ECOGRAPHY

Research

Uncertainty matters: ascertaining where specimens in natural history collections come from and its implications for predicting species distributions

Arnald Marcer, Arthur D. Chapman, John R. Wieczorek, F. Xavier Picó, Francesc Uribe, John Waller and Arturo H. Ariño

A. Marcer (https://orcid.org/0000-0002-6532-7712) ✉ (arnald.marcer@uab.cat), CREAF, Bellaterra (Cerdanyola del Vallès), Catalonia, Spain and Univ. Autònoma de Barcelona, Bellaterra (Cerdanyola del Vallès), Catalonia, Spain. – A. D. Chapman, Australian Biodiversity Information Services, Melbourne, VIC, Australia. – J. R. Wieczorek, Univ. of California, Berkeley, CA, USA. – F. Xavier Picó (https://orcid.org/0000-0003-2849-4922), Estación Biológica de Doñana (EBD), Consejo Superior de Investigaciones Científicas (CSIC), Sevilla, Spain. – F. Uribe (https://orcid.org/0000-0002-0832-6561), Museu de Ciències Naturals, Barcelona, Catalonia, Spain. – J. Waller, GBIF, Copenhagen, Denmark. – A. H. Ariño (https://orcid.org/0000-0003-4620-6445), Inst. for Biodiversity and Environmental Research (BIOMA) and DATAI, Univ. de Navarra, Pamplona, Spain and Museo de Ciencias de la Univ. de Navarra, Pamplona, Spain.

Ecography
2022: e06025
doi: 10.1111/ecog.06025
Subject Editor: Alice C. Hughes
Editor-in-Chief: Miguel Araújo
Accepted 9 May 2022

Natural history collections (NHCs) represent an enormous and largely untapped wealth of information on the Earth's biota, made available through GBIF as digital preserved specimen records. Precise knowledge of where the specimens were collected is paramount to rigorous ecological studies, especially in the field of species distribution modelling. Here, we present a first comprehensive analysis of georeferencing quality for all preserved specimen records served by GBIF, and illustrate the impact that coordinate uncertainty may have on predicted potential distributions. We used all GBIF preserved specimen records to analyse the availability of coordinates and associated spatial uncertainty across geography, spatial resolution, taxonomy, publishing institutions and collection time. We used three plant species across their native ranges in different parts of the world to show the impact of uncertainty on predicted potential distributions. We found that 38% of the 180+ million records provide coordinates only and 18% coordinates and uncertainty. Georeferencing quality is determined more by country of collection and publishing than by taxonomic group. Distinct georeferencing practices are more determinant than implicit characteristics and georeferencing difficulty of specimens. Availability and quality of records contrasts across world regions. Uncertainty values are not normally distributed but peak at very distinct values, which can be traced back to specific regions of the world. Uncertainty leads to a wide spectrum of range sizes when modelling species distributions, potentially affecting conclusions in biogeographical and climate change studies. In summary, the digitised fraction of the world's NHCs are far from optimal in terms of georeferencing and quality mainly depends on where the collections are hosted. A collective effort between communities around NHC institutions, ecological research and data infrastructure is needed to bring the data on a par with its importance and relevance for ecological research.

NORDIC SOCIETY OIKOS

www.ecography.org

Keywords: ecological niche modelling (ENM), ecological research, GBIF, georeferencing, natural history collections, preserved specimens, species distribution modelling (SDM), uncertainty

Introduction

Natural history collections (NHCs) held in museums, botanical gardens and similar institutions constitute an enormous wealth of irreplaceable information on the Earth's biota. Such information provides crucial baseline knowledge for understanding the present and past distributions of biodiversity on Earth and how it may be affected by global change (Allmon 1994, Shaffer et al. 1998, Boakes et al. 2010, Lister 2011, Bradley et al. 2014, Meineke et al. 2018a). NHCs are especially valuable in research as they are evidence based and constitute the primary source of historical data on biodiversity, going back hundreds of years with broad taxonomic and geographic coverage. These data provide information about species associations and community assemblages through both space and time (James et al. 2018). As such, they constitute a window to the evolutionary processes taking place in response to environmental change (Boakes et al. 2010, Pyke and Ehrlich 2010, Holmes et al. 2016). NHCs have provided invaluable data for studies in fields such as species distribution modelling (SDM) (Gaubert et al. 2006, Mateo et al. 2010), ecophysiology (DeLeo et al. 2019, Tseng and Pari 2019), global change (Lang et al. 2019, Denney and Anderson 2020), phenology (Hart et al. 2014, Kiat et al. 2019), ecological interactions (Kido and Hood 2019), invasion biology (Crawford and Hoagland 2009, Jorissen et al. 2020), conservation (Drew et al. 2017, Lughadha et al. 2019) and public health and safety (Suarez and Tsutsui 2004, Komar et al. 2005).

Despite their immense potential, NHC data are still rather underused in eco-evolutionary and global change research and their potential is still largely untapped (Meineke et al. 2018b, Andrew et al. 2019, Bartomeus et al. 2019), even though their use has been steadily growing during recent decades (Lavoie 2013, Nelson and Ellis 2018, GBIF Secretariat 2021a, Heberling et al. 2021). In large measure, this recent growth has been enabled by monumental digitisation efforts by an increasing number of individual institutions, supported by large-scale biological data mobilisation projects such as DISSCO (Distributed System of Scientific Collections) in Europe (<www.dissco.eu>), iDigBio in North America (<www.idigbio.org>) and the Atlas of Living Australia (ALA) (<www.ala.org.au>) together with the rest of its enabled living atlases (<https://living-atlases.gbif.org>). The Global Biodiversity Information Facility (GBIF, <www.gbif.org>) serves as the major gateway for accessing all of these digitised biological collections. Databasing and giving free and open access to NHCs is a critical step for leveraging the scientific value of these data (Krishtalka and Humphrey 2000, Boakes et al. 2010, Cicero et al. 2017). As of March 2021, GBIF provided access to over 180M digital records of natural history collection specimens. Easy access to these data has represented an increase in use of GBIF-mediated data in scientific journals from 52 articles in 2008 to over 743 in 2020 (GBIF Secretariat 2021a). This trend is expected to continue given that only a fraction of the estimated one to three billion global specimen holdings (Ariño 2010, Marcer et al. 2021b) has been digitised.

In a process called retrospective georeferencing, coordinates are assigned to specimens by interpreting the textual location descriptions, written in their associated tags (Chapman and Wieczorek 2020). This is a critical, difficult, time-consuming and potentially error-prone process, which must be conducted with care (Chapman 2005, Guralnick et al. 2006, Hill et al. 2009, Chapman and Wieczorek 2020). This is because a specimen's locational data, along with its taxonomic identification and time of collection, constitutes the crucial information that determines its value in ecological research. The newly interpreted digital spatial representation can be used to extract environmental information from where the specimen lived using readily available data (e.g. WorldClim ver. 2 (Fick and Hijmans 2017, <www.worldclim.org>) and CHELSA (Karger et al. 2017, <https://chelsa-climate.org>), for global climate data; or the Copernicus Global Land Service, <https://land.copernicus.eu/global>, for land cover data). According to a recent survey of NHCs worldwide (784 respondents from 73 countries; Marcer et al. 2021b), 44% of the reported collections have reached 50% digitisation, of which approximately 60% have coordinates for less than 50% of their records, despite the fact that most specimens carry tagged information on the location where they were collected (Beaman et al. 2004).

However, a pair of coordinates does not suffice to georeference a specimen. Optimally, the result of the georeferencing process would be coordinate-based geometries precisely defined with its corresponding reference system and uncertainty (Wieczorek et al. 2004, Chapman and Wieczorek 2020) or more informative uncertainty imprints based on probability distributions (Guo et al. 2008). Explicit knowledge of uncertainty around the location of a specimen is needed to adequately approximate the relation of an organism to its habitat (but see Smith et al. 2021). Otherwise, erroneous inferences can lead to committing substantial errors (Chapman and Wieczorek 2020). GBIF's data schema provides three specific Darwin Core dataset elements that relate to this information, namely *geodeticDatum*, *coordinatePrecision* and *coordinateUncertaintyInMeters* (<https://dwc.tdwg.org>, Wieczorek et al. 2012). Yet, despite the digitisation programs and the availability of standards and protocols for georeferencing (Wieczorek et al. 2004, 2012, Chapman and Wieczorek 2006, 2020, Bloom et al. 2018), as of March 2021, uncertainty was provided by only one third of the preserved specimen records aggregated in GBIF that have

coordinates. Therefore, NHC data become very difficult to incorporate into robust research frameworks, which require clear-cut knowledge on the reliability of the records' location. By exploring the scientific literature it is relatively easy to find SDM articles where either coordinate uncertainty is not taken into account, or it is not mentioned as part of the data filtering process, or the number of coordinate decimals is used as a surrogate for coordinate uncertainty instead of the actual coordinate uncertainty (McMichael et al. 2014, Biber-Freudenberger et al. 2016, Wicaksono et al. 2017, MacDonald et al. 2020, Mayani-Parás et al. 2021).

The main objective of this study was to provide a first global comprehensive view on the state of the georeferencing of the world's NHCs which is also a measure of the completeness of digitisation programmes or initiatives. We analysed the differences in georeferenced data across a range of factors, including NHC holding institutions, the countries in which they are held, the date and location of the collection event, and higher taxonomy. In addition, we framed the results in the context of ecological research and the implications for it. To this end, we provided three modelling examples to illustrate the importance of considering uncertainty when inferring species distributions from NHC data. These examples correspond to three plant species chosen from very distinct parts of the world to illustrate varying degrees of environmental heterogeneity driven by their respective latitudinal range and their topographic variation.

