Macroecology in the age of Big Data – Where to go from here?

Rafael Wüest, Niklaus Zimmermann, Damaris Zurell, Jake Alexander, Susanne Fritz, Christian Hof, Holger Kreft, Signe Normand, Juliano Sarmento Cabral, Enikö Székely, et al.

Received: 7 February 2019 | Revised: 8 May 2019 | Accepted: 9 May 2019

DOI: 10.1111/jbi.13633

PERSPECTIVE

Journal of Biogeography | WILEY

Macroecology in the age of Big Data – Where to go from here?

Rafael O. Wüest[1] | Niklaus E. Zimmermann[1] | Damaris Zurell[2] | Jake M. Alexander[3] | Susanne A. Fritz[4,5] | Christian Hof[6] | Holger Kreft[7] | Signe Normand[8] | Juliano Sarmento Cabral[9] | Eniko Szekely[10] | Wilfried Thuiller[11] | Martin Wikelski[12,13] | Dirk Nikolaus Karger[1]

[1]Swiss Federal Research Institute WSL, Birmensdorf, Switzerland

[2]Department of Geography, Humboldt University Berlin, Berlin, Germany

[3]Institute of Integrative Biology, ETH Zurich, Zürich, Switzerland

[4]Senckenberg Biodiversity and Climate Research Centre, Frankfurt (Main), Germany

[5]Institute of Ecology, Evolution & Diversity, Goethe-University, Frankfurt (Main), Germany

[6]Terrestrial Ecology Research Group, Technical University of Munich, Freising, Germany

[7]Biodiversity, Macroecology & Biogeography, Faculty of Forest Sciences and Forest Ecology, University of Göttingen, Göttingen, Germany

[8]Ecoinformatics & Biodiversity & Center for Biodiversity on a Changing world, Aarhus, Denmark

[9]Ecosystem Modeling, CCTB University of Würzburg, Würzburg, Germany

[10]Swiss Data Science Center, ETH Zurich and EPFL, Lausanne, Switzerland

[11]Evolution, Modeling and Analysis of Biodiversity (EMABIO), Laboratoire d'Ecologie Alpine (LECA), Université Grenoble Alpes, Grenoble, France

[12]Max-Planck-Institut for Ornithology, Vogelwarte Radolfzell, Radolfzell, Germany

[13]Centre for the Advanced Study of Collective Behaviour, University of Konstanz, Konstanz, Germany

Correspondence
Dirk Nikolaus Karger, Swiss Federal Research Institute WSL, Zürcherstrasse 111, 8903 Birmensdorf, Switzerland.
Email: dirk.karger@wsl.ch

Funding information
European Union's Horizon 2020 research and innovation programme, Grant/Award Number: 678841; French National fundier Agence Nationale pour la Recherche, Grant/Award Number: ANR-18-EBI4–0009; Swiss National Science Foundation; SNF, Grant/Award Number: 310030L-170059; ANR, Grant/Award Number: ANR-16-CE93-004; German Science Foundation, Grant/Award Number: ZU 361/ and 1-1; Villum Foundation; DFG, Grant/Award Number: 422037984

Handling Editor: Christine Meynard

Abstract

Recent years have seen an exponential increase in the amount of data available in all sciences and application domains. Macroecology is part of this "Big Data" trend, with a strong rise in the volume of data that we are using for our research. Here, we summarize the most recent developments in macroecology in the age of Big Data that were presented at the 2018 annual meeting of the Specialist Group Macroecology of the Ecological Society of Germany, Austria and Switzerland (GfÖ). Supported by computational advances, macroecology has been a rapidly developing field over recent years. Our meeting highlighted important avenues for further progress in terms of standardized data collection, data integration, method development and process integration. In particular, we focus on (a) important data gaps and new initiatives to close them, for example through space- and airborne sensors, (b) how various data sources and types can be integrated, (c) how uncertainty can be assessed in data-driven analyses and (d) how Big Data and machine learning approaches have opened new ways of investigating processes rather than simply describing patterns. We discuss how Big Data opens up new opportunities, but also poses new challenges to macroecological research. In the future, it will be essential to carefully assess data quality, the reproducibility of data compilation and analytical methods, and the communication of uncertainties. Major progress in the field will depend on the definition

of data standards and workflows for macroecology, such that scientific quality and integrity are guaranteed, and collaboration in research projects is made easier.

KEYWORDS

Biogeography, conference overview, data science, Linnean shortfall, machine learning, Macroecology, remote sensing, space-borne ecology, Wallacean shortfall

1 | INTRODUCTION

Data analysis through Big Data and machine learning methods is used in a wide range of applications in our daily lives. Big Data are defined by the five Vs: volume, variety (heterogeneity of sources, unstructured data), velocity (speed of data generation and collection), veracity (uncertainty and data quality) and value. As the first four Vs are self-explanatory, the value of Big Data refers to what scientific knowledge we can extract from it. The applications of Big Data range from personalized medicine, movie recommendations, personalized advertisements in social media, to transportation scheduling, online shopping suggestions or even self-driving cars. But of what value are Big Data in macroecological research? This was the topic of the 2018 meeting of the Specialist Group Macroecology of the Ecological Society of Germany, Austria, and Switzerland (GfÖ) "Macro2018" held at the Swiss Federal Research Institute WSL in Birmensdorf, Switzerland. The aim of this paper was to synthesize, and provide an overview of, the current developments in macroecology with respect to Big Data that were highlighted during this conference, and to flag potential pitfalls and opportunities arising from these developments.

Over the last decades, we have seen a massive increase in the amount of publicly available data that are relevant for macroecology. Such data can include environmental, genetic, trait, distribution, movement, co-occurrence or population demographic data. Figure 1 illustrates this increase in data availability using the example of occurrence information in the Global Biodiversity Information Facility (GBIF; www.gbif.org) and animal tracking data in Movebank (www.movebank.org). As more data become available, publications on Big Data have increased correspondingly in most natural and life sciences (Figure 1). In fact, within the broad field of ecological and biogeographic research, Big Data tools (from generation, maintenance, integration of multiple data sets, to actual analyses) appear to have been most often used in macroecological publications (Figure 1). Due to the large amount of data assembled, many data collection initiatives now offer unprecedented possibilities to search for general patterns and mechanisms in the global distribution of biodiversity (Bruelheide et al., 2018). This development has been partly assisted by the fact that many journals have adopted a data sharing policy and increasingly demand that authors make their data publicly available, thus moving towards open science. In addition, a new category of papers has appeared, where data and their metadata, as well as their generation protocols, can be published.

Progress in the acquisition of large, publicly available data sets has been accompanied by the development of artificial intelligence, novel statistical tools and adapted computer platforms that are able to manage and analyse Big Data beyond the capacity of a single computer. The term "Big Data" is often used synonymously with "large data" in macroecology. Volume is, however, only one of the five Vs. In other words, "Big Data" do not only refer to data sets that are large in size. Instead, the four remaining Vs (variety, velocity, veracity and value) imply that Big Data analyses require specific methods that are suitable to analyse data of large volume that stem from heterogeneous, autonomous sources and aim at exploring the complex and rapidly evolving relationships among data (Kambatla, Kollias, Kumar, & Grama, 2014; Xindong et al., 2014). The difference between large and Big Data can be illustrated by a thought experiment: imagine a univariate data set of millions of data points that, despite being large, can be assessed with classical methodology such as frequentist or Bayesian statistics. Such data are not necessarily Big Data, but rather large data. What makes large data big is the fact that the data are heterogeneous, collected from a multitude of sources that each had their own aim for gathering the data. The sheer complexity of these heterogeneous data requires purpose built methodologies and careful assessment of uncertainties, while opening up a multitude of questions that could be answered with it.

In their horizon-scanning paper for macroecology, Beck et al. (2012) identified four major challenges in macroecological research: integration of historical contingencies, explicit consideration of processes, aggregation of large high-quality data sets on a global scale and the advancement of statistical methods tailored to the needs of macroecology. Contributions at the GfÖ macroecology meeting 2018 explicitly linked aspects of the latter three of these major challenges to Big Data. We specifically cover the topics that emerged from the meeting in the four first sections of this paper (a) aggregation of large data sets, (b) data harmonization and integration, (c) uncertainty propagation and bias in data, and (d) the explicit consideration of processes. Each section outlines recent advances in relation to Big Data, and, where appropriate, links to the macroecological challenges outlined by Beck et al. (2012). In the last section (the future of Big Data in macroecology), we postulate that the huge amounts of information that are now becoming available to macroecologists – be it environmental, taxonomic, biogeographic, trait or phylogenetic information – open up new possibilities for macroecological research in the upcoming decade, if the major challenges regarding data processing, quality control and analyses can be met.

