Daru, B.H.

Article type: Points of View

GreenMaps: a Tool for Addressing the Wallacean Shortfall in the Global Distribution of Plants

Barnabas H. Daru*

Affiliation:

Department of Life Sciences, Texas A&M University-Corpus Christi, Corpus Christi, TX 78412, USA

**Correspondence to: Barnabas H. Daru (barnabas.daru@tamucc.edu)*

Running head: GreenMaps to address Wallacean shortfall

GreenMaps to address Wallacean shortfall for plants

Abstract

The exponential growth of species occurrence data can facilitate dynamic biodiversity analyses. However, raw biodiversity data alone should not be used indiscriminately due to inherent sampling biases, impediments that contribute to Wallacean shortfall (i.e. the paucity of species' geographic information). It has been suggested that Wallacean shortfall is a common phenomenon across taxa, however, there is no global assessment geared toward overcoming this impediment for plants, despite the fundamental role of plants in ecosystem stability, food security and biodiversity conservation. Here, I present GreenMaps, a new tool that will permit a rapid initial assessment of the Wallacean shortfall for plants by building base maps of species' predicted distributions upon which citizen science participation could contribute to spatial validation of the actual range occupied by species. The initial stages of GreenMaps have now been accomplished, providing a massive dataset of modeled range maps for over 194,000 vascular plant species. This will make it the largest and only global assessment of geographic distributions for plant species at scales relevant to research and conservation. Ultimately, GreenMaps will interface with a mobile application to enable volunteers from any region of the world to validate predicted species distributions to be used for the generation of new and improved global map of plant distributions.

Key Words: Citizen Science Validation; GreenMaps; Plant Biogeography; Species Distribution Modeling; Vascular Plants

Daru, B.H.

1.0 Shortfalls in Knowledge of the Geography of Plants

Understanding patterns, underlying processes and principles governing biodiversity at large scales requires an accurate knowledge of species geographic distributions. The increasing availability of biodiversity data from multiple heterogeneous sources, including museum collections, observation records, biological literature and local databases, has led to their compilation into major data hubs such as the Global Biodiversity Information Facility (GBIF; Edwards et al. 2000), iNaturalist (iNaturalist 2019), and iDigBio (www.idigbio.org), that can facilitate macroecological analyses (Díaz et al. 2016; Pollock et al. 2017; Zanne 2018; Daru et al. 2019). These datasets are often point occurrences of where a species has been documented as present based on a voucher specimen in a museum/herbarium or reliable sighting in the field. However, the raw occurrence data alone should not be used indiscriminately due to prevailing sampling biases, gaps and uncertainties inherent in this type of data (Lira-Noriega et al. 2007; Meyer et al. 2016; Stropp et al. 2016; Daru et al. 2018). Another disadvantage of raw occurrence data is that they do not extrapolate beyond relatively few known locations to make inferences about the potential geographic range occupied by a species. Such impediment is often called Wallacean shortfall, the paucity of information on species distributions at a global scale (Hortal et al. 2015), is more acute for less-charismatic groups such as invertebrates and plants, and can induce strong limitations in conservation and spatial planning (Hortal et al. 2008; Boakes et al. 2010). For example, existing plant occurrence information is heavily skewed towards better sampled or more accessible areas and regions where digitisation is more advanced (e.g. North America, Western Europe), whereas some of the world's most biodiverse regions, for example, Central Asia, Africa or Amazonia, remain heavily under-represented in botanical sampling (Bush and Lovejoy 2007; Meyer et al. 2016). Progress in addressing some of the key questions in

GreenMaps to address Wallacean shortfall for plants

biodiversity science, such as how diversity is distributed in space, how it changes across time, and how it contributes to the resilience of ecosystems to global change, depends heavily on our understanding of accurate species distributions. It is therefore critical to address the Wallacean shortfall in species distributions by improving the global coverage and representativeness of biodiversity data.

In order to reduce the Wallacean shortfall in species distributions, there have been recent efforts incorporating niche-based models for a variety of taxonomic groups. Examples include efforts from Aquamaps (Kaschner et al. 2013) for marine species, and Lifemapper (Beach et al. 2015) and Map of Life (MoL, <http://www.mol.org>; Jetz et al. 2012) for terrestrial vertebrates. Because these initiatives complemented ecological data (e.g. climate and functional traits) with fine-scale aggregated data (e.g. specimen records, surveys, and experiments) to model species distributions, they create strong synergies that bridge the gap between fine-scale precision and global representativeness (König et al. 2019). So far, however, there is no solution that incorporates modeled distributions and online mobilisation for plants at a global scale in a generic way. Plants are foundation species, sustaining food chains and driving terrestrial ecosystem productivity (Loreau et al. 2001; Zak et al. 2003). Consequently, plant distribution often underlies the biogeographic histories of other organisms, and thus a global distribution map for plants is critical for predicting and conserving diversity at other trophic levels.

Vascular plants form a very diverse taxonomic group comprising about 300,750 species worldwide (Christenhusz and Chase 2014; APG IV 2016; Christenhusz et al. 2017). Plants occur across all types of biomes, from rainforests to savannas. Some are cosmopolitan, spanning borders e.g. the Cogon grass (*Imperata cylindrica*), whereas some species occur in very restricted distributions of a few hectares such as the Olympic violet (*Viola flettii*) endemic to the

Daru, B.H.

Olympic Mountains of Washington in United States. The overall ecosystem services provided by plants are estimated to be worth at least US\$16–54 trillion per year worldwide (Costanza et al. 2007). These services include provisioning (e.g. food and medicines), regulation of ecosystem processes (e.g. trophic regulation, water purification), cultural (firewood, ornamental) and supporting services (e.g. primary productivity), but yet estimates of plant geographic distributions at a global scale rest largely on extrapolations. Species distribution models (SDMs) provide an unbiased and easily interpretable estimate of improving representativeness and coverage of species distributions (Peterson et al. 2011).

