Type of paper: ***Concepts***

# Big data for forecasting global change impacts on plant communities

Janet Franklin[1]

Josep M. Serra-Diaz[1,2]

Alexandra D. Syphard[3]

Helen M. Regan[4]

[1]School of Geographical Sciences and Urban Planning, Arizona State University, PO Box 875302, Tempe, AZ 85287-5302 USA. E-mail: janet.franklin@asu.edu

[2]Harvard Forest, Harvard University, 324 N. Main St., Petersham, MA 01366

USA. E-mail: serradiaz@fas.harvard.edu

[3]Conservation Biology Institute, 10423 Sierra Vista Ave., La Mesa, CA 91941 USA. E-mail: asyphard@yahoo.com

[4]Department of Biology, University of California, Riverside, CA 92521 USA. E-mail: helen.regan@ucr.edu

**Keywords**: database; environmental maps; geospatial; informatics; remote sensing; species occurrences; uncertainty; vegetation

**THIS IS THE FINAL VERSION OF THE PAPER PUBLISHED AS:**

Franklin, J., Serra-Diaz, J.M., Syphard, A.D., Regan, H.M., 2017, Big data for forecasting global change impacts on plant communities. *Global Ecology and Biogeography* 26(1):6-17. DOI: 10.1111/geb.12501

## ABSTRACT

**Aim** Plant distributions and vegetation dynamics underpin key global phenomena including biogeochemical cycling, ecosystem productivity and terrestrial biodiversity patterns. Aggregated and remotely collected 'big' data are required to forecast global change effects on plant communities. We synthesize advances in developing and exploiting big data in global change plant ecology, and identify challenges to their effective use in global change studies.

**Location** Global.

**Methods** We explored databases, catalogs and registries with respect to their accessibility, geographical and taxonomic extent, sample bias, and other types of uncertainty, from both a user and contributor perspective. We identified four kinds of big data needed to predict global change impacts on plant populations and communities using spatially explicit models: remotely sensed and other environmental maps, species occurrence records, community composition (plots), and species traits, especially demographic.

**Results** Digital environmental maps, including remotely sensed data, is the most mature class of big data discussed herein whereby protocols for archiving, discovering, and analyzing them have developed over three decades. Species locality records are being aggregated into databases that are easy to search and access, and while methods for addressing uncertainties are a major research focus, better spatial representation is still needed. Plot data from inventories have tremendous potential for monitoring and modeling global change impacts on plant communities but tend to be restricted to forests or concentrated in certain geographical areas. Ongoing efforts to aggregate plot and trait data from multiple sources are challenged by their heterogeneous coverage, attributes and protocols, and lack of data standards.

**Main Conclusions** Future goals include developing systematic frameworks for selecting geospatial data, improving tools for assessing the quality of species occurrence records, and, increased aggregation and discoverability of plot and trait data. Aggregated data collected by scientists, not sensors, provide more meaningful insights when data collectors are involved in analysis.

## INTRODUCTION

Big data refers to data sets so large or complex that they challenge established methods for data capture, curation, storage, analysis, and transfer (Lynch, 2008). Big data have been used to study vegetation distributions and dynamics for more than three decades, but recent and rapid global change, as well as new avenues for interdisciplinary research, increase dependence on them. Remotely sensed data have been large in volume since the dawn of the earth resources remote sensing era in the 1970s. In the 1980s, geographers coped with the challenges of developing nascent Geographic Information Systems that could handle NASA's MODIS-mission with its terabyte streams of data (Smith *et al.*, 1987), and by the 2000's, terabytes had become petabytes.

In addition to data that are big in volume, big data also refers to a type of data analysis, called predictive analytics, often applied to data derived from modern digital technology, such as internet searches and social media (Lazer *et al.*, 2014). Predictive analytics encompasses approaches used for decades in environmental and geographical sciences such as machine learning, statistical learning and data mining (Skidmore *et al.*, 2011). With the field of biodiversity informatics now well established (Soberon & Peterson, 2004), new directions in data analytics, and a culture of bringing multiple disciplines and approaches to bear on 'wicked' problems, the time is ripe for plant ecologists to adopt evolving techniques that will go far beyond computer science departments in order to address pressing global change research problems.

Understanding plant distributions and vegetation dynamics driven by global change requires multiple lines of evidence and integrated frameworks that link environmental variables to ecological processes and it relies on data from many sources (Franklin *et al.*, 2016): 1) environmental maps, 2) biodiversity (species occurrence) records and maps, 3) plant community composition data (vegetation plots), and 4) species demographic parameters, as well as other species traits. As we have confronted in our own work, all of these data types are necessary to predict separate and combined impacts of major global change drivers – climate change, land use change, altered disturbance regimes and invasive species – on plant communities at large spatial extents (Franklin, 2010b; Regan *et al.*, 2012; Syphard *et al.*, 2013; Franklin *et al.*, 2014). Not all big data used in plant ecology—sourced from pixels vs. polygons vs. plots vs. pressed plants—are the same; different data are subject to different pre- and post-processing techniques and uncertainties (Regan *et al.*, 2002). We highlight the challenges faced when producing, using and integrating big data sets to monitor and forecast global change impacts on vegetation.

In the following sections we outline these four types of big data needed to address global change questions in plant ecology, identify key uncertainties, discuss challenges associated with contributing to big data, and suggest ways to improve the discovery, screening and use of big data to inform conservation and planning in a rapidly changing world.

## BIG PLANT DATA FOR GLOBAL CHANGE ECOLOGY

### Digital Environmental Maps

We highlight selected types of environmental maps that are needed to monitor or forecast global change effects on vegetation. These include satellite-measured vegetation properties, thematic maps of vegetation, soils and land cover, digital terrain data, and climate maps. A comprehensive discussion of geospatial data and analytics used in plant ecology is beyond the scope of this paper and reviewed elsewhere (Wilson & Gallant, 2000; Kerr & Ostrovsky, 2003; Franklin, 2010a; Skidmore *et al.*, 2011; Thrasher *et al.*, 2013).

Global products from the National Aeronautics and Space Administration's (NASA) Moderate Resolution (250 m or coarser) Imaging Spectrometer (MODIS) data (Justice *et al.*, 1998), available beginning in 1999, include maps of land cover (Friedl *et al.*, 2010), leaf area index, and vegetation indices. A global map of forest cover change (2000-2013) has recently developed from high-resolution 30-m Landsat imagery (Hansen *et al.*, 2013). New developments in remote sensing show great promise for filling information gaps in understanding vegetation properties and processes including biodiversity, biogeochemical cycling and community dynamics (Schimel *et al.*, 2013; Schimel *et al.*, 2015; Shugart *et al.*, 2015).

Although these data products support monitoring of earth system properties, studies forecasting global change effects on plant communities are often carried out at regional scales and rely on topographic, soil, vegetation type, and land cover maps to characterize baseline conditions. For example, a landscape simulation model of vegetation disturbance and succession (Franklin *et al.*, 2005), relying on statistical models of baseline distributions for focal species (Franklin, 2002), required digital soil maps, vegetation type maps, elevation models, interpolated climate and modeled potential solar radiation and topographic moisture (Franklin, 1998). Mapped projections of future land use change (Syphard *et al.*, 2005) and climate change (Scheller & Mladenoff, 2005) are also needed to forecast vegetation dynamics under global change scenarios.

Use of climate maps for projecting the effects of climate change on ecosystems has been growing. Historical climate, e.g., monthly temperature and precipitation, has been mapped by spatial interpolation of weather station records at continental and global scales (Hijmans *et al.*, 2005; Daly, 2006), and these datasets are widely used in ecological modeling. Earth's future climate is modeled using general circulation models of the oceans and atmosphere, also called global climate models (GCMs). These global simulations are conducted at a spatial resolution (1 degree cells or larger) that is too coarse for most studies of plant community response (a case of underspecification; Table 1), and are thus spatially downscaled to incorporate regional and local scale climate variability (Beaumont *et al.*, 2008). Statistical downscaling uses either the delta method (calculating the difference between a future and baseline modeled climate and adding that difference to a high resolution current climate map) or pattern scaling (also called weather typing) which yields more accurate climate maps than the delta method (Flint & Flint, 2012). Nevertheless, dynamical downscaling through nesting a regional climate model within the

simulations from a global model may be required to capture regional processes affecting vegetation patterns (Hall, 2014). Uncertainty in climate models has been scrutinized in detail because of the profound implications of their forecasts. Gridded data from downscaled climate models can include hourly, daily or monthly raster maps of temperature, precipitation, humidity, and wind speed, and can quickly grow to hundreds of gigabytes.