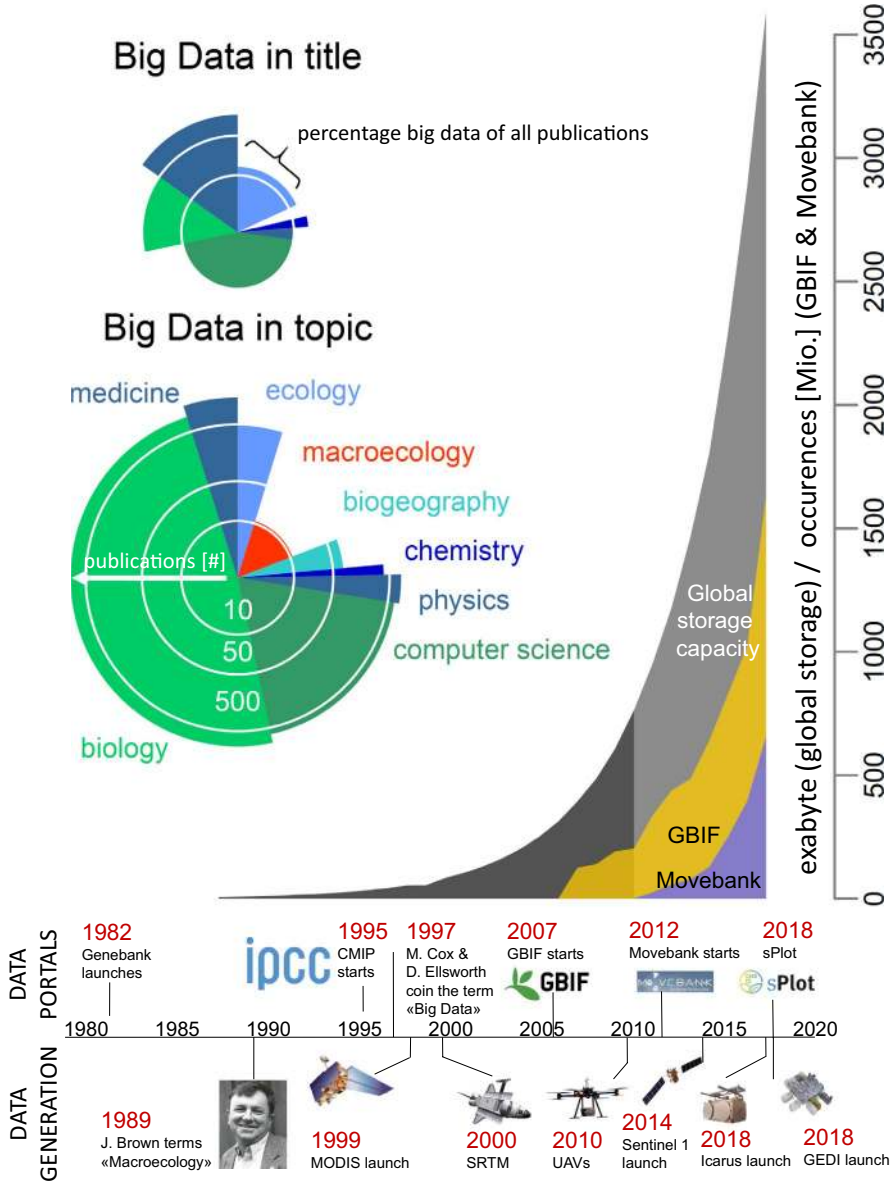

FIGURE 1 Timeline of Big Data in macroecology (bottom panel) and its imprint on the publication record (top panel). Pie charts show the amount of publications with the term "Big Data" in either the title, or the topic, in combination with several scientific disciplines (colours are categorical and represent the respective discipline), from a search on ISI Web of Knowledge over the years. The scale is similar for both pie charts. The width of each slice gives the percentage of publications that carry the term Big Data in relation to the overall amount of publications in the respective fields. The timeline at the bottom indicates the creation of major data portals, as well as key events with respect to sources generating Big Data. Global storage capacity (1 exabyte = 10^9GB) is taken from (Hilbert & López, 2011) which provide data until 2011. Times after 2011 (light grey) are extrapolated [Colour figure can be viewed at wileyonlinelibrary.com]

2 | AGGREGATION OF LARGE DATA SETS

Major new sources of data for macroecological research that have become available in recent years helped to fill three major gaps up to now: gaps across spatial scales (the "scale shortfall"), gaps in the biomes covered (the "Wallacean shortfall") and gaps in the number of taxa covered (the "Linnean shorfall", Beck et al., 2012; Hortal et al., 2015). The developments that we outline below have contributed to overcoming these shortfalls and to all aspects of Big Data, but especially so to volume, variety and velocity.

Improvements in the availability of data at high spatial resolution (i.e. grain size ≤1 km^2) at large extents (continental to global) have been relatively recent and are linked to the use of remotely sensed, space-borne environmental, animal and plant data. Well-known examples of reducing the scale shortfall include time-series for a multitude of vegetation indices from MODIS (e.g. Huete et al., 2002), the SENTINEL imagery complementing LANDSAT, and high-resolution digital elevation models with global coverage delivered by SRTM and now TanDEM_X (Figure 1). Every new generation of sensors thereby leads to higher spatial, temporal or thematic resolution. And there's more to come. For example, the recent launch of the ICARUS (International Cooperation for Animal Research Using Space, Figure 1) antenna on the International Space Station marks the beginning of a new era in animal tracking, generating enormous amounts of data every minute. Applications of animal tracking data range from forecasting earthquakes (Mai et al., 2018) to investigating the spread of diseases by animal vectors (www.icarus.mpg.de). The even more recent launch of the GEDI mission (Global Ecosystem Dynamics Investigation) uses light detection and ranging (LiDAR) to track changes in tree canopies at a global scale and is expected to produce about 10 billion cloud-free observations during its 24-month mission length.

The increase in spatial resolution over large extents is not limited to space-borne sensors but is paralleled by airborne systems.

Even though airborne campaigns provide data with higher resolution, they typically cover smaller geographical extents. Airborne LiDAR campaigns at national scales provide data on vegetation and terrain structure of approximately 1x1m resolution and have been found to explain a considerable amount of the local variation in the diversity of different taxonomic groups (e.g. Clawges, Vierling, Vierling, & Rowell, 2008; Thers et al., 2017; Zellweger et al., 2016). Additionally, airborne systems provide important information on microclimate quantification (Zellweger, Frenne, Lenoir, Rocchini, & Coomes, 2018). UAVs (unmanned aerial vehicles) equipped with, for example, visual, multispectral, thermal or LiDAR sensors, provide data with spatial resolutions of up to a few millimetres. Typically, higher spatial resolution comes at the expense of lower geographical coverage as it generally depends on flight height. However, such ultra-high-resolution data captured on demand allow researchers to investigate microclimatic differentiation at the scale of single plant individuals within an entire landscape (Figure 2; Cruzan et al., 2016; Cunliffe, Brazier, & Anderson, 2016; Tang & Shao, 2015; Tay, Erfmeier, & Kalwij, 2018; Wich & Koh, 2018). Orthophotos created from multiple images can additionally be used to detect vegetation composition by means of machine learning (pattern recognition algorithms) based on training data, but once trained these algorithms are able to gather community composition data over large spatial extents (sensu Waser, Ginzler, Kuechler, Baltsavias, & Hurni, 2011). Up to now, many of these techniques are limited to relatively small areas (approximately 1 km^2) and generally applied in easily accessible landscapes, where UAVs can be conveniently operated, and plant or animal species can be clearly distinguished.