Material and methods

Dataset preparation and analyses

On 11 March 2021, we downloaded all preserved specimen records from NHCs (filtered by basisOfRecord = 'PRESERVED_SPECIMEN') held at GBIF (derived dataset GBIF.org, data available at <https://doi.org/10.15468/dd.9ched4>), which we stored in an sqlite database (ver. 3.29.0, <www.sqlite.org>) for querying purposes. We kept fields relevant for the objectives of our analysis, namely: *gbifID*, *occurrenceID*, *institutionCode*, *datasetName*, *hasCoordinate*, *decimalLongitude*, *decimalLatitude*, *coordinateUncertaintyInMeters*, *coordinatePrecision*, *verbatimCoordinateSystem*, *verbatimSRS*, *georeferencedDate*, *georeferenceProtocol*, *hasGeospatialIssues*, *issue*, *eventDate*, *continent*, *countryCode*, *kingdom*, *phylum*, *class*, *order*, *family*, *genus* and *acceptedScientificName*. Among these, the values of the following fields are interpreted by GBIF from the original data provided by the sources: *hasCoordinate*, *decimalLatitude*, *decimalLongitude*, *coordinateUncertaintyInMeters*, *coordinatePrecision*, *hasGeospatialIssues*, *issue*, *countryCode*, *kingdom*, *phylum*, *class*, *order*, *family*, *genus*. When dealing with preserved specimens by country of collection and by publishing country, we created a subset of data in order to make them fit for graphical illustration while preserving representativity of the whole. In both cases, we used the criterion of the minimum list of countries which together represented 80% of all records.

In order to visualise spatial resolutions typically used in SDM studies, we classified values in the *coordinateUncertaintyInMeters* field into a new *UNC.CAT* field with categories resulting from binning uncertainty values using the following cutoff points: < 1, < 10, < 100, < 250 m, < 1, < 5, < 10, < 50, < 100, ≥ 100 km. In this study, we considered data fit for use at a given resolution (grid size) when uncertainty was less than or equal to that grid size. Records at longitude = 0 and latitude = 0 were filtered out as this is a known source of error (Zizka et al. 2019). We assessed uncertainty via the explicit *coordinateUncertaintyInMeters* field, excluding records with implicit uncertainty information through the *footprintWKT* field (0.26% of records). We will refer to subsets of records as: *records without coordinates*, *records with coordinates* for those records with only coordinates and *records with uncertainty* for those records with both coordinates and uncertainty. As there is no information on the reliability of uncertainty measures themselves as mediated by GBIF, we assume that the given uncertainties are bonafide estimates of true uncertainty given the resources available to georeferencers. All analyses were done with the R statistical computing environment ver. 4.0.3. (<www.r-project.org>).

Species distribution modelling

To illustrate the importance of considering uncertainty, we modelled the potential distribution of three plant species across their native ranges in different parts of the world: *Rhododendron groenlandicum* (northern Canada and Greenland), *Guazuma ulmifolia* (from northern Mexico to northern Argentina) and *Eucalyptus gongylocarpa* (southwestern Australia). For each species, the number of available occurrences with both coordinates and uncertainty were: 1000, 585 and 325, respectively. We filtered these to one per 30 arc-second grid cell, selecting always the occurrence with the least uncertainty. After this process, the number of occurrences available for the models were 726, 477 and 207, respectively. For predictors, we used the set of bioclimatic variables in WorldClim ver. 2 (Fick and Hijmans 2017) and the percentage of tree cover (Hansen et al. 2013). These models are to illustrate the issues with coordinate uncertainty. Full models would use a larger set of predictors and a more systematic approach (Williams et al. 2012). Potential multicollinearity in predictor variables did not represent a problem since the objective of this modelling exercise was only illustrative with respect to predictions. However, to optimise the number of predictors used, we automatically selected a subset of predictors for each species (Supporting information) with a variance inflation factor below 10 (Dormann et al. 2013). The selection was done using the function *vifstep* from the *usdm* R package ver. 1.1-18 (Naimi et al. 2014). All predictor variables were used at 30 arc-seconds resolution (approx. 1 km).

In order to measure the effects of uncertainty on predictions of potential distributions, we generated 500 possible occurrence datasets taking into account the positional uncertainty of occurrences. For each occurrence in each dataset and

following a uniform probability distribution, we randomly selected one grid cell from within its uncertainty boundary and extracted the cell's environmental values for all predictors. We then randomly split each dataset into 80% occurrences for model training and 20% for testing model accuracy with the area under the receiver-operator curve (AUC) which is fit for our modelling settings (Merow et al. 2013); i.e. same landscape and background sample for each species. Background samples consisted of a single selection of 10 000 random points per modelled species, i.e. models were trained with the same background selection. These were sampled from their respective native areas which correspond to the areas on which predictions were made. Finally, we used the whole set of occurrences and background to generate the 500 models and their predictions using Maxent ver. 3.4.1 within the R *dismo* package ver. 1.3-3 (Hijmans et al. 2020). We chose Maxent with its default parameters (Phillips et al. 2006) for our illustrative modelling example since it is the tool most widely-used (Santini et al. 2021) by the SDM community. Since our goal was only to predict species' potential distributions, we let Maxent select the predictive features automatically as is usually done in machine learning approaches (Phillips et al. 2006, Elith et al. 2011, Merow et al. 2013). In order to ascertain the variability in the predicted modelling ranges, we binarised Maxent continuous models using the maximum sum of specificity and sensitivity (maxSSS) which is a good method when using presence-only data (Liu et al. 2013). To avoid reporting the variability based on extreme cases, i.e. minimum and maximum predicted ranges, we also provide the 5th and 95th percentiles of the predictions ordered by range area. Finally, to further explore the influence of uncertainty on predictions of potential distributions, we analogously prepared 500 additional datasets per species and per a selected set of maximum uncertainty thresholds, i.e. only occurrences below each threshold were used for modelling. The thresholds were chosen based on peaks of reported uncertainty values in GBIF data. We chose 3536 m as the starting threshold and then we selected three more, each a factor of 2 × of the previous one, i.e. 7071, 14 142 and 28 284.

Results

Overall numbers and trends

We retrieved over 180 million records for preserved specimen records available via GBIF at the time of download (Table 1). Records without coordinates formed the largest part (43.62%), followed by records with coordinates (38.23%) and records with coordinates and uncertainty measures (18.15%). These numbers reflect a steep increase in aggregated records in GBIF. Since 2015, more than 80 million specimen records have been added (Fig. 1). Georeferencing efforts increased, too; in January 2015, 38 505 358 records had coordinates, representing 38.2% of a total of 100 913 930. In 2021, this number increased to 103 625 307 records, representing 56.4% of a total of 183 795 180. However, this positive trend is not followed by the data on uncertainty. In 2015, 15 668 937 records had uncertainty information, representing a total of 40.7% of the records with coordinates, while in March 2021, despite the absolute number increasing to 33 352 275, the percentage went down to 32.2% of the records with coordinates. These results indicated an accelerating pace of coordinate digitisation, but at the expense of leaving out crucial uncertainty information.

Table 1. Number and percentages of preserved specimen records split between those without coordinates, records with coordinates and records with uncertainty, downloaded from GBIF on 11 March 2021.

Subset	No.	Percentage
Total	183 795 180	100.0
Records without coordinates	80 169 873	43.62
Records with coordinates	70 273 032	38.23
Records with uncertainty	33 352 275	18.15

Distribution of georeferencing and uncertainty

At a global level

The distribution of records with coordinates was uneven between and within continents (Fig. 2a). The highest densities of records occurred in Europe, the southern half of North America, Central America, scattered parts of South America (e.g. the north–east and the Atlantic coast), southern Africa and Madagascar, Southeast Asia including Japan and Australia. The lowest densities mainly corresponded to Siberia and parts of central Asia, the Sahara and Sahel in Africa, the Arabian desert, Greenland and Antarctica. Oceans showed a pattern ranging from very low densities in areas far from the continents, towards higher densities close to the coasts. The northern Atlantic Ocean and the Southern Ocean south of Australia showed much higher densities than the rest of the oceans. Linear patterns corresponding to oceanic expeditions can be observed.

We found a contrasting distribution of records with uncertainty with respect to records with coordinates, with georeferencing practices clearly differing among regions with respect to their degree of completeness (Fig. 2b). Clusters of records with uncertainty in percentages above 80% (blue colour) can be observed in southern Alaska, eastern Iberian Peninsula, the Alps, Belgium, Finland, Norway, northeastern Russia, northern India and Australia. The Southern Ocean also showed high percentages of records with uncertainty, but corresponding to relatively lower numbers of specimens. The United States also stood out as an area of relatively high density of records with uncertainty, except for parts of the south and midwest. At the lower end of the distribution, several regions stood out: Central America, Canada and parts of the eastern United States, South America, Africa, most of Asia, parts of Europe and the North Atlantic.

