Received: 15 July 2022 | Revised: 5 April 2023 | Accepted: 13 April 2023

DOI: 10.1111/geb.13693

Global Ecology and Biogeography | A Journal of Macroecology

WILEY

METHOD

A guide to the processing and standardization of global palaeoecological data for large-scale syntheses using fossil pollen

Suzette G. A. Flantua[1] | Ondrej Mottl[1] | Vivian A. Felde[1] | Kuber P. Bhatta[2] | Hilary H. Birks[1] | John-Arvid Grytnes[2] | Alistair W. R. Seddon[1] | H. John B. Birks[1,3]

[1]Department of Biological Sciences, University of Bergen and Bjerknes Centre for Climate Research, Bergen, Norway

[2]Department of Biological Sciences, University of Bergen, Bergen, Norway

[3]Environmental Change Research Centre, University College London, London, UK

Correspondence
Suzette G. A. Flantua, Department of Biological Sciences, University of Bergen and Bjerknes Centre for Climate Research, Bergen, Norway.
Email: s.g.a.flantua@gmail.com

Funding information
Horizon 2020 Framework Programme, Grant/Award Number: 741413; Trond Mohn Stiftelse & University of Bergen, Grant/Award Number: TMS2022STG03

Handling Editor: Moriaki Yasuhara

Abstract

Aim: Palaeoecological data are crucial for comprehending large-scale biodiversity patterns and the natural and anthropogenic drivers that influence them over time. Over the last decade, the availability of open-access research databases of palaeoecological proxies has substantially increased. These databases open the door to research questions needing advanced numerical analyses and modelling based on big-data compilations. However, compiling and analysing palaeoecological data pose unique challenges that require a guide for producing standardized and reproducible compilations.

Innovation: We present a step-by-step guide of how to process fossil pollen data into a standardized dataset compilation ready for macroecological and palaeoecological analyses. We describe successive criteria that will enhance the quality of the compilations. Though these criteria are project and research question-dependent, we discuss the most important assumptions that should be considered and adjusted accordingly. Our guide is accompanied by an R-workflow—called *FOSSILPOL*—and corresponding R-package—called *R-Fossilpol*—that provide a detailed protocol ready for interdisciplinary users. We illustrate the workflow by sourcing and processing Scandinavian fossil pollen datasets and show the reproducibility of continental-scale data processing.

Main Conclusions: The study of biodiversity and macroecological patterns through time and space requires large-scale syntheses of palaeoecological datasets. The data preparation for such syntheses must be transparent and reproducible. With our *FOSSILPOL* workflow and R-package, we provide a protocol for optimal handling of large compilations of fossil pollen datasets and workflow reproducibility. Our workflow is also relevant for the compilation and synthesis of other palaeoecological proxies and as such offers a guide for synthetic and cross-disciplinary analyses with macroecological, biogeographical and palaeoecological perspectives. However, we emphasize that expertise and informed decisions based on palaeoecological knowledge

Suzette G. A. Flantua, Ondrej Mottl, Vivian A. Felde, and Kuber P. Bhatta are joint first authors.

remain crucial for high-quality data syntheses and should be strongly embedded in studies that rely on the increasing amount of open-access palaeoecological data.

KEYWORDS
data processing workflow, fossil pollen data, *FOSSILPOL*, large-scale syntheses, macroecology, Neotoma Paleoecology Database, palaeoecology, R package, *R-Fossilpol*

1 | INTRODUCTION

Macroecological patterns observed today are shaped through time and space by environmental, evolutionary and biotic processes operating across a wide range of scales (e.g. Antonelli et al., 2018; Divíšek et al., 2020; Jackson & Blois, 2015). Investigating how these present-day patterns developed through time provides important clues to the relative importance of the processes underlying the patterns. Globally, ecosystems carry the legacy of several thousand years of anthropogenic impact (e.g. Ellis et al., 2021; Ellis & Ramankutty, 2008; Mottl et al., 2021; Stephens et al., 2019). By jointly assessing interactions and past legacies of natural and anthropogenic drivers of ecosystem changes, one can decipher macroecological processes that have led to the contemporary global distribution of biodiversity and ecosystems (e.g. Jackson, 2007; Nolan et al., 2018).

