainty* can tremendously underestimate local *uncertainty*, as demonstrated by substantially higher mean percentages of excluded records across grid cells (moderate filter: 22.3% (*sd*: 26.0); strict filter: 58.6% (*sd*: 35.7); Fig. 3C-D).

Large areas of relatively low *geographical uncertainty* were in Europe, the western United States, Southern Africa, Japan, New Zealand and parts of Australia (Fig. 3C-D). Records not fulfilling the strictest *geographical uncertainty* criteria were common in tropical regions, but also in remote non-tropical regions, including Alaska, temperate Asia, and Western Australia (Fig. 3D). Imprecise sampling locations for those regions may be related to a lack of high-quality maps and more sparsely distributed settlements, which often serve as geographical reference points, particularly in older records. However, *geographical uncertainty* may also be created at the time of data mobilization. For instance, in Australia, differences in *geographical uncertainty* closely mirrored administrative boundaries, reflecting different mobilization policies of Australian state departments, which contributed 53.8% of Australian records (Fig. 3C). At the time of downloading our records (Jan 2014), certain Australian datasets were mobilized into the GBIF network via intermediaries that deliberately generalized location coordinates of any potentially sensitive information. Mobilization pathways have since changed and generalizations are now restricted to much lower percentages of Australian records (e.g., for species threatened by illegal collecting; Klazenga & Vaughan (2014)).

Geographical uncertainty of records appeared similar across major plant groups (Fig. 3I), but there were several notable changes through time. Older sampling locations were not generally reported with lower precision (Murphey *et al.* 2004), although such patterns could be observed in several regions, like Spain or south-eastern Australia (Fig. S4O-T). Instead, there were two major periods during which global *geographical uncertainty* increased, in both cases likely reflecting increased explorations of tropical and remote regions, one between 1860 and 1910, coinciding with the second wave of European colonial expansion, and one between 1940 and 1965 (Fig. 3J; Fig. S4B-D; Fig. 3C-D). The steady decrease in *geographical uncertainty* since 1965 may reflect increasing availability of high-quality maps, and later of GPS technology.

Temporal uncertainty

Early-collected records represent vital information about past biota. However, if this is the only available information, they also inherit greater *temporal uncertainty* regarding species' continuing presences near sampling locations, as distributions may respond to environmental change (Thuiller *et al.* 2008) and biological processes (Schurr *et al.* 2012). Therefore, many applications like conservation planning or SDMs that link DAI with modern habitat data usually require modern occurrence information (Boitani *et al.* 2011).

86.3% of globally represented species had at least one record collected after 1970 and 72.4% had records collected after 1990. Using these dates as filters for excluding records would cause an average loss of, respectively, 32.0 and 61.8% of species per grid cell (Fig. 3E-F). Regions where most species had records collected after 1990 include continuously well-sampled north-western Europe, but also areas where most species were only recorded during recent surveys, such as Benin, the circum-Tibetan mountain ranges, or Indochina (Fig. 3F, Fig. S4F). In contrast, much of Arctic Canada, central Africa, Iraq, eastern India, Myanmar and Java were characterized by very old information, as most recorded species did not even have records collected after 1970 (Fig. 3E). Local reasons for spatio-temporal changes in sampling activity may include shifting funding priorities (Ahrends *et al.* 2011a), arising security concerns (Brito *et al.* 2013), or lowered botanical appeal of environmentally degraded regions (Boakes *et al.* 2010). Whatever the reasons, it is important to detect and account for such spatio-temporal biases and uncertainties. Mean percentages of excluded species also

varied three-fold across major plant groups (5.4%-15.10%; moderate filter; Fig. 3I), showcasing potential taxonomic biases introduced by temporal filters.

Combined uncertainty

Combining filters to minimize *uncertainty* in all three dimensions lead to substantial trade-offs for *coverage* (compare Feeley & Silman (2010); Boitani *et al.* (2011)). 78.9% of all species in our dataset had no record that passed all strict filters; 52.2% of species had no record passing all moderate filters. *Uncertainty* was even more apparent in geographical patterns: North-western Europe was the only larger regions where typically ≥80% of species in a 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 parts of Europe, Benin, Indochina, and central and south-eastern Australia (Fig. 3H).

Given such pervasive levels of data *uncertainty*, it is very likely that species identities and their environmental associations are frequently misinterpreted (Feeley & Silman 2010; Jansen & Dengler 2010; Naimi *et al.* 2014). 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).

Relationships between different aspects of occurrence information

Pairwise Spearman rank correlations across 9 variables of occurrence information mostly yielded weak to moderate associations in space ($|r_S|$=0.00-0.86, median=0.23; Fig. S5). Different *coverage* aspects correlated moderately to highly (r_S=0.63-0.86), which was expected as *coverage* of any dimension is numerically constrained by the number of available records (correlations with record number: 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* aspects showed no or only weak correlations, the only high correlation being that between *temporal* and combined *uncertainty* (r_S=0.75). Most metrics correlated poorly with quantities of mobilized raw records (Fig. S5), 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 9 variables accounted for 69.8% of the variance (Fig. 4; note percentages accounted for by each axis in A-C). Plotting ordination site scores on a world map characterized regions in terms of their multidimensional data limitations (Fig. 4D; Weigelt *et al.* (2013)). The most important axis (38%) mainly separated regions of high *taxonomic* and *geographical coverage*, e.g., in Europe (r_S=0.86/0.85; Fig. 4A-B/D), from regions where a long time has passed since the last recording activities, e.g., in Central Africa and South Asia (r_S=-0.85; 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 characterized 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), characterizing, e.g., Borneo.

The above patterns and analyses 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 stand out for their high *taxonomic coverage* but also show the highest levels of *taxonomic uncertainty*. This multidimensionality of limitations in occurrence information should be considered in research and conservation applications, as well as in future assessments of data limitations.

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 distribution estimates could 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*.

Assuming species distributions could be estimated from 10 sampling locations (Rivers *et al*. 2011) and 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 compared to the IUCN Red List (i.e., ignoring national red lists; as of Aug 2014). However, this number would drop to only 1,921 or 0.5% for *uncertainty*-sensitive methods requiring ≥200 locations (Feeley & Silman 2011). Similarly, depending on methods' data requirements, distribution estimates might be feasible for 0.1-15.7% of 'Threatened' plants, and for 0.1-6.6% of all plants for each of three twenty-year periods since 1950. While these figures demonstrate considerable potential for DAI applications, this potential is geographically highly biased (Fig. 5). For instance, DAI-based monitoring of distributional changes since 1950 might be feasible for 386-3,682 European but only 0-26 Pacific plant species (Fig. 5).