Digital environmental maps are the big data with which the global change research community has the most experience (30-40 years). A fundamental difference between digital environmental maps and species, plot or trait data aggregated from multiple sources (discussed in the following sections) is that environmental maps are usually produced 'top down' from a uniform data source and modeling procedure. Remotely sensed imagery, the raw data used for many environmental maps, may have a large number of bands (of the electromagnetic spectrum) and pixels, but each observation has well characterized properties. All spatial data are subject to uncertainty (Table 1), but characterizing this uncertainty may be more feasible for remotely sensed data than it is for aggregated datasets. Indeed, great effort has been put into validation and error assessment of geospatial data (Hunter & Goodchild, 1997; Pontius, 2000; Tian *et al.*, 2002), which also have well-established standards for metadata (e.g., ISO 19115-2:2009) and data exchange (http://www.opengeospatial.org/docs/is).

The challenge to ecologists requiring digital environmental maps for evaluating global change impacts on plant communities is not screening individual observations but rather determining whether the maps are fit for purpose in terms of the variable represented (proximal versus distal to the response – for example, interpolated temperature versus elevation), spatial resolution, and temporal resolution. Analyses of vegetation patterns and dynamics are sensitive to the spatial and temporal scale of the input data (Syphard & Franklin, 2004; Franklin *et al.*, 2013; Serra-Diaz *et al.*, in press). Selecting the many types of geospatial data required for ecological research is generally done on a case-by-case basis, as part of the research design process. Frameworks (Phinn *et al.*, 2003) and guidelines (Kennedy *et al.*, 2009) for systematically identifying remotely sensed and other geospatial data and processing methods to address a particular problem can improve the performance of environmental models (Bennett *et al.*, 2013).

An example of a federated database designed to provide a single access point to many types of Earth observational data is DataOne (see also Michener, 2015), a recently developed "distributed cyberinfrastructure" (Appendix S1 in Supporting Information lists information about databases). Data are submitted to DataOne via nodes that include existing data warehouses such as the US Geological Survey (USGS) and Oak Ridge National Laboratory Distributed Archive (ORNL DAAC; data produced by NASA's Terrestrial Ecology Program). We have found that environmental data can be challenging to discover in DataOne, and in some cases it might be easier for analysts to go directly to familiar data providers (USGS, DAAC).

## Species Occurrence Data

Species occurrence data, along with digital environmental maps, are required for species distribution modeling (SDM) to forecast species range shifts (Thuiller, 2004; Franklin *et al.*, 2013; Serra-Diaz *et al.*, 2014) under global change scenarios. Furthermore, when moving beyond correlative models to forecast ecological responses to global change (Franklin, 2010b) SDM is still required to set baseline conditions for process-based simulation models of plant population (Keith *et al.*, 2008) and community dynamics (Syphard *et al.*, 2007). Biodiversity records of species localities and ranges are digitized and georeferenced by natural history museums and herbaria (Graham *et al.*, 2004), as well as conservation organizations, research agencies, and citizen science projects, and aggregated into regional and global databases (Guralnick *et al.*, 2007). An extensive research methodology has developed, initially focused on specimen collection records, to address widespread uncertainty problems of locational underspecification, geocoding errors, and taxonomic changes, errors or ambiguity (Table 1), and to develop data standards (Thessen & Patterson, 2011).

Specimen locality records are particularly prone to errors in spatial coordinates (Belbin *et al.*, 2013) and analyses that use them are sensitive to these errors (Guisan & Rahbek, 2011). Tools have been developed for georeferencing (e.g., Guralnick *et al.*, 2006) and error screening (Hijmans *et al.*, 2012). Taxonomic misidentification and changes in nomenclature as well as changes in the concepts represented by species names (Franz & Peet, 2009; Wiser, 2016) also lead to data uncertainty for ecologists using aggregated species occurrence data (Graham *et al.*, 2004; Boyle *et al.*, 2013). The Plant List (http://www.theplantlist.org/) and Tropicos (http://www.tropicos.org/) are key sources of taxonomic information, and Taxonomic Name Resolution Service (TNRS; tnrs.iplantcollaborative.org/) is a widely used web-based tool for resolving plant names.

In the biodiversity informatics community, data aggregators are assembling and maintaining these archives, focusing on interoperability, workflows, licensing, engaging, discoverability, and exposing biological data. In response to a critique of taxonomic errors found in a biodiversity database (Mesibov, 2013), Belbin *et al.* (2013) asserted that publishing digital data reveals inherent problems; nevertheless, data aggregators would rather flag potential errors and expose the flag than delete data.

There are several sources of aggregated species occurrence data particularly relevant to plant ecology. The Global Biodiversity Information Facility (GBIF; Table 2, Appendix S1) is the largest and most widely known source of species records (Costello *et al.*, 2013). GBIF makes globally distributed databases of biological specimens and other species observations interoperable, with an emphasis on a comprehensive catalog of scientific names (Edwards *et al.*, 2000). Information systems are interoperable when they can communicate and exchange data. Institutions and organizations can publish their biodiversity databases through GBIF and thereby become part of a distributed data network (Guralnick *et al.*, 2007). Participants use the same data standards (Wieczorek *et al.*, 2012), and therefore searches yield results in a common format. The development of the Darwin Core data exchange standards (http://rs.tdwg.org/dwc/) was essential for the

large-scale aggregation and exchange of specimen data, enabling the development of GBIF and similar initiatives.

The total number of species records grew from about 550 to 640 million just between Jun 2015 and Jan 2016. At recent check (29 Jan 2016), there are almost 133 million observations of 719 thousand species in the Kingdom Plantae. GBIF includes growing numbers of direct observations (for plants these are often from plots, relevés or checklists) and fossil localities (see Table 2 for example). With growing heterogeneity of data sources, screening with respect to source is increasingly important to confirm that the observations are fit for purpose.

Other aggregators of plant occurrences include regional networks from which specimen locality data can be downloaded (see examples in Appendix S1). These consortia are often among those institutions feeding data to GBIF, and yet we have found that regional networks may include records that are not in GBIF (Fig. 1). Regional scale analyses should therefore be based on thorough exploration and screening of all available sources of species data.

Research-oriented organizations also aggregate and provide information about species distributions limited to certain taxonomic groups, ecosystems, or geographical regions to address specific research questions. For example, the Botanical Information and Ecology Network (BIEN) is a data portal for information about species occurrence records, plant traits, and phylogeny for all plant species in the New World (Appendix S1), with many specimen observations carefully screened from GBIF (http://bien.nceas.ucsb.edu) and others from well vetted sources. Discrete BIEN databases (version 2 and 3) were established at different time periods. BIEN data have been used to evaluate the effect of sampling bias on analyzing diversity patterns and their drivers from aggregated species data (Engemann *et al.*, 2015) and support the finding that lack of sampling in a region (Fig. 1) could not be completely overcome by bias correction methods.

Some conservation organizations and consortia provide web-based information about species distributions (Appendix S1). For example, the IUCN Red List of Threatened Species documents the conservation status of many species, and Encyclopedia of Life (eol.org) is rich in information about taxa; both can display distribution maps for some species, which are drawn from GBIF and other sources. The Group on Earth Observations Biodiversity Observation Network (GEO-BON) is a "global partnership to help collect, manage, analyze, and report data relating to the status of the world's biodiversity" (Scholes *et al.*, 2008). GEO-BON is developing Essential Biodiversity Variables (EBV) for monitoring, including species distributions and community composition (Pereira *et al.*, 2013). The members of GEO-BON, governments and organizations, are establishing a Global Earth Observing System of Systems (GEOSS), whose web-based data portal (geoportal.org) has links to GBIF as a source of species occurrence records. In summary, a growing number of regional and global biodiversity research and monitoring initiatives link to GBIF to populate their species occurrence data.

**Plant Community Data**

Plant community plot data (relevés, forest inventory) are used in large scale assessments of vegetation dynamics (reviewed in Franklin *et al.*, 2016; Wiser, 2016) because they often include information on species distribution (presence and absence), abundance (cover, basal area, density) and vegetation structure (height of strata). When sampling is repeated (e.g., forest inventory) these data also provide information about establishment, growth, survival and mortality. This structure and demographic information is required to identify large-scale trends in plant communities, including the effects of global change (Thomas *et al.*, 2010; Zhu *et al.*, 2014; Serra-Diaz *et al.*, 2015), and to establish parameters and initial conditions for simulating community dynamics under global change scenarios (Schumacher *et al.*, 2006; Scheller *et al.*, 2007; Keane *et al.*, 2013). Plot databases can be rich in information required for detecting and forecasting global change effects on vegetation, and so we discuss this data type in some detail.

Plant community data can be roughly divided into two types, databases collected using standard protocols, often by a single institution, versus those aggregated from multiple sources, often collected using non-standardized protocols. Examples of the former are forest inventories conducted by state and national agencies (Appendix S1). While historically established to inventory timber resources, forest inventory programs offer repeated monitoring, often on a decadal cycle. Plot locations are typically determined with a probability-designed sample, e.g., stratified systematic, and so these data do not suffer from the sampling bias common in opportunistic and aggregated observations. Plot data provide information on species presence, absence and co-occurrence. However, inventories are usually only conducted for forest vegetation, and data for non-tree plant species may be lacking. Another characteristic of these long-term inventory databases is that protocols tend to change through time. At each resurvey new measurements may be incorporated and others may be dropped, and users must be cognizant of the lineage of the database they are using.