Global biodiversity currently faces a plethora of threats related to climate change, land use, habitat conversion and species invasions. These interacting threats can have both direct and indirect effects on species distribution and ecosystem functioning (e.g. Poloczanska et al., 2013; Chaudhary et al., 2021; Chen et al., 2011; IPCC, 2021; Lenoir et al., 2020; Lenoir & Svenning, 2013; Parmesan & Yohe, 2003). Improved knowledge of how drivers and processes have affected changes in species distributions and biodiversity patterns through time is of paramount importance for addressing questions related to the ongoing biodiversity crisis and understanding how to mitigate the deleterious effects of human impact on environment and ecosystems (Fordham et al., 2020). Solutions to the biodiversity crisis will require integrated approaches, using information from a range of data sources, spanning historical observations, experiments and computational models (Daniau et al., 2019; Dawson et al., 2011).

'Palaeoecological records' (Box 1) can, with careful interpretation, provide an important source of information about ecosystem responses to natural (e.g. glacial–interglacial climatic variations, geological events, evolutionary processes, fire) and anthropogenic (e.g. extensive agriculture, fire, forest clearance, pollution) drivers over a range of spatial and temporal scales. Fossil pollen records are the most commonly used resource in 'palaeoecology' (Box 1) for elucidating the history of terrestrial vegetation dynamics. Global compilations of such palaeoecological records can be useful for cross-scale macroecological and biogeographical studies and also help to evaluate the relative impacts of long-term evolutionary or climate-driven changes in ecosystems against changes resulting from human impact in the past few thousand years (e.g. Finsinger et al., 2017; Mottl et al., 2021; Nolan et al., 2018). With the increasing number of records from a range of 'proxies' (Box 1), palaeoecology has now an ever-increasing capacity to uncover Earth system dynamics across multiple taxonomic groups and beyond the time frame of directly observed data. Examples include identifying systems with high ecosystem resilience (e.g. Buma et al., 2019; Davies et al., 2018; Willis et al., 2010), comparison of present-day ecological models to environmental conditions outside the range of those existing today (e.g. Svenning et al., 2011; Williams et al., 2007), assessing the role of humans in driving compositional turnover (e.g. Nogué et al., 2021; Woodbridge et al., 2020), guiding conservation efforts (Barnosky et al., 2017; Dietl et al., 2015), assessing temporal variability in ecosystem services (Jeffers et al., 2011, 2015); and climate reconstructions (e.g. Chevalier et al., 2020; Hébert et al., 2022), among many others.

With the rapid growth of palaeoecological datasets in the public domain (e.g. Neotoma Paleoecological Database, Williams et al., 2018, https://www.neotomadb.org/, 'Neotoma' hereafter; PANGAEA, https://www.pangaea.de/; Data Publisher for Earth & Environmental Science), there is greater opportunity for macroecologists to expand their temporal scales of analyses. The potential is especially high for fossil pollen data, which have a long history of community data assembly and curation, recently aided by a series of data mobilization efforts (e.g. Latin America in progress), Africa (Ivory et al., 2020; Runge et al., 2021) and the Indo-Pacific (in progress). Such data assemblages allow a deepened understanding of vegetation dynamics across various spatial and temporal scales. See, for instance, the recent continental analyses on rates of vegetation change (Mottl et al., 2021), community novelty (Staples et al., 2022), ecosystem properties of Asian vegetation (Herzschuh, 2020) and latitudinal gradients across Europe (Giesecke et al., 2019).