Most distribution modeling 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), cutoffs 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 standardization 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. Sampled taxonomic re-assessments of original material (Scott & Hallam 2002; Ahrends *et al.* 2011b) 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, which could additionally be accounted for in analyses.

Our analyses demonstrate that after decades of intensive data mobilization, 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.* 2014a). The multidimensionality of data limitations also implies flaws in the accuracy of distribution datasets that are derived from primary biodiversity records, such as checklists, range maps, and atlas data. This is exemplified by the many WSCP-listed species that are recorded in regions adjacent to their supposedly correct native ranges. Botanical inventorying will never be complete and severe data gaps will likely persist for decades to come, as evident in slow progress towards regional floras (Paton 2013). Meeting GSPC targets on plant conservation seems unlikely without substantial increases in funding and personnel allocated to data collection, curation and mobilization. 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 prioritized (Meyer *et al.* 2015).

Towards more effective improvement of DAI

Our analyses provide an important first step towards prioritizing efforts to enhance global DAI on plant occurrences. Distinguishing between 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 mobilizing respective occurrence datasets seems like an obvious priority. To maximize 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) environmental dissimilarity to well-sampled regions (Sousa-Baena *et al.* 2014a), and opportunities for continuing or closing gaps in long time series (Johnson *et al.* 2011). Relevant collections for such targeted data mobilization may be identified through metadata digitization (Berendsohn & Seltmann 2010),

while identifying socio-economic drivers of information gaps can help prioritize key activities likely to have a large impact (Yang *et al.* 2014; Meyer *et al.* 2015). Specialized 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 analyzing 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 prioritization of efforts to collect and mobilize 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 Tim Robertson and Janet Scott for assistance with data retrieval, and IUCN red list 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 harmonization, C.M. performed the analyses and wrote the first draft of the manuscript, and all authors contributed substantially to revisions.

References

Ahrends, A., Burgess, N.D., Gereau, R.E., Marchant, R., Bulling, M.T., Lovett, J.C., *et al.* (2011a). Funding begets biodiversity. *Divers. Distrib.*, 17, 191–200.

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

Atlas of Living Australia. (2015). Spatial Portal [WWW Document]. URL http://spatial.ala.org.au/.

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.

Bebber, D.P., Carine, M.A., Wood, J.R.I., Wortley, A.H., Harris, D.J., Prance, G.T., *et al.* (2010). Herbaria are a major frontier for species discovery. *Proc. Natl. Acad. Sci. U. S. A.*, 107, 22169–71.

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

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.

Brito, J.C., Godinho, R., Martínez-Freiría, F., Pleguezuelos, J.M., Rebelo, H., Santos, X., *et al.* (2013). Unravelling biodiversity, evolution and threats to conservation in the Sahara-Sahel. *Biol. Rev. Camb. Philos. Soc.*, 89, 215–231.

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.

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.

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.

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.

Engemann, K., Enquist, B.J., Sandel, B., Boyle, B., Jørgensen, P.M., Morueta-Holme, N., *et al.* (2015). Limited sampling hampers "big data" estimation of species richness in a tropical biodiversity hotspot. *Ecol. Evol.*, 5, 807–20.

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., 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.

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.

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., 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.

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.

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.

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.

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.

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*, Early View. DOI: 10.1038/nature14910.

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

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.

Lobo, J.M. (2008). Database records as a surrogate for sampling effort provide higher species richness estimations. *Biodivers. Conserv.*, 17, 873–881.

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. (in press). *Nat. Commun.*

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.

Naimi, B., Hamm, N.A.S., Groen, T.A., Skidmore, A.K. & Toxopeus, A.G. (2014). Where is positional uncertainty a problem for species distribution modelling? *Ecography*, 37, 191–203.

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 [WWW Document]. *Index Herb. A Glob. Dir. Public Herb. Assoc. Staff*. URL http://sciweb.nybg.org/science2/IndexHerbariorum.asp.

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. URL http://www.R-project.org/.

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: *Conserv. Biogeogr.* (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.

Schulman, L., Toivonen, T. & Ruokolainen, K. (2007). Analysing botanical collecting effort in Amazonia and correcting for it in species range estimation. *J. Biogeogr.*, 34, 1388–1399.

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. (2014a). Completeness of digital accessible knowledge of the plants of Brazil and priorities for survey and inventory. *Divers. Distrib.*, 20, 369–381.

Sousa-Baena, M.S., Garcia, L.C. & Peterson, A.T. (2014b). Knowledge behind conservation status decisions: Data basis for "Data Deficient" Brazilian plant species. *Biol. Conserv.*, 173, 80–89.

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

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 accessed Apr 2014. [WWW Document]. URL http://tnrs.iplantcollaborative.org/.

TPL. (2014). The Plant List. Version 1.1; Published on the Internet; http://www.theplantlist.org/ (accessed 1st January 2014). [WWW Document]. URL http://www.theplantlist.org/.

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.

WCSP. (2013). World Checklist of Selected Plant Families. Facilitated by the Royal Botanic Gardens, Kew. [WWW Document]. URL http://apps.kew.org/wcsp/.

Weigelt, P., Jetz, W. & Kreft, H. (2013). Bioclimatic and physical characterization of the world's islands. *Proc. Natl. Acad. Sci. U. S. A.*, 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.

Yesson, C., Brewer, P.W., Sutton, T., Caithness, N., Pahwa, J.S., Burgess, M., *et al.* (2007). How global is the global biodiversity information facility? *PLoS One*, 2, e1124.