By country

A total of 24 countries represented the minimum set with 80% of all preserved land specimen records (Fig. 3a). South Korea, the Netherlands, Brazil, Japan and Indonesia are the

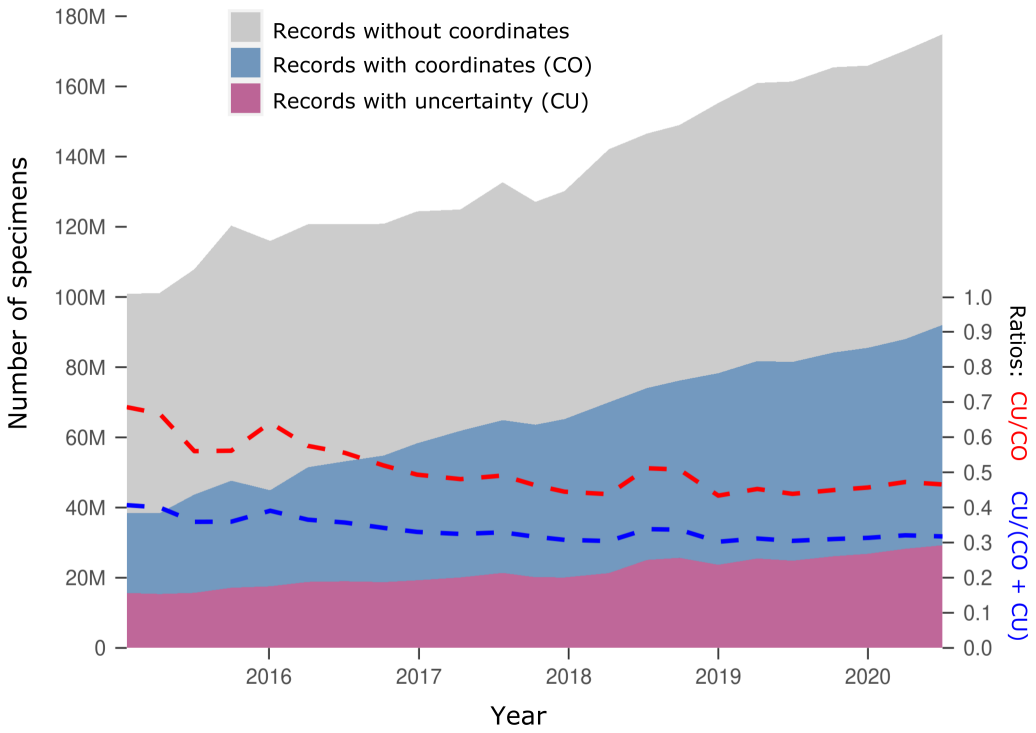

Figure 1. Temporal trend of the number of GBIF's preserved specimen records. In grey are records without coordinates, in blue records with coordinates (CO), and in magenta records with uncertainty (CU). The blue dashed line shows the ratio of CU to CO + CU, a decreasing trend for uncertainty. The red line shows the ration of CU to CO. The ratio of records with coordinates that also have uncertainty is diminishing over time.

countries with over half of their records without coordinates. On the other hand, Costa Rica, Australia, Finland, Canada, Switzerland and Norway had the highest rates of georeferencing, all above 80%. With respect to uncertainty, countries for which more than half of their specimen records provided uncertainty were: Switzerland, Finland, Norway and Australia. Despite the United States not reaching 50%, its 25.9% of records with uncertainty was noteworthy given its very large number of preserved specimens. The remainder of countries in Fig. 3a had less than a fourth of their records with uncertainty, except for Spain with a 34.6%. Costa Rica stood out with the highest percentage of records with coordinates combined with the lowest percentage of records with uncertainty. In absolute terms, the United States, Australia, Brazil, Mexico and Canada represented the areas where most specimens have been collected (Supporting information), which is in part determined by the size of these countries. However, other large countries, e.g. Russia and China, showed substantially fewer numbers of digitised collected specimens.

By publishing country

Eighty percent of the total number of preserved specimens were kept in the institutions of only 13 countries (Supporting information). Institutions in the United States, with a total of 62 623 334 (34.1%) records hold the highest number of preserved specimens, followed by Australian institutions with 12 536 148 (6.8%), and United Kingdom institutions with 10 811 413 (5.9%) (Fig. 3b). Over 90% of the records from institutions in Costa Rica, Australia and Mexico are records with coordinates. Institutions in Norway and Australia have the highest prevalence of assigning uncertainty to coordinates.

By taxonomy

The great majority, 92.7%, of digitised preserved specimen records were from just two kingdoms: Animalia with 46.7% of records and Plantae with 46.0%. Fungi were a distant third with 3.8% (Fig. 4a) and the remaining kingdoms comprised 3.5% of the database. The Animalia kingdom had the highest percentage of records with coordinates (66.8%) and records with uncertainty (21.4%), followed by Plantae with 47.1% and 15.5%, respectively, and Fungi with 54.9% and 21.9%, respectively. The taxonomic schema for kingdoms and phyla used in this work is that of the GBIF Backbone Taxonomy (GBIF Secretariat 2021b).

The same pattern is evident at the phylum level (Fig. 4b). For each kingdom, just two or three phyla represent the great majority of specimen records and the percentage of records with coordinates and records with uncertainty are similar to those seen at the kingdom level. Tracheophyta and Bryophyta represent 91.7% and 4.6% of all plant records, respectively. Arthropoda, Chordata and Mollusca made up 53, 33.4 and 8.7% of all animal records, respectively. Ascomycota and Basidiomycota represent 62.5% and 36.5% of all Fungi records, respectively.

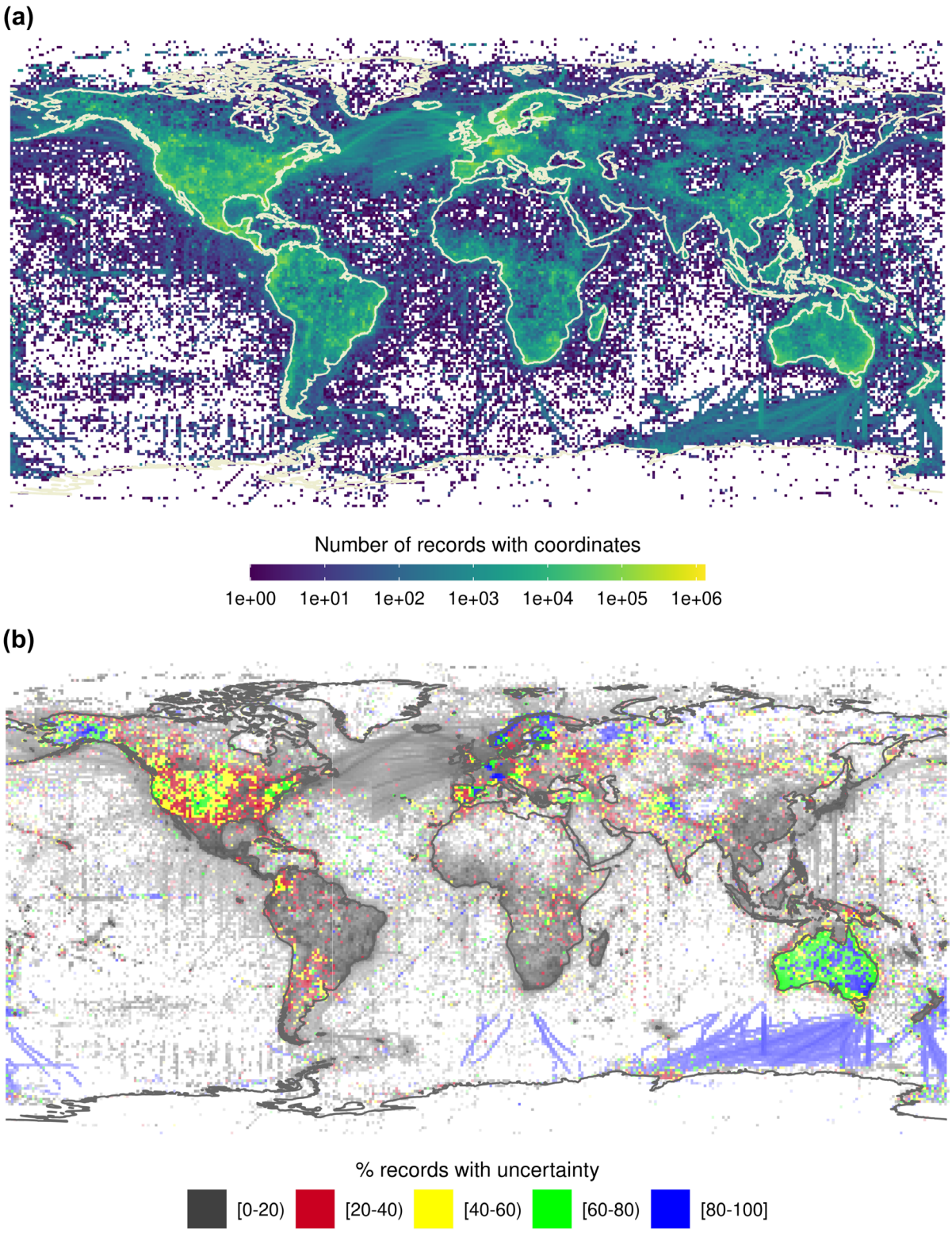

Figure 2. Global distribution of preserved specimens with number of coordinates in one degree by one degree grid squares. (a) Number of records with coordinates (colour indexing is represented on a log scale). (b) Percentage categories for records with uncertainty (colour hue represents percentage categories while colour intensity represents number of records per square degree, i.e. darker colours represent high numbers).