Open-access data increase research opportunities but also come at a risk of uninformed use that leads to erroneous interpretation (Dillon et al., 2023; Jackson, 2012). Palaeoecologists working regularly with fossil data are familiar with the vast heterogeneity hidden in a large palaeoecological 'dataset compilation' (Box 1). Taxonomic uncertainties and differences in the temporal resolution of data (e.g. Birks & Birks, 1980; Dillon et al., 2023; Prentice, 1988; Rull, 2020; Webb, 1993) are crucial issues, but the heterogeneous character of fossil pollen data is manifold. Pollen assemblages are influenced by processes of dispersal and sedimentation, environmental settings, pollen identification expertise of researchers, and collection and processing methodologies, which are usually defined by the initial research questions for records.

BOX 1 Key palaeoecological terms used in the paper, *FOSSILPOL* workflow and *R-Fossilpol*.

Age-depth model: an algorithm used to estimate the age-depth relationship for a series of stratified palaeoenvironmental data points (e.g. depths within a core or stratigraphic profile), whose relative chronological relationships are known but for which only a limited amount of absolute chronological information is available from the age controls. Age-depth models are used to make estimates of ages for depths not directly associated with an age control or to resolve discrepancies among age controls (adapted from Grimm et al., 2014). Note that 'age model' is sometimes used as a shorthand synonym for 'age-depth model'.

Archive: refers to a geological sedimentary deposit that contains the physical material from which palaeoecological or palaeoclimatic proxies are extracted. Such archives can include ice cores, speleothems or sedimentary deposits (lake sediments and peat bogs). Multiple palaeoecologial and palaeoenvironmental proxies (macro-, microfossil, charcoal, etc.) can be obtained from one archive.

Bayesian age model: an age model that provides fully probabilistic estimates of the uncertainties in sample ages via the application of Bayes' theorem. Bayesian models rely upon prior assumptions about e.g. sediment accumulation rates, stratigraphic superposition and thus monotonicity of ages. Programmes that implement Bayesian age models include *Bacon* (Blaauw & Christen, 2011), *OxCal* (Bronk Ramsey, 2001) and *Bchron* (Haslett & Parnell, 2008). Age controls may be 'uncalibrated radiocarbon years' (see definition) or calendar ages with uncertainties. These age models produce calibrated or calendar ages, and they can automatically deal with most cases of outlying dates (adapted from Grimm et al., 2014).

Before present (BP): by convention, most radiocarbon dates are reported as BP where 'present' is 1950 CE; however, the exact definition of 'the present' can vary among palaeoecological and palaeoclimatic papers. Therefore, the precise definition of 'present' is important to specify at all times.

Calibration curve: is used to convert uncalibrated radiocarbon years (uncalibrated ^{14}C BP, see definition) to calendar years (calibrated years before present, cal yr BP or cal yr B2K (2000 CE)). Depending on the location of the record and if the locality is marine or not, it is important to use the appropriate calibration curve. The radiocarbon calibration curve is empirically derived and is regularly updated as new observations are collected. At this time, IntCal21 is the standard calibration curve, replacing the previous IntCal13. See Reimer et al. (2020) and Hogg et al. (2020) for the latest curves for radiocarbon calibration.

Chronology control point(s): an estimate of absolute age, often with a specified uncertainty, for a level within a core or stratigraphic profile that is used to constrain an age model for that core or profile. Also called 'age control' (adapted from Grimm et al., 2014).

Chronology: a series of estimated ages and associated uncertainty estimates for levels in a stratigraphic record. Such estimates usually derive from an age-depth model and its associated age controls (adapted from Grimm et al., 2014).

Chronology control table: a table that contains all the chronology control points. Includes depth, uncalibrated age of radiocarbon date and age error. *Clam* and *Bacon* require additional columns related to the reservoir effect or calibration curve used. See *clam* and *Bacon* manuals.