PeerJ PrePrints

Figures

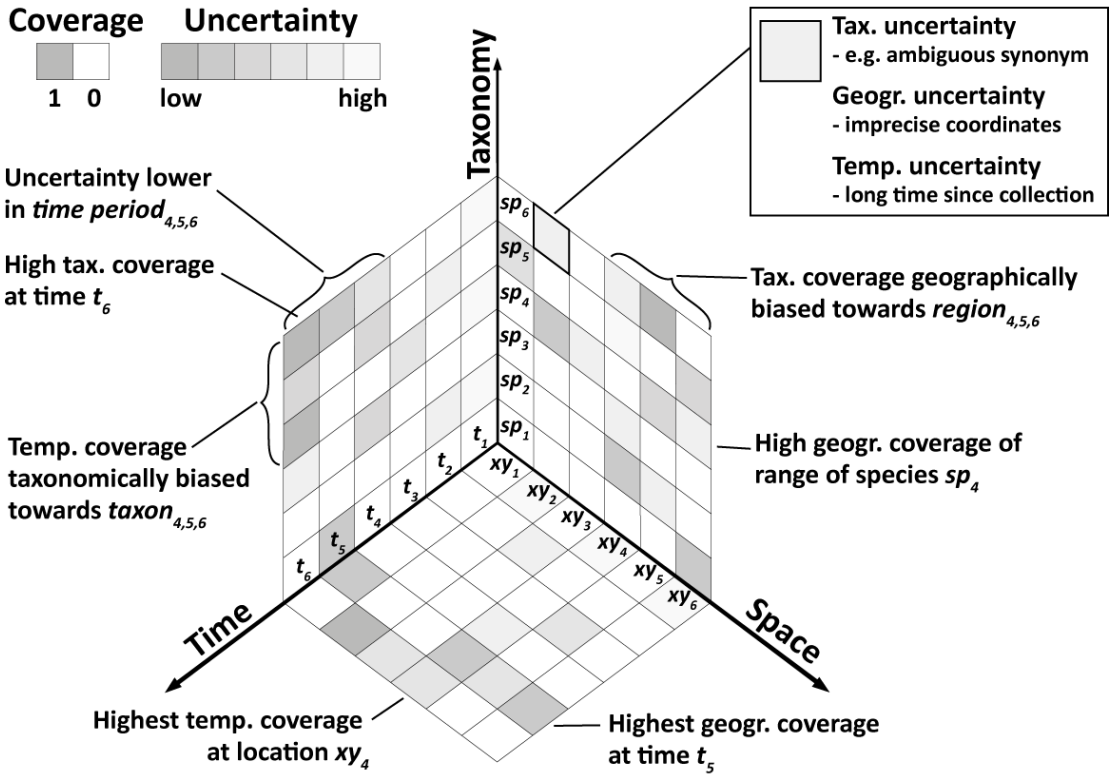

Figure 1. Framework for analyzing limitations in occurrence information along taxonomic, geographical and temporal dimensions. Occurrence records cover different species (sp_1, sp_2, …), different locations (xy_1, xy_2, …) and different points in time (t_1, t_2, …). Planes of cells illustrate spread 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: i) *coverage* of the three dimensions with information (grey cells), and ii) *uncertainty* regarding the interpretation of information on the three dimensions (shade of cells). Integrating across cells in one dimension summarizes 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., center left: *temporal coverage* taxonomically biased because species of $taxon_{4,5,6}$ have systematically higher *coverage*, compared to $taxon_{1,2,3}$).