To date, studies utilising SDMs to improve coverage of plant occurrences have focused mostly at the regional or country scale despite species distributions spanning political and regional borders, calling for a truly global assessment of species distributions (Jetz et al. 2019). To accurately depict the geographic range for every vascular plant species represents great challenges in terms of computation and global coverage at the species level. For instance, an examination of GBIF reveals that available point occurrences covered most plant species at only seven or fewer unique locations (Meyer et al. 2016), too few for estimating robust species distribution models (Guisan et al. 2007; Feeley and Silman 2011) or generating geographic range maps (Gaston and Fuller 2009; Rivers et al. 2011), and thus might require additional filtering, human labor and continuous update. I proposed to use climate as reliable predictor of plant distributions at broad scales, consistent with Hutchinson’s multidimensional niche concept that every species is limited by a number of environmental factors (Hutchinson 1957), coupled with well-established relationships between plant species richness and climate (Currie 1991; Olson et al. 2001; Wiens and Donoghue 2004; Fine 2015). Thus, I define the probability of occurrence of a species as the set of favourable environmental conditions that allow a species to survive and

GreenMaps to address Wallacean shortfall for plants

reproduce, and populations to maintain their numbers (**Fig. 1**). A global analysis of this nature should identify common links between species' ecological preferences and potential biomes that are most susceptible to changes in community composition under global change (Hooper et al. 2005; Cadotte et al. 2009; Barnosky et al. 2012; Lewis and Maslin 2015).

Here, I develop GreenMaps, a novel cyberinfrastructure that permits a rapid initial assessment of the Wallacean shortfall for plants by building base maps of species' predicted distributions based on SDMs upon which citizen science participation could contribute to spatial validation of the actual range occupied by species. Ultimately, GreenMaps will interface with a mobile application to enable volunteers from any region of the world to validate predicted species distributions to be used for the generation of new and improved global map of plant distributions at scales relevant to research and decision-making. Specifically, I address three objectives. 1) Design and development of GreenMaps, toward addressing the Wallacean shortfall in plant distribution. These requires high performance computing strategies for species distribution modeling spanning 300,000 species of vascular plants at a global scale. 2) Design and development of easy to use mobile application for GreenMaps to be used by biodiversity scientists and citizen scientists alike for spatial validation of the actual range occupied by species. 3) Assess how plant diversity is distributed in space, and explore the tempo and mode of plant diversification on mountains.

2.0 Design and Development of GreenMaps: How it Works

GreenMaps consists of two primary types of data - GreenMaps archive data and citizen-defined maps (**Fig. 1**). GreenMaps archive data consists of global model-based species distributions along with occurrence records and carefully vetted taxonomic names (species, genus, family,

Daru, B.H.

order and clade) for all vascular plants at a global scale. Citizen-defined maps consist of an interface in which experts and citizen scientists can review, edit and approve these predicted range maps. The main features of the tool are described below.

In the first stage of GreenMaps, I generated the base maps of predicted species' distributions for citizen science validation. To this end, I obtained occurrence records from specimen records, online repositories (e.g. iNaturalist, GBIF, iDigBio, records from the primary literature, and personal observations). These records were thoroughly cleaned to reconcile old names/synonyms to current taxon concepts e.g. The Plant List (www.theplantlist.org), and/or to remove duplicates and records with doubtful or imprecise localities (e.g. points in the sea and records with inverted coordinates). Applying such cleaning filters wiped out most available occurrence records, resulting in too few records for the vast majority of plant species. To improve the utility of species with too few records for SDMs, I created buffers of 1-2 km radius as spatial offsets in Poisson point process model around each point (see Merow et al. 2016) and randomly (Poisson) distributed points (adding up to 20) within the boundaries of the buffers (a good SDM requires 10-200 unique records) (Kadmon et al. 2003; Rivers et al. 2011). These points were used in addition to the cleaned dataset as inputs for the species distribution modelling (**Fig. 1**).

In a second step, I imposed two stringent spatial filters to restrict the area of species to their known native ranges and to prevent erroneous records and predictions in areas that might contain suitable habitat but are not occupied by a species. 1) *APGfamilyGeo*: I bounded every species distributions based on expert drawn occurrence polygons ("expert maps") of global plant family geographic distributions (Heywood 1993; APG IV 2016). Although expert maps have limitations such as encompassing areas of unsuitable habitat within the range boundaries (Merow

GreenMaps to address Wallacean shortfall for plants

et al. 2017), they ultimately represent excellent resources for delimiting the boundaries beyond which a species is not expected to occur. This spatial filter was applied to constrain species predictions during the modeling process. 2) *GeoEigenvector*: These are orthogonal variables representing spatial relationships among cells in a grid, encompassing the geometry of the study area at various scales (Diniz-Filho et al. 2005). The software SAM (Spatial Analysis in Macroecology; Rangel et al. 2010) was used to generate a pairwise geographical connectivity matrix among grid cells as input to establish a truncation distance for the eigenvector-based spatial filtering, resulting in a total of 150 spatial filters. These filters were then resampled to the same resolution as the environmental variables and were included with the bioclimatic variables for the SDM.

Third, I combined the occurrence records with 19 georeferenced bioclimatic variables at a resolution of 30 arc-second from WorldClim (Hijmans et al. 2005). These variables correlate with key physical and biological attributes of many plant species at a scale close to *in situ* records (Jetz et al. 2019) such as precipitation, temperature, primary productivity, plant functional traits, elevation (Hansen et al. 2013; Tuanmu and Jetz 2015; Jetz et al. 2016; Wilson and Jetz 2016), and sometimes the actual niche realised by species directly (Fretwell and Trathan 2009; He et al. 2015; Asner and Martin 2016). Environmental data can also be derived from MODIS, LandSat, NLCD, PRISM, and EarthEnv.