A second data gap, the Wallacean shortfall, is linked to the fact that not every biome is equally well represented by publicly available species distribution data from sources like GBIF (https://www.gbif.org, 2018) or BIEN (Enquist, Condit, Peet, Schildhauer, & Thiers, 2016). While Europe and North America are rather well represented, some countries (e.g. Russia, see Figure 3a) or entire biomes such as

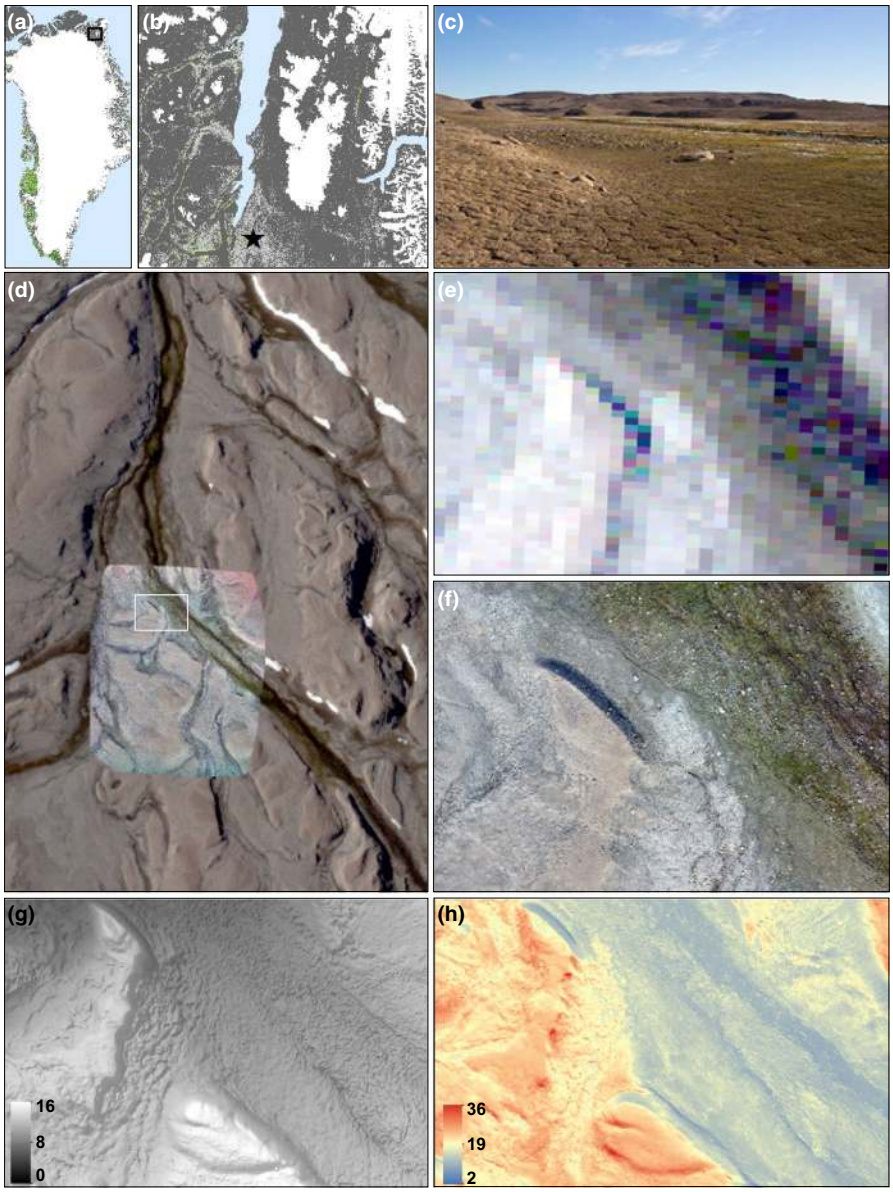

FIGURE 2 Example comparison of remotely sensed environmental data derived from satellite and UAV (Unmanned Aerial Vehicle) remote sensing at a site (80.5029, −23.6274) in North Eastern Greenland (a), indicated with a star in (b; corresponding to the inlet in a). The site harbours dry and wet tundra developed on permafrost soils (c). Overlap between WorldView-3 satellite imagery and a UAV derived orthomosaic (ca. 2 × 2 cm pixel size) is shown in (d). A comparison of the spatial resolution of the WorldView-3 (e) and UAV based data (f) is illustrated by zooming to the white inlet area in (d). Information on terrain structure (g) and microclimatic variation (h) illustrated with a digital surface model (g, in relative height difference, ca. 2 × 2 cm pixel size) and a thermal orthomosaic (h, in absolute temperature, ca. 15 × 15 cm pixel size). UAV footage and data from UAS4Ecology Lab, Aarhus University. The vegetation classes in (a) and (b) are based on Karami et al. (2018): barren, dark grey; abrasion plateau, light grey; fen areas, yellow; light to dark green, represent dry and wet tundra as well as tall shrub vegetation [Colour figure can be viewed at wileyonlinelibrary.com]

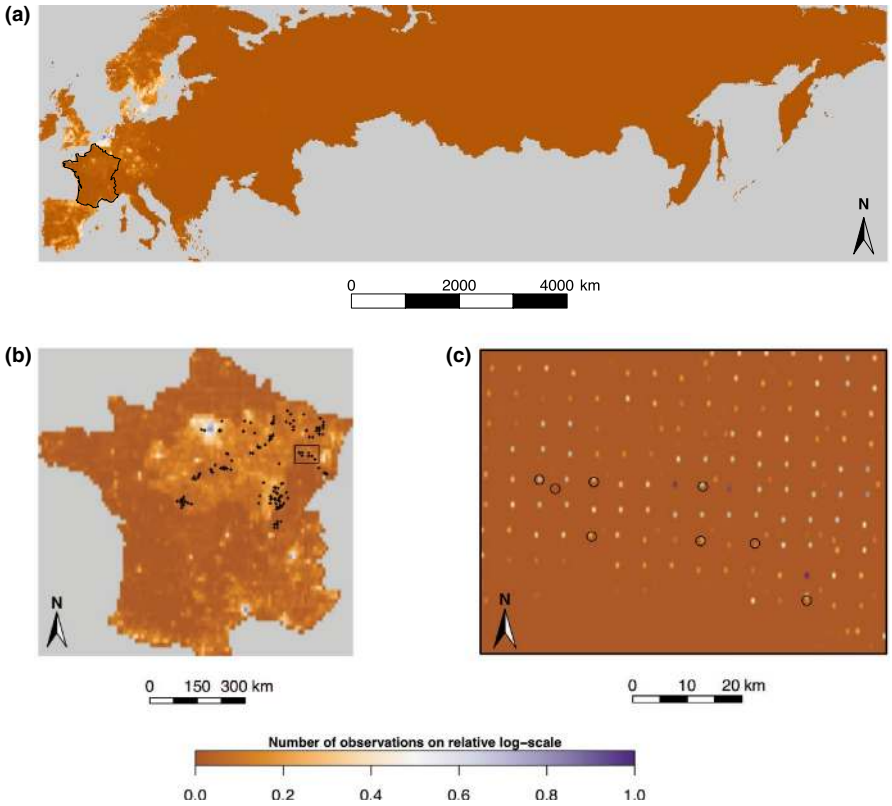

FIGURE 3 a, Number of observations from the BIEN data portal on a relative logarithmic scale in Europe (35–70°N). b, Number of observations in France on a relative logarithmic scale with the observations of *Carex bohemica* indicated as black dots. c, A close-up of the region defined by the black square in (b) shows that most observational data are recorded on a regular grid (ca. 5 km resolution), indicating that the data may stem from an atlas that would not represent point observations (BIEN version 4.1, accessed 5th Nov. 2018) [Colour figure can be viewed at wileyonlinelibrary.com]

the diverse tropics are under-represented (Beck et al., 2012; Meyer, Weigelt, Weigelt, & Kreft, 2016). Bulk collections such as the recently launched GIFT (Global Inventory of Floras and Traits; Weigelt et al., 2019) represent important steps towards filling these data gaps by assembling species lists (checklists) for regions across the globe. Moreover, such aggregated information about where species occur might help to assess biases in data sets based on point locality information (Meyer, Weigelt, et al., 2016).

A third major shortcoming is that many taxonomic groups are poorly represented or entirely absent from currently available data sets (the Linnean shortfall). The meeting highlighted that macroecologists have worked hard in recent years to aggregate data on undersampled taxa that have been collected by specialized taxonomists or field biologists. The fact that these data collectors are opening up their archives and make their data available is catalysing efforts to overcome the Linnean shortfall. For example, a biogeographic assessment of saprotrophic and ectomycorrhizal fungi (Andrew et al., 2018) was based on a Big Data integration effort (Andrew et al., 2017), boosting data availability for such an under-represented group as fungi. Furthermore, data integration and synthesis on marine taxa have allowed researchers to globally synthesize marine diversity (Tittensor et al., 2010), or assess the biogeography of copepod traits (Brun, Payne, & Kiørboe, 2016). Yet, many organism groups remain underexplored at the continental to global scale, and even well-known groups like mammals show significant data gaps (Jones et al., 2009; Meyer, Jetz, Guralnick, Fritz, & Kreft, 2016).

In summary, we have seen major advances in the recent past in aggregating data from various sources to overcome the data gaps. However, despite these major advances, there is still a long way to go, in particular because data gathering that contributes to Big Data in macroecology in itself is not sufficient to achieve a better understanding of macroecological patterns and processes.