PeerJ PrePrints

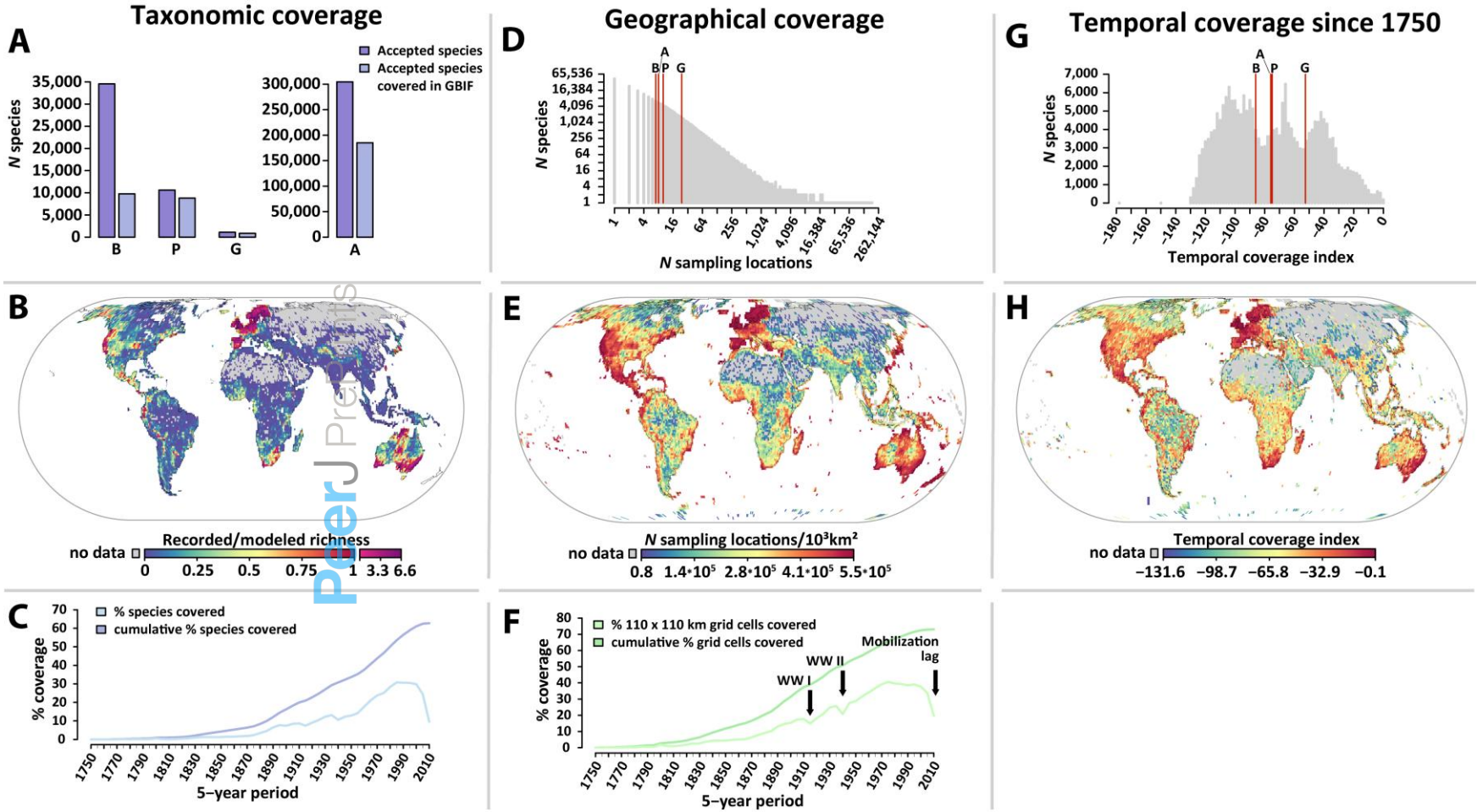

Figure 2. Global variation in occurrence information *coverage*. A) *Taxonomic coverage* of major plant groups (accepted (TPL 2014) vs. recorded species; B – bryophytes, P – pteridophytes, G – gymnosperms, A - angiosperms); B) Geographical variation across 12,100 km grid cells of *taxonomic coverage* for vascular plants (recorded/modeled richness; Kreft & Jetz, (2007)); values >1 indicate higher recorded than modeled richness; C) Percentages of species *covered* within, and up to, five-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, five-year periods. G) *Temporal coverage* 1750-2010 of G) species and H) cells; small negative values denote high *coverage*. Red bars in D/G: 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 mobilization).

PeerJ PrePrints

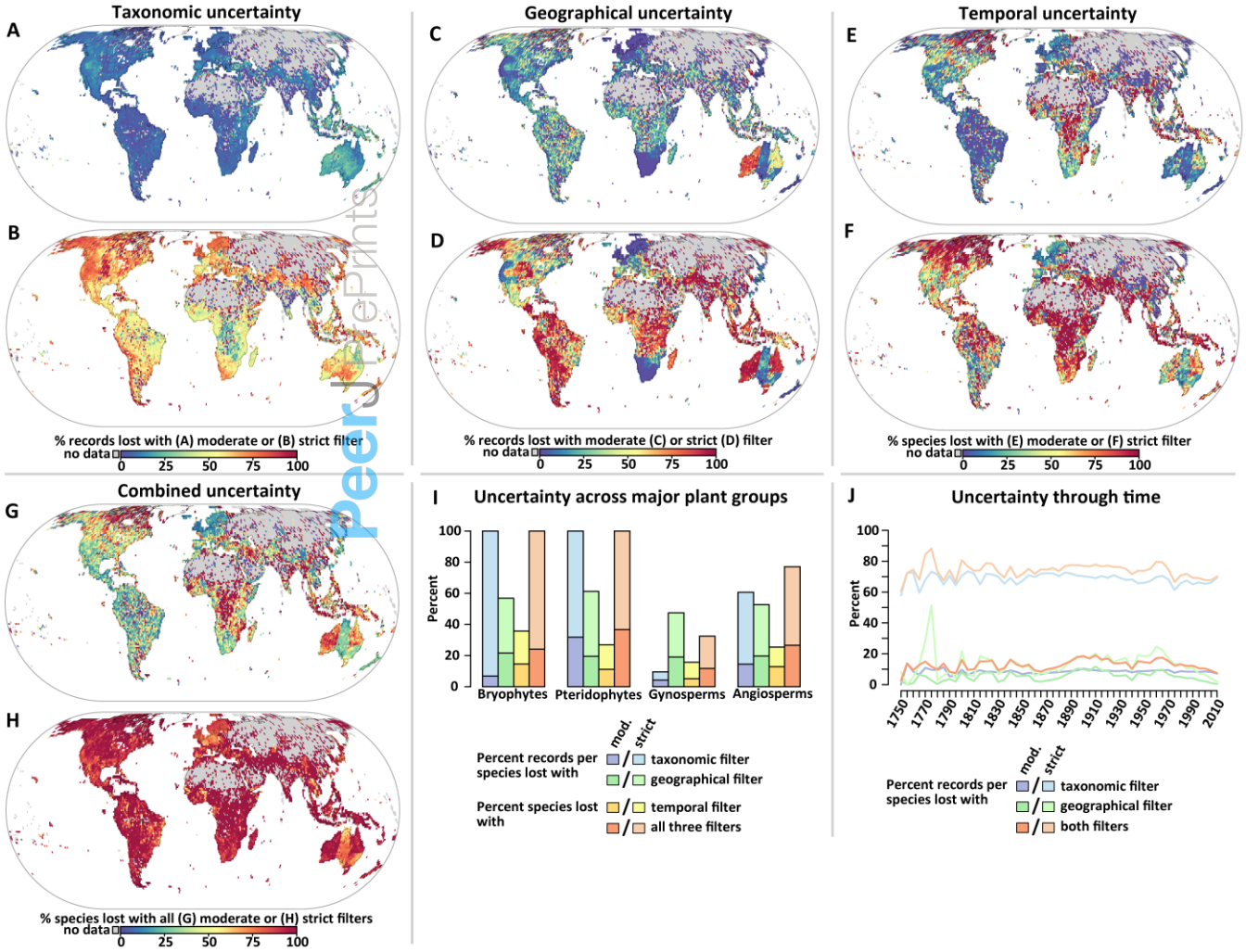

Figure 3. Global variation in occurrence information *uncertainty*. Geographical patterns across 12,100 km grid cells of percentages of records excluded by A) moderate and B) strict *taxonomic uncertainty* filtering, by C) moderate and D) strict *geographical uncertainty* filtering; geographical patterns 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 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 five-year periods between 1750 and 2010 of percentages of records excluded by taxonomic, geographical and the two combined filters.