The algorithm MaxEnt (Phillips et al. 2006) was used as the basis for modeling the distributions of plants not only because it works with presence-only data but it is effective and sensitive to differences in modelling settings and can accommodate large datasets that span regional and global scales (Shcheglovitova and Anderson 2013; Syfert et al. 2013; Warren and Seifert 2011). MaxEnt estimates a species' probability distribution that has maximum entropy,

Daru, B.H.

subject to a set of constraints based on knowledge of environmental conditions at the known occurrence locations (Phillips et al. 2006). I integrated cloud-based statistical workflows via the High Performance Computing Clusters of Texas A&M University-Corpus Christi (<https://hpc.tamucc.edu/>) implemented in R (R Core Team 2018) and modeled the potential distribution of 194,235 terrestrial vascular plant species using the presence only data and 10,000 background points. I used the equal training sensitivity and specificity threshold (Liu et al. 2005) to create presence-absence maps. Model performances were evaluated using the Area Under the Receiver Operator Curve (AUC; Bewick et al. 2004). The final outputs from MaxEnt consisted of a modeled range map (stored in raster format at grid cell resolution of 0.5 degree equivalent to 50 km) and the observed point occurrences (**Fig. 1**). A modeled range map resembles a species' fundamental niche, its environmental suitability, its likelihood of being collected, and its native range area of occupancy as known from its family's expert drawn geographic range (Heywood 1993; APG IV 2016).

A total of 194,235 modeled range maps, representing 194,235 species and 222 families have been generated and are available at a resolution of $0.5^{\circ} \times 0.5^{\circ}$ (**Fig. 2**), making it the largest and only global assessment of geographic distributions for plants at the species-level. Based on these base maps alone, I found that the geographic distribution of plant species is very uneven, with higher diversity in the tropics peaking in Amazonia, Andes, southern Africa, Madagascar, Himalaya-Hengduan mountains, Southeast Asia, Southern Europe and Central America, mirroring global latitudinal diversity gradient (**Fig. 2**). My predicted distributions contrast sharply to observed patterns of sampling biases in plant occurrence records (Meyer et al. 2016), lending strong support and power to the modeled distributions. For example, my modeled maps highlight well-known plant diversity hotspots that have essentially no virtual collections

GreenMaps to address Wallacean shortfall for plants

available e.g. Amazonia and Himalaya-Hengduan Mountains. On the other hand, a point map (stored in CSV format), displays the locations of the georeferenced occurrence of a species following data cleaning. A third layer of GreenMaps archive data corresponds to map of geographic biases in botanical sampling density. Here, I used the observed occurrence points to map geographic biases in sampling density (**Fig. 1**), defined as grid cells of excessive (hotspots) or insufficient (coldspots) collection (Hijmans et al. 2000; Daru et al. 2018). The initial milestones of GreenMaps have now been implemented and include modeled range maps, point maps and bias maps for 194,235 species (222 families) and will be stored on a web portal as base maps for citizen science validation. Code for running the species distribution models, mapping hotspots and coldspots of sampling density, generating species richness, and endemism is publicly available in a new R package *bioregion* (Daru and Schliep 2019).

2.1 Next Steps for GreenMaps

Two major steps are required for the full implementation of GreenMaps: 1) generating additional modeled base maps for the missing 106,515 species (out of the estimated 300,750 species; Christenhusz and Chase 2014; APG IV 2016; Christenhusz et al. 2017). I expect that the majority of the missing species will likely already be rare, since a local distribution is one of the major determinants in their escaping detection so far; and 2) the implementation of citizen-defined maps to ensure that the modeled range maps are technically valid and representative of the species native range area. Because the raw point occurrences for most plant species are too few for estimating robust species distribution models or generating geographic range maps (Guisan et al. 2007; Feeley and Silman 2011), citizen science participation is a relatively inexpensive form of contributing data to overcome the gaps in botanical sampling. Conceptually,

Daru, B.H.

the accuracy of a good SDM depends on the quality of the input occurrence records and the area restrictions imposed for a species (Phillips et al. 2006). Thus, GreenMaps provides volunteers access to an interactive mode that allows them to edit maps via a mobile app or web browser. In this mode, volunteers can exclusively choose species of interest based on specific traits and change the mapping parameters using pencil and eraser tools. For example, volunteers can explore which trees are common in an area, which species are succulent, herbs, or epiphytes; or by taxonomic group such as ferns, gymnosperms, monocots, orchids, grasses etc. Since almost any information can nowadays be geoTagged with appropriate geographic location (Goodchild 2007; Sui 2008), I implemented the option that allows volunteers to collect ancillary data such as pictures, recordings, and DNA sequences, upload them and metadata will be populated automatically.

Another promising feature of GreenMaps is that users can validate the occurrence of a species from any $0.5^\circ \times 0.5^\circ$ pixel on the globe by voting whether or not a species is actually present or absent in the area as predicted by the modeled distributions. For instance, GreenMaps predicts the occurrence of 6118 plant species (representing 125 families, 39 orders) in the state of Texas, USA, with areas of high predicted plant species richness in the southwest running from Nuevo Laredo in Mexico to Big Bend National Park Texas (**Fig. 3a**). When I weighted each grid cell by the number of species actually observed in the State, I found a mismatch between areas of high predicted richness versus observed richness, suggesting that the sampling of plant species in Texas is non-random, but clustered around big cities such as Austin, San Antonio, Houston, Dallas and Corpus Christi (**Fig. 3b**). Indeed, only 25.5% (2023 out of 7924) of observed species distributions overlap the geographic ranges of species modeled by GreenMaps (**Fig. 3c**). This overlap increased to 67.7% when assessed at the family level, with 12-42% of families unique to

GreenMaps to address Wallacean shortfall for plants

either modeled distributions or observations (**Fig. 3d**). Thus, more observations are required in places that are under-sampled, but predicted to have more diversity. Using rarefaction, the observed occurrences did not fully predict plant species diversity in Texas (**Fig. 3e**), further highlighting inherent sampling biases in plant occurrence records (Meyer et al. 2016; Daru et al. 2018). I also found a tendency of the observed occurrences to concentrate in particular clades as opposed to more dispersed phylogenetic distributions for the modeled distributions (**Fig. 4**), lending support to the utility of species distribution models in overcoming impediments of sampling biases. Thus, GreenMaps will be designed by implementing these features (observed sampling density vs predicted distributions, phylogenetic representation) to realize a truly robust assessment of plant diversity at scales relevant to research.