3 | DATA HARMONIZATION AND INTEGRATION

The huge efforts to overcome shortcomings in macroecological data in the past years have led to an exponential increase in data becoming available from data portals (Figure 1) such as GBIF (https://www.gbif.org, 2018), MOVEBANK (Wikelski & Kays, 2018), TRY (Kattge et al., 2011), sPLOT (Dengler et al., 2014), BIEN (Enquist et al., 2016) and GIFT (Weigelt et al., 2019). A crucial question is how to integrate this steadily increasing amount of data to obtain meaningful scientific inference.

König et al. (2019) showed that existing data portals and infrastructure predominantly focus on the disaggregated end of the data spectrum (e.g. point occurrences, individual trait measurements), while the wealth of often highly curated information in aggregated data, like species checklists or taxonomic monographs, often remains scattered across different platforms. Integrating information on the distribution of species with functional and/or physiological data in a phylogenetic context is not easy, but key for advancing research at the macroecological scale (Pearse et al., 2018). A recent example of this is the integration of plant trait data from TRY with that of

the plant community database sPlot (Bruelheide et al., 2018) to detect environmental drivers of community trait composition at the community level. To achieve this, it was necessary to overcome the common problem of integrating data from multiple spatial scales (e.g. dealing with varying taxonomies). While vegetation plots are often assessed at a very small spatial grain (usually below 1ha), environmental variables are globally only available at much coarser resolutions (ca. 1 km2; eg. WorldClim, Hijmans, Cameron, Parra, Jones, & Jarvis, 2005; or CHELSA, Karger et al., 2017a, 2017b), which raises the issue of detecting the relevant environmental drivers of community composition or dynamics.

Data integration not only requires data portals that collect and provide integrated data, but also the development of integrative methodologies to cope with heterogeneous data from various sources and of different quality. The last decade has seen an increasing use of such integrative methods. Examples include occupancy models, which use information from repeated observations at each site to estimate detectability and were mainly developed to solve the problems created by imperfect detectability (Kéry, Guillera-Arroita, Lahoz-Monfort, Guillera-Arroita, & Lahoz-Monfort, 2013). Also, joint species distribution models (JSDMs) are increasingly used to model species jointly, rather than using data only from single species, which should support models on species distributions by using information from co-occurrence patterns (Clark, Gelfand, Woodall, & Zhu, 2014; e.g. Ovaskainen & Soininen, 2011; Pollock et al., 2014; Zurell et al., 2019). Furthermore, an increasing number of packages in the free statistical computing software R is devoted to facilitate workflows and data standardization, and these are heavily used by the scientific community (e.g. *Taxonstand*, Cayuela, Cerda, Albuquerque, & Golicher 2012, *CoordinateCleaner*, Zizka, 2019).

The meeting highlighted that statistical methods employed in macroecology are steadily moving towards more integrative methods. For example, Tobler et al. (unpublished) started integrating the two above-mentioned methodologies: a combination of joint modelling of species with models that account for imperfect detection increased the robustness of predicting occurrences. Including intra-specific trait variation and joint trait modelling can improve the prediction of traits along environmental gradients when extrapolating outside the observed environmental range (Wüest, Münkemüller, Lavergne, Pollock, & Thuiller, 2018). Löbel, Mair, Lönnell, Schröder, and Snäll (2018) used a combination of ensemble SDMs and hybrid fourth-corner models based on data from public data sources to identify critical response traits of dead wood inhabiting bryophytes to climate change. Finally, Hof et al. (2018) outlined opportunities and pitfalls in approaches that combine data on land-use and climate change to forecast biodiversity dynamics.

In sum, macroecology continues to be a rapidly progressing subdiscipline of ecology, both in gathering and providing integrated data and in developing new methodologies to integrate heterogeneous data from various sources.

4 | UNCERTAINTY PROPAGATION AND BIAS IN DATA

With the rapid aggregation, integration and harmonization of data, the proliferation and propagation of uncertainties have become an important focus of macroecological research. One of the major challenges in the coming years will be: how can we address problems related to bias and uncertainties in data?

In the recent past, macroecology has learned from adjacent fields such as climate science. In both fields, ensembles are used to embrace uncertainties that arise from using different models – different climate models in climate sciences (see, e.g. Palmer, 2000; Tebaldi & Knutti, 2007), and different species distribution models (SDMs) as well as random subsets of initial distribution data in macroecology (see, e.g. Araújo & New, 2007; Thuiller, 2003). Ensembles in macroecology aim at accessing the uncertainties stemming from a range of variable conditions such as input data, type, structure and complexity of the models, their parameters, or climate scenarios (Dormann et al., 2018, 2012). Research throughout the last decade has indicated that most uncertainty in correlative macroecological biodiversity models stems from the various modelling algorithms that are involved (e.g. Buisson, Thuiller, Csajus, Lek, & Grenouillet, 2010). Yet, uncertainty also originates from the biodiversity measure investigated (Thuiller, Guéguen, Renaud, Karger, & Zimmermann, 2019), and many important uncertainty sources, such as the number of and correlation between variables, as well as the response shape complexity in SDMs, are only being explored now (Brun et al., this volume).

Another source of uncertainty is linked to data: all data sets contain limitations of some sort. While the original data owners (ecologists that collect data in situ) know their data well, macroecologists accessing such data through data portals may not be aware of all data properties and resulting potentials and limitations when using these data. For example, if a macroecologist were to download occurrence data of *Carex bohemica* for France from BIEN through the BIEN R-package (Enquist et al., 2016), she might be happy to find 245 georeferenced observations (Figure 3b). She might be concerned about the coordinate precision, but will diagnose that the WGS84 coordinates are given to more than four digits after the decimal separator and may conclude that the precision of the coordinates is very high (<15 m); certainly high enough to extract climatic data from CHELSA that are available at a resolution of 30 arc sec (ca. 1 km^2; Karger et al., 2017a, 2017b). What macroecologists could easily overlook is that the data might not represent a precise point observation. The inspection of the BIEN data for all species in France reveals that in certain areas, most observations appear to be collected in a regular grid, suggesting that the observations could well be from an Atlas-type collection. This implies that the coordinates associated with the records in fact do not represent a point location but rather represent an area with a footprint of ca. 5 × 5 km. The metadata as provided by the BIEN R-package, however, do not give any indication about the type of observation (all data accessed through the BIEN R-package in November 2018, BIEN version 4.1; Enquist et al., 2016).

In conclusion, inference in macroecology should be based on an appropriate consideration of biases and uncertainties stemming from heterogeneous input data sets. Ideally, best practice standards should be developed and constantly refined (Araújo et al., 2019). With an increasing heterogeneity in data, macroecological inference can only be robust if we consider bias and uncertainties in the data as well. This requires not only that researchers comply with metadata standards at the data gathering stage (such as the Darwin Core, Darwin Core Task Group, 2009), but to also deliver standardized metadata to the end-users of data portals. It also requires that data users properly consider which data are appropriate for their specific research question.

5 | THE EXPLICIT CONSIDERATION OF PROCESSES

Linking process to patterns has been announced as one of the major challenges in macroecology (Beck et al., 2012). Traditionally, macroecology has focused on identifying patterns and has inferred the processes that may have generated the observed patterns (McGill, 2019). Causal inference and projection of the inferred relationships to novel conditions, however, is often hampered by the fact that most methodologies are correlational (e.g. Brown, 1999; Dormann, 2007). How can Big Data approaches contribute to the ongoing quest for identifying processes rather than only documenting patterns in macroecology?

Several approaches are used and were discussed during the meeting to overcome problems in inferring processes from the pattern, with varying success. A first approach is sometimes termed experimental macroecology (Alexander, Diez, Hart, & Levine, 2016) and suggests that macroecology moves forward by experimentally testing assumptions and inferring causal relationships. Transplant experiments of entire communities to mimic expected climatic changes are one example of such an experiment (Alexander, Diez, & Levine, 2015). The challenge here is the massive difference in scale at which experiments are conducted (usually local) and at which macroecological processes operate (regional to global, Currie, 2019). Experiments alone are unlikely to be able to bridge this scale gap, but this might be achieved by combining experimental and observational data, for example, in an Approximate Bayesian Computation framework (Pearse et al., 2018).