PeerJ PrePrints

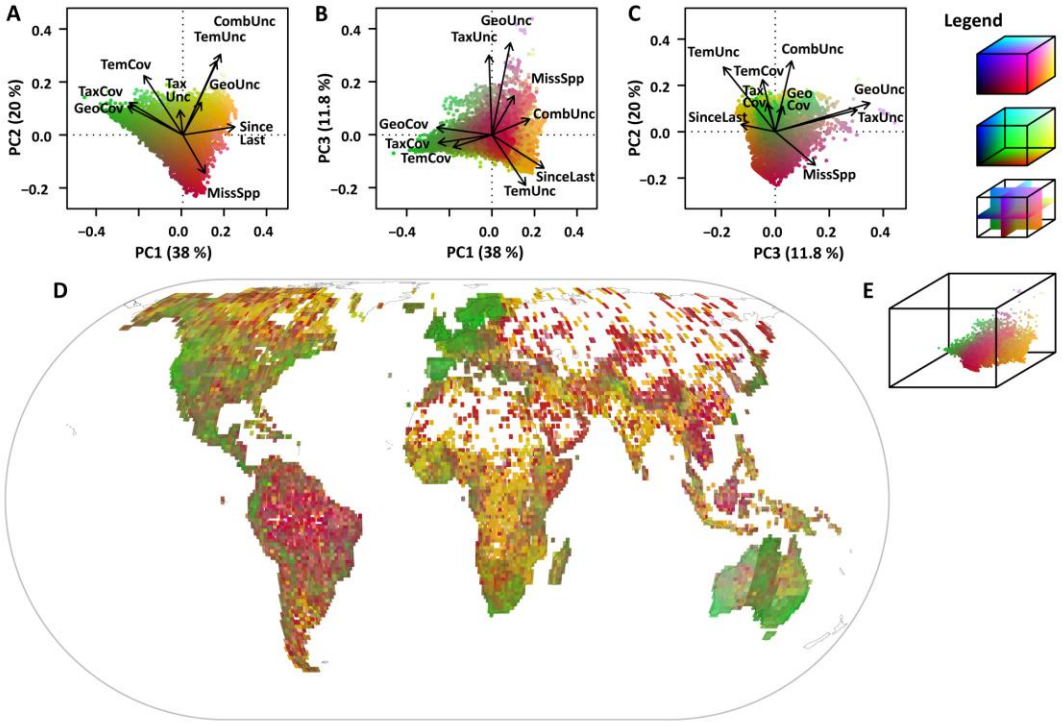

Figure 4. Principal component analysis (PCA) of 9 metrics of plant occurrence information across 12,100 km grid cells with ≥1 record. A-C) Biplots of the first three PCA axes. D) Global map of ordination site scores; similar colors denote regions characterized by similar information metrics. Colors refer to a red–green–blue (RGB) color space (legend) projected onto the E) 3D PCA space (Weigelt *et al.* 2013). **TaxCov**: *taxonomic coverage* (recorded/modeled 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 closests 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.

PeerJ PrePrints

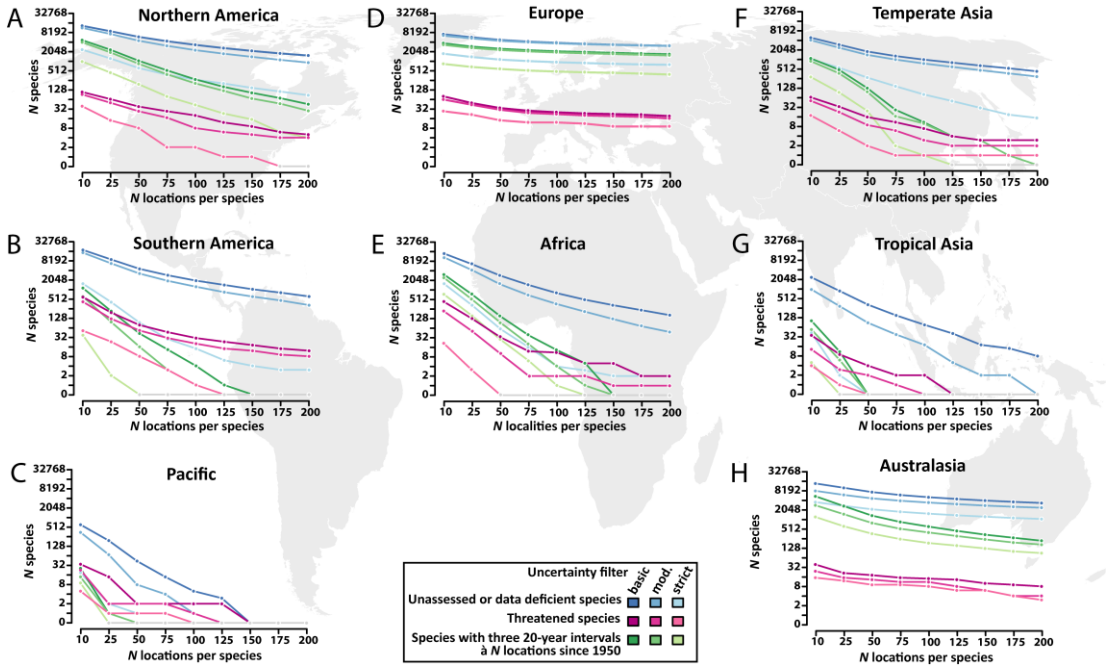

Figure 5. Global trade-offs between plant occurrence information *coverage* and *uncertainty*. Shown are numbers of species whose distributions could be estimated with hypothetical methods, depending on those methods' minimum requirements (10 to 200 sampling locations; Rivers *et al.* (2011); Feeley & Silman (2011a)) 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 colors: species that are either un-assessed or 'Data Deficient' on the International Union for the Conservation of Nature's Red List (2014). Violet colors: species with Red List categories 'Vulnerable', 'Endangered' or 'Critically Endangered'. Green colors: species for which the indicated number of sampling locations exists in each of three twenty-year periods since 1950. Color 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).

PeerJ PrePrints

Supplementary information
Multidimensional biases, gaps and uncertainties in global plant occurrence information

Carsten Meyer, Patrick Weigelt and Holger Kreft

SI 1. Treatment of taxonomic information.

The datasets downloaded via GBIF contained 119,058,280 raw records (Fig. S1A). We first cleaned verbatim scientific names strings (Fig. S1B), by excluding name strings that would not be reliably linkable to accepted species names. For instance, we excluded records that were not identified to species level (e.g. 'sp. nov.' or '*Sorbus* sp.', 'ined.', etc.) or where it was implied that the species identification was doubtful (e.g. 'cf.', 'aff.', 'à confirmer', '*Sorbus* ?*arnoldiana*', '*Oxyanthus* sp. possibly *unilocular*', etc.). We further excluded hybrids and cultivated forms (e.g. 'x', '<->', 'hybr.', 'hort.', 'cult.', 'var. "Ballerina"', etc.). We corrected wrong capitalizations of letters, and removed random punctuations and signs. These steps reduced 2,206,831 verbatim name strings to 1,552,901 interpretable names, including accepted species and subspecies names, synonyms, and spelling variants with or without author information.