2.2 Design and Development of a Mobile-based Application of GreenMaps

When incorporated in a mobile app with the locations setting set to TRUE, GreenMaps can provide users with three possible outcomes for each query: “present” (when the predicted species actually occur in an area), “absent” (when no matches to the predicted species are found in the area), and “rare/introduced” when the species observed in an area are different to the predicted ones. Volunteers can also edit the input occurrence point maps by removing outliers or erroneous records such as points corresponding to big cities, botanical gardens or herbarium buildings, or by manually entering geographic coordinates of occurrence points (as a CSV file), if a species is validated to occur in its predicted range. User input will be checked and when expert-validated will be used to generate new maps that incorporate these changes. Because the goal of GreenMaps is to generate high quality range maps that can facilitate dynamic analyses in macroecology and complement efforts of major existing data hubs, information collected through

Daru, B.H.

this tool will be saved in Darwin Core format and continuously synchronized with existing repositories such as the GBIF and iNaturalist.

3.0 Assessment of Plant Distributions in Space and Plant Diversification on Mountains

The products from GreenMaps will not only address the Wallacean shortfall for plants, but offer exciting opportunities for biodiversity research, ecosystem monitoring and conservation prioritization. For instance, to understand how plant diversity is distributed in space, I used the preliminary result to map hotspots of species weighted endemism (Crisp et al. 2001; Laffan and Crisp 2003)—species richness inversely weighted by species ranges—by varying the analysis along global, regional and local scales at country level (**Fig. 5**). In a global setting, the key plant endemism hotspots are chiefly located in the tropics and include Mesoamerica, Andes, Amazonia, southern Africa, Madagascar, Himalaya-Hengduan mountains, and southeast Asia (**Fig. 5a**). At the continental scale, hotspots of species endemism were less spatially clumped and more dispersed into new locations including Western Europe, Southeast Asia, and Southwestern Australia (**Fig. 5b**). In parallel, some regions which emerged hotspots at the global scale including the Atlantic forest of Brazil and Madagascar disappeared in the continental scale analysis (**Fig. 5b**). At the national scale, spatial patterns of species endemism became more widespread across countries, clustering more at the political borders of countries (**Fig. 5c**). Thus, these data can provide relevant natural history information at the evaluated scale (global, continental, country or biome), highlighting regions where conservation measures will have the most impact. As a next step, the detailed range maps can be paired with a supertree for all the vascular species derived using the function *phylobuilder* from our new R package bioregion (Daru and Schliep 2019), to explore phylogenetic endemism (PE). PE is the phylogenetic

GreenMaps to address Wallacean shortfall for plants

equivalent of species endemism, and measures the degree to which a substantial proportion of phylogenetic diversity is restricted to an area (Rosauer et al. 2009). Although both WE and PE capture different facets of endemism and are increasingly considered crucial for conservation prioritization (Daru et al. 2019), PE is less sensitive to changes in taxonomic knowledge and captures additional information that can facilitate more informed conservation decisions.

A global distribution of plants can provide insights into the tempo and mode of plant diversification in high elevations. Mountain ranges harbour disproportionately rich and unique biodiversity (Hoorn et al. 2013; Quintero and Jetz 2018), but face larger threats, particularly because they serve as buffers for species against global climate change (Sandel et al. 2011; González-Varo et al. 2017). Several explanations for the rich biodiversity found in mountains propose that orogeny creates conditions favouring rapid *in situ* speciation of resident lineages (Hoorn et al. 2013; Wen et al. 2014; Favre et al. 2015; Schwery et al. 2015; Hughes 2016; Lagomarsino et al. 2016), however, global assessment of mountain plant diversity is lacking. A comprehensive global geographic map of plants can provide insights into the processes underlying the assembly and maintenance of mountain biodiversity and their evolutionary origins. For example, we can use detailed distribution maps and phylogenetic relationships to test the relationship between elevation, species richness and speciation rate for plants across the major mountain systems of the world that represent theatres of spectacular evolutionary events such as the Drakensberg in southern Africa, Eastern Arc, Hengduan, Himalayas, and Andes. If species richness gradients are controlled by differences in speciation, extinction and migration – perhaps because species at higher elevations tend to have lower speciation rates or higher extinction rates (Kozak and Wiens 2010) – then we should observe a decrease in diversification rates and lower species richness with elevation. I also hypothesize that mountain age (time since

Daru, B.H.

uplift), seasonality, primary productivity and climate change will explain substantial variation in the species richness and diversification rates of mountain species. The modeled range maps can be combined with detailed phylogenetic relationships (e.g. Zanne et al. 2014; Smith and Brown 2018) to estimate species richness along elevation for plants at a global scale. Speciation rates can be estimated using BAMM (Rabosky et al. 2017), a Bayesian framework for reconstructing complex evolutionary dynamics from phylogenetic trees. To the extent that speciation rates exhibit a signal, the effect of other variables including mountain age and climate can be tested.

Other outstanding questions remain: 1) What are the effects of reduced area and increased isolation of habitats?, 2) Where are the hotspots of plant diversity, endangerment and endemism?, 3) Where will alien exotic plants spread to in future?, 4) Where should we locate protected area networks and conservation corridors to safeguard plant diversity? 5) Where will native plant species disperse to under alternative scenarios of climate change? 6) How have anthropogenic activities e.g. habitat conversion and urbanization change the geography of nature?, and 7) Which geographic locations to target for undiscovered species?

4.0 Incentivizing citizen science contribution and outreach

Two common challenges of citizen science campaigns include attracting and retaining a wide range of volunteers from all over the world, and ensuring that good quality data is generated (Fritz et al. 2009). The GreenMaps campaign will use a combination of educational feedback (regarding natural history, rarity, endemism), and quizzes, to incentivize volunteer participation.