Classical age model: an age-depth model in which a curve or line is fitted to a series of age-depth points with no prior assumptions about sediment accumulation rate or monotonicity of ages (Blaauw, 2010). If the age controls are radiocarbon dates, they may be calibrated or uncalibrated. Calibration of radiocarbon dates should be undertaken before the curve is fitted, and outliers can be rejected a priori or after producing the model. Common classical algorithms include linear interpolation, linear or polynomial regression and various splines. Many classical age models do not provide an estimate of the errors for interpolated ages. Though *clam* (Blaauw, 2010) provides error estimates, these are only a single 'best' error estimate, while the true age uncertainty is not provided (in a Bayesian age model, an age uncertainty distribution is given; adapted from Grimm et al., 2014).

Coring: most common means of obtaining a sedimentary archive from lakes or mires. Coring can be done from a filled-in depression in the landscape or from a floating platform or surface ice in a lake, using hand-driven equipment to more advanced equipment for deep drilling.

Dataset compilation: a suite of palaeoecological datasets that ideally are processed, standardized and harmonized in a consistent manner, based on a set of user-based criteria that are stated and reproducible.

Depositional environment: the environmental context that produced the sedimentary archive from where the record was taken. The main categories of depositional environments in Neotoma are archaeological, biological, estuarine, lacustrine, palustrine and terrestrial, with numerous sub-categories. Fossil pollen records have been analysed from a wide range of depositional environments (Chevalier et al., 2020). The fossil pollen spectrum differs between different depositional environments due to taphonomy (see definition), which is therefore an important criterion to consider for multi-site data syntheses.

Estuarine: depositional environments with both terrestrial and marine influences.

Harmonization region(s): the geographical region to which the harmonization table should be applied. When records are sourced from a broad spatial domain (e.g. continental to global) we recommend delimiting homogeneous regions and creating a separate harmonization table for each region.

BOX 1 Continued.

Harmonization table: a table with a column 'A' for all unharmonized, original pollen-type names from all records and a corresponding column 'B' that assigns each pollen-type and spore-type name to a higher harmonized taxon name. The taxonomic harmonization from columns A to B is applied consistently to all records within a multi-data synthesis. However, large geographical regions should have their own harmonization table as the different flora in each region can lead to different mappings of plant taxon names onto pollen morphotypes. Such a table can contain different columns of harmonization (B, C, D, etc) depending on the criteria applied (e.g. as in Giesecke et al., 2019 for the European Pollen Database levels = MHVar2, http://www.europeanpollendatabase.net/data/downloads/image/EPD_P_VARS_high3.csv).

Lacustrine: depositional environments from existing or ancient lakes.

Levels: refers to the depths from which samples were taken for palaeoecological analysis, for instance, pollen, diatoms, phytoliths, etc.

Other datasets: applies here to datasets that are not derived from Neotoma, including private (exclusive) datasets and those from other data sources, e.g. PANGAEA and datasets in publications. For overall reproducibility following FAIR guidelines, we underline the high scientific value of limiting or avoiding private datasets and use Neotoma or PANGAEA for open-access research.

Palaeoecology: (i) 'the ecology of the past … ideally, palaeoecology could be defined as the study and understanding of the relationships between past organisms and the environment in which they lived. In practice, palaeoecology is largely concerned with the reconstruction of past ecosystems' (Birks & Birks, 1980), (ii) 'the branch of ecology that studies (the) past (of) ecological systems and their trends in time using fossils and other proxies' (Rull, 2010).

Palaeoecological record: A time series at a single location consisting of all samples of the same type of palaeoecological proxy, ordered in temporal sequence. For example, all pollen samples from a single core retrieved from a lake would be a single pollen record.

Palustrine: depositional environments from inland, wetland settings including vegetated and nonriverine systems, such as bogs and swamps.

Pollen morphotypes: refers to the physical characteristics of pollen grains that are used to identify and consequently classify them into morphotypes. These characteristics include the shape, size, wall structure and surface ornamentation of the grain, and they are unique to each pollen type. Palynologists observe these traits under a microscope to identify the plant family, genus or species that produced the pollen. Not all morphotypes clearly map onto these levels, as some morphotypes comprise multiple genera, e.g. Ostrya/Carpinus as a distinction is currently not possible.