Aggregating plot data from multiple sources raises a number of challenges, including measurement protocol and plot size differences (Otypková & Chytry, 2006), taxonomic name resolution (Cayuela *et al.*, 2012), and data ownership agreements (Zimmerman, 2008). Furthermore, plot datasets often have different data structures and component elements with different names and different meanings to these names. The need for a data exchange standard is a major impediment to establishing large vegetation data repositories. Wiser *et al*. (2011) proposed Veg-X as an exchange standard for plot-based plant community data that, if adopted, will facilitate truly big data analysis of vegetation in support of global change research.

Aggregated plot data can be described in three categories – those that support an administrative purpose, those that have been assembled to address a particular research objective, and data registries that contain information about datasets. An example of the first type is VegBank, (Appendix S1), an archive of plot data for quantitative classification in support of the US National Vegetation Classification System (Franklin *et al.*, 2012). The purpose of VegBank "is to allow plant ecologists to submit and share data[…] which will

provide a permanent record of plots which define communities." Because these data are accessible to anyone, they can yield information on community composition and vegetation structure needed for global change research on plant community dynamics. Individuals can contribute plot data but this requires some experience with database informatics; thus, most contributors have been agencies with a mandate to archive large sets of publically funded data. The European Vegetation Archive (EVA, Chytrý *et al.*, 2016) is a centralized database with more than 1 million vegetation plots and serves a similar purpose to VegBank, to support quantitative vegetation classification. EVA data can be requested for other research objectives (Appendix S1) and therefore is an important source of plant community data for global change studies.

A number of research networks host vegetation plot data focused on a particular region or question (Appendix S1). For example, there are several with overlapping objectives and complex history of data inheritance in the western hemisphere. SALVIAS aggregates local vegetation inventories, mainly for the New World tropics (Enquist & Boyle, 2012). BIEN 2 has aggregated species occurrence information from more than 300,000 New World vegetation plots (the vast majority from FIA, VegBank and SALVIAS); however, abundance data are not yet included (but are planned for the future). SALVIAS plots are openly available, while BIEN will become publically available once the research consortium has addressed its research objectives. Also in the western hemisphere, but more narrowly focused on Neotropical seasonally dry forests, DryFlor (Appendix S1) is a research and conservation network that has recently released an open-access database assembled from vegetation plots and inventories, with taxonomic ambiguities resolved, providing information of species presence and absence throughout this threatened biome (e.g., Fig. 1).

There are other significant efforts to assemble vegetation plot data to address large-scale research questions. sPlot is a research consortium compiling global vegetation plot data, linked to the plant trait database TRY (next section) in support of a global analysis of plant traits (Appendix S1). sPlot data can only be used by members of the consortium to address the global-scale research questions defined by the group. The Center for Tropical Forest Science–Forest Global Earth Observatory (CTFS-ForestGEO), is a global research network of large (10-50 ha) forest plots, where all trees are measured every 5 years, starting as early as 1981, using a standard protocol. Network-wide comparisons address global change effects on plant population and ecosystem processes (Anderson-Teixeira *et al.*, 2015).

As an alternative to an aggregated database, the Global Index of Vegetation-Plot Databases (GIVD) (Dengler *et al.*, 2011) is a data registry for vegetation plot data. Two hundred and thirty seven databases are registered (Appendix S1); only 18 are available online (and these include SALVIAS and FIA), but more are available on request (accessed 21 Feb 2016). Registered datasets range in size from very large, including the Dutch National Vegetation Database and US FIA (each more than a half a million plots), to very small, for example, 13 plots from Papua New Guinea. It is not possible to search by vegetation type (for example 'dry forest').

## Species Traits

Monitoring and modeling plant communities under global change requires information about plant traits, including morphological (e.g. life form), physiological (e.g. shade tolerance) and demographic (e.g. survival rates). A rapidly expanding area of plant ecology relying on big trait data is functional biogeography (Violle *et al.*, 2014). The largest and most widely used trait database available is TRY (Kattge *et al.*, 2011), assembled as part of an international effort to integrate plant trait data, including anatomical, morphological, biochemical, physiological and phenological characteristics of plants. We counted 1,103 traits and 103,829 species as of 1 Sep 2015 (Appendix S1). These large numbers may hide gaps — not all species are equally sampled in trait space or in geographical space. Sandel et al. (2015) reported that as of 4 Jun 2014 "about a quarter of the world's plant species are represented, and for those, the trait matrix is 1.5% filled. A well-studied trait, specific leaf area (SLA), has [information for] roughly 3.5% of global plant diversity."

It may not be coincidental that the name of the database – TRY, "not an acronym, rather an expression of sentiment" (Kattge *et al.*, 2011) – reflects key challenges of big data in plant ecology. Sparse and biased data are typical; beyond big, coverage and depth are needed. The community is implementing methods to resolve biases (Sandel *et al.*, 2015) and improve coverage via data imputation and machine learning techniques tailored to the characteristics of trait databases, such as trait correlations and phylogenetic signals (Schrodt *et al.*, 2015).

Furthermore, while TRY strives to assemble information on plant traits in support of functional ecology and biogeography, demographic parameters are not included among those traits, but are nonetheless crucial for projecting population dynamics under global change scenarios (e.g., Regan *et al.*, 2012). Although complete sets of demographic data for parameterizing population models are notoriously difficult to find for many plant species, a couple of databases exist that provide pre-constructed matrix population models or compile data useful for population model construction. COMPADRE version 3.0 (Salguero-Gómez *et al.*, 2015) is an open repository for plant demographic data collated into matrix population models (MPMs) in a standardized format (Appendix S1). MPMs are the most widely used population model structures for plant species as they can accommodate distinct growth stages and complex life histories (Crone *et al.*, 2011). The database contains the demographic and associated data available (as of 25 Sep 2015), regularly compiled from the literature by a dedicated digitization team following strict, documented protocols to ensure consistency and standardization. Individual MPMs are included for each season, year, study population and treatment. Extensive ancillary data are organized into seven general categories: taxonomy, plant architecture, source, details of the study, geolocation, and population model. Studies span six continents, the bulk of them from North America and Europe, with the vast majority focusing on herbaceous perennials. Limitations of the database for population modeling include the absence of information regarding density dependence and seed dispersal, plant responses to disturbances such as fire, and the lack of stochastic models (however the latter could be derived from individual matrices if multiple

years are included in the database). Hence, supplemental demographic and life history data from the literature or field studies are necessary in many cases to construct useful population models that can project impacts of global change.

The Global Population Dynamics Database (GPDD, NERC Centre for Population Biology, 2010) is a collection of population time series data for more than 5000 animal and plant species (Appendix S1); ninety percent of the data pertain to animals. Population data range from counts of population abundance to estimates of density or temporal coverage from annual to weekly to relative periods. Data are compiled from journal articles, books, online repositories, and unpublished datasets. The GPDD contains only time series that include at least 10 records, usually annual population counts from unmanipulated studies. The GPDD scores data quality through a qualitative ranking ranging from 1 (low) to 5 (high). As with COMPADRE, consistency of data entry and interpretation is achieved via a dedicated data entry team. Extensive life history and demographic data are not included, which limits this database's utility for population modeling; however, stage-based time series are presented for some species from which demographic parameters for structured population models could be calculated, and trends of populations and growth rates can be derived from the time series data. Density dependence is not explicitly included, nor is dispersal or response to disturbance. Hence, the time series extracted from the GPDD would also need to be supplemented with published or field data to construct models representing complex life histories. For these reasons, population models of herbaceous perennial plants may be the most amenable to population model construction with the time series extracted from the GPDD.

## CONTRIBUTING BIG PLANT ECOLOGY DATA

In addition to 'consuming' big data, ecologists are also increasingly encouraged or required to publish, register or archive original data used in scientific publications in support of reproducible science and data synthesis (Thessen & Patterson, 2011; Michener, 2015). For individual researchers with small datasets, however, it can be challenging to contribute data to aggregated archives because of the time involved in formatting data to meet specified standards (although this challenge has been overcome by COMPADRE, DryFlor and other initiatives that dedicate personnel for data discovery and entry).

Online appendices have become a common way to include data and other supplemental material in support of published research, but they are not subjected to the same peer review or editorial scrutiny as the paper itself. Nor are they permanently archived or openly available, but rather are subjected to the journal publishers' access restrictions (Costello *et al.*, 2013). As one example of the problems that can arise when supplemental data are not rigorously evaluated, a global database of physical and climate characteristics of ~18,000 islands was published as supplementary material (Weigelt *et al.*, 2013), but these data contained numerous errors of island names, rendering them unusable for the kinds of analyses they intended to support.