Educational feedback constitutes a key element in quality control (Fritz et al. 2017). GreenMaps archive data is designed with an in-built system that allows participants to learn about centers of high conservation potential near their locality in real-time e.g. centers of

GreenMaps to address Wallacean shortfall for plants

endemism (**Fig. 5**) or identify regions of sampling bias in botanical collecting (see **Fig. 1**). Areas of endemism, where species with exceptionally small geographic ranges concentrate, are targets for conservation because they capture facets of biodiversity not represented elsewhere (Myers et al. 2000; Ceballos and Ehrlich 2006; Rosauer et al. 2009). From any 0.5 degree grid cell on the planet, volunteers can get instant feedback on the number of plant species predicted to occur versus actual observations around them (e.g. see **Fig. 3**) relative to the national, continental or global flora. Similarly, users can get instant feedback on geographic biases in botanical sampling density. This approach is vital to ensure that volunteers do not waste time in areas that are oversampled, but make effective use of their efforts, and allows them to see where botanical sampling data alone disagree with the predicted range maps of plant species. In the short-term, these approaches can lead to stronger appreciation of data as scientific output.

Mobile learning is another way to incentivize student participation in the GreenMaps campaign as it has the potential to extend education beyond the traditional classroom (Hlodan 2010). In the short-term, biology quizzes and game elements such as those used for most computer games e.g. *Pokemon*, could be added to GreenMaps to make participation more engaging. The long-term goals will be that such approach can improve student appreciation of natural history and can potentially lead to more interest in considering a career in biodiversity science.

Ensuring good quality data is crucial. I propose to implement a feature in a similar way to *Amazon* star ratings that allows entries to be monitored by volunteers through votes and comments by anyone who disagrees with them or filter out records with low star ratings (see **Fig. 1**). The data generated will be publically available and serve as a useful resource for science (e.g. biogeography) and conservation managers.

Daru, B.H.

5.0 Concluding remarks

By integrating expert maps, point occurrence records, species distribution models and citizen science validation, this framework will enable wider use of species geographic distribution information particularly for less charismatic groups, such as plants with very sparse and biased occurrence data, but which play key crucial roles for diversity at other trophic levels as well as human health and well-being. Modelling allows predictions of species geographic distributions based on environmental conditions and expert-delineated range maps. Cloud-scale computing can significantly make the predictions faster particularly when the number of species is large, spanning more than 194,000 species, and the number of point occurrences per species are too few for estimating robust species distribution models. In turn, citizen science input is recorded along with ancillary data and geolocations for future use to generate new and improve global maps of plant diversity. Thus, GreenMaps combines strengths from heterogeneous data sources to improve predictions of plant species geographic distributions in a statistical way, a significant step toward addressing the Wallacean shortfall and can facilitate dynamic biodiversity analyses.

Acknowledgements

I specially thank Klaus Schliep, Peter C. le Roux, Carolyn Weaver, Israel Loera, Tammy L. Elliott, Emily Meineke for valuable comments and discussion during the early formation of this paper. My appreciation goes to Texas A&M University-Corpus Christi for logistic support.

GreenMaps to address Wallacean shortfall for plants

REFERENCES

- APG IV. 2016. An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG IV. Bot. J. Linn. Soc. 181:1–20.
- Asner G.P., Martin, R.E. 2016. Spectranomics: emerging science and conservation opportunities at the interface of biodiversity and remote sensing. Glob. Ecol. Conserv. 8:212–219.
- Barnosky A.D., Hadly E.A., Bascompte J., Berlow E.L., Brown J.H., Fortelius M., Getz W.M., Harte J., Hastings A., Marquet P.A., Martinez N.D., Mooers A., Roopnarine P., Vermeij G., Williams J.W., Gillespie R., Kitzes J., Marshall C., Matzke N., Mindell D.P., Revilla, E., Smith A.B. 2012. Approaching a state shift in Earth’s biosphere. Nature 486:52–58.
- Beach J.H., Stewart A.M., Grady C.J., Cavner, J.A. 2015. *Lifemapper* (Computational services and software for species distribution modeling and biodiversity pattern analysis <http://www.lifemapper.org>).
- Bewick V., Cheek L., Ball, J. 2004. Statistics review 13: Receiver operating characteristics curves. Critical Care 8:508–512.
- Boakes E.H., McGowan P.J., Fuller R.A., Chang-qing D., Clark N.E., O'Connor K., Mace G.M. 2010. Distorted views of biodiversity: spatial and temporal bias in species occurrence data. PLoS Biol. 8:e1000385.
- Bush M.B., Lovejoy T.E. 2007. Amazonian conservation: pushing the limits of biogeographical knowledge. J. Biogeogr. 34:1291–1293.
- Cadotte M.W., Cavender-Bares J., Tilman D., Oakley T.H. 2009. Using phylogenetic, functional and trait diversity to understand patterns of plant community productivity. PLoS ONE 4:e5695.
- Ceballos G., Ehrlich, P.R. 2006. Global mammal distributions, biodiversity hotspots, and conservation. Proc. Natl Acad. Sci. USA 103:19374–19379.
- Christenhusz M.J., Chase M.W. 2014. Trends and concepts in fern classification. Ann. Bot. 113:571–594.
- Christenhusz M.J.M., Fay M.F., Chase, M.W. 2017. *Plants of the world. An illustrated encyclopedia of vascular plants* (Kew Publishing, Royal Botanic Gardens, Kew UK).
- Costanza R., d'Arge R., de Groot R., Farber S., Grasso M., Hannon B., Limburg K., Naeem S., O'Neill R.V., Paruelo J., Raskin R.G., Sutton P., van den Belt M. 2007. The value of the world’s ecosystem services and natural capital. Nature 387:253–260.

Daru, B.H.