Processing: the data steps needed for data 'integration' (Michener & Jones, 2012; Nieto-Lugilde et al., 2021). This is not to be confused with the preliminary collecting and describing of palaeoecological data.

Proxy: palaeoecological proxies include fossil pollen, plant macrofossils (e.g. seeds), phytoliths, diatoms, charcoal, nonpollen palynomorphs, charcoal, sediment and water chemistry and ancient DNA.

Radiocarbon calibration: the process of converting radiocarbon dates into calendar years. Radiocarbon age estimates (^{14}C yr BP) can be converted to calendar years using a calibration curve (see definition) and expressed as calibrated years before present (cal yr BP; BP = 1950).

Redeposition: the process by which fossils that originated at one time may be remobilized from their current sedimentary context, transported some distance and then redeposited and preserved with fossils originating from a different space and time (Birks & Birks, 1980).

Research infrastructure: facilities, resources and related services that are used by the scientific community to conduct robust and reliable research in their respective fields. This concept includes major scientific equipment or sets of instruments; knowledge-based resources such as collections, archives or structures for scientific information; information and communication technology-based infrastructure such as grid, computing, software and communication; or any other entity of a semi-permanent nature essential to achieve reproducible research (adapted from Nieto-Lugilde et al., 2021).

Shapefile: a geospatial vector data format that stores information on the location, shape and attributes of points, lines or polygons.

Source area: the likely spatial area represented by the pollen assemblages in the record.

Taphonomy: the geological and biotic processes acting on biological remains after the end of life until their deposition in their current sedimentary archive, e.g. transportation and dispersal, deposition (sedimentation) and preservation.

Uncalibrated radiocarbon years: by convention, all radiocarbon labs report radiocarbon dates in uncalibrated ages. These uncalibrated ages are based upon the simplifying assumption that the amount of radiocarbon in the atmospheric reservoir has always been constant at 1950 CE values. Calibration curves correct for the reality that the amount of atmospheric radiocarbon changes over time. These radiocarbon measures are expressed in uncalibrated radiocarbon years before present ^{14}C yr BP (written as ^{14}C BP). These are not suitable to be used as an expression of time as these need to be converted (calibrated) to true calendar years (see Radiocarbon calibration).

To fully utilize the potential of fossil pollen data in open-access multidisciplinary research, it is imperative to establish guidelines to address the challenges of working with heterogeneous data and the sources of error and uncertainty, particularly when multiple sources are used. Such guidelines should cover correct terminology, sets of inferences for appropriate data analysis and best practices to enhance data quality (Blois, 2012; Brovkin et al., 2021; Fordham et al., 2020). Collaboration among disciplinary experts is critical for knowledge exchange, and user-friendly guidelines can greatly enhance this process for interdisciplinary data analysts (Jackson, 2012). Without such guidelines, there is a risk of overconfidence in existing databases and a failure to identify underlying inconsistencies and pitfalls (Blois, 2012). Additionally, these guidelines are useful for integrating palaeoecological data into open software and workflows, which promote open-research practices by ensuring full disclosure of data-handling and statistical pipelines.