Alternatives to online supplemental material are repositories like Dryad and Pangaea (Appendix S1) that accept datasets associated with particular published papers. Dryad originated from the efforts of journals and scholarly societies in ecology and evolutionary biology to develop common policy for data archiving. The dataset is assigned a Digital Object Identifier (DOI), making the data citable. Dryad has a modest fee for submitting small datasets (Appendix S1) and is not suitable for very large datasets. Pangaea is another data publisher with an emphasis on Earth science. A search for 'pollen assemblage,' for example, yielded 5459 datasets (accessed 1 Jul 2015). While Dryad will publish data in any format, Pangaea converts data to a machine-independent format. Dryad and Pangaea are easy to use for depositing small data sets, but the very diverse kinds of data deposited, while they support reproducible science, do not support data synthesis. They do not integrate disparate datasets of a similar type into common formats.

Data Basin is an example of an online data sharing website that explicitly archives spatial datasets in addition to other types of ecological and environmental data (Appendix S1). Data Basin also supports a web-based mapping and collaboration platform with visualization, mapping, and analysis tools to support networks of scientists. All uploaded data require standard geographical metadata information, and the website rates the quality of the data based on whether they were peer-reviewed.

Kervin et al. (2013) conducted a case study of 53 data papers published in Ecological Society of America's *Ecological Archives* from 2004-2012; almost three quarters of the papers lacked sufficient descriptions of data collection methods, and half did not describe data checking and screening (quality control and assurance) procedures in detail. They concluded that these common errors make it difficult for data re-users to discover the data or judge their usefulness for analysis, thereby recommending additional archiving of scientific workflows and scripts (R, Python) with the metadata to provide a record of data transformations. We anticipate that effective data publishing and sharing will become universal in plant ecology as it becomes easier to do (Tenopir *et al.*, 2011).

Peters et al. (2014) propose a knowledge, learning, analysis system (KLAS) for ecology that borrows cyber-infrastructure (CI) concepts from other fields such as genomics. The proposed system would, in a Google-like fashion, learn from queries to the system, and become smarter over time at suggesting which data (from federated databases) and analytical tools the user needs to answer their question. They too recommend that analytical tools and derived data products should be archived by researchers and "discoverable" to future researchers. While many fields, including bioinformatics, economics, and geocomputation are clamoring for reproducible science (Lazer *et al.*, 2014), Peters et al. emphasize supporting subsequent research and collaboration over reproducibility (verification) of published work, and did not address the thorny issues of attribution, authorship and ownership.

## BIG CHALLENGES

Environmental maps represent diverse and complex types of big data needed to detect and predict global change effects on plant communities, but fortunately this data type is also the furthest along in terms of standards, metadata, and error models. Given the enormous volume and complexity of these data, a systematic framework for selecting geospatial data that are most appropriate for global change studies would be very useful.

Although it can be tempting to fully exploit the millions of biodiversity records from federated databases, ignorance of their limitations can lead to wrong conclusions about, e.g., patterns of species diversity. For example, range maps with erroneous records, or diversity maps based on stacked species models, overestimate richness (Dubuis *et al.*, 2011). Species occurrence data are particularly prone to observation error, sampling bias, and underspecification (Table 1, Regan *et al.*, 2002). Careful data screening is required to select data that are fit for purpose (e.g., Fig. 1, Table 2) and appropriate analytic tools for the types of uncertainty in the data. Data screening and cleaning (or scrubbing) of large and heterogeneous datasets is not trivial or for the feint of heart.

Vegetation plot data are prone to uncertainties due to vague concepts, ambiguity and measurement error (Table 1). However, vegetation inventories based on probability-designed samples overcome problems of spatial bias, and repeated measurements may provide demographic data in addition to data depicting structure and composition. Inventories are typically restricted to forest vegetation. Data registries or aggregated databases that are devoted to a particular data type help scientists narrow down and discover the data they need. Wide adoption of data exchange standards for vegetation plot data will allow community data to be more comprehensively aggregated and better serve global change research (Wiser, 2016).

Issues of attribution, authorship and ownership of contributed data are still being worked out by the ecological research community (Michener, 2015). Publishing biodiversity data would rewards data contributors and could improve data accuracy and usefulness (Costello *et al.*, 2013). A code of conduct protecting data and those who collect it will allow long-term ecological research to thrive in a new era of public data archiving (Mills *et al.*, 2015). Data attribution, along with continued improvements in biodiversity informatics infrastructure, may lead to more enthusiastic data sharing and data-driven discovery (Thessen & Patterson, 2011).

We encourage skepticism about the claim that, for big data analyses in ecology, the size of the dataset can overcome problems in individual data points that represent noise. If errors are not random but rather systematic, the data are biased, undermining the conclusions. Strategies are being developed for reducing the effect of bias on analyses using aggregated species occurrence (Phillips *et al.*, 2009; Beck *et al.*, 2014) and vegetation plot data (Lengyel *et al.*, 2011; Wiser & De Cáceres, 2013). Furthermore, revealing the data, in addition to supporting reproducible science, allows the community to identify and correct errors.

Research in plant ecology and other fields is still often carried out by small groups of collaborators using original data, and a lot of insightful research gets done in the context of

this framework. The pressing need for global change research on plant community dynamics necessitates interactions among interdisciplinary collaborators, using a mixture of original and mined data and models. Research networks, working groups, and data aggregators have developed to address this need. But legacy datasets still abound in plant ecology and have not yet been captured to their full potential in service of global change science. Research networks will promote solid science when those who designed the studies and collected the data, and are therefore familiar with nuances of the data, are involved in analyzing them and interpreting the results when they are aggregated into larger datasets.

**Acknowledgements**

This work was supported in part by National Science Foundation grants DEB-0824708 to JF and HMR, EF-1065826 to JF, DEB-1353301 and 2014-SGR-1491 to JMS-D, EF-1065864 to ADS and EF-1065753 to HMR. It was inspired by a symposium at the Ecological Society of America annual meeting (2015) on Global Ecology in the Era of Big Data, and we thank the organizers and participants for sharing their ideas and criticisms. The manuscript was improved by the comments of B. Boyle, S. Wiser and two anonymous reviewers.

**References**

Anderson-Teixeira, K.J., Davies, S.J., Bennett, A.C., Gonzalez‐Akre, E.B., Muller‐Landau, H.C., Joseph Wright, S., Abu Salim, K., Almeyda Zambrano, A.M., Alonso, A. & Baltzer, J.L. (2015) CTFS‐ForestGEO: a worldwide network monitoring forests in an era of global change. *Global Change Biology*, **21**, 528-549.

Beaumont, L.J., Hughes, L. & Pitman, A.J. (2008) Why is the choice of future climate scenarios for species distribution modelling important? *Ecology Letters*, **11**, 1135-1146.

Beck, J., Böller, M., Erhardt, A. & Schwanghart, W. (2014) Spatial bias in the GBIF database and its effect on modeling species' geographic distributions. *Ecological Informatics*, **19**, 10-15.

Belbin, L., Daly, J., Hirsch, T., Hobern, D. & La Salle, J. (2013) A specialist's audit of aggregated occurrence records: An 'aggregator's' perspective. *ZooKeys*, 67.

Bennett, N.D., Croke, B.F., Guariso, G., Guillaume, J.H., Hamilton, S.H., Jakeman, A.J., Marsili-Libelli, S., Newham, L.T., Norton, J.P. & Perrin, C. (2013) Characterising performance of environmental models. *Environmental Modelling & Software*, **40**, 1-20.

Boyle, B., Hopkins, N., Lu, Z., Garay, J.A.R., Mozzherin, D., Rees, T., Matasci, N., Narro, M.L., Piel, W.H. & Mckay, S.J. (2013) The taxonomic name resolution service: an online tool for automated standardization of plant names. *BMC Bioinformatics*, **14**, 16.

Cayuela, L., Granzow‐de la Cerda, Í., Albuquerque, F.S. & Golicher, D.J. (2012) Taxonstand: An R package for species names standardisation in vegetation databases. *Methods in Ecology and Evolution*, **3**, 1078-1083.

Chytrý, M., Hennekens, S.M., Jiménez‐Alfaro, B., Knollová, I., Dengler, J., Jansen, F., Landucci, F., Schaminée, J.H., Aćić, S. & Agrillo, E. (2016) European Vegetation Archive (EVA): an integrated database of European vegetation plots. *Applied Vegetation Science*, **19**, 173-180.

Costello, M.J., Michener, W.K., Gahegan, M., Zhang, Z.-Q. & Bourne, P.E. (2013) Biodiversity data should be published, cited, and peer reviewed. *Trends in Ecology & Evolution*, **28**, 454-461.

Crone, E.E., Menges, E.S., Ellis, M.M., Bell, T., Bierzychudek, P., Ehrlén, J., Kaye, T.N., Knight, T.M., Lesica, P. & Morris, W.F. (2011) How do plant ecologists use matrix population models? *Ecology Letters*, **14**, 1-8.

Daly, C. (2006) Guidelines for assessing the suitability of spatial climate data sets. *International Journal of Climatology*, **26**, 707-721.