Another approach to establish causality is to explicitly model processes and compare the emerging results with observed patterns. The call for process-based models in macroecology is not new (e.g. Brown, 1999), and the continuing diminution of computational limitations has made it possible to include various processes into macroecological analyses. These processes include, for example, physiology-related mechanisms (Kearney & Porter, 2004), microevolutionary dynamics of populations via explicit simulation of the genetic architecture of phenotypes (Schiffers et al., 2014), metapopulation dynamics via explicit simulation of dispersal and local demography across changing environment in distribution models (Juliano S Cabral & Schurr, 2010; Zurell et al., 2016), metacommunity dynamics via inclusion of resource competition and other biotic interactions (Juliano Sarmento Cabral & Kreft, 2012; Münkemüller et al., 2012), macroevolutionary processes (Aguilée, Gascuel, Lambert, & Ferriere, 2018; Cabral, Wiegand, & Kreft, 2019; Jõks & Pärtel, 2018; Rangel et al., 2018) and plate tectonics (Descombes et al., 2018; Leprieur et al., 2016). The current trend in mechanistic macroecology is to include the manifold processes into an integrative modelling framework (Cabral, Valente, & Hartig, 2017; Leidinger & Cabral, 2017; Methorst, Böhning-Gaese, Khaliq, & Hof, 2017; Pontarp et al., 2018; Thuiller et al., 2013; Urban et al., 2016). Indeed, a discussion group at the meeting focusing on mechanistic simulation models attracted many participants integrating various processes in their models and highlighted several motivations in going mechanistic, for example, better theoretical explorations, generalization of concepts and system understanding; unfeasibility of doing similar experiments at macroecological scales; providing feedback to empiricists on hypothesized patterns (i.e. models as hypothesis generator), developing more informative metrics and data requirements; properly linking process to data/Big Data and thus data-constrained process inference. The high number of participants in the discussion group concerning mechanistic models was unprecedented in earlier meetings. Hence, the trend towards mechanistic macroecology will likely continue as macroecologists and biogeographers increase their interest and efforts in representing explicitly eco-evolutionary and environmental mechanisms in simulation models while Big Data becomes more and more informative on constraining these mechanisms.

Another promising avenue to study macroecological processes is to tap into novel information by extracting data from sources that have been underexplored in macroecology. Pinkert et al (unpubl.) presented a study on species interactions from images obtained from GOOGLE™ picture searches, by detecting within the images whether the proboscis of a butterfly was in contact with a potential food plant or not. Automatization of such image processing can be achieved using pattern recognition algorithms from image analysis (e.g. Weinstein, 2015) and might make the detection of biotic interactions from large picture searches through web search engines possible. Although this is a promising approach, it might easily be biased towards very few enigmatic or attractive taxa, such as butterflies on plants with attractive flowers. An approach without such a strong observer bias is the detection of species using GOOGLE™ street view data (Rousselet et al., 2013, although this approach may lead to a spatial and urban bias instead). Finally, such data mining approaches might make it possible to track the process of invasive species spreading along major traffic routes (Nobis et al. unpublished).

Taken together, macroecology is just starting to investigate processes rather than simply documenting patterns. Experimental macroecology is still the most challenging of the three approaches described above, mainly due to the extremely large spatial extents usually investigated in macroecology. Modelling processes and comparing them with observations is certainly becoming more popular, as computing power increases. Literally observing processes in large databases, as shown in the example of pollination in images available

on the web, represents a frontier whose potential application to macroecology is currently hard to foresee.

6 | THE FUTURE OF BIG DATA IN MACROECOLOGY

The meeting showed that Big Data tools and machine learning approaches have already demonstrated the potential to advance the field of macroecology. However, in order to utilize the full potential of Big Data approaches, the macroecology community needs to make a coordinated effort to address the major challenges outlined below. We have highlighted examples where Big Data have already contributed to addressing some of the well-known challenges that macroecology is facing (Beck et al., 2012; Juliano Sarmento Cabral et al., 2017; Pearse et al., 2018). With or without the help of Big Data, these challenges remain.

6.1 | Challenges

One of the future challenges is certainly how to integrate the massive amounts of data that are available in the ecological literature, including data that are gathered for a specific study but have never been integrated with any of the common data portals. We have outlined above how designated web search algorithms, in combination with pattern recognition methods, could be used to gather data on macroecological patterns. Given the vast amount of data available today, the challenge is how we can make use of these data in an efficient way to answer macroecological questions. Another challenge is associated with bulk collections such as GIFT, which provide access to floristic data that are usually based on inventories at the level of administrative units, and which face the challenging problem of integrating biological with environmental data. Areas for which the biological data are available (usually political entities) are often environmentally heterogeneous, and generating the relevant environmental parameters at this scale is a major hurdle that needs to be overcome (see Keil, Belmaker, Wilson, Unitt, & Jetz, 2013 for an example). Furthermore, efforts towards data integration should be converted into a mechanistic understanding of underlying causal relationships, where we see the use of Big Data to constrain parameters of mechanistic models as a promising research avenue.

Given the challenges of error propagation and uncertainties in data, we foresee the need for defining standards in the coming years. Establishing a common set of essential variables (Kissling et al., 2018), standards for biodiversity assessments (Araújo et al., 2019) and a common glossary of terms to facilitate data exchange (the Darwin Core standard, https://dwc.tdwg.org/) are steps in the right direction. They may, however, hardly be achievable across the wide range of applications and data in macroecological research, as data are usually collected with very specific objectives rather than the intention to publish them in a data portal. One challenge is to find appropriate algorithms and transfer functions that account for the specific limitations of data such as inaccuracies in geographic coordinates (e.g. Figure 3). However, even more important is a thorough documentation of data: metadata must not only be carefully established by the data owners, but also be transferred to data portals, and not be ignored by the user of the portals.

Planning and setting up new macroecological research should not only consider the problems that arise from the integration of heterogeneous data sources, the means of analysing them, and the underlying scientific hypotheses. Instead, a carefully planned macroecological study should also account for the computational challenges that it is facing. While data management plans have become a requirement from several funding agencies for submitted research proposals, planning computational demands are usually not required, even though the lack of considering computational demands could lead to failure of the proposed research.

6.2 | Opportunities

Big Data have already proven their value for macroecology by facilitating increasingly complex and integrative analyses. We suggest that macroecologists use their creativity to explore new ways to collect and integrate data. A look outside macroecology, such as at climatology, image analysis or macroevolution, may help in tackling some of the challenges. Modern algorithms that extract data from huge online databases, such as social media or commercial internet platforms, are already integrating data of various sources in a human-readable way without much input from humans (Kambatla et al., 2014). However, algorithms generally subsumed under the term "deep learning" (deep in the sense that these algorithms use many more hidden layers in their networks than the traditional neural networks with one to three hidden layers) have hardly been used by the macroecological community up to now, despite their large potential to solve some of the important data gathering and integration challenges that macroecologists will face in the years to come. A discussion group at the meeting with data scientists from the Big Data community has shown that macroecologists are well prepared to cope with the methodological challenges of such algorithms: many machine learning techniques, such as dimension reduction, clustering or classification to name just a few, are already routinely applied in macroecology.

We are convinced that our community should (and will) use the newest data-driven methods, computational tools and technologies available to tackle scientific questions. Deep learning, a subfield of machine learning, has lately emerged as a very powerful tool for data analysis that works best when large amounts of data, such as image, text or audio, are available. Deep learning techniques such as convolutional neural networks (CNNs) or long short-term memory networks (LSTMs) can be extremely accurate in classification or prediction tasks. However, they are prone to overfitting and are more difficult to interpret than standard statistical and machine learning techniques such as linear regression or decision trees, making them more difficult to use in critical applications when transparency is required (Lecun, Bengio, & Hinton, 2015).

Finally, the discussion group also showed that macroecologists should not be shy to contact data scientists to jointly formulate the needs of the community, and either identify available methods or develop new techniques that can solve the problem at hand. There is a strong potential for interdisciplinary collaboration between macroecologists and data scientists, and data scientists can help when standard techniques cannot be applied off-the-shelf to specific problems encountered in macroecology.

ACKNOWLEDGEMENTS

We are grateful to the macroecology specialist group of the Ecological Society of Germany, Austria and Switzerland (GfÖ) for funding and organizing the 2018 meeting "Macroecology in the age of Big Data." We are grateful to all who participated in the meeting and stimulated discussions with their contributions. D.N.K, D.Z., and N.E.Z. acknowledge funding from the Swiss Data Science Center (SDSC), part of the project SPEEDMIND. J.M.A received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 678841. W.T., D.N.K. and N.E.Z. received funding from the French Agence nationale de la recherche (ANR-18-EBI4–0009) and the Swiss National Science Foundation (20BD21_184131/1), part of the 2018 Joint BiodivERsA-Belmont Forum call (project 'FutureWeb'). N.E.Z. & W.T. further acknowledged support from the SNF/ANR grant 310030L-170059/ANR-16-CE93-004. D.Z. received funding from the German Science Foundation DFG (grant ZU 361/1-1). S.N. considers it a contribution to her Carlsberg Distinguished Associated Professor Fellowship. The illustrated worldview data were funded by the Villum Foundation. M.W. acknowledges the DFG Centre of Excellence 2117 "Centre for the Advanced Study of Collective Behaviour" (ID: 422037984).