In this paper, we aim to use our combined knowledge of palaeoecology and large-scale data analysis to create a guide and workflow intended for a general audience of macroecologists, biogeographers and others. We present *FOSSILPOL*, a workflow to process efficiently global fossil pollen data, to create a standardized dataset compilation suitable for macroecological and palaeoecological research, supported by an R-package called R-*Fossilpol*. Our approach offers a stepwise, reproducible and transparent procedure for synthesizing complex and heterogeneous fossil pollen data (Box 1). The workflow provides a detailed protocol that guides the data analyst through the data processing procedure, helping them to overcome the challenges of working with such data. By using our approach, the data analyst can ensure that the entire data processing procedure is transparent and easily reviewed. Moreover, the workflow can be easily included in any publication to facilitate reproducibility. Our specific recommendations for data processing are further reinforced by the detailed guidelines in the *Step-by-step guide to data processing* on the *FOSSILPOL* website (https://hope-uib-bio.github.io/FOSSILPOL-website/). In addition, we provide a Glossary (Box 1) for key palaeoecological terms used here and a list of introductory readings in Appendix S1. Our overall goal is to expand the overall contribution of palaeoecological data to macroecological research, and thereby contribute to the advancement of understanding of the drivers and processes of biodiversity and ecosystem dynamics over time and space (Dietl et al., 2015; Rapacciuolo & Blois, 2019; Rull, 2010, 2012; Willis & MacDonald, 2011).

2 | KEY CONCEPTS AND STEPS IN LARGE-SCALE PALAEOECOLOGICAL DATA ANALYTICS

We summarize the key concepts involved in large-scale analyses of late Quaternary palaeoecological datasets that primarily consist of radiometrically (e.g. ^{14}C, ^{210}Pb) dates and fossil pollen records. When compiling and synthesizing such data, standardized terminology aids interdisciplinary understanding and downstream workflow decisions. Therefore, we introduce key terms (Box 1) related to obtaining palaeoecological 'records', 'depositional environments', 'chronologies', 'taxonomic harmonisation' and filtering of records and samples, and provide specific examples for fossil pollen data. We highlight common standards and issues that we strongly recommend considering during the steps of large-scale palaeoecological data analytics (Figure 1). Although the focus here is on fossil pollen, the general procedure and underlying concepts are similar for many other late Quaternary palaeoecological proxies retrieved from sedimentary 'archives' (Box 1).

2.1 | Obtaining a fossil pollen dataset

A compilation of fossil pollen records consists of multiple datasets that each represent a single time series of multivariate data represented by individual 'pollen morphotypes', each of which represents one or more plant species (Box 1). The series of procedures needed to obtain an individual fossil pollen record are outlined by e.g. Birks and Birks (1980), Daniau et al. (2019), Fægri and Iversen (1964) and Fægri et al. (1989). In brief, a sediment record is obtained by various extraction methods such as 'coring' (Box 1) lake sediments or mires, excavating a soil section or collecting sediments from cave walls and archaeological sites, among many others. Wetland environments, including lakes and bogs, accumulate anaerobic sediment continuously and are suitable for the 'preservation' (Box 1) of pollen originating from the surrounding local and landscape-level vegetation ('source area', Box 1) through time.

A sediment record is sampled at measured depth intervals, usually at regular intervals (e.g. every 2 cm), but records sampled at irregular intervals are also common. The sample is chemically treated to isolate pollen from the background sediment matrix, the residue is mounted on microscope slides. Pollen types are systematically identified and counted until the desired number of pollen grains is tallied (Fægri et al., 1989). A suitable set of pollen types are chosen for analysis and used as the basis of the sum for calculating relative pollen percentages (see Section 2.6). All pollen grains and spores are identified by a pollen analyst (although machine-learning methods are in development, e.g. Sevillano et al., 2020) to morphotypes corresponding to the finest taxonomic level possible, e.g. species, genus or family. The finest possible taxonomic level depends on the availability of pollen and spore morphological characteristics for identification, the palynological skills and experience of the pollen analyst, the available reference material and the degree of preservation of the grains (Birks & Birks, 1980). 'Levels' (i.e. depths, Box 1) of interest are dated, usually by radiocarbon dating (see Section 2.4). From these ages, an 'age-depth model' (Box 1) of sediment accumulation is constructed, and calendar (calibrated) ages are inferred for all levels from the model. (For further information about palynological methods, see Appendix S1).