Dengler, J., Jansen, F., Glockler, F., Peet, R.K., De Cáceres, M., Chytry, M., Ewald, J., Oldeland, J., Lopez-Gonzalez, G., Finckh, M., Mucina, L., Rodwell, J.S., Schaminee, J.H.J. & Spencer, N. (2011) The Global Index of Vegetation Plot Databases (GIVD): a new resource for vegetation science. *Journal of Vegetation Science*, **22**, 582-597.

Dubuis, A., Pottier, J., Rion, V., Pellissier, L., Theurillat, J.P. & Guisan, A. (2011) Predicting spatial patterns of plant species richness: a comparison of direct macroecological and species stacking modelling approaches. *Diversity And Distributions*, **17**, 1122-1131.

Edwards, J.L., Lane, M.A. & Nielsen, E.S. (2000) Interoperability of biodiversity databases: biodiversity information on every desktop. *Science*, **289**, 2312-2314.

Elith, J., Burgman, M.A. & Regan, H.M. (2002) Mapping epistemic uncertainties and vague concepts in predictions of species distribution. *Ecological Modelling*, **157**, 313-329.

Engemann, K., Enquist, B.J., Sandel, B., Boyle, B., Jørgensen, P.M., Morueta‐Holme, N., Peet, R.K., Violle, C. & Svenning, J.C. (2015) Limited sampling hampers "big data" estimation of species richness in a tropical biodiversity hotspot. *Ecology and evolution*, **5**, 807-820.

Enquist, B. & Boyle, B. (2012) SALVIAS – the SALVIAS vegetation inventory database. *Biodiversity & Ecology*, **4**, 288.

Flint, A.L. & Flint, L.E. (2012) Downscaling future climate scenarios to fine scales for hydrologic and ecologic modeling and analysis. *Ecological Processes*, **1**

Franklin, J. (1998) Predicting the distribution of shrub species in southern California from climate and terrain-derived variables. *Journal of Vegetation Science*, **9**, 733-748.

Franklin, J. (2002) Enhancing a regional vegetation map with predictive models of dominant plant species in chaparral. *Applied Vegetation Science*, **5**, 135-146.

Franklin, J. (2010a) *Mapping species distributions: spatial inference and prediction*. Cambridge University Press, Cambridge, UK.

Franklin, J. (2010b) Moving beyond static species distribution models in support of conservation biogeography. *Diversity and Distributions*, **16**, 321-330.

Franklin, J., Regan, H.M. & Syphard, A.D. (2014) Linking spatially explicit species distribution and population models to plan for the persistence of plant species under global change. *Environmental Conservation*, **41**, 97-107.

Franklin, J., Syphard, A.D., He, H.S. & Mladenoff, D.J. (2005) The effects of altered fire regimes on patterns of plant succession in the foothills and mountains of southern California. *Ecosystems*, **8**, 885-898.

Franklin, J., Serra-Diaz, J.M., Syphard, A.D. & Regan, H.M. (2016) Global change and terrestrial plant community dynamics. *Proceedings of the National Academy of Science, USA*, **113**, 3725-3734.

Franklin, J., Davis, F.W., Ikagami, M., Syphard, A.D., Flint, A., Flint, L. & Hannah, L. (2013) Modeling plant species distributions under future climates: how fine-scale do climate models need to be? *Global Change Biology* **19**, 473-483.

Franklin, S., Faber-Langendoen, D., Jennings, M., Keeler-Wolf, T., Loucks, O., Peet, R., Roberts, D. & McKerrow, A. (2012) Building the United States National Vegetation Classification. *Annali di Botanica*, **2**, 1-9.

Franz, N. & Peet, R. (2009) Perspectives: Towards a language for mapping relationships among taxonomic concepts. *Systematics and Biodiversity*, **7**, 5-20.

Friedl, M.A., Sulla-Menashe, D., Tan, B., Schneider, A., Ramankutty, N., Sibley, A. & Huang, X. (2010) MODIS Collection 5 global land cover: Algorithm refinements and characterization of new datasets. *Remote Sensing of Environment*, **114**, 168-182.

Graham, C.H., Ferrier, S., Huettman, F., Moritz, C. & Peterson, A.T. (2004) New developments in museum-based informatics and applications in biodiversity analysis. *Trends in Ecology & Evolution*, **19**, 497-503.

Guisan, A. & Rahbek, C. (2011) SESAM–a new framework integrating macroecological and species distribution models for predicting spatio‐temporal patterns of species assemblages. *Journal of Biogeography*, **38**, 1433-1444.

Guralnick, R.P., Hill, A.W. & Lane, M. (2007) Towards a collaborative, global infrastructure for biodiversity assessment. *Ecology Letters*, **10**, 663-672.

Guralnick, R.P., Wieczorek, J., Beaman, R., Hijmans, R.J. & BioGeomancer Working Group (2006) BioGeomancer: automated georeferencing to map the world's biodiversity data. *Plos Biology*, **4**, e381.

Hall, A. (2014) Projecting regional change: How accurate are regional projections of climate change derived from downscaling global climate model results? *Science*, **346**, 1461-1462.

Hansen, M.C., Potapov, P.V., Moore, R., Hancher, M., Turubanova, S., Tyukavina, A., Thau, D., Stehman, S., Goetz, S. & Loveland, T. (2013) High-resolution global maps of 21st-century forest cover change. *Science*, **342**, 850-853.

Hijmans, R.J., Phillips, S., Leathwick, J. & Elith, J. (2012) dismo: Species distribution modeling. *R package version 0.7-17*,

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

Hunter, G.J. & Goodchild, M.F. (1997) Modeling the uncertainty of slope and aspect estimates derived from spatial databases. *Geographical Analysis*, **29**, 35-49.

Justice, C.O., Vermote, E., Townshend, J.R.G., Defries, R., Roy, D.P., Hall, D.K., Salomonson, V.V., Privette, J.L., Riggs, G., Strahler, A., Lucht, W., Myneni, R.B., Knyazikhin, Y., Running, S.W., Nemani, R.R., Wan, Z.M., Huete, A.R., van Leeuwen, W., Wolfe, R.E., Giglio, L., Muller, J.P., Lewis, P. & Barnsley, M.J. (1998) The Moderate Resolution Imaging Spectroradiometer (MODIS): Land remote sensing for global change research. *Ieee Transactions on Geoscience and Remote Sensing*, **36**, 1228-1249.

Kattge, J., Diaz, S., Lavorel, S., Prentice, I., Leadley, P., Bönisch, G., Garnier, E., Westoby, M., Reich, P.B. & Wright, I. (2011) TRY–a global database of plant traits. *Global Change Biology*, **17**, 2905-2935.

Keane, R.E., Cary, G.J., Flannigan, M.D., Parsons, R.A., Davies, I.D., King, K.J., Li, C., Bradstock, R.A. & Gill, M. (2013) Exploring the role of fire, succession, climate, and weather on landscape dynamics using comparative modeling. *Ecological Modelling*, **266**, 172-186.

Keith, D.A., Akçakaya, H.R., Thuiller, W., Midgley, G.F., Pearson, R.G., Phillips, S.J., Regan, H.M., Araujo, M.B. & Rebelo, T.G. (2008) Predicting extinction risks under climate

change: coupling stochastic population models with dynamic bioclimatic habitat models. *Biology Letters*, **4**, 560-563.

Kennedy, R.E., Townsend, P.A., Gross, J.E., Cohen, W.B., Bolstad, P., Wang, Y. & Adams, P. (2009) Remote sensing change detection tools for natural resource managers: Understanding concepts and tradeoffs in the design of landscape monitoring projects. *Remote Sensing of Environment*, **113**, 1382-1396.

Kerr, J.T. & Ostrovsky, M. (2003) From space to species: ecological applications for remote sensing. *Trends in Ecology & Evolution*, **16**, 299-305.

Kervin, K.E., Michener, W.K. & Cook, R.B. (2013) Common errors in ecological data sharing. *Journal of eScience Librarianship*, **2**, 1.

Lazer, D., Kennedy, R., King, G. & Vespignani, A. (2014) The parable of Google Flu: traps in big data analysis. *Science*, **343**

Lengyel, A., Chytrý, M. & Tichý, L. (2011) Heterogeneity‐constrained random resampling of phytosociological databases. *Journal of Vegetation Science*, **22**, 175-183.

Lynch, C. (2008) Big data: How do your data grow? *Nature*, **455**, 28-29.

Mesibov, R. (2013) A specialist's audit of aggregated occurrence records. *ZooKeys*, 1.

Michener, W.K. (2015) Ecological data sharing. *Ecological Informatics*, **29**, 33-44.

Mills, J.A., Teplitsky, C., Arroyo, B., Charmantier, A., Becker, P.H., Birkhead, T.R., Bize, P., Blumstein, D.T., Bonenfant, C. & Boutin, S. (2015) Archiving Primary Data: Solutions for Long-Term Studies. *Trends in Ecology & Evolution*, **30**, 581-589.

NERC Centre for Population Biology (2010) The Global Population Dynamics Database v2.0. http://www.sw.ic.ac.uk/cpb/cpb/gpdd.html.