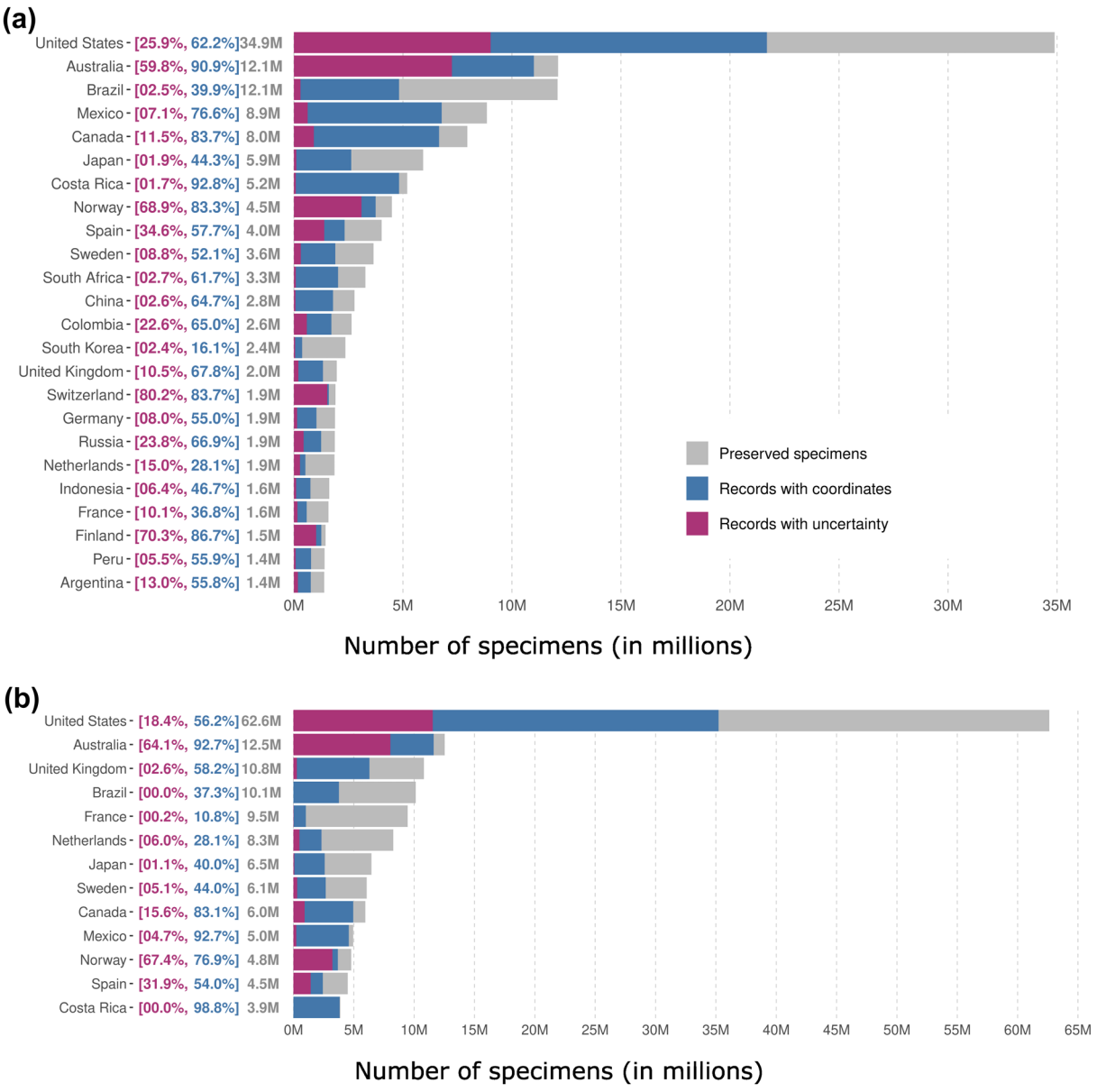

Figure 3. Preserved specimens by (a) country of collection, and (b) publishing country. The minimum set of countries representing at least 80% of the total number of specimens are shown. Number of specimens are expressed in millions (grey numbers) and records with coordinates and records with uncertainty are expressed as percentages of the total (dark blue and red numbers, respectively).

By collection year

Preserved specimen records in GBIF date back to the start of the 17th century, although in very sparse numbers until the late 19th century, when numbers started to rapidly increase (Fig. 5a). Before the 19th century, there were a total of 53 754 registered records, of which 38 624 were records without coordinates, 12 498 were records with coordinates, and 2632 were records with uncertainty. Starting in the 19th century, digitised specimens increased until the end of the 20th century when records in GBIF levelled off. The first two decades of the 21st century showed a sharp decrease in the number of records. Until the 21st century, we observed a lag in the increase of records with coordinates with respect to the total of digitised records. The 21st century showed a large increase in records with coordinates with respect to the total. Records with uncertainty slightly increased until the end of the 20th century, but diminish again in the 21st century. The number of records per unit of time peaked at the end of the 20th century and then started declining , even with a large percentage of records that still need to be digitized (Cocks et al. 2020, Hardisty et al. 2020). This may be due to a combination of reasons, among them an overall decline in collection activity across the world (Bradley et al. 2014, Gardner et al. 2014, Tewksbury et al. 2014), the delay in

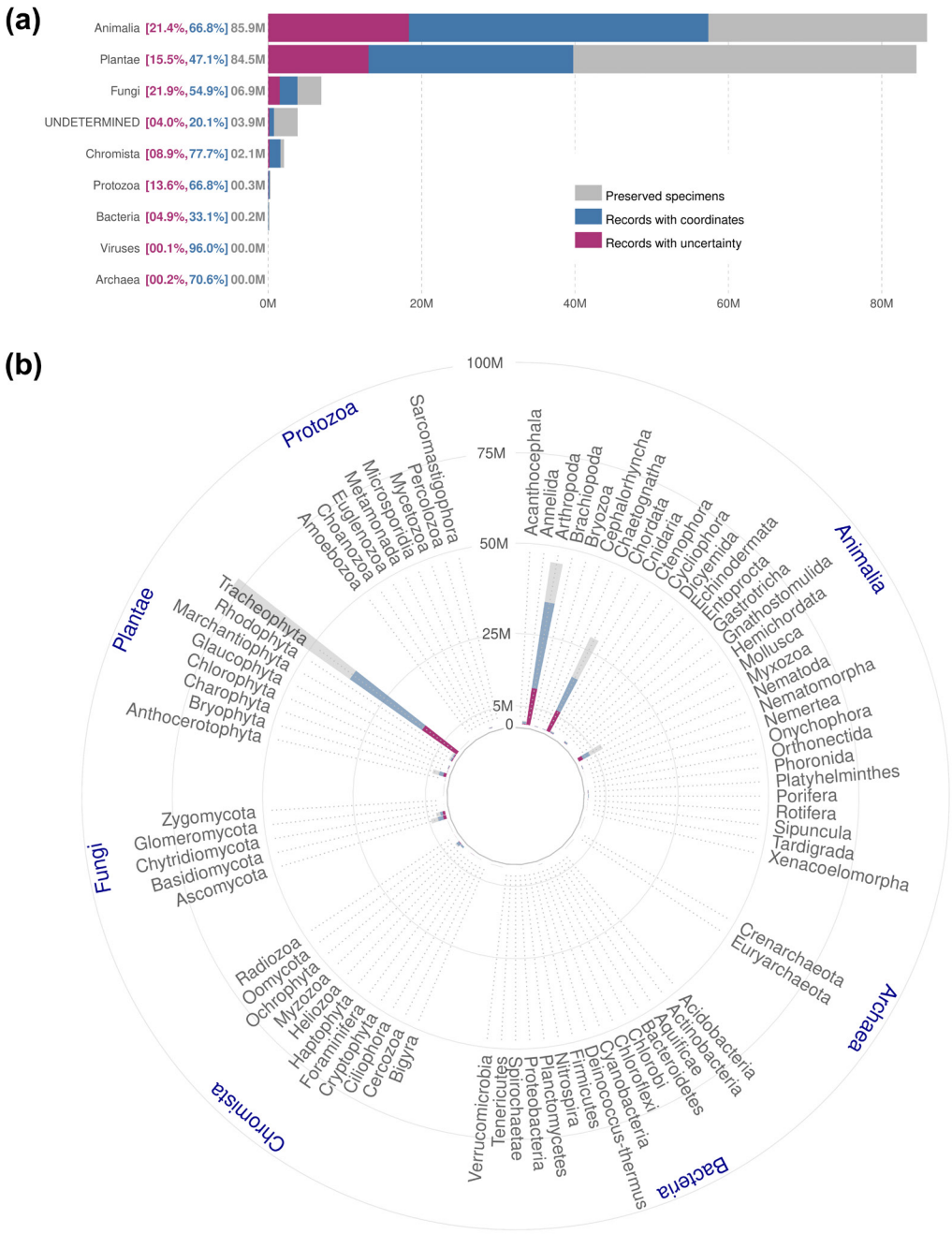

Figure 4. Taxonomic distribution of total number of specimen records. (a) By kingdom, (b) by phylum (viruses and records with unknown kingdom are not represented). Numbers of specimens are expressed in millions (grey numbers) and records with coordinates and records with uncertainty are expressed as percentages of the total (dark blue and red numbers, respectively).

getting new and backlogged material catalogued, digitised and published through GBIF (Gaiji et al. 2013), a global decline in the number of taxonomists and staff in collections institutions and resourcing issues (Noss 1996, Ferreira et al. 2016), or even that the specimens that are easier to georeference have been added at a faster pace than the more difficult ones requiring more intensive manual work.

Records with lower uncertainties belonged to specimens collected more recently (Fig. 5b), which would be in accordance with a relatively greater ease in the interpretation of labels than those collected in earlier centuries (e.g. recent labels are more likely to be typed rather than handwritten), as well as the use of GPS-capable devices in the field. Another pattern was that, despite this observable improvement with respect to uncertainty in more recent records, the variability remained high with uncertainties in the last century encompassing all values from very high to very low. A decrease in records belonging to the higher