We then performed the taxonomic standardization and validation. The basis for our taxonomic treatment was the comprehensive taxonomic information provided via *The Plant List* (TPL 2014) and iPlant's *Taxonomic Name Resolution Service* (TNRS 2014). In cases of conflicting information, we gave TPL priority. First, we compared genus names against genus names listed in TPL or TNRS. Then, we corrected misspelled genus names where we were confident regarding the true genus (doubtful cases were excluded). We then compared each name string to all possible scientific names listed under that genus in TPL. For each resulting pair of verbatim name and TPL-listed scientific name, we calculated the orthographic distance between species epithets and between the entire name strings (e.g. including author information), using an approximate string matching algorithm (generalized Levenshtein distance; using the adist function in R). This algorithm counts the total number of changes that have to be applied to one string in order to match another, and related that number to the entire length of the string. We then linked verbatim names via the best-matching TPL-listed name to the respective accepted species. For names that could not be matched to TPL-listed names or were not resolved to accepted species, we repeated these steps using taxonomic information from TNRS. We excluded all verbatim names that did not match names treated by TPL or TNRS as accepted names with no more than 25% orthographic distance, either directly or through a synonym. Overall, cleaning and validation led to an exclusion of 242,043 verbatim names strings (Fig. S1E); All remaining 1,964,788 verbatim name strings (89%) converged to 367,703 accepted species. These were further reduced to 229,218 accepted species (Fig. S1I) by applying our basic geographical filter (see Methods).

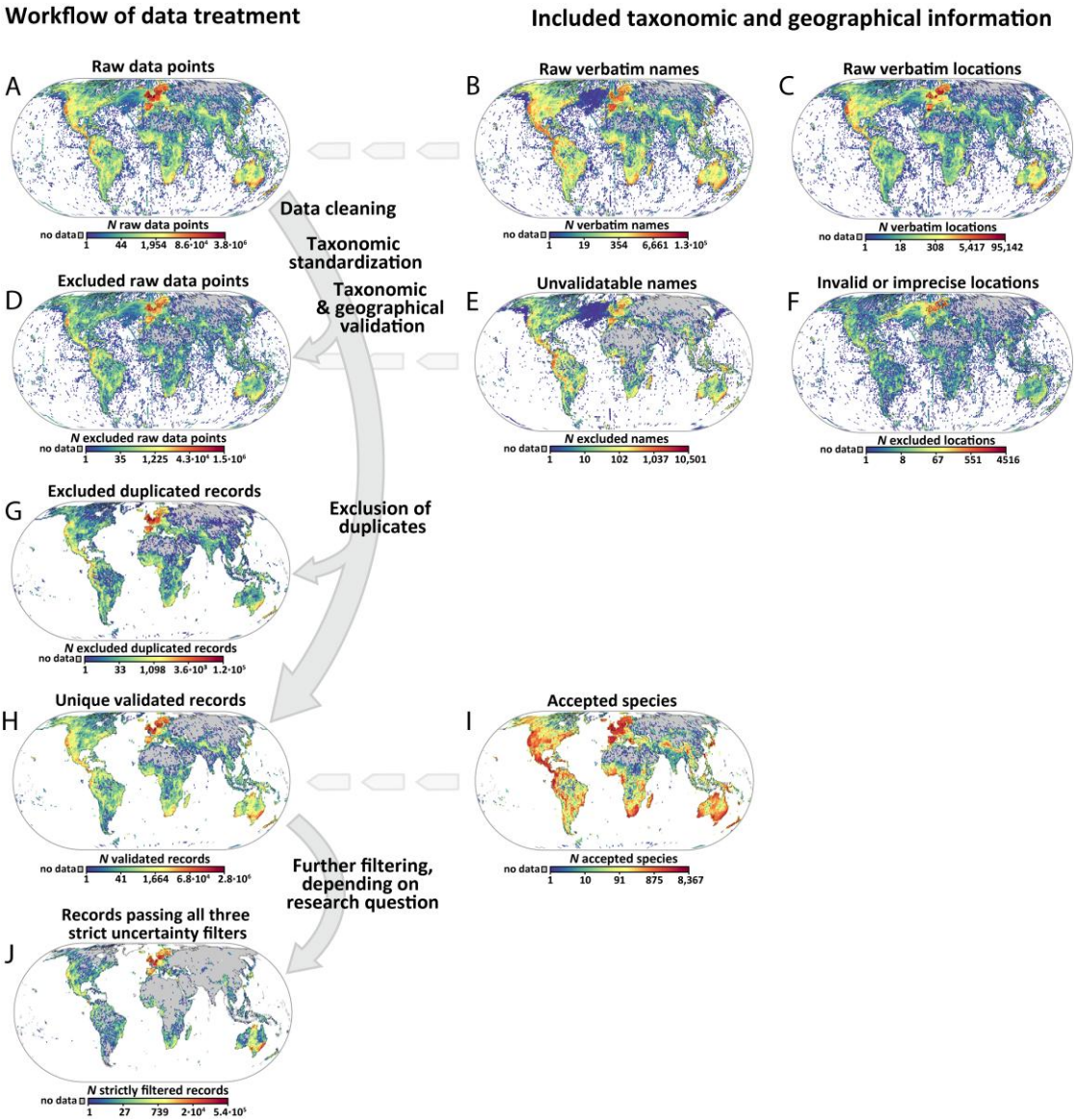

Figure S1. Workflow from raw mobilized data to usable occurrence records. Maps show spread of occurrence information for land plants, as mobilized via the Global Biodiversity Information Facility (GBIF), across 12,100 km grid cells. We retrieved (A) 119,058,280 raw records via GBIF, including (B) 2,206,831 verbatim name strings and (C) 4,314,752 verbatim coordinate combinations. Data cleaning, taxonomic standardization and taxonomic and geographical validation led to the exclusion of (D) 38,228,372 raw records, including (E) 242,043 unvalidatable name strings and (F) 252,550 invalid or imprecise coordinate combinations. Remaining validated records were reduced to unique records, which led to the exclusion of (G) 24,899,514 duplicated species-location-year-month-combinations and left (H) 55,929,317 unique validated records including (I) 229,218 accepted species. Depending on research question, further filtering might be necessary; e.g., applying our strict taxonomic, geographical and temporal filters (see *Methods*) would leave (J) 9,295,847 strictly filtered records. For details, see *Methods* and *SI 1*.