Otypková, Z. & Chytry, M. (2006) Effects of plot size on the ordination of vegetation samples. *Journal of Vegetation Science*, **17**, 465-472.

Pereira, H.M., Ferrier, S., Walters, M., Geller, G., Jongman, R., Scholes, R., Bruford, M.W., Brummitt, N., Butchart, S. & Cardoso, A. (2013) Essential biodiversity variables. *Science*, **339**, 277-278.

Peters, D.P.C., Havstad, K.M., Cushing, J., Tweedie, C., Fuentes, O. & Villanueva-Rosales, N. (2014) Harnessing the power of big data: infusing the scientific method with machine learning to transform ecology. *Ecosphere*, **5**, art67.

Phillips, S.J., Dudik, M., Elith, J., Graham, C.H., Lehmann, A., Leathwick, J. & Ferrier, S. (2009) Sample selection bias and presence-only distribution models: implications for background and pseudo-absence data. *Ecological Applications*, **19**, 181-197.

Phinn, S.R., Stow, D.A., Franklin, J., Mertes, L.A.K. & Michaelsen, J. (2003) Remotely sensed data for ecosystem analyses: Combining hierarchy theory and scene models. *Environmental Management*, **31**, 429-441.

Pontius, R.G. (2000) Quantification error versus location error in comparison of categorical maps. *Photogrammetric Engineering and Remote Sensing*, **66**, 1011-1016.

Regan, H.M., Colyvan, M. & Burgman, M.A. (2002) A taxonomy and treatment of uncertainty for ecology and conservation biology. *Ecological Applications*, **12**, 618-628.

Regan, H.M., Syphard, A.D., Franklin, J., Swab, R.M., Markovchick, L., Flint, A.L., Flint, L.E. & Zedler, P.H. (2012) Evaluation of assisted colonization strategies under global change for a rare, fire-dependent plant. *Global Change Biology*, **18**, 936-947.

Salguero-Gómez, R., Jones, O.R., Archer, C.R., Buckley, Y.M., Che-Castaldo, J., Caswell, H., Hodgson, D., Scheuerlein, A., Conde, D.A., Brinks, E., de Buhr, H., Farack, C., Gottschalk, F., Hartmann, A., Henning, A., Hoppe, G., Römer, G., Runge, J., Ruoff, T., Wille, J., Zeh, S., Davison, R., Vieregg, D., Baudisch, A., Altwegg, R., Colchero, F., Dong, M., de Kroon, H., Lebreton, J.-D., Metcalf, C.J.E., Neel, M.M., Parker, I.M., Takada, T., Valverde, T., Vélez-Espino, L.A., Wardle, G.M., Franco, M. & Vaupel, J.W. (2015) The compadre Plant Matrix Database: an open online repository for plant demography. *Journal of Ecology*, **103**, 202-218.

Sandel, B., Gutiérrez, A.G., Reich, P.B., Schrodt, F., Dickie, J. & Kattge, J. (2015) Estimating the missing species bias in plant trait measurements. *Journal of Vegetation Science*, **26**, 828-838.

Scheller, R.M. & Mladenoff, D.J. (2005) A spatially interactive simulation of climate change, harvesting, wind, and tree species migration and projected changes to forest composition and biomass in northern Wisconsin, USA. *Global Change Biology*, **11**, 307-321.

Scheller, R.M., Domingo, J.B., Sturtevant, B.R., Williams, J.S., Rudy, A., Gustafson, E.J. & Mladenoff, D.J. (2007) Design, development, and application of LANDIS-II, a spatial landscape simulation model with flexible temporal and spatial resolution. *Ecological Modelling*, **201**, 409-419.

Schimel, D., Pavlick, R., Fisher, J.B., Asner, G.P., Saatchi, S., Townsend, P., Miller, C., Frankenberg, C., Hibbard, K. & Cox, P. (2015) Observing terrestrial ecosystems and the carbon cycle from space. *Global Change Biology*, **21**, 1762-1776.

Schimel, D.S., Asner, G.P. & Moorcroft, P. (2013) Observing changing ecological diversity in the Anthropocene. *Frontiers In Ecology And The Environment*, **11**, 129-137.

Scholes, R., Mace, G., Turner, W., Geller, G., Jürgens, N., Larigauderie, A., Muchoney, D., Walther, B. & Mooney, H. (2008) Toward a global biodiversity observing system. *Science*, **321**, 1044-1045.

Schrodt, F., Kattge, J., Shan, H., Fazayeli, F., Joswig, J., Banerjee, A., Reichstein, M., Bönisch, G., Díaz, S., Dickie, J., Gillison, A., Karpatne, A., Lavorel, S., Leadley, P., Wirth, C.B., Wright, I.J., Wright, S.J. & Reich, P.B. (2015) BHPMF – a hierarchical Bayesian approach to gap-filling and trait prediction for macroecology and functional biogeography. *Global Ecology and Biogeography*, n/a-n/a.

Schumacher, S., Reineking, B., Sibold, J. & Bugmann, H. (2006) Modeling the impact of climate and vegetation on fire regimes in mountain landscapes. *Landscape Ecology*, **21**, 539-554.

Serra-Diaz, J.M., Franklin, J., Dillon, W.W., Syphard, A.D., Davis, F.W. & Meentemeyer, R.K. (2015) California forest show early indications of both range shifts and local persistence under climate change. *Global Ecology & Biogeography*, **25**, 164-175.

Serra-Diaz, J.M., Franklin, J., Ninyerola, M., Davis, F.W., Syphard, A.D., Regan, H.M. & Ikegami, M. (2014) Bioclimatic velocity: the pace of species exposure to climate change. *Diversity and Distributions*, **20**, 169-180.

Serra-Diaz, J.M., Franklin, J., Sweet, L.C., McCullough, I., Syphard, A.D., Regan, H.M., Flint, L.E., Flint, A.L., Dingman, J.R., Moritz, M.A., Redmond, K., Hannah, L. & Davis, F.W. (in press) Averaged 30-year climate change projections mask opportunities for species establishment. *Ecography* doi: 10.1111/ecog.02074

Shugart, H.H., Asner, G.P., Fischer, R., Huth, A., Knapp, N., Le Toan, T. & Shuman, J.K. (2015) Computer and remote-sensing infrastructure to enhance large-scale testing of individual-based forest models. *Frontiers In Ecology And The Environment*, **13**, 503-511.

Skidmore, A.K., Franklin, J., Dawson, T.P. & Pilesjö, P. (2011) Geospatial tools address emerging issues in spatial ecology: A review and commentary on the Special Issue. *International Journal of Geographical Information Science*, **25**, 337-365.

Smith, T.R., Menon, S., Star, J.L. & Estes, J.E. (1987) Requirements and principles for the implementation and construction of large-scale geographic information systems. *International Journal of Geographical Information System*, **1**, 13-31.

Soberon, J. & Peterson, A.T. (2004) Biodiversity informatics: managing and applying primary biodiversity data. *Philosophical Transactions of the Royal Society of London Series B-Biological Sciences*, **359**, 689-698.

Syphard, A.D. & Franklin, J. (2004) Spatial aggregation effects on the simulation of landscape pattern and ecological processes in southern California plant communities. *Ecological Modelling*, **180**, 21-40.

Syphard, A.D., Clarke, K.C. & Franklin, J. (2005) Using cellular automaton model to forecast the effects of alternative scenarios of urban growth on habitat fragmentation in southern California. *Ecological Complexity*, **2**, 185-203.

Syphard, A.D., Clarke, K.C. & Franklin, J. (2007) Simulating fire frequency and urban growth in southern California coastal shrublands, USA. *Landscape Ecology*, **22**, 431-445.

Syphard, A.D., Regan, H.M., Franklin, J. & Swab, R. (2013) Does functional type vulnerability to multiple threats depend on spatial context in Mediterranean-climate regions? . *Diversity and Distributions*, **19**, 1263-1274.

Tenopir, C., Allard, S., Douglass, K., Aydinoglu, A.U., Wu, L., Read, E., Manoff, M. & Frame, M. (2011) Data Sharing by Scientists: Practices and Perceptions. *Plos One*, **6**, e21101.

Thessen, A.E. & Patterson, D.J. (2011) Data issues in the life sciences. *ZooKeys*, 15.

Thomas, R.Q., Canham, C.D., Weathers, K.C. & Goodale, C.L. (2010) Increased tree carbon storage in response to nitrogen deposition in the US. *Nature Geoscience*, **3**, 13-17.

Thrasher, B., Xiong, J., Wang, W., Melton, F., Michaelis, A. & Nemani, R. (2013) Downscaled climate projections suitable for resource management. *Eos, Transactions American Geophysical Union*, **94**, 321-323.

Thuiller, W. (2004) Patterns and uncertainties of species' range shifts under climate change. *Global Change Biology*, **10**, 2020-2027.