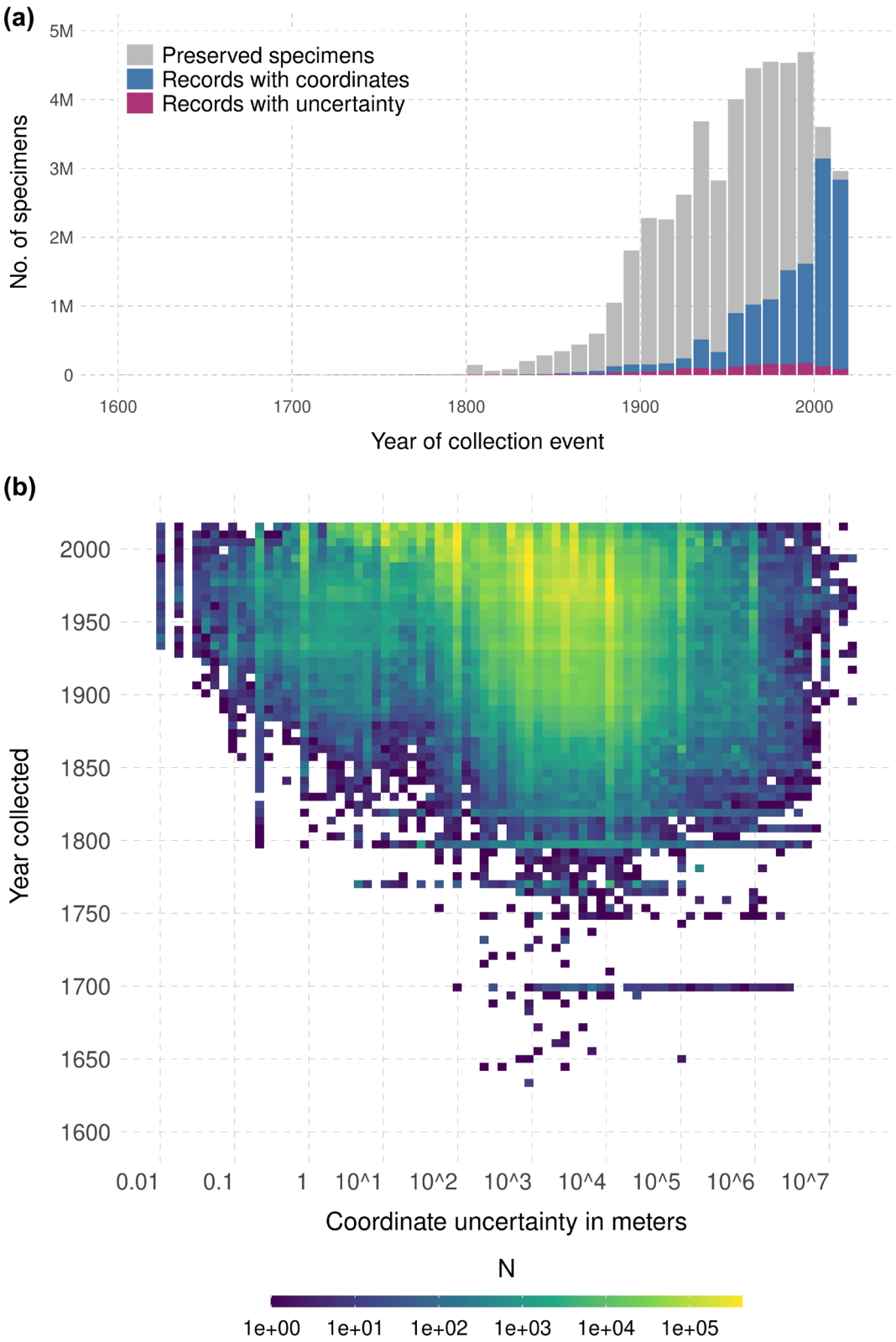

Figure 5. (a) Distribution of overall number of records, records with coordinates, and records with uncertainty by collecting year. (b) Density of coordinate uncertainty values per collecting year.

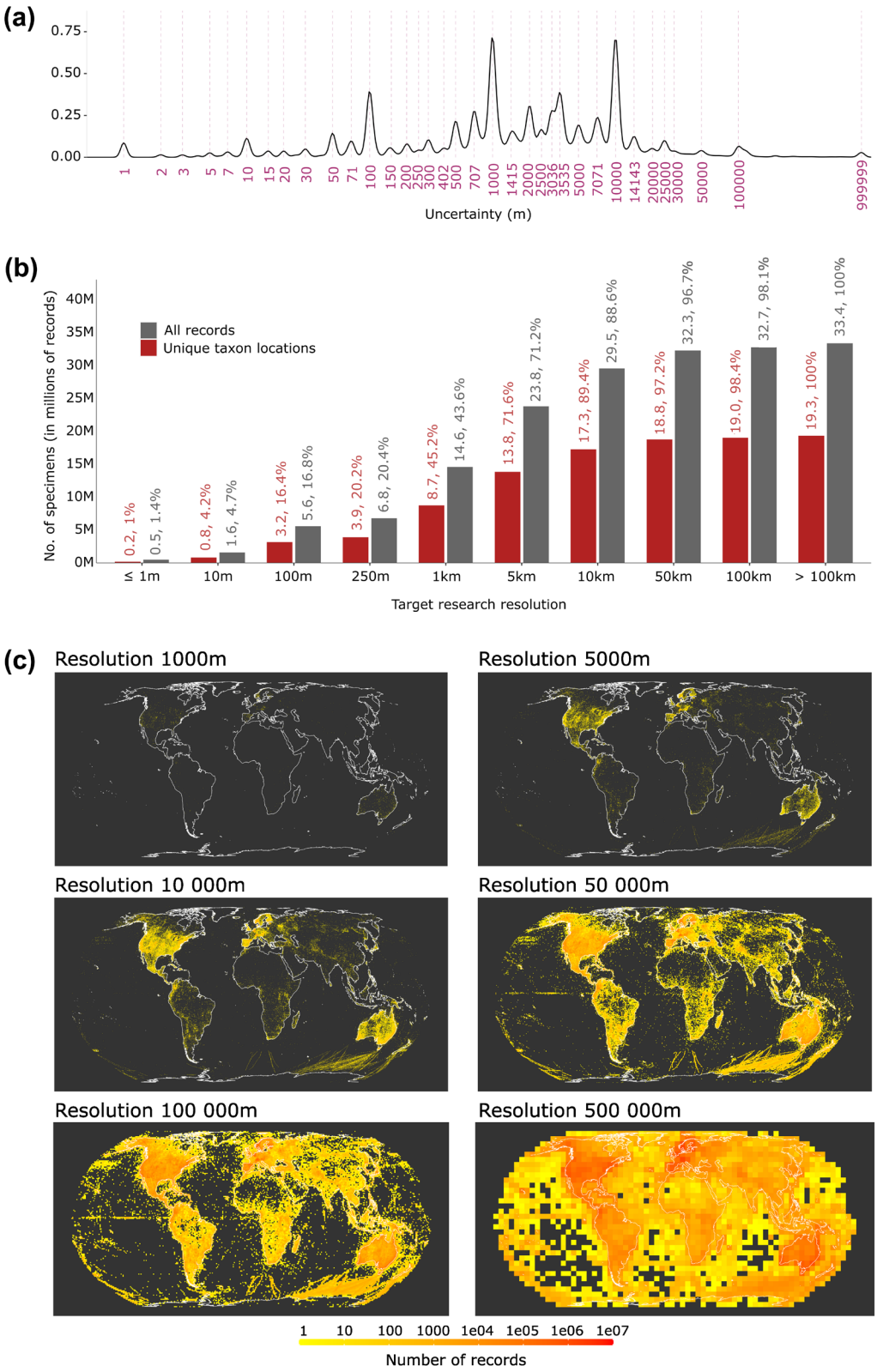

Figure 6. Uncertainty values distribution. (a) Density distribution for uncertainty values for all records with uncertainty. (b) Cumulative distribution of available records fit for research at different spatial resolutions. (c) Density maps of available records fit for research at different spatial resolutions.

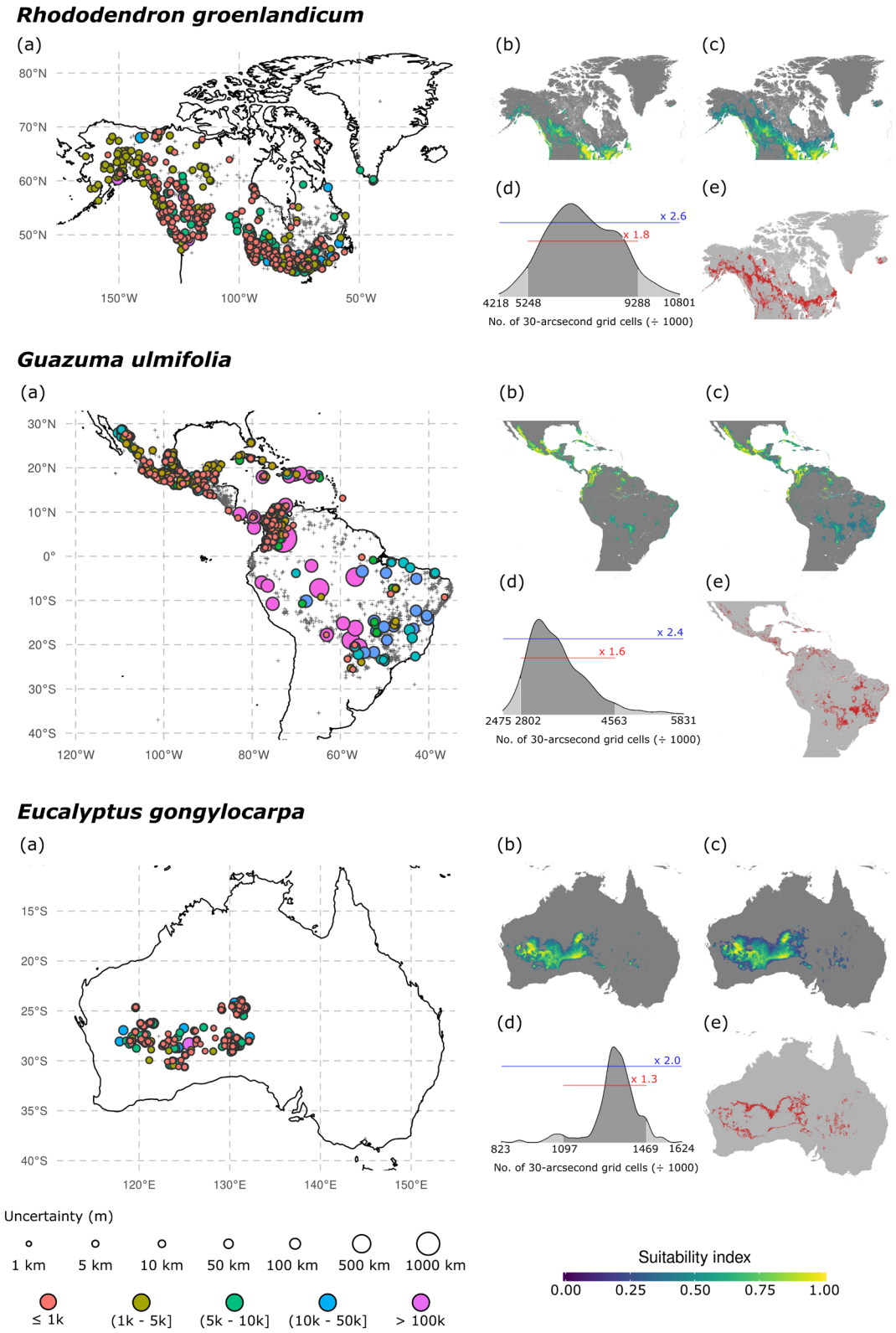

Figure 7. Potential distribution of the three studied species as affected by coordinate uncertainty in preserved specimens. (a) Occurrences with uncertainty boundaries. Small grey crosses represent records with coordinates (no uncertainty) while coloured circles represent classes of uncertainty range. (b) 5th percentile predicted range. (c) 95th percentile predicted range. (d) Distribution of predicted range areas for the 500 simulations. The red line expresses the incremental factor in range area of the 95th percentile with respect to the 5th percentile. The blue red line expresses the incremental factor in range area between the minimum and maximum simulated ranges. (e) Map of differences in predicted potential distributions between the 5th and 95th percentiles.

uncertainty categories was not seen as we looked at increasingly more recent records. This would be in accordance with retrospective georeferencing along with the georeferences that originated in digital form coming from the capture of GPS information.