ORCID

Rafael O. Wüest https://orcid.org/0000-0001-6047-1945
Niklaus E. Zimmermann https://orcid.org/0000-0003-3099-9604
Damaris Zurell https://orcid.org/0000-0002-4628-3558
Jake M. Alexander https://orcid.org/0000-0003-2226-7913
Susanne A. Fritz https://orcid.org/0000-0002-4085-636X
Christian Hof https://orcid.org/0000-0002-7763-1885
Holger Kreft https://orcid.org/0000-0003-4471-8236
Signe Normand https://orcid.org/0000-0002-8782-4154
Juliano Sarmento Cabral https://orcid.org/0000-0002-0116-220X
Wilfried Thuiller https://orcid.org/0000-0002-5388-5274
Dirk Nikolaus Karger https://orcid.org/0000-0001-7770-6229

REFERENCES

Aguilée, R., Gascuel, F., Lambert, A., & Ferriere, R. (2018). Clade diversification dynamics and the biotic and abiotic controls of speciation and extinction rates. *Nature Communications*, *9*(1), 1–13. https://doi.org/10.1038/s41467-018-05419-7

Alexander, J. M., Diez, J. M., Hart, S. P., & Levine, J. M. (2016). When Climate Reshuffles Competitors: A Call for Experimental Macroecology. *Trends in Ecology & Evolution*, *31*(11), 831–841. https://doi.org/10.1016/j.tree.2016.08.003

Alexander, J. M., Diez, J. M., & Levine, J. M. (2015). Novel competitors shape species' responses to climate change. *Nature*, *525*(7570), 515–518. https://doi.org/10.1038/nature14952

Andrew, C., Halvorsen, R., Heegaard, E., Kuyper, T. W., Heilmann-Clausen, J., Krisai-Greilhuber, I., … Kauserud, H. (2018). Continental-scale macrofungal assemblage patterns correlate with climate, soil carbon and nitrogen deposition. *Journal of Biogeography*, *45*(8), 1942–1953. https://doi.org/10.1111/jbi.13374

Andrew, C., Heegaard, E., Kirk, P. M., Bässler, C., Heilmann-Clausen, J., Krisai-Greilhuber, I., … Kauserud, H. (2017). Big data integration: Pan-European fungal species observations' assembly for addressing contemporary questions in ecology and global change biology. *Fungal Biology Reviews*, *31*(2), 88–98. https://doi.org/10.1016/j.fbr.2017.01.001

Araújo, M. B., Anderson, R. P., Márcia Barbosa, A., Beale, C. M., Dormann, C. F., Early, R., … Rahbek, C. (2019). Standards for distribution models in biodiversity assessments. *Science Advances*, *5*(1), eaat4858. https://doi.org/10.1126/sciadv.aat4858

Araújo, M. B., & New, M. (2007). Ensemble forecasting of species distributions. *Trends in Ecology & Evolution*, *22*(1), 42–47. https://doi.org/10.1016/j.tree.2006.09.010

Beck, J., Ballesteros-Mejia, L., Buchmann, C. M., Dengler, J., Fritz, S. A., Gruber, B., … Dormann, C. F. (2012). What's on the horizon for macroecology? *Ecography*, *35*(8), 673–683. https://doi.org/10.1111/j.1600-0587.2012.07364.x

Brown, J. H. (1999). Macroecology: Progress and Prospect. *Oikos*, *87*(1), 3. https://doi.org/10.2307/3546991

Bruelheide, H., Dengler, J., Purschke, O., Lenoir, J., Jiménez-Alfaro, B., Hennekens, S. M., … Jandt, U. (2018). Global trait–environment relationships of plant communities. *Nature Ecology & Evolution*, *2*(12), 1906. https://doi.org/10.1038/s41559-018-0699-8

Brun, P., Payne, M. R., & Kiørboe, T. (2016). Trait biogeography of marine copepods - an analysis across scales. *Ecology Letters*, *19*(12), 1403–1413. https://doi.org/10.1111/ele.12688

Philipp Brun, Wilfried Thuiller, Yohann Chauvier, Loïc Pellissier, Rafael O. Wüest, Zhiheng Wang, Niklaus E. Zimmermann (in review) Model complexity affects species distribution projections under climate change. *Journal of Biogeography*, this volume.

Buisson, L., Thuiller, W., Csajus, N., Lek, S., & Grenouillet, G. (2010). Uncertainty in ensemble forecasting of species distribution. *Global Change Biology*, *16*(4), 1145–1157. https://doi.org/10.1111/j.1365-2486.2009.02000.x

Cabral, J. S., & Kreft, H. (2012). Linking ecological niche, community ecology and biogeography: Insights from a mechanistic niche model. *Journal of Biogeography*, *39*(12), 2212–2224. https://doi.org/10.1111/jbi.12010

Cabral, J. S., & Schurr, F. M. (2010). Estimating demographic models for the range dynamics of plant species. *Global Ecology and Biogeography*, *19*(1), 85–97. https://doi.org/10.1111/j.1466-8238.2009.00492.x

Cabral, J. S., Valente, L., & Hartig, F. (2017). Mechanistic simulation models in macroecology and biogeography: State-of-art and prospects. *Ecography*, *40*(2), 267–280. https://doi.org/10.1111/ecog.02480

Cabral, J., Wiegand, K., & Kreft, H. (2019). Interactions between ecological, evolutionary, and environmental processes unveil complex dynamics of insular plant diversity. *Journal of*, Biogeography.

Cayuela, L., la Cerda, Í. G., Albuquerque, F. S., & Golicher, D. J. (2012). taxonstand: An r package for species names standardisation in vegetation databases. *Methods in Ecology and Evolution*, *3*(6), 1078–1083. https://doi.org/10.1111/j.2041-210X.2012.00232.x

Clark, J. S., Gelfand, A. E., Woodall, C. W., & Zhu, K. (2014). More than the sum of the parts: Forest climate response from joint species

distribution models. *Ecological Applications*, *24*(5), 990–999. https://doi.org/10.1890/13-1015.1

Clawges, R., Vierling, K., Vierling, L., & Rowell, E. (2008). The use of airborne lidar to assess avian species diversity, density, and occurrence in a pine/aspen forest. *Remote Sensing of Environment*, *112*(5), 2064–2073. https://doi.org/10.1016/j.rse.2007.08.023

Cruzan, M. B., Weinstein, B. G., Grasty, M. R., Kohrn, B. F., Hendrickson, E. C., Arredondo, T. M., & Thompson, P. G. (2016). Small unmanned aerial vehicles (micro-UAVs, drones) in plant ecology. *Applications in Plant Sciences*, *4*(9), 1600041. https://doi.org/10.3732/apps.1600041

Cunliffe, A. M., Brazier, R. E., & Anderson, K. (2016). Ultra-fine grain landscape-scale quantification of dryland vegetation structure with drone-acquired structure-from-motion photogrammetry. *Remote Sensing of Environment*, *183*, 129–143. https://doi.org/10.1016/j.rse.2016.05.019

Currie, D. J. (2019). Where Newton might have taken ecology. *Global Ecology and Biogeography*, *28*(1), 18–27. https://doi.org/10.1111/geb.12842

Darwin Core Task Group. (2009). Darwin Core (Kampmeier G, review manager). Biodiversity Information Standards (TDWG).

Dengler, J., Bruelheide, H., Purschke, O., Chytrý, M., Jansen, F., Hennekens, S. M., … the sPlot Consortium (2014). sPlot – the new global vegetation-plot database for addressing trait-environment relationships across the world's biomes. In L. Mucina, J. N. Price, & J. Kalwij (Eds.), *Biodiversity and vegetation: Patterns*, *processes*, *conservation* (p. 90). Perth: Kwongan Foundation.