- 413 Crisp M.D., Laffan S., Linder H.P., Monro, A. 2001. Endemism in the Australian flora. J.
414 Biogeogr. 28:183–198.
- 415 Currie D.J. 1991. Energy and large-scale patterns of animal-and plant-species richness. Am. Nat.
416 137:27–49.
- 417 Daru B.H., Schliep K. 2019. *bioregion: an R package for mapping biodiversity and conservation*
418 (R package version 0.1.0. <https://github.com/darunabas/bioregion>).
- 419 Daru B.H., le Roux P.C., Gopalraj J., Park D.S., Holt B.G., Greve M. 2019. Spatial overlaps
420 between the global protected areas network and terrestrial hotspots of evolutionary
421 diversity. Glob. Ecol. Biogeogr. 28:757–766.
- 422 Daru B.H., Park D.S., Primack R.B., Willis C.G., Barrington D.S., Whitfeld T.J.S., Seidler T.G.,
423 Sweeney P.W., Foster D.R., Ellison A.M., Davis C.C. 2018. Widespread sampling biases
424 in herbaria revealed from large-scale digitization. New Phytol. 217:939–955.
- 425 Díaz S., Kattge J., Cornelissen J.H., Wright I.J., Lavorel S., Dray S., Reu B., Kleyer M., Wirth
426 C., Prentice I C., Garnier E., Bönsch G., Westoby M., Poorter H., Reich P.B., Moles
427 A.T., Dickie J., Gillison A.N., Zanne A.E., Chave J., Wright S.J., Sheremet'ev S.N.,
428 Jactel H., Baraloto C., Cerabolini B., Pierce S., Shipley B., Kirkup D., Casanoves F.,
429 Joswig J.S., Günther A., Falczuk V., Rüger N., Mahecha M.D., Gorné L.D. 2016. The
430 global spectrum of plant form and function. Nature 529:167–171.
- 431 Diniz-Filho, J.A.F., Bini L.M. 2005. Modelling geographical patterns in species richness using
432 eigenvector- based spatial filters. Glob. Ecol. Biogeogr. 14:177–185.
- 433 Edwards J.L., Lane M.A., Nielsen E.S. 2000. Interoperability of biodiversity databases:
434 biodiversity information on every desktop. Science 289:2312–2314.
- 435 Favre A., Päckert M., Pauls S.U., Jähnig S.C., Uhl D., Michalak I., Muellner-Riehl A.N. 2015.
436 The role of the uplift of the Qinghai-Tibetan Plateau for the evolution of Tibetan biotas.
437 Biol. Rev. 90:236–253.
- 438 Feeley K.J., Silman M.R. 2011. Keep collecting: accurate species distribution modelling requires
439 more collections than previously thought. Divers. Distrib. 17:1132–1140.
- 440 Fine P.V. 2015. Ecological and evolutionary drivers of geographic variation in species diversity.
441 Annu. Rev. Ecol. Evol. Syst. 46:369–392.
- 442 Fretwell P.T., Trathan P.N. 2009. Penguins from space: faecal stains reveal the location of
443 emperor penguin colonies. Glob. Ecol. Biogeogr. 18:543–552.

GreenMaps to address Wallacean shortfall for plants

444 Fritz S., McCallum I., Schill C., Perger C., Grillmayer R., Achard F., Kraxner F., Obersteiner M.
 445 2009. Geo-Wiki.Org: The use of crowdsourcing to improve global land cover. Remote
 446 Sens. 1:345–354.

447 Fritz S., See L., Perger C., McCallum I., Schill C., Schepaschenko D., Duerauer M., Karner M.,
 448 Dresel C., Laso-Bayas J.C., Lesiv M., Moorthy I., Salk C.F., Danylo O., Sturn T.,
 449 Albrecht F., You L., Kraxner F., Obersteiner M. 2017. A global dataset of crowdsourced
 450 land cover and land use reference data. Sci. Data 4:170075.

451 Gaston K.J., Fuller R.A. 2009. The sizes of species' geographic ranges. J. Appl. Ecol. 46:1–9.

452 González-Varo J.P., López-Bao J.V., Guitián, J. 2017. Seed dispersers help plants to escape
 453 global warming. Oikos 126:1600–1606.

454 Goodchild M.F. 2007. Citizens as sensors: The world of volunteered geography. GeoJournal
 455 69:211–221.

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

459 Hansen M.C., Potapov P.V., Moore R., Hancher M., Turubanova S.A., Tyukavina A., Thau D.,
 460 Stehman S.V., Goetz S.J., Loveland T.R., Kommareddy A., Egorov A., Chini L., Justice
 461 C.O., Townshend J.R. 2013. High-resolution global maps of 21st-century forest cover
 462 change. Science 342:850–853.

463 He K.S., Bradley B.A., Cord A.F., Rocchini D., Tuanmu M., Schmidtlein S., Turner W.,
 464 Wegmann M., Pettorelli N. 2015. Will remote sensing shape the next generation of
 465 species distribution models? Remote Sens. Ecol. Conserv. 1:4–18.

466 Heywood V.H. 1993. *Flowering Plants of the World* (Batsford, London).

467 Hijmans R.J., Cameron S.E., Parra J.L., Jones P.G., Jarvis A. 2005. Very high resolution
 468 interpolated climate surfaces for global land areas. Int. J. Climatol. 25:1965–1978.

469 Hijmans R.J., Garrett K.A., Huaman Z., Zhang D.P., Schreuder M., Bonierbale M. 2000.
 470 Assessing the geographic representation of genebank collections: the case of the Bolivian
 471 wild potatoes. Conserv. Biol. 14:755–1765.

472 Hlodan O. 2010. Mobile learning anytime, anywhere. BioScience 60:682.

473 Hooper D.U., Chapin F.S., Ewel J.J., Hector A., Inchausti P., Lavorel S., Lawton J.H., Lodge
 474 D.M., Loreau M., Naeem S., Schmid B., Setälä H., Symstad A.J., Vandermeer J., Wardle

Daru, B.H.

475 D.A. 2005. Effects of biodiversity on ecosystem functioning: a consensus of current
476 knowledge. *Ecol. Monogr.* 75:3–35.

477 Hoorn C., Mosbrugger V., Mulch A., Antonelli A. 2013. Biodiversity from mountain building.
478 *Nat. Geosci.* 6:154–154.

479 Hortal J., de Bello F., Diniz-Filho J.A.F., Lewinsohn T.M., Lobo J.M., Ladle R.J. 2015. Seven
480 shortfalls that beset large-scale knowledge of biodiversity. *Annu. Rev. Ecol. Evol. Syst.*
481 46:523–549.

482 Hortal J., Jiménez-Valverde A., Gómez J.F., Lobo J.M., Baselga A. 2008. Historical bias in
483 biodiversity inventories affects the observed environmental niche of the species. *Oikos*
484 117:847–858.