Distribution of uncertainty values

The distribution of uncertainty values clearly deviated from any normal distribution, being multimodal with peaks at very specific values (Fig. 6a). Three values clearly stood out as the most reported: 10 000, 1000 and 100 m. At a second level, values 3535, 2000, 707 and 500 m were also very apparent. Values ranged from as low as 1 cm to values higher than 1000 km and, along this range, density clearly peaked at very specific values (e.g. 1, 10, 50, 71, 100, …). Some peaks can be clearly attributed to specific georeferencing practices carried out in specific regions (Supporting information); e.g. 402 m (~ 1/4 of a mile) was almost exclusive to the United States (Supporting information), 1415 m (half the diagonal of a square with side equal to 1 km) and 14 143 m (half the diagonal of a square with side equal to 10 km) were mostly assigned in Finland (Supporting information) and 30 000 m was mostly concentrated in Australia (Supporting information). Peaks such as 14 135, 7071, 707, 143 and 71 m related to default estimates of uncertainty calculated using grid diagonal metrics; e.g. 7071 and 707 may correspond to half the diagonal of square grids of sizes 5000 and 500 m, respectively. The dependence of uncertainty values on georeferencing practices was also observed for density distributions of assigned uncertainties by publishing institutions in each country (Supporting information). Different countries can be distinguished by different profiles. We found similar patterns by kingdom, in this case related to georeferencing practices among distinct research communities (Supporting information). However, patterns by taxonomy were much less specific than patterns by publishing countries.

Availability of records for research at different resolutions

The number of available records at different resolutions changes substantially if we consider unique taxon-location combinations, i.e. only one record per taxon and locality (Fig. 6b). This pattern seems mostly contributed by kingdom Animalia (Supporting information). Nonetheless, in both cases, the number of records that were fit for use showed a sharp decrease from resolutions of 10 km down to 1 m. There were a total of 14.6 million records (8.7 million unique taxon-location combinations) which were fit for use at 1 km resolution, a frequently used target resolution for regional level studies. Although this seems like a large number of records, we need to consider that this is the total number of records for the whole world and all species. The same pattern of decline of the number of species available at ever finer resolutions was observed when separated by kingdom (Supporting information), except for Fungi, which showed a much smoother decline in numbers towards finer resolutions. When looking at the records from a global spatial view, the higher numbers of plant records over animals was observable (Supporting information). Global densities of records for animals and plants at fine 1 and 5 km resolutions were similar, except for some regions of clear predominance of plants over animals such as the Iberian Peninsula, Scandinavia, parts of central Europe, Russia and Australia. For animals, the Southern Ocean between Australia and Antarctica showed a higher density of fine resolution records. On the other hand, Fungi have much more sparse records but two zones stand out with higher densities of fine resolution records: southern Scandinavia and Slovenia. Scattered locations in the Iberian Peninsula, the United States and Australia were also perceivable.

Effects of uncertainty on the prediction of potential distributions

Average predictive performance among the 500 models for each species and uncertainty threshold was good to excellent (Supporting information) according to AUC (Swets 1988). The predicted potential distribution for our three example species covered a wide range of sizes when uncertainty was taken into account. The 5th and 95th percentile distributions (Fig. 7b, c) from the 500 simulations for each species differ by a factor of 1.8 in the case of *R. groenlandicum*, 1.6 in the case of *G. ulmifolia* and 1.3 in the case of *E. gongylocarpa* (Fig. 7d, e). These differences are the result of the variation introduced in each dataset when choosing a random occurrence point within the uncertainty boundaries around each occurrence (Fig. 7a). The larger the circle, i.e. uncertainty boundary, the higher the possibility of picking occurrence locations with markedly different environmental predictor values. This is particularly true when the occurrence lies in a locally diverse zone such as a mountainous area (Supporting information). We estimated the predictors' value range, i.e. differences between minimum and maximum values for each occurrence at different uncertainty values (Supporting information). Higher uncertainties had a much wider dispersion of differences between minima and maxima within uncertainty boundaries than occurrences with low uncertainty. Also, for all common predictors, *E. gongylocarpa* shows the least dispersion of values (Supporting information), which corresponds with the least variation in predicted ranges (Fig. 7d). The waste in the number of records that a conservative and robust modelling entails could also be appreciated. For *R. groenlandicum* we could only use 1000 records with uncertainty out of a total 2228 with coordinates, for *G. ulmifolia* it was 585 out of 7382 and for *E. gongylocarpa* 325 out of 385. Finally, there is still considerable variation in the predicted ranges even when using data which has been limited in occurrence uncertainty. Predicted range sizes between the 5th and 95th percentiles for all species and maximum uncertainty thresholds vary in percentage from 32% up to 89% (Supporting information).

Discussion

We analysed more than 180 million records from GBIF to provide a first detailed analysis of the world's preserved specimen records in terms of georeferencing quality across continents, taxonomy, publishing country and year of collection. Georeferencing quality is crucial for ecological research as it allows to rigorously retrieve the environmental conditions of where the species lived and the specimen was collected. We illustrated this with the three plant species by showing the effects that the incorporation of uncertainty in species distribution modelling can have in predicted species ranges and the waste in records which can not be incorporated into modelling due to incomplete georeferencing. The three species were selected from three different parts of the world to illustrate varying degrees of environmental heterogeneity driven by their respective latitudinal range and their topographic variation.

Digitisation efforts make it possible to easily retrieve NHC data via GBIF. Yet, most of the occurrence records of the world's NHCs remain in analogue form. Despite this relatively large number of available records, not all are ready to use in ecological research that requires information on the environment in which these specimens lived. Only 104 million records have coordinates, and only 33 million also have information on coordinate uncertainty. The trends (Fig. 1) indicate the prospects are also not good for a short term improvement. While georeferencing seems to keep pace with digitisation, the ratio of records with coordinate uncertainty to coordinate-only diminishes. This trend probably reflects the prioritisation of quantity over completeness in a rush to bring online as many digital specimens as fast as possible. Georeferencing is difficult to automate and labour intensive to do well. Rate estimates for complete, high quality georeferencing, considering median-type specimens in terms of complexity and optimal conditions of access to digital cartographic resources, range between 16.6 sites per hour per georeferencer (Wieczorek et al. 2004) and 30 when using tools that are more recent, such as GeoLocate (<www.geo-locate.org/default.html>) (Wieczorek pers. comm.). A back-of-the-envelope calculation with these rates would mean somewhere between five months and more than three years to georeference one million specimens with a team of 10 georeferencers, depending on the degree of reuse of already georeferenced sites (Supporting information). This gives an idea of the huge task ahead to fully georeference the world's billions of preserved specimens (Ariño 2010, Marcer et al. 2021a). In addition, the disquieting situation of the current state of georeferences in the existing global digital dataset does not make things any better. Currently, incorrect or incomplete georeferences can only be corrected at the source, and corrected records have to be uploaded again to GBIF. At present, there are no stable identifiers for occurrence records, without which there is no way to annotate occurrence records by a third party system in a stable or lasting way (John Waller, pers. comm.). On another level, GBIF holds over 1.6 billion records of human observations, mainly from community science initiatives such as the Cornell Lab of Ornithology (<www.birds.cornell.edu>), representing the great majority of records and a fast growing source of data on species occurrences which may help in increasing the number of records available for biogeographical studies. However, the nature of these data is quite different and care needs to be taken when using them.

Coordinates do not suffice to know confidently and rigorously the environmental conditions of a specimen's preferred habitat (Gábor et al. 2019). In fact, the knowledge of the degree of uncertainty with which these coordinates have been determined is crucial to determine the fitness of data for a particular research objective. We also need explicit documentation of the reference system, including the datum, on which the coordinates are based. Without this information, there is a risk of committing substantial errors. For example, coordinates are interpreted by GBIF as decimal degrees in WGS84. If the coordinate reference system is not documented or recognizable in the original record, GBIF automatically assigns WGS84 and flags the corresponding record as datum assumed. This is an important issue to be considered when using GBIF-mediated data since the incorrect assumption of the datum can potentially lead to significant additional uncertainty (Chapman and Wieczorek 2020, Konowalik and Nosol 2021). In some cases, such as records using old datums, the error derived from misspecifying the datum can amount to values over 5 km in certain parts of the world (Chapman and Wieczorek 2020, Wieczorek and Wieczorek 2021).