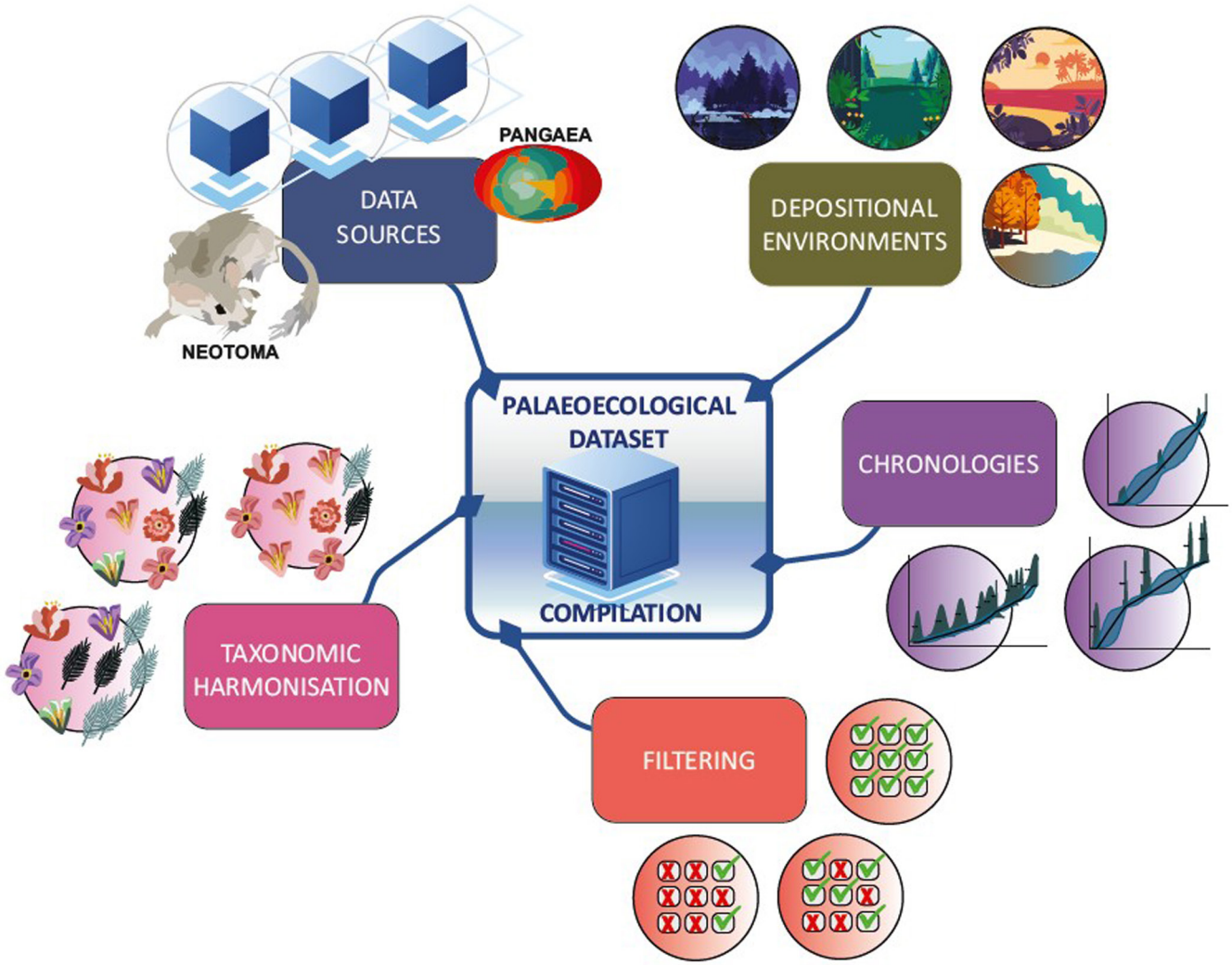

FIGURE 1 Essential data processing components needed to create a standardized, harmonized, palaeoecological dataset compilation before macro-scale data analysis. Note that each component consists of selecting appropriate datasets and samples based on user-defined criteria guided by the research questions. Such criteria influence the outcome of the analyses obtained from the dataset compilation. Therefore, careful documentation of these criteria is pivotal for data quality control and reproducibility. Vector credits: Dataset compilation and data sources: Design by fullvector/Freepik; Flowers: Design by rawpixel.com/Freepik. Depositional environments: Design by All-free-download.com.