485 Hughes C.E. 2016. The tropical Andean plant diversity powerhouse. *New Phytol.* 210:1152–
486 1154.

487 Hutchinson G.E. 1957. Concluding Remarks. *Cold Spring Harbor Symposia on Quantitative*
488 *Biology* 22:415–427.

489 iNaturalist. 2019. *iNaturalist* (<https://www.inaturalist.org> [accessed 14 May 2019]).

490 Jetz W., Cavender-Bares J., Pavlick R., Schimel D., Davis F.W., Asner G.P., Guralnick R.,
491 Kattge J., Latimer A.M., Moorcroft P., Schaepman M.E., Schildhauer M.P., Schneider
492 F.D., Schrod F., Stahl U., Ustin S.L. 2016. Monitoring plant functional diversity from
493 space. *Nat. Plants* 2:16024.

494 Jetz W., McGeoch M.A., Guralnick R., Ferrier S., Beck J., Costello M.J., Fernandez M., Geller
495 G.N., Keil P., Merow C., Meyer C., Muller-Karger F.E., Pereira H.M., Regan E.C.,
496 Schmeller D.S., Turak E. 2019. Essential biodiversity variables for mapping and
497 monitoring species populations. *Nature Ecol. Evol.* 3:539–551.

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

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

502 Kaschner K., Rius-Barile J., Kesner-Reyes K., Garilao C., Kullander S.O., Rees T., Froese R.
503 2013. AquaMaps: Predicted range maps for aquatic species version 08/2013
504 (<http://www.aquamaps.org>).

GreenMaps to address Wallacean shortfall for plants

- 505 König C., Weigelt P., Schrader J., Taylor A., Kattge J., Kreft H. 2019. Biodiversity data
506 integration—the significance of data resolution and domain. *PLoS Biol.* 17:e3000183.
- 507 Kozak K.H., Wiens J.J. 2010. Niche conservatism drives elevational diversity patterns in
508 Appalachian salamanders. *Am. Nat.* 176:40–54.
- 509 Laffan S.W., Crisp M.D. 2003. Assessing endemism at multiple spatial scales, with an example
510 from the Australian vascular flora. *J. Biogeogr.* 30:511–520.
- 511 Lagomarsino L.P., Condamine F.L., Antonelli A., Mulch A., Davis C.C. 2016. The abiotic and
512 biotic drivers of rapid diversification in Andean bellflowers (Campanulaceae). *New*
513 *Phytol.* 210:1430–1442.
- 514 Lewis S.L., Maslin M.A. 2015. Defining the Anthropocene. *Nature* 519:171–180.
- 515 Lira-Noriega A., Soberón J., Navarro-Sigüenza A.G., Nakazawa Y., Peterson A.T. 2007. Scale
516 dependency of diversity components estimated from primary biodiversity data and
517 distribution maps. *Divers. Distrib.* 13:185–195.
- 518 Liu C., Berry P.M., Dawson T.P., Pearson R.G. 2005. Selecting thresholds of occurrence in the
519 prediction of species distributions. *Ecography* 28:385–393.
- 520 Loreau M., Naeem S., Inchausti P., Bengtsson J., Grime J.P., Hector A., Hooper D.U., Huston
521 M.A., Raffaelli D., Schmid B., Tilman D., Wardle D.A. 2001. Biodiversity and
522 ecosystem functioning: Current knowledge and future challenges. *Science* 294:804–808.
- 523 Merow C., Allen J.M., Aiello-Lammens M.E., Silander J.A. 2016. Improving niche and range
524 estimates with Maxent and point process models by integrating spatially explicit
525 information. *Glob. Ecol. Biogeogr.* 25:1022–1036.
- 526 Merow C., Wilson A.M., Jetz W. 2017. Integrating occurrence data and expert maps for
527 improved species range predictions. *Glob. Ecol. Biogeogr.* 26:243–258.
- 528 Meyer C., Weigelt P., Kreft H. 2016. Multidimensional biases, gaps and uncertainties in global
529 plant occurrence information. *Ecol. Lett.* 19:992–1006.
- 530 Myers N., Mittermeier R.A., Mittermeier C.G., da Fonseca G.A.B., Kent J. 2000. Biodiversity
531 hotspots for conservation priorities. *Nature* 403:853–858.
- 532 Olson D.M., Dinerstein E., Wikramanayake E.D., Burgess N.D., Powell G.V.N., Underwood
533 E.C., D’amico J.A., Itoua I., Strand H.E., Morrison J.C., Loucks C.J., Allnutt T.F.,
534 Ricketts T.H., Kura Y., Lamoreux J.F., Wettengel W.W., Hedao P., Kassem K.R. 2001.
535 Terrestrial ecoregions of the worlds: a new map of life on Earth. *Bioscience* 51:933–938.

Daru, B.H.

536 Peterson A.T., Soberón J., Pearson R.G., Anderson R.P., Martínez-Meyer E., Nakamura M.,
537 Araújo M.B. 2011. in *Monographs in Population Biology* (vol. 49) (Princeton University
538 Press, Princeton, NJ).

539 Phillips S.J., Anderson R.P., Schapire R.E. 2006. Maximum entropy modeling of species
540 geographic distributions. *Ecol. Modell.* 190:231–259.

541 Pollock L.J., Thuiller W., Jetz W. 2017. Large conservation gains possible for global biodiversity
542 facets. *Nature* 546:141–144.

543 Quintero I., Jetz W. 2018. Global elevational diversity and diversification of birds. *Nature*
544 555:246–250.

545 R Core Team. 2018. *R: A language and environment for statistical computing* (R Foundation for
546 Statistical Computing, Vienna, Austria. URL <https://www.R-project.org/>).

547 Rabosky D.L., Mitchell J.S., Chang J. 2017. Is BAMM flawed? Theoretical and practical
548 concerns in the analysis of multi-rate diversification models. *Syst. Biol.* 66:477–498.

549 Rangel T.F., Diniz-Filho J.A.F., Bini L.M. 2010. SAM: A comprehensive application for spatial
550 analysis in macroecology. *Ecography* 33:46–50.