Not appropriately documenting the location information of a specimen during the georeferencing process leads to a waste of effort as the records cannot be confidently used in research, despite some concrete use case scenarios (Smith et al. 2021). In case of doubt, the lack of metadata may leave the final user with the only choice of discarding the record if the given coordinate and uncertainty cannot be checked against original source and methods. Having correct georeferences with their associated metadata is especially relevant in the field of species distribution or environmental niche modelling, a fast-growing ecological research area (Pecchi et al. 2019). Models depend on the assumption that the occurrences' coordinates truly reflect the habitat where the modelled species lives and are used to extract that information from other spatial data layers. When the uncertainty is larger than the target resolution of the study, point estimates of environmental conditions are not sufficient. At any given predictor resolution, a coordinate with an uncertainty larger than the resolution will certainly encompass more than one raster grid cell. In other words, the larger the uncertainty in relation to the resolution, the larger the number of grid cells and the larger the differences between possible values (Fig. 7 and Supporting information). This will be especially true in environmentally diverse landscapes such as mountainous areas, while in extensive topographically homogeneous areas (e.g. large

parts of the Amazon basin) uncertainty will result in a lesser effect due to the higher homogeneity of climate surfaces (Supporting information). This can result in pronounced differences between the different range estimates that result from models using different combinations of environmental values that come from the many different possible configurations of occurrences.

In our modelling examples, with 500 simulations we detected notable differences between the 5th and 95th percentiles of predicted ranges for each species. The least difference represented a factor of increase of 1.3 for *E. gongylocarpa* and the maximum difference represented a factor of 1.8 for *R. groenlandicum* (Fig. 7d). Moreover, limiting the degree of uncertainty in occurrence data does not necessarily result in less variation in predicted potential distributions (Supporting information). Even using occurrences limited to a maximum uncertainty of 3536 m may still result in considerable variation in predicted distributions. These results indicate that not taking uncertainty into consideration may profoundly mislead biogeographical, conservation or global change studies. On another level, a different source of error which may affect biogeographical studies is the pervasive bias that exists in species observation data from repositories such as GBIF (Hughes et al. 2021), and, in the case of species distribution modelling from NHC data, especially the spatial bias of collected specimens (Phillips et al. 2009). Finally, the SDM results presented in this work need to be carefully interpreted as they are simplified examples of the effects of spatial uncertainty on SDMs. Real case studies of species distributions should follow a systematic approach for the selection of relevant environmental predictors in order to deliver robust predictive models (Williams et al. 2012).

Clear and detailed guidelines for quality georeferencing have long existed (Chapman and Wieczorek 2006, 2020). However, our exploration clearly shows that GBIF-mediated data are clearly not on a par with them, as shown recently (Marcer et al. 2021a, b). Specimens with coordinates and with both coordinates and uncertainty are not evenly spread across the world. Some areas stand out as rich with uncertainty information, mainly in North America, Europe and Australia (Fig. 2a, b). Only Australia, Switzerland, Norway and Finland have over half of their records documented with uncertainty (Fig. 3). The United States stands out because of its absolute number of records. We observe similar percentages of records with uncertainty among taxonomic kingdoms. Georeferencing quality seems more related to different georeferencing practices between country cultures, when georeferencing occurred, than taxonomic communities; i.e. communities of georeferencers are more different between countries than between taxonomic kingdoms (Supporting information).

The overall distribution of uncertainty values is multimodal (Fig. 6a), peaking at very specific values which can be traced back to localised practices (see the examples given in the Results section and in the Supporting information, e.g. q, u, y). As a result, the availability of records at different resolutions that can be used in ecological research studies is very dependent on the geographic region (Fig. 6b, c). The available records at different resolutions and their global spatial pattern is similar for the major taxononomic kingdoms, i.e. Animalia, Plantae and Fungi (Supporting information). Another potential factor in determining georeferencing quality is the year of the collection event. For example, a centuries old handwritten specimen tag is more difficult to interpret than a modern tag as it may refer to places that changed name and are more difficult to georeference. However, although this is the case for coordinate-only records (Fig. 5a), it is not the case for records with uncertainty. Although there is a tendency of having lower uncertainty values in more recent specimens, probably due to the use of GPS technologies, the range of uncertainty values is well spread over the range of collecting years (Fig. 5b).

Conclusions

In summary, this study represents a first exploration of the global effort spent on georeferencing the world's preserved specimens in NHCs and highlights the existing gap between currently available data and hoped-for full georeferencing information which can inform about the value for use in ecological research. This gap hinders the potential of global digitisation efforts for research and diminishes the return on investment in georeferencing projects. We advocate for NHC institutions to embrace and document uncertainty by following best practices of georeferencing, even at the expense of diminishing the rate of sharing records with spatial information. We also encourage the ecological research community to include uncertainty when downloading data from digital data repositories and to take it always into consideration when modelling species distributions. We also suggest to provide feedback to NHC institutions when downloaded datasets have been corrected or improved. To facilitate this, though, it is necessary to make an overhaul of the existing digital infrastructure in order to allow repatriation of data improved by users back to the original databases, benefitting both researchers and the NHC community (Cicero et al. 2017).

Acknowledgements – We are grateful to all the NHCs institutions and GBIF for the enormous collective effort in making the data on preserved specimens readily available on the Internet. Without their contribution no study of this kind would have been possible. We want to give special thanks to the personnel at these institutions, which, despite the prevalent underfunding, make it possible to undertake the gargantuan effort of digitising the world's billions of collected specimens. We also wish to thank five anonymous reviewers whose comments have led to an improved manuscript.

Funding – This research has been supported by grants PID2019-104135GB-I00 from the Agencia Estatal de Investigación (AEI) of Spain and the Fondo Europeo de Desarrollo Regional (FEDER, UE), and EU Cost Action MOBILISE (code reference CA17106) and the Agència de Gestió d'Ajuts Universitaris i de Recerca 2017 SGR 1006.

Author contributions

Arnald Marcer: Conceptualization (lead); Data curation (lead); Formal analysis (lead); Methodology (lead); Writing – original draft (lead); Writing – review and editing (lead). **Arthur D. Chapman**: Conceptualization (equal); Methodology (equal); Writing – original draft (supporting); Writing – review and editing (equal). **John R. Wieczorek**: Conceptualization (equal); Methodology (equal); Writing – original draft (supporting); Writing – review and editing (equal). **F. Xavier Picó**: Conceptualization (equal); Methodology (equal); Writing – original draft (supporting); Writing – review and editing (equal). **Francesc Uribe**: Conceptualization (equal); Methodology (equal); Writing – original draft (supporting); Writing – review and editing (equal). **John Waller**: Conceptualization (equal); Data curation (lead); Methodology (equal); Writing – original draft (supporting); Writing – review and editing (equal). **Arturo H. Arino**: Conceptualization (equal); Methodology (equal); Writing – original draft (supporting); Writing – review and editing (equal).

Transparent peer review

The peer review history for this article is available at <https://publons.com/publon/10.1111/ecog.06025>.

Data availability statement

Data are available from the Zenodo Digital Repository: <https://doi.org/10.5281/zenodo.5052596>, (derived dataset GBIF.org (6 July 2021)) (Marcer et al. 2022).

Supporting information

The Supporting information associated with this article is available with the online version.

References

Allmon, W. D. 1994. The value of natural history collections. – Curator Museum J. 37: 83–89.

Andrew, C. et al. 2019. Fungarium specimens: a largely untapped source in global change biology and beyond. – Phil. Trans. R. Soc. B 374: 20170392.

Ariño, A. H. 2010. Approaches to estimating the universe of natural history collections data. – Biodivers. Inform. 7: 81–92.

Bartomeus, I. et al. 2019. Historical collections as a tool for assessing the global pollination crisis. – Phil. Trans. R. Soc. B 374: 20170389.

Beaman, R. et al. 2004. Determining space from place for natural history collections: in a distributed digital library environment. – D-Lib Magaz. 10. <http://dlib.org/dlib/may04/beaman/05beaman.html>.

Biber-Freudenberger, L. et al. 2016. Future risks of pest species under changing climatic conditions. – PLoS One 11: e0153237.

Bloom, T. D. S. et al. 2018. Why georeferencing matters: introducing a practical protocol to prepare species occurrence records for spatial analysis. – Ecol. Evol. 8: 765–777.

Boakes, E. H. et al. 2010. Distorted views of biodiversity: spatial and temporal bias in species occurrence data. – PLoS Biol. 8: e1000385.

Bradley, R. D. et al. 2014. Assessing the value of natural history collections and addressing issues regarding long-term growth and care. – BioScience 64: 1150–1158.

Chapman, A. D. 2005. Principles and methods of data cleaning – primary species and species occurrence data, ver. 1.0. – Report for the Global Biodiversity Information Facility, <www.gbif.org/document/80528>.

Chapman, A. D. and Wieczorek, J. R. (eds) 2006. Guide to best practices for georeferencing. – GBIF Secretariat.

Chapman, A. D. and Wieczorek, J. R. 2020. Georeferencing best practices. – GBIF Secretariat.

Cicero, C. et al. 2017. Biodiversity informatics and data quality on a global scale 1. – In: Webster, M. S. (ed.), The extended specimen. emerging frontiers in collections-based ornithological research. Taylor & Francis, pp. 201–218.

Cocks, N. et al. 2020 Technical capacities of digitisation centres within ICEDIG participating institutions. – Res. Ideas Outcomes 6: e55522.

Crawford, P. H. C. and Hoagland, B. W. 2009. Can herbarium records be used to map alien species invasion and native species expansion over the past 100 years? – J. Biogeogr. 36: 651–661.

DeLeo, V. L. et al. 2019. Effects of two centuries of global environmental variation on phenology and physiology of *Arabidopsis thaliana*. – Global Change Biol. 26: 523–538.

Denney, D. A. and Anderson, J. T. 2020. Natural history collections document biological responses to climate change: a commentary on DeLeo et al. 2019. – Global Change Biol. 26: 340–342.

Derived dataset GBIF.org (6 July 2021) filtered export of GBIF occurrence data, <https://doi.org/10.15468/dd.9ched4>.

Dormann, C. F. et al. 2013. Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. – Ecography 36: 27–46.