PeerJ PrePrints

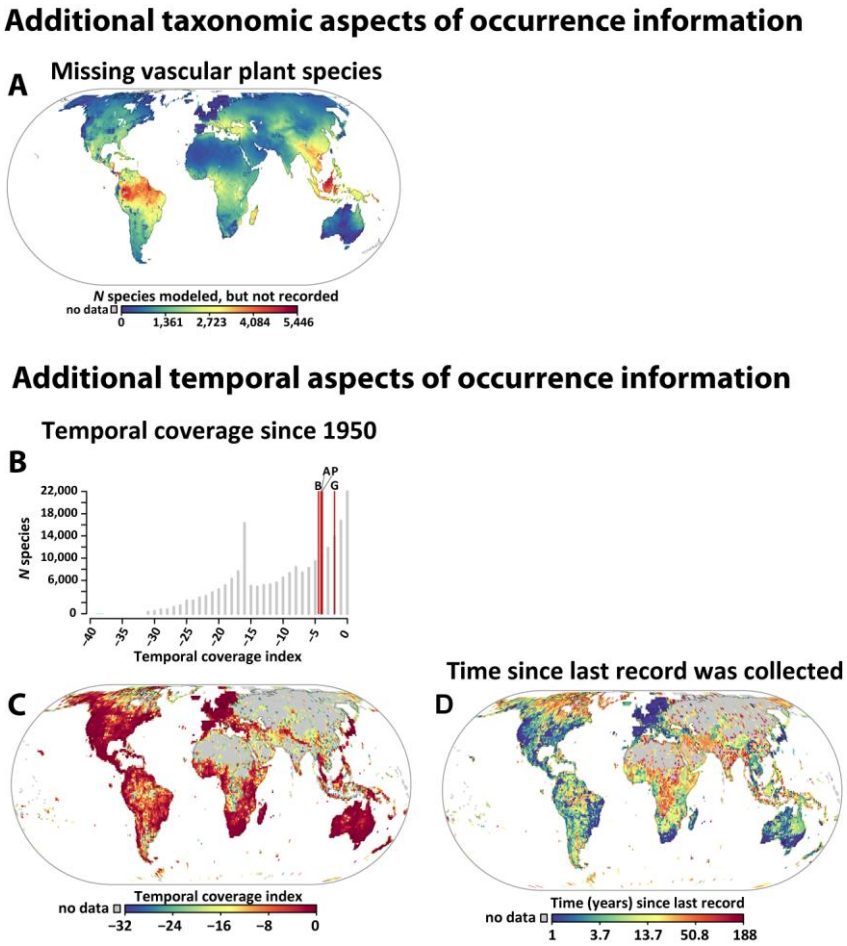

Figure S2. Non-*covered* species, *temporal coverage* of recent decades and record age in global DAI of plant occurrences. A) Geographical variation across 12,100 km grid cells in that portion of modeled vascular plant richness (Kreft & Jetz 2007) that was missing from mobilized occurrence information. B) Frequency distribution across land plant species in scores of *temporal coverage* since 1950, calculated as the mean minimum Euclidean distance between all possible months between 1950 and 2010 to their respective closest months with records. C) Geographical variation in *temporal coverage* since 1950. D) Geographical variation in the time (in years) since the last mobilized record has been collected.

PeerJ PrePrints

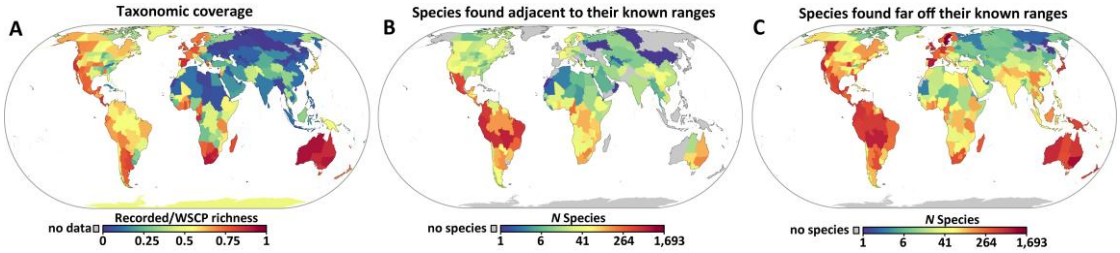

Figure S3. *Taxonomic coverage* of native and non-native species for selected families of seed plants. Species records for a subset of the global seed plant flora (105,031 species, c. 34% of all) were geographically validated against 'botanical country' checklists sourced from the World Checklist of Selected Plant Families (WCSP, 2013)). A) *Taxonomic coverage* of native seed plant species of selected families, based on geographically validated records. B-C) Number of species represented by occurrence records outside their known native ranges: B) Species recorded immediately adjacent to their native ranges; C) Species recorded far off their native ranges. Botanical countries are level-3 regions of the Biodiversity Information Standards (TDWG). Color scales are the same in B-C.