Tian, Y., Woodcock, C.E., Wang, Y., Privette, J.L., Shabanov, N.V., Zhou, L., Zhang, Y., Buermann, W., Dong, J. & Veikkanen, B. (2002) Multiscale analysis and validation of the MODIS LAI product: I. Uncertainty assessment. *Remote Sensing of Environment*, **83**, 414-430.

Violle, C., Reich, P.B., Pacala, S.W., Enquist, B.J. & Kattge, J. (2014) The emergence and promise of functional biogeography. *Proceedings of the National Academy of Sciences*, **111**, 13690-13696.

Weigelt, P., Jetz, W. & Kreft, H. (2013) Bioclimatic and physical characterization of the world's islands. *Proceedings of the National Academy of Sciences*, **110**, 15307-15312.

Wieczorek, J., Bloom, D., Guralnick, R., Blum, S., Döring, M., Giovanni, R., Robertson, T. & Vieglais, D. (2012) Darwin Core: An evolving community-developed biodiversity data standard. *Plos One*, **7**, e29715.

Wilson, J. & Gallant, J. (2000) *Terrain analysis: principles and applications*. John Wiley & Sons, New York.

Wiser, S.K. (2016) Achievements and challenges in the integration, reuse and synthesis of vegetation plot data. *Journal of Vegetation Science*, **in press**, 10.111/jvs.12419.

Wiser, S.K. & De Cáceres, M. (2013) Updating vegetation classifications: an example with New Zealand's woody vegetation. *Journal of Vegetation Science*, **24**, 80-93.

Wiser, S.K., Spencer, N., De Cáceres, M., Kleikamp, M., Boyle, B. & Peet, R.K. (2011) Veg‐X–an exchange standard for plot‐based vegetation data. *Journal of Vegetation Science*, **22**, 598-609.

Zhu, K., Woodall, C.W., Ghosh, S., Gelfand, A.E. & Clark, J.S. (2014) Dual impacts of climate change: forest migration and turnover through life history. *Global Change Biology*, **20**, 251-264.

Zimmerman, A.S. (2008) New knowledge from old data the role of standards in the sharing and reuse of ecological data. *Science, Technology & Human Values*, **33**, 631-652.

## SUPPORTING INFORMATION

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

**Appendix S1** Databases, registries, repositories and archives with plant species, community, trait and environmental information.

## BIOSKETCH

The authors have collaborated extensively on integrating data and models to forecast the effects of global change on plant populations, species and communities. They share an interest in spatial ecology, landscape ecology and conservation biogeography.

Table 1. Types of uncertainty affecting ecological data and models (from Elith *et al.*, 2002; Regan *et al.*, 2002).

| **Type of Uncertainty** | **Examples** | **Potential Solution** |
|---|---|---|
| Measurement / observation error | Incorrect species identifications; Incorrect coordinates of species or plot; Unbiased error in population counts, estimation of cover, size, growth; Inaccuracies in digital environmental maps | Repeated measurements; Data screening and scrubbing; Intervals; Statistical analysis; Develop new maps |
| Bias/Systematic error | Roadless areas are undersampled; Public land oversampled; Subjectively located observations; Sensor miscalibration | Bias correction; Modeling methods robust to bias; Minimizing bias in sampling design |
| Model uncertainty | Different image processing algorithms, terrain analyses, climate, species distribution, population, or community models, yield different results with same inputs | Ensemble modeling; Use models best suited for available data; Model validation; Revision of theory based on observation |
| Subjective judgment | Expert opinion of demographic and life history parameters in a population or community model | Collect more data; Use more experts; Degrees of belief; Subjective probabilities |
| Ambiguity | Taxonomic ambiguity; Canopy cover could refer to projected foliar cover or projected canopy outline; Definitions of canopy strata may vary | Standardize meaning of terms |
| Vague concepts | Seedling versus sapling; Habitat suitability; Forest types; Ecotone; Vegetation classification | Sharp delineation, fuzzy sets and logic, and other non-classical or alternative logics (see Regan *et al.*, 2002 for details) |

| Type of Uncertainty | Examples | Potential Solution |
|---|---|---|
| Underspecification | Digital environmental maps available at resolution too coarse to capture ecological processes; Historical species occurrences reported with large positional uncertainty | Conduct analysis at coarser spatial scale; Specify all available data |

Table 2. Example of sources ("Basis of Record") of plant species occurrence records in a Global Biodiversity Information Facility (GBIF) based on species searches for five tree species in the family Pinaceae that are endemic to the California Floristic Province. While natural history collections (PRESERVED_SPECIMEN) make up the majority, a growing number of human observations (including citizen science) and fossil records are being incorporated. LIVING_SPECIMEN records may refer to individuals in botanical gardens or plantations well outside the species' native range. For these species, some preserved specimen records (shown in parentheses) were actually found to be living specimens (or preserved records from those specimens) cultivated far outside the native ranges (determined from "countrycode"). This is especially apparent for the widely cultivated and, in some places, invasive *P. radiata*.

| **Basis of Record** | *Pinus coulteri* D. Don[1] | *Pinus sabiniana* Douglas ex D. Don[2] | *Pinus balfouriana* Balf. [3] | *Pinus muricata* D. Don[4] | *Pinus radiata* D. Don[5] |
|---|---|---|---|---|---|
| FOSSIL_SPECIMEN | 16 | 21 | 17 | 28 | 9 |
| HUMAN_OBSERVATION + OBSERVATION | 42 | 101 | 5 | 11 | 6846 |
| LIVING_SPECIMEN | 3 (9) | 0 (16) | 0 (1) | 4 (2) | 3 (320) |
| PRESERVED_SPECIMEN | 424 | 600 | 6 | 12 | 535 |
| UNKNOWN | 24 | 20 | 367 | 645 | 47 |
| **Total** | **509** | **742** | **395** | **86** | **6965** |

[1] GBIF.org (28th January 2016) GBIF Occurrence Download http://doi.org/10.15468/dl.iou7qq

[2] GBIF.org (28th January 2016) GBIF Occurrence Download http://doi.org/10.15468/dl.7s6cd2

[3] GBIF.org (28th January 2016) GBIF Occurrence Download http://doi.org/10.15468/dl.gwjevc

[4] GBIF.org (28th January 2016) GBIF Occurrence Download http://doi.org/10.15468/dl.skzrzg

[5] GBIF.org (28th January 2016) GBIF Occurrence Download http://doi.org/10.15468/dl.znqbsw

FIGURE LEGEND

Figure 1. Search of occurrence records conducted for *Bursera simaruba* (L.) Sarg. (Burseraceae), a widespread tree of Neotropical dry forest; 752 records from BIEN 2 in gold, primarily from Mexico and Central America, and 3042 records from GBIF (both accessed 3 Jun 2015) in red that include many more records for Florida, The Bahamas, Hispaniola, Puerto Rico, and Columbia; 123 records from DryFlor (blue; accessed 2 Feb 2016) fill gaps in northern South America and the Lesser Antilles. Note that 2965 occurrences in BIEN 3 (bien3.org) now include many records for Florida, but still few or none in the Bahamas, Greater and Lesser Antilles, and northern South America (accessed 14 Apr 2016).

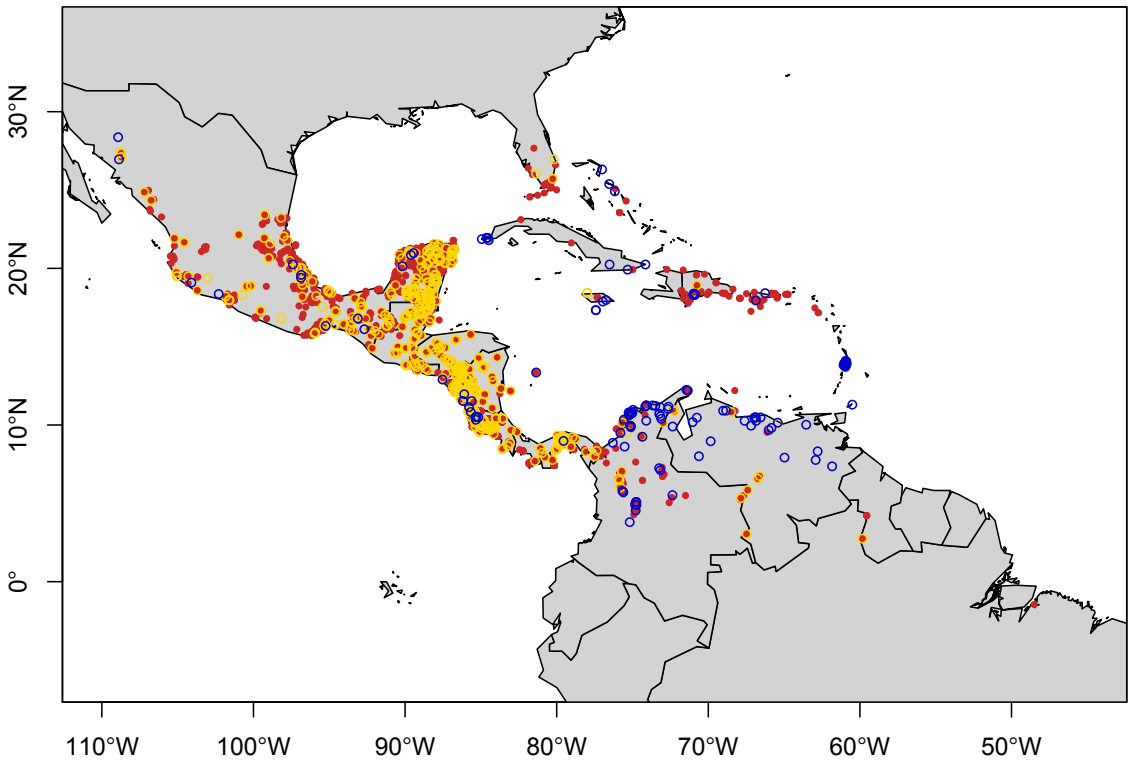

Table S1. Databases, registries, repositories and archives containing spatially explicit plant species, community, trait and environmental information.