Drew, J. A. et al. 2017. Digitization of museum collections holds the potential to enhance researcher diversity. – Nat. Ecol. Evol. 1: 1789–1790.

Elith, J. et al. 2011. A statistical explanation of MaxEnt for ecologists. – Divers. Distrib. 17: 43–57.

Ferreira, C. et al. 2016 Hail local fieldwork, not just global models. – Nature 534: 326.

Fick, S. E. and Hijmans, R. J. 2017. WorldClim 2: new 1-km spatial resolution climate surfaces for global land areas. – Int. J. Climatol. 37: 4302–4315.

Gábor, L. et al. 2019. The effect of positional error on fine scale species distribution models increases for specialist species. – Ecography 43: 256–269.

Gaiji, S. et al. 2013 Content assessment of the primary biodiversity data published through GBIF network: status, challenges and potentials. – Biodivers. Inform. 8: 94–72.

Gardner, J. L. et al. 2014 Are natural history collections coming to an end as time-series? – Front. Ecol. Environ. 12: 436–438.

Gaubert, P. et al. 2006. Natural history collections and the conservation of poorly known taxa: ecological niche modeling in central African rainforest genets (Genetta spp.). – Biol. Conserv. 130: 106–117.

GBIF Secretariat 2021a. GBIF Science Review 2020. – <https://doi.org/10.35035/bezp-jj23>.

GBIF Secretariat 2021b. GBIF Backbone Taxonomy. – Checklist dataset <https://doi.org/10.15468/39omei>, accessed via GBIF.org 11 March 2021.

Guo, Q. et al. 2008. Georeferencing locality descriptions and computing associated uncertainty using a probabilistic approach. – Int. J. Geogr. Inform. Sci. 22: 1067–1090.
Guralnick, R. et al. 2006. BioGeomancer: automated georeferencing to map the world's biodiversity data. – PLoS Biol. 4: e381.
Hansen, M. C. et al. 2013. High-resolution global maps of 21st-century forest cover change. – Science 342: 850–853, <http://earthenginepartners.appspot.com/science-2013-global-forest>.
Hardisty, A. et al. 2020 Conceptual design blueprint for the DiSSCo digitization infrastructure – DELIVERABLE D8.1. – Res. Ideas Outcomes 6: e54280.
Hart, R. et al. 2014. Herbarium specimens show contrasting phenological responses to Himalayan climate. – Proc. Natl Acad. Sci. USA 111: 10615–10619.
Heberling, J. M. et al. 2021. Data integration enables global biodiversity synthesis. – Proc. Natl Acad. Sci. USA 118: e2018093118.
Hijmans, R. J. et al. 2020. dismo: species distribution modeling. – R package ver. 1.3-3. <https://rspatial.org/raster/sdm>.
Hill, A. W. et al. 2009. Location, location, location: utilizing pipelines and services to more effectively georeference the world's biodiversity data. – BMC Bioinform. 10: S3.
Holmes, M. W. et al. 2016. Natural history collections as windows on evolutionary processes. – Mol. Ecol. 25: 864–881.
Hughes, A. C. et al. 2021. Sampling biases shape our view of the natural world. – Ecography 44: 1259–1269.
James, S. A. et al. 2018. Herbarium data: global biodiversity and societal needs for novel research. – Appl. Plant Sci. 6: e1024.
Jorissen, M. W. P. et al. 2020. Historical museum collections help detect parasite species jumps after tilapia introductions in the Congo Basin. – Biol. Invas. 22: 2825–2844.
Karger, D. N. et al. 2017. Climatologies at high resolution for the earth's land surface areas. – Sci. Data 4: 170122.
Kiat, Y. et al. 2019. Feather moult and bird appearance are correlated with global warming over the last 200 years. – Nat. Commun. 10: 2540.
Kido, A. and Hood, M. E. 2019. Mining new sources of natural history observations for disease interactions. – Am. J. Bot. 107: 3–11.
Komar, O. et al. 2005. West Nile virus survey of birds and mosquitoes in the Dominican Republic. – Vector-Borne Zoon. Dis. 5: 120–126.
Konowalik, K. and Nosol, A. 2021. Evaluation metrics and validation of presence-only species distribution models based on distributional maps with varying coverage. – Sci. Rep. 11: 1482.
Krishtalka, L. and Humphrey, P. S. 2000. Can natural history museums capture the future? – BioScience 50: 611.
Lang, P. L. M. et al. 2019. Using herbaria to study global environmental change. – New Phytol. 221: 110–122.
Lavoie, C. 2013. Biological collections in an ever changing world: herbaria as tools for biogeographical and environmental studies. – Perspect. Plant Ecol. Evol. Syst. 15: 68–76.
Lister, A. M. 2011. Natural history collections as sources of long-term datasets. – Trends Ecol. Evol. 26: 153–154.
Liu, C. et al. 2013. Selecting thresholds for the prediction of species occurrence with presence-only data. – J. Biogeogr. 40: 778–789.
Lughadha, E. N. et al. 2019. The use and misuse of herbarium specimens in evaluating plant extinction risks. – Phil. Trans. R. Soc. B 374: 20170402.
MacDonald, Z. G. et al. 2020. Gene flow and climate-associated genetic variation in a vagile habitat specialist. – Mol. Ecol. 29: 3889–3906.
Marcer, A. et al. 2021a. Quality issues in georeferencing: from physical collections to digital data repositories for ecological research. – Divers. Distrib. 27: 564–567.
Marcer, A. et al. 2021b. Natural history collections georeferencing survey report. Current georeferencing practices across institutions worldwide. – Zenodo, <https://doi.org/10.5281/zenodo.4644529>.
Marcer, A. et al. 2022. GBIF data for: Uncertainty matters: ascertaining where specimens in natural history collections come from and its implications for predicting species distributions. – Zenodo Digital Repository, <https://doi.org/10.5281/zenodo.5052596>.
Mateo, R. G. et al. 2010. Profile or group discriminative techniques? Generating reliable species distribution models using pseudo-absences and target group absences from natural history collections. – Divers. Distrib. 16: 84–94.
Mayani-Parás, F. et al. 2021. Cumulative habitat loss increases conservation threats on endemic species of terrestrial vertebrates in Mexico. – Biol. Conserv. 253: 108864.
McMichael, C. H. et al. 2014. Bamboo-dominated forests and pre-Columbian earthwork formations in south-western Amazonia. – J. Biogeogr. 41: 1733–1745.
Meineke, E. K. et al. 2018a. Biological collections for understanding biodiversity in the Anthropocene. – Phil. Trans. R. Soc. B 374: 20170386.
Meineke, E. K. et al. 2018b. The unrealized potential of herbaria for global change biology. – Ecol. Monogr. 88: 505–525.
Merow, C. et al. 2013. A practical guide to MaxEnt for modeling species' distributions: what it does, and why inputs and settings matter. – Ecography 36: 1058–1069.
Naimi, B. et al. 2014. Where is positional uncertainty a problem for species distribution modelling. – Ecography 37: 191–203.
Nelson, G. and Ellis, S. 2018. The history and impact of digitization and digital data mobilization on biodiversity research. – Phil. Trans. R. Soc. B 374: 20170391.
Noss, R. F. 1996. Editorial: The naturalists are dying off. – Conserv. Biol. 10: 1–3.
Pecchi, M. et al. 2019. Species distribution modelling to support forest management. A literature review. – Ecol. Model. 411: 108817.
Phillips, S. J. et al. 2006. Maximum entropy modeling of species geographic distributions. – Ecol. Model. 190: 231–259.
Phillips, S. J. et al. 2009. Sample selection bias and presence-only distribution models: implications for background and pseudo-absence data. – Ecol. Appl. 19: 181–197.
Pyke, G. H. and Ehrlich, P. R. 2010. Biological collections and ecological/environmental research: a review, some observations and a look to the future. – Biol. Rev. 85: 247–266.
Santini, L. et al. 2021 Assessing the reliability of species distribution projections in climate change research. – Divers. Distrib. 27: 1035–1050.
Shaffer, H. et al. 1998. The role of natural history collections in documenting species declines. – Trends Ecol. Evol. 13: 27–30.
Smith, A. B. et al. 2021. Imprecisely georeferenced specimen data provide unique information on species' distributions and environmental tolerances: don't let the perfect be the enemy of the good. – bioRxiv , doi: 10.1101/2021.06.10.447988.
Suarez, A. V. and Tsutsui, N. D. 2004. The value of museum collections for research and society. – BioScience 54: 66.
Swets, J. A. 1988 Measuring the accuracy of diagnostic systems. – Science 240: 1285–1293.

Tewksbury, J. J. et al. 2014. Natural history's place in science and society. – Bioscience 64: 300–310.
Tseng, M. and Pari, S. S. 2019. Body size explains interspecific variation in size-latitude relationships in geographically widespread beetle species: body size-latitude relationships in Coleoptera. – Ecol. Entomol. 44: 151–156.
Wicaksono, C. Y. et al. 2017. Contracting montane cloud forests: a case study of the Andean alder (*Alnus acuminata*) and associated fungi in the Yungas. – Biotropica 49: 141–152.
Wieczorek, C. and Wieczorek, J. 2021. Georeferencing calculator. – <http://georeferencing.org/georefcalculator/gc.html>, accessed 4 October 2021.
Wieczorek, J. R. et al. 2004. The point-radius method for georeferencing locality descriptions and calculating associated uncertainty. – Int. J. Geogr. Inform. Sci. 18: 745–767.
Wieczorek, J. R. et al. 2012 Darwin core: an evolving community-developed biodiversity data standard. – PLoS One 7: e29715.
Williams, K. J. et al. 2012 Which environmental variables should I use in my biodiversity model? – Int. J. Geogr. Inform. Sci. 26: 2009–2047.
Zizka, A. et al. 2019. CoordinateCleaner: standardized cleaning of occurrence records from biological collection databases. – Methods Ecol. Evol. 10: 744–751.