551 Rivers M.C., Taylor L., Brummitt N.A., Meagher T.R., Roberts D.L., NicLughadha E. 2011.
552 How many herbarium specimens are needed to detect threatened species? *Biol. Conserv.*
553 144:2541–2547.

554 Rosauer D., Laffan S.W., Crisp M.D., Donnellan C., Cook L.G. 2009. Phylogenetic endemism:
555 A new approach for identifying geographical concentrations of evolutionary history. *Mol.*
556 *Ecol.* 18:4061–4072.

557 Sandel B., Arge L., Dalsgaard B., Davies R.G., Gaston K.J., Sutherland W.J., Svenning J.C.
558 2011. The influence of Late Quaternary climate-change velocity on species endemism.
559 *Science* 334:660–664.

560 Schwery O., Onstein R.E., Bouchenak-Khelladi Y., Xing Y., Carter R.J., Linder H.P. 2015. As
561 old as the mountains: The radiations of the Ericaceae. *New Phytol.* 207:355–367.

562 Shcheglovitova M., Anderson R.P. 2013. Estimating optimal complexity for ecological niche
563 models: A jackknife approach for species with small sample sizes. *Ecol. Modell.* 269:9–
564 17.

565 Smith S.A., Brown J.W. 2018. Constructing a broadly inclusive seed plant phylogeny. *Am. J.*
566 *Bot.* 105:302–314.

GreenMaps to address Wallacean shortfall for plants

- Stropp J., Ladle R.J., M. Malhado A.C., Hortal J., Gaffuri J., H. Temperley W., Olav Skøien J.,
Mayaux P. 2016. Mapping ignorance: 300 years of collecting flowering plants in Africa.
Glob. Ecol. Biogeogr. 25:1085–1096.
- Sui D.Z. 2008. The wikification of GIS and its consequences: or Angelina Jolie's new tattoo and
the future of GIS. Comp. Env. Urb. Sys. 32:1–5.
- Syfert M.M., Smith M.J., Coomes D.A. 2013. The effects of sampling bias and model
complexity on the predictive performance of MaxEnt species distribution models. PLoS
ONE 8:e55158.
- Tuanmu M.-N., Jetz W. 2015. A global, remote sensing-based characterization of terrestrial
habitat heterogeneity for biodiversity and ecosystem modelling. Glob. Ecol. Biogeogr.
24:1329–1339.
- Warren D.L., Seifert S.N. 2011. Ecological niche modeling in Maxent: The importance of model
complexity and the performance of model selection criteria. Ecol. Appl. 21:335–342.
- Wen J., Zhang J.Q., Nie Z.L. 2014. Zhong, Y. & Sun, H. Evolutionary diversifications of plants
on the Qinghai-Tibetan Plateau. Front. Genet. 5:1–16.
- Wiens J.J., Donoghue M.J. 2004. Historical biogeography, ecology and species richness. Trends
Ecol. Evol. 19:639–644.
- Wilson A.M., Jetz W. 2016. Remotely sensed high-resolution global cloud dynamics for
predicting ecosystem and biodiversity distributions. PLoS Biol. 14:e1002415.
- Zak D.R., Holmes W.E., White D.C., Peacock A.D., Tilman D. 2003. Plant diversity, soil
microbial communities, and ecosystem function: Are there any links? Ecology 84:2042–
2050.
- Zanne A.E. 2018. Functional biogeography of angiosperms. Life at the extremes. New Phytol.
218:1697–1709.
- Zanne A.E., Tank D.C., Cornwell W.K., Eastman J.M., Smith S.A., FitzJohn R.G., McGlenn
D.J., O'Meara B.C., Moles A.T., Reich P.B., Royer D.L., Soltis D.E., Stevens P.F.,
Westoby M., Wright I.J., Aarssen L., Bertin R.I., Calaminus A., Govaerts R., Hemmings
F., Leishman M.R., Oleksyn J., Soltis P.S., Swenson N.G., Warman L., Beaulieu J.M.
2014. Three keys to the radiation of angiosperms into freezing environments. Nature
506:89–92.

Daru, B.H.

FIGURE LEGENDS

Fig. 1 | Overview of the workflow of GreenMaps showing archive data and citizen-defined maps. Products from the first phase of GreenMaps (1 to 4) have been accomplished, providing model-based predictions for 194,235 species. Phase 2 will involve citizen science validation of the species predictions.

Fig. 2 | The geographic distributions of global patterns of vascular plant species richness. The map shows predicted distributions for 194,235 species (221 families) based on species distribution models within grid cells at a resolution of $0.5^{\circ} \times 0.5^{\circ}$, ranging from deep blue (minimum 1 species) to deep red (maximum 17,884 species).

Fig. 3 | Overlap and mismatch between predicted and observed distributions of plants in Texas, USA. a, Predicted species richness ($n = 6118$ species). **b,** Observed species richness ($n = 3829$ species). **c,** Venn diagram of overlap and incongruence in the number of species predicted versus observed (predicted species $\cup$ observed species = 7924 species). **d,** Overlap and incongruence in the number of families predicted to occur versus families actually observed in Texas USA (predicted families $\cup$ observed families = 167 families). **e,** Observed occurrence-based rarefaction sampling curves (blue lines) for species richness for the vascular plants of Texas, USA. Each blue line represents the rarefaction curve for different locations within Texas.

GreenMaps to address Wallacean shortfall for plants

Fig. 4 | Phylogenetic representations of observed versus predicted geographic distributions of vascular plants in Texas USA. Iconic plant clades are indicated with images. The phylogenetic tree is summarized to family level to aid visualization.

Fig. 5 | Variation in spatial extent of hotspots of species weighted endemism for vascular plants of the world across successive spatial extents (national, continental, and global). Hotspots are defined as the grid cells with the highest 2.5% of species richness (indicated in red), and 5% hotspots in yellow.

Daru, B.H.

Fig. 1

GreenMaps to address Wallacean shortfall for plants

Fig. 2

Daru, B.H.

Fig. 3

GreenMaps to address Wallacean shortfall for plants

Fig. 4

Daru, B.H.

Fig. 5