Descombes, P., Gaboriau, T., Albouy, C., Heine, C., Leprieur, F., & Pellissier, L. (2018). Linking species diversification to palaeo-environmental changes: A process-based modelling approach. *Global Ecology and Biogeography*, *27*(2), 233–244. https://doi.org/10.1111/geb.12683

Dormann, C. F. (2007). Promising the future? Global change projections of species distributions. *Basic and Applied Ecology*, *8*(5), 387–397. https://doi.org/10.1016/j.baae.2006.11.001

Dormann, C. F., Calabrese, J. M., Guillera-Arroita, G., Matechou, E., Bahn, V., Bartoń, K., … Hartig, F. (2018). Model averaging in ecology: A review of Bayesian, information-theoretic, and tactical approaches for predictive inference. *Ecological Monographs*, *88*(4), 485–504. https://doi.org/10.1002/ecm.1309

Dormann, C. F., Schymanski, S. J., Cabral, J., Chuine, I., Graham, C., Hartig, F., … Singer, A. (2012). Correlation and process in species distribution models: Bridging a dichotomy. *Journal of Biogeography*, *39*(12), 2119–2131. https://doi.org/10.1111/j.1365-2699.2011.02659.x

Enquist, B. J., Condit, R., Peet, B., Schildhauer, M., Thiers, B., & BIEN, working group. (2016). Cyber infrastructure for an integrated botanical information network to investigate the ecological impacts of global climate change on plant biodiversity. *PeerJ Preprints*, *4*, e2615v2. https://doi.org/10.7287/peerj.preprints.2615v2

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*(15), 1965–1978. https://doi.org/10.1002/joc.1276

Hilbert, M., & López, P. (2011). The world's technological capacity to store, communicate, and compute information. *Science*, *332*(6025), 60–65.

Hof, C., Voskamp, A., Biber, M. F., Böhning-Gaese, K., Engelhardt, E. K., Niamir, A., … Hickler, T. (2018). Bioenergy cropland expansion may offset positive effects of climate change mitigation for global vertebrate diversity. *Proceedings of the National Academy of Sciences*, *201807745*, https://doi.org/10.1073/pnas.1807745115

Hortal, J., de Bello, F., Diniz-Filho, J. A. F., Lewinsohn, T. M., Lobo, J. M., & Ladle, R. J. (2015). Seven Shortfalls that Beset Large-Scale Knowledge of Biodiversity. *Annual Review of Ecology*, *Evolution*, *and Systematics*, *46*(1), 523–549. https://doi.org/10.1146/annurev-ecolsys-112414-054400

https://www.gbif.org. (2018). GBIF: The Global Biodiversity Information Facility. What is GBIF? Retrieved from https://www.gbif.org/what-is-gbif.

Huete, A., Didan, K., Miura, T., Rodriguez, E., Gao, X., & Ferreira, L. (2002). Overview of the radiometric and biophysical performance of the MODIS vegetation indices. *Remote Sensing of Environment*, *83*(1–2), 195–213. https://doi.org/10.1016/S0034-4257(02)00096-2

Jõks, M., & Pärtel, M. (2018). Plant diversity in Oceanic archipelagos: Realistic patterns emulated by an agent-based computer simulation. *Ecography*, *42*(4), 740–754. https://doi.org/10.1111/ecog.03985

Jones, K. E., Bielby, J., Cardillo, M., Fritz, S. A., O'Dell, J., Orme, C. D. L., … Purvis, A. (2009). PanTHERIA: A species-level database of life history, ecology, and geography of extant and recently extinct mammals. *Ecology*, *90*(9), 2648–2648. https://doi.org/10.1890/08-1494.1

Kambatla, K., Kollias, G., Kumar, V., & Grama, A. (2014). Trends in big data analytics. *Journal of Parallel and Distributed Computing*, *74*(7), 2561–2573. https://doi.org/10.1016/j.jpdc.2014.01.003

Karami, M., Westergaard-Nielsen, A., Normand, S., Treier, U. A., Elberling, R. W., & Hansen, B. U. (2018). A phenology-based approach to the classification of Arctic tundra ecosystems in Greenland. *ISPRS Journal of Photogrammetry and Remote Sensing*, *146*(September), 518–529. https://doi.org/10.1016/j.isprsjprs.2018.11.005

Karger, D. N., Conrad, O., Böhner, J., Kawohl, T., Kreft, H., Soria-Auza, R. W., … Kessler, M. (2017a). Climatologies at high resolution for the earth's land surface areas. *Scientific Data*, *4*, 170122.

Karger, D. N., Conrad, O., Böhner, J., Kawohl, T., Kreft, H., & Soria-Auza, R. W., … Kessler, M. (2017b). Climatologies at high resolution for the earth's land surface areas. *Dryad Digital Repository*, https://doi.org/10.5061/dryad.kd1d4.

Kattge, J., Díaz, S., Lavorel, S., Prentice, I. C., Leadley, P., Bönisch, G., … Wirth, C. (2011). TRY - a global database of plant traits. *Global Change Biology*, *17*(9), 2905–2935. https://doi.org/10.1111/j.1365-2486.2011.02451.x

Kearney, M., & Porter, W. P. (2004). Mapping the fundamental niche: Physiology, climate, and the distribution of a nocturnal lizard. *Ecology*, *85*(11), 3119–3131. https://doi.org/10.1890/03-0820

Keil, P., Belmaker, J., Wilson, A. M., Unitt, P., & Jetz, W. (2013). Downscaling of species distribution models: A hierarchical approach. *Methods in Ecology and Evolution*, *4*(1), 82–94. https://doi.org/10.1111/j.2041-210x.2012.00264.x

Kéry, M., Guillera-Arroita, G., Lahoz-Monfort, J. J., Guillera-Arroita, G., & Lahoz-Monfort, J. J. (2013). Analysing and mapping species range dynamics using occupancy models. *Journal of Biogeography*, *40*(8), 1463–1474. https://doi.org/10.1111/jbi.12087

Kissling, W. D., Ahumada, J. A., Bowser, A., Fernandez, M., Fernández, N., García, E. A., … Hardisty, A. R. (2018). Building essential biodiversity variables (EBVs) of species distribution and abundance at a global scale. *Biological Reviews*, *93*(1), 600–625. https://doi.org/10.1111/brv.12359

König, C., Weigelt, P., Schrader, J., Taylor, A., Kattge, J., & Kreft, H. (2019). Biodiversity data integration–The significance of data resolution and domain. *PLOS Biology*, *17*(3), e3000183. https://doi.org/10.1371/journal.pbio.3000183

Lecun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436–444. https://doi.org/10.1038/nature14539

Leidinger, L., & Cabral, J. S. (2017). Biodiversity dynamics on Islands: Explicitly accounting for causality in mechanistic models. *Diversity*, *9*(3), 1–17. https://doi.org/10.1109/ICMA.2017.8015786.

Leprieur, F., Descombes, P., Gaboriau, T., Cowman, P. F., Parravicini, V., Kulbicki, M., … Pellissier, L. (2016). Plate tectonics drive tropical reef biodiversity dynamics. *Nature Communications*, *7*, 11461. https://doi.org/10.1038/ncomms11461

Löbel, S., Mair, L., Lönnell, N., Schröder, B., & Snäll, T. (2018). Biological traits explain bryophyte species distributions and responses to

forest fragmentation and climatic variation. *Journal of Ecology*, *106*(4), 1700–1713. https://doi.org/10.1111/1365-2745.12930

Mai, P. M., Wikelski, M., Scocco, P., Catorci, A., Keim, D., Pohlmeier, W., & Fechteler, G. (2018). Monitoring pre-seismic activity changes in a domestic animal collective in Central Italy. *Geophysical Research Abstracts*, *20*, EGU2018–19348.

McGill, B. J. (2019). The what, how and why of doing macroecology. *Global Ecology and Biogeography*, *28*(1), 6–17. https://doi.org/10.1111/geb.12855

Methorst, J., Böhning-Gaese, K., Khaliq, I., & Hof, C. (2017). A framework integrating physiology, dispersal and land-use to project species ranges under climate change. *Journal of Avian Biology*, *48*(12), 1532–1548. https://doi.org/10.1111/jav.01299

Meyer, C., Jetz, W., Guralnick, R. P., Fritz, S. A., & Kreft, H. (2016). Range geometry and socio-economics dominate species-level biases in occurrence information. *Global Ecology and Biogeography*, *25*(10), 1181–1193. https://doi.org/10.1111/geb.12483

Meyer, C., Weigelt, P., & Kreft, H. (2016). Multidimensional biases, gaps and uncertainties in global plant occurrence information. *Ecology Letters*, *19*(8), 992–1006. https://doi.org/10.1111/ele.12624

Münkemüller, T., de Bello, F., Meynard, C. N., Gravel, D., Lavergne, S., Mouillot, D., … Thuiller, W. (2012). From diversity indices to community assembly processes: A test with simulated data. *Ecography*, *35*(5), 468–480. https://doi.org/10.1111/j.1600-0587.2011.07259.x

Ovaskainen, O. O., & Soininen, J. (2011). Making more out of sparse data: Hierarchical modeling of species communities. *Ecology*, *92*(2), 289–295. https://doi.org/10.1890/10-1251.1