2.2 | Selecting data sources

The first decision in the compilation of multiple fossil pollen records is the selection of the data sources (Figure 1). Since the 1970s, there have been several initiatives to establish palaeoecological research infrastructure, which serves as the foundation for current open community data efforts. Currently, there are two major and active open-access research infrastructures for fossil pollen data (Chevalier et al., 2020), namely Neotoma and PANGAEA. Neotoma is a collaborative effort between a consortium of project leaders and institutions around the world, currently containing more than 42,000 datasets and 19,000 unique sites from more than 6900 authors (updated Jan-2022; Williams et al., 2018). As of 27th March 2023, Neotoma contains 6129 fossil pollen datasets from across 5048 sites (obtained from Neotoma Explorer; https://apps.neotomadb.org/explorer/). PANGAEA aims at archiving and distributing georeferenced data related to Earth system research, including studies on the biosphere, atmosphere and cryosphere. At present, PANGAEA contains over 408,000 datasets from diverse disciplines, including palaeoecology and palaeontology.

Although there are overlapping datasets in Neotoma and PANGAEA, they differ in terms of the structure and prestandardization of the datasets. For instance, Neotoma has assigned expert data stewards drawn from the communities of scientists that generate these data, who check for data quality and consistency and who have been trained to upload data to Neotoma following established procedures. The procedure and quality control for uploaded datasets to PANGAEA depend on individual data contributors. In addition, the Neotoma upload procedure requires all data to follow fixed formats and units, which supports a higher level of standardization

for the end-user accessing the raw data from the database. Though the *FOSSILPOL* workflow and its related *R-Fossilpol* package can handle other sources of data, here we use Neotoma as the default source.

The key factor in the use of data from multiple sources is consistency. Common inconsistencies in data formatting among sources include: (i) differences in the file naming convention among datasets; (ii) differences in the format of geospatial coordinates; (iii) differences in units of depth measures (e.g. metres or centimetres); (iv) pollen data stored as percentages instead of raw pollen counts; and (v) differences in age units in chronology data (e.g. Before Present, Box 1) versus AD or CE, or years (yr) versus thousands of years (kyr). Although aware of these inconsistencies, pollen data analysts might still want to use a variety of data sources, as there are numerous datasets that are not or cannot be available in open-access research infrastructures for privacy or legal reasons (e.g. nationally funded infrastructure that request exclusivity of data). We will refer to any data from a source other than Neotoma as 'other' (Box 1) further on, as such data will need to be cleaned and reformatted before being compatible with Neotoma.

The next step in data acquisition is selecting the spatiotemporal scope. Due to depositional conditions and coring/extraction methods, many records only span a few thousand years. For historical reasons, the continents of Europe and North America have higher data coverage in Neotoma than the Southern Hemisphere, although ongoing data mobilization campaigns are reducing this disparity (Figure 2a). In addition, there is a rapid drop-off in the number of available records toward periods older than 10,000 yr BP (Figure 2b; see Europe, Giesecke et al., 2014; Latin America, Flantua et al., 2015, 2016; North America, Williams et al., 2004), in part because many of the lakes and mires in this compilation began to accumulate sediment only after the last deglaciation.

2.3 | Selecting depositional environments

The depositional environment from which a record was recovered (e.g. lake, bog, fen, cave, colluvial fan) should be considered when using fossil pollen data (Figure 1), because taphonomic processes (Taphonomy, Box 1; Appendix S1) may modify the biological and spatial signal in a record (Chevalier et al., 2020; Daniau et al., 