| Database Organization URL | Geography | Taxa, Entities | Description | Availability | Notes |
|---|---|---|---|---|---|
| **Environmental maps** | | | | | |
| DataOne www.dataone.org | Global | Many kinds of environmental data, including maps, imagery, plots | Distributed cyberinfrastructure supporting open, persistent, robust, and secure access to easily discovered Earth observational data | Anyone can search; access varies by dataset | Contribute data only via nodes |
| Data Basin http://databasin.org/ | Global to local | Environmental maps | A mapping and analysis platform supporting environmental stewardship | Open. Sign up for user account. | Can download spatial data and maps |
| **Species Occurrence Data +/- Range Maps** | | | | | |
| The Global Biodiversity Information Facility (GBIF) http://gbif.org | Global | Animals and plants; occurrences | International open data infrastructure, funded by governments | Open. Sign up for user account. | Widely used federated database for species occurrence records |
| California Consortium of Herbaria http://ucjeps.berkeley.edu/consortium/ | California | Plants; occurrences | Gateway to California vascular plant specimens housed in participant herbaria | Open. Sign up for user account. | Specimen records |
| Southwest Environmental Information Network (SEINet) http://swbiodiversity.org/ | Southwestern USA | Plants; occurrences | Gateway to distributed data resources of interest to the environmental research community | Open. Sign up for user account. | Specimen records and other species observations |
| South African National Biodiversity Institute SANBI http://www.sanbi.org/ | South Africa | Plants and animals; occurrences, maps | Coordinates research, and monitors and reports on the state of biodiversity in South Africa | Open | Species occurrences, checklists, species atlases, vegetation map |
| Atlas of Living Australia https://www.ala.org.au/ | Australia | Plants and animals; occurrences, maps | Includes web mapping tools; for research, environmental monitoring, conservation planning, education, and biosecurity | Open | >55m records, >12,000 users; includes photographs, sound recordings, molecular data, links to literature |

| Database Organization URL | Geography | Taxa, Entities | Description | Availability | Notes |
|---|---|---|---|---|---|
| REMIB http://www.conabio.gob.mx/remib_ingles/doctos/remib_ing.html | Mexico | Animals and plants | One of the first aggregated specimen databases | Open access once accept data use agreement | Specimen records. Contains many small collections not in GBIF |
| SpeciesLink http://splink.cria.org.br/ | Brazil | Animals and plants | Integrate species and specimen data available in natural history museums, herbaria and culture collections, making it openly and freely available on the Internet | Open access | Specimen records. ~3.5 m georeferenced records, 450,000 species (accessed 14 April 2016) |
| Botanical Information and Ecology Network (BIEN) http://bien.nceas.ucsb.edu http://bien3.org | New World | Plants; occurrences, plots, traits | Research network *for the integration, access, and discovery of botanical information for all plants in the New World* | Currently by permission; scheduled to open late 2016 | Occurrences primarily from GBIF; Plot data include SALVIAS (including Gentry plots), CTFS (Panama only) and VegBank (public data only), plus several smaller datasets. Trait database partially overlapping with TRY. Synonymized taxonomy |
| **Vegetation Plots** | | | | | |
| Forest Inventory and Analysis http://www.fia.fs.fed.us/ | USA | Plots; repeated surveys | Provides information needed to assess America's forests | Open | Plot location accuracy degraded to protect privacy |
| VegBank http://vegbank.org | USA plus other contributed data | Plots | The vegetation plot database of the Ecological Society of America's Panel on Vegetation Classification | Open | About 75,000 plots (accessed 14 Jul 2015) including from state natural heritage programs and some federal agencies |
| LandCare National Vegetation Survey (NVS) Data Bank http://www.landcareresearch.co.nz/resources/data/national-vegetation-survey-nvs | New Zealand | Plots | Exceptional due to high density coverage for most of New Zealand | By request | Plots collected using standard protocols therefore sampling methods and format highly uniform |
| SALVIAS http://salvias.net/ Synthesis and Analysis of Local Vegetation Inventories Across Scales | Primarily New World tropics | Plots | Web-based utility for compiling data on diverse aspects of plant organismal biology | Open. Sign up for user account | Database of vegetation inventories from around the world, with emphasis on the New World tropics, including Gentry plots |

| Database Organization URL | Geography | Taxa, Entities | Description | Availability | Notes |
|---|---|---|---|---|---|
| Alwyn H. Gentry Forest Transect Dataset http://www.mobot.org/MOBOT/Research/gentry/transect.shtml | Global | Plots | Data from individual transects available to the research and conservation communities | Open | 226 plots collected by Alwyn H. Gentry (held by Missouri Botanical Garden) |
| DryFlor http://elmer.rbge.org.uk/dryflor/ | New World | P/A, derived from plots, inventories | First comprehensive dataset of the flora of neotropical dry forests across their full range | Open | Neotropical dry forest floristic data for woody plants compiled in an open-access database |
| European Vegetation Archive http://euroveg.org/eva-database [European vegetation archive, 2016] | Europe | Plots | centralized database developed by the IAVS Working Group European Vegetation Survey | By permission, depending on data provide(s) | |
| sPlot http://www.idiv-biodiversity.de/en/sdiv/workshops/workshops-2013/splot/splot-database | Global | Plots | V**egetation-plot database covering all biomes of the world;** can only be used to address the sPlot Working Group's research questions | Members approved by Steering Committee | Uses versioning system for releases of database to members |
| Center for Tropical Forest Science – Forest Global Earth Observatory (CTFS-ForestGEO) http://www.forestgeo.si.edu/ | Global | Plots | Repeated tree censuses in large forest plots, some since 1981, 59 sites | Members; other researchers can request data | Tree demographic data; Common measurement protocols |
| Global Index of Vegetation Plot Databases http://www.givd.info/ | Global | Plot registry | Metadatabase providing overview of vegetation data worldwide, allow researchers to retrieve data | Members | 237 databases with ~3.6m vegetation plots registered |
| **Traits** | | | | | |
| TRY https://www.try-db.org/ | Global | Plants; traits | a global archive of curated plant traits; > 5 million trait records for >1,100 traits of 2.2 million individual plants; >100,000 species | Largely open | Accessible through a portal that can be queried; users can upload or e-mailing contributed data |
| Botanical Information and Ecology Network (BIEN) http://bien3.org | New World | Plants; occurrences, plots, traits | See above | Currently by permission | Trait database partially overlapping with TRY |
| COMPADRE http://www.compadre-db.org/Compadre/Home | Global | Plants; Demographic data that can | A curated database of matrix population models describing population dynamics of a | Open | Does not include density dependence, dispersal, stochasticity or disturbance |

| Database Organization URL | Geography | Taxa, Entities | Description | Availability | Notes |
|---|---|---|---|---|---|
| | | be structure into a matrix population model (MPM) | given study x species x population x period x treatment combination. Mean MPMs also provided. | | response usually included in plant population models |
| Global Population Dynamics Database (GPPD) https://www.imperial.ac.uk/cpb/gpdd2/gpdd.aspx | Global | Animals, plants; Population counts for over 5000 species, mostly animals | Time series data for at least 10 years | Open. Sign up for user account. | Does not include density dependence, dispersal, stochasticity or disturbance response usually included in plant population models |
| **Data Repositories** | | | | | |
| Dryad Digital Repository http://datadryad.org/ | Global | Any type of data associated with published paper | A **curated** resource that makes the data underlying scientific publications **discoverable, freely reusable**, and **citable** | Open. Sign up for user account. | Does not impose file format restrictions; encourages ASCII or HTML; $80-90 to submit data <10 GB |
| Pangaea www.pangaea.de | Global | Any type of earth or life science data | Data publisher for earth and environmental science | Open. User account for data under moratorium. | Converts data to a uniform machine-independent format; donations accepted |

Plots = vegetation plots, typically consisting of records of species abundance within fixed areas. P/A indicates only species presence/absence are provided.
GB = Gigabyte