Palmer, T. N. (2000). Predicting uncertainty in forecasts of weather and climate. *Reports on Progress in Physics*, *63*(2), 71–116. https://doi.org/10.1088/0034-4885/63/2/201

Pearse, W. D., Barbosa, A. M., Fritz, S. A., Keith, S. A., Harmon, L. J., Harte, J., … Davies, T. J. (2018). Building up biogeography: Pattern to process. *Journal of Biogeography*, *45*(6), 1223–1230. https://doi.org/10.1111/jbi.13242

Pollock, L. J., Tingley, R., Morris, W. K., Golding, N., O'Hara, R. B., Parris, K. M., … McCarthy, M. A. (2014). Understanding co-occurrence by modelling species simultaneously with a Joint Species Distribution Model (JSDM). *Methods in Ecology and Evolution*, *5*(5), 397–406. https://doi.org/10.1111/2041-210X.12180

Pontarp, M., Bunnefeld, L., Cabral, J. S., Etienne, R. S., Fritz, S. A., Gillespie, R., … Hurlbert, A. H. (2018). The latitudinal diversity gradient: Novel understanding through mechanistic eco-evolutionary models. *Trends in Ecology & Evolution*, *34*(3), 211–223. https://doi.org/10.1016/j.tree.2018.11.009

Rangel, T. F., Edwards, N. R., Holden, P. B., Diniz-Filho, J. A. F., Gosling, W. D., Coelho, M. T. P., … Colwell, R. K. (2018). Modeling the ecology and evolution of biodiversity: Biogeographical cradles, museums, and graves. *Science*, *361*(6399), 1–13. https://doi.org/10.1126/science.aar5452

Rousselet, J., Imbert, C.-E., Dekri, A., Garcia, J., Goussard, F., Vincent, B., … Rossi, J.-P. (2013). Assessing species distribution using google street view: A pilot study with the pine processionary moth. *PLoS ONE*, *8*(10), e74918. https://doi.org/10.1371/journal.pone.0074918

Schiffers, K., Schurr, F. M. F. M., Travis, J. M. J. M. J., Duputié, A., Eckhart, V. M. V. M., & Lavergne, S., … Holt, R. D. R. D. (2014). Landscape structure and genetic architecture jointly impact rates of niche evolution. *Ecography*, *37*(12), 1218–1229. https://doi.org/10.1111/ecog.00768.

Tang, L., & Shao, G. (2015). Drone remote sensing for forestry research and practices. *Journal of Forestry Research*, *26*(4), 791–797. https://doi.org/10.1007/s11676-015-0088-y

Tay, J. Y. L., Erfmeier, A., & Kalwij, J. M. (2018). Reaching new heights: Can drones replace current methods to study plant population dynamics? *Plant Ecology*, *219*(10), 1139–1150. https://doi.org/10.1007/s11258-018-0865-8

Tebaldi, C., & Knutti, R. (2007). The use of the multi-model ensemble in probabilistic climate projections. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *365*(1857), 2053–2075. https://doi.org/10.1098/rsta.2007.2076

Thers, H., Brunbjerg, A. K., Læssøe, T., Ejrnæs, R., Bøcher, P. K., & Svenning, J.-C. (2017). Lidar-derived variables as a proxy for fungal species richness and composition in temperate Northern Europe. *Remote Sensing of Environment*, *200*, 102–113. https://doi.org/10.1016/j.rse.2017.08.011

Thuiller, W. (2003). BIOMOD - optimizing predictions of species distributions and projecting potential future shifts under global change. *Global Change Biology*, *9*(10), 1353–1362. https://doi.org/10.1046/j.1365-2486.2003.00666.x

Thuiller, W., Guéguen, M., Renaud, J., Karger, D. N., & Zimmermann, N. E. (2019). Uncertainty in ensembles of global biodiversity scenarios. *Nature Communications*, *10*(1), 1446. https://doi.org/10.1038/s41467-019-09519-w

Thuiller, W., Münkemüller, T., Lavergne, S., Mouillot, D., Mouquet, N., Schiffers, K., & Gravel, D. (2013). A road map for integrating eco-evolutionary processes into biodiversity models. *Ecology Letters*, *16*, 94–105. https://doi.org/10.1111/ele.12104

Tittensor, D. P., Mora, C., Jetz, W., Lotze, H. K., Ricard, D., Berghe, E. V., & Worm, B. (2010). Global patterns and predictors of marine biodiversity across taxa. *Nature*, *466*(7310), 1098–1101. https://doi.org/10.1038/nature09329

Urban, M. C., Bocedi, G., Hendry, A. P., Mihoub, J.-B., Peer, G., Singer, A., … Travis, J. M. J. (2016). Improving the forecast for biodiversity under climate change. *Science*, *353*(6304), aad8466-aad8466. https://doi.org/10.1126/science.aad8466

Waser, L. T., Ginzler, C., Kuechler, M., Baltsavias, E., & Hurni, L. (2011). Semi-automatic classification of tree species in different forest ecosystems by spectral and geometric variables derived from Airborne Digital Sensor (ADS40) and RC30 data. *Remote Sensing of Environment*, *115*(1), 76–85. https://doi.org/10.1016/j.rse.2010.08.006

Weigelt, P., Koenig, C., & Kreft, H. (2019). GIFT - A global inventory of floras and traits for macroecology and biogeography. *Journal of Biogeography*, https://doi.org/10.1111/jbi.13623.

Weinstein, B. G. (2015). MotionMeerkat: integrating motion video detection and ecological monitoring. *Methods in Ecology and Evolution*, *6*(3), 357–362. https://doi.org/10.1111/2041-210X.12320.

Wich, S. A., & Koh, L. P. (2018). *Conservation Drones: Mapping and Monitoring Biodiversity*. Oxford: Oxford University Press.

Wikelski, M., & Kays, R. (2018). Movebank: Archive, analysis and sharing of animal movement data. Hosted by the Max Planck Institute for Ornithology: Retrieved from www.movebank.org.

Wüest, R. O., Münkemüller, T., Lavergne, S., Pollock, L. J. L. J., & Thuiller, W. (2018). Integrating correlation between traits improves spatial predictions of plant functional composition. *Oikos*, *127*(3), 472–481. https://doi.org/10.1111/oik.04420

Xindong, W. U., Zhu, X., Gong-Qing, W. U., Wei Ding, W. U., Zhu, X., … Ding, W. (2014). Data mining with big data. *IEEE Transactions on Knowledge and Data Engineering*, *26*(1), 97–107. https://doi.org/10.1109/TKDE.2013.109

Zellweger, F., Baltensweiler, A., Ginzler, C., Roth, T., Braunisch, V., Bugmann, H., & Bollmann, K. (2016). Environmental predictors of species richness in forest landscapes: Abiotic factors versus vegetation structure. *Journal of Biogeography*, *43*(6), 1080–1090. https://doi.org/10.1111/jbi.12696

Zellweger, F., Frenne, P. D., Lenoir, J., Rocchini, D., & Coomes, D. (2018). Advances in microclimate ecology arising from remote sensing. *Trends in Ecology & Evolution*, *34*(4), 327–341. https://doi.org/10.1016/j.tree.2018.12.012

Zizka, A., Silvestro, D., Andermann, T., Azevedo, J., Ritter, C. D., Edler, D., … Antonelli, A. (2019). CoordinateCleaner: Standardized cleaning of occurrence records from biological collection databases. *Methods in Ecology and Evolution*, *10*(5), 744–751. https://doi.org/10.1111/2041-210X.13152

Zurell, D., Thuiller, W., Pagel, J., Cabral, J. S., Münkemüller, T., Gravel, D., … Zimmermann, N. E. (2016). Benchmarking novel approaches for modelling species range dynamics. *Global Change Biology*, *22*(8), 2651–2664. https://doi.org/10.1111/gcb.13251

Zurell, D., Zimmermann, N. E., Gross, H., Baltensweiler, A., Sattler, T., & Wüest, R. O. (2019). Testing species assemblage predictions from stacked and joint species distribution models. *Journal of Biogeography*.

BIOSKETCH

Rafael O Wüest explores the generation, preservation, and future fate of the diverse facets of biodiversity. He assesses community structure and assembly, models diversity along environmental gradients, and analyses how evolution and biogeography contribute to shape biodiversity patterns across scales.

How to cite this article: Wüest RO, Zimmermann NE, Zurell D, et al. Macroecology in the age of Big Data – Where to go from here? *J Biogeogr*. 2020;47:1–12. https://doi.org/10.1111/jbi.13633