PeerJ PrePrints

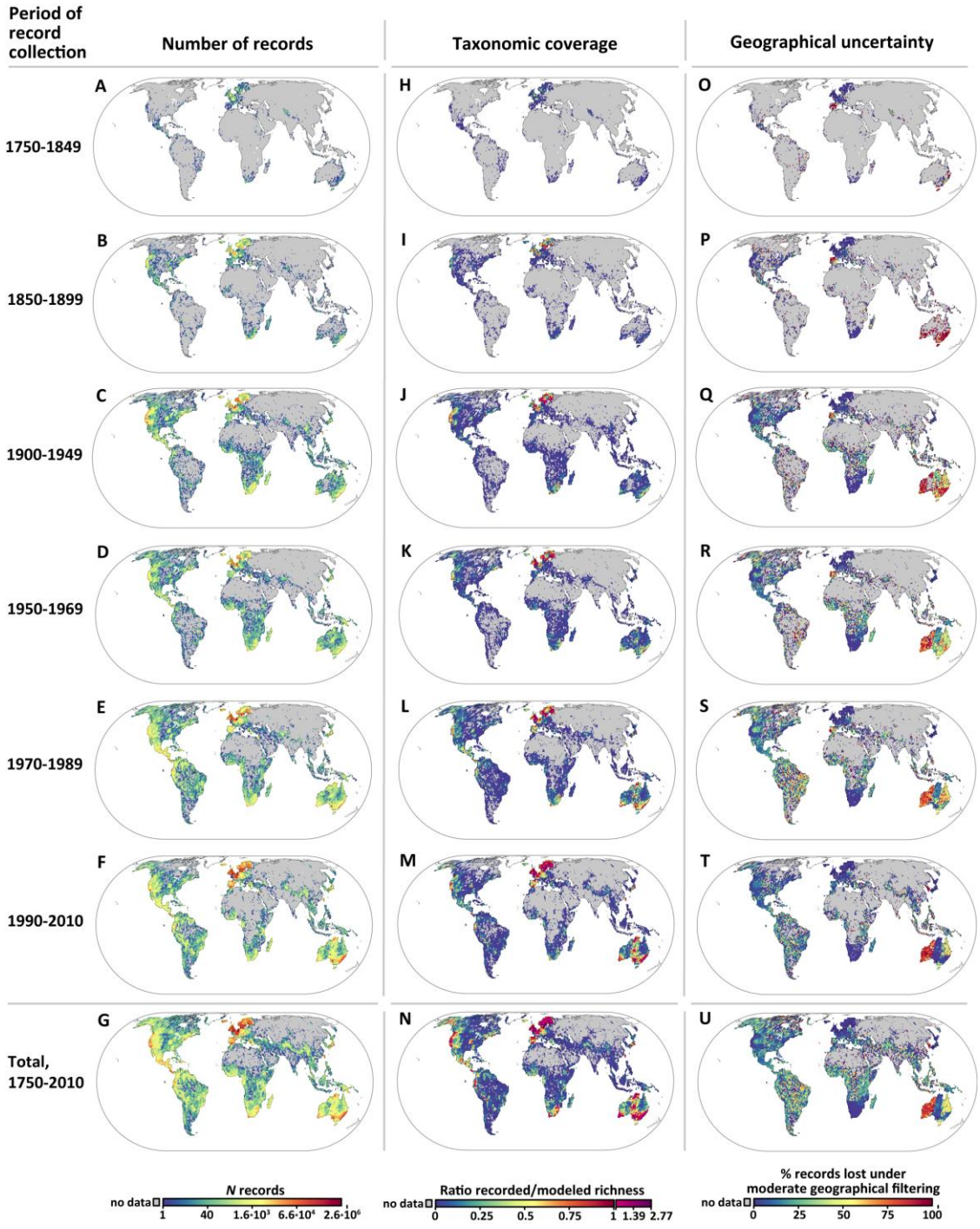

Figure S4. Spatio-temporal patterns in digital accessible occurrence information. Maps show geographical patterns of three exemplary aspects of vascular plant occurrence information across five time periods between 1750 and 2010 in, and for the entire time span. (A-G) Record number; (H-N) *Taxonomic coverage* for vascular plants (recorded richness (GBIF) / modeled richness (Kreft & Jetz, 2007)); values >1 mean larger recorded than modeled richness; note that mobilized records include non-native species whereas the model predicts native species richness; (O-U) Percentages of records excluded by moderate *geographical uncertainty* filtering (see *Methods*). Color scales are the same in (A-F), (G-L), (M-R).

PeerJ PrePrints

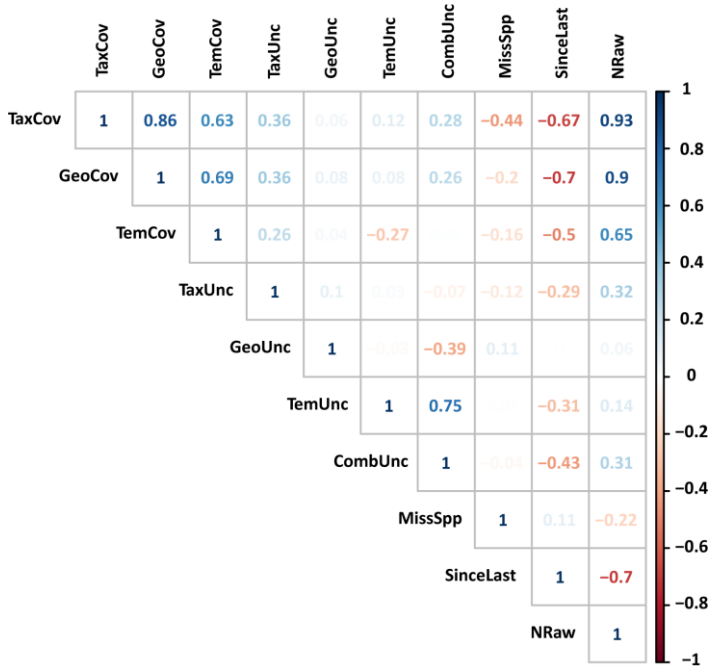

Figure S5. Relationships between 9 metrics of occurrence information and the number of raw data. Pairwise Spearman-rank correlations between geographical patterns of different occurrence information metrics at the level of 12,100 km grid cells. **TaxCov**: *taxonomic coverage*, calculated as the ratio between recorded richness and richness modeled by (Kreft & Jetz 2007); **GeoCov**: *geographical coverage*, estimated as the number of sampling locations per 10^4 km² land area; **TempCov**: *temporal coverage* since 1750, estimated as the mean minimum Euclidean distance between all possible months between 1750 and 2010 to their respective closests month with records; **TaxUnc**: percentage of records lost under moderate taxonomic filtering; **GeoUnc**: percentage of records lost under moderate geographical filtering; **TempUnc**: percentage of species lost under moderate temporal filtering; **CombUnc**: percentage of species lost with all three moderate filters applied (see *Methods* for information on filters). **MissSpp**: number of species that are not recorded but expected based on the environment-richness; **SinceLast**: Time (in years) since the last mobilized record was recorded. **NRaw:** number of raw data mobilized via GBIF, included to test whether this simple surrogate is a good indicator of different occurrence information metrics. All correlations based on z-transformed variables.

PeerJ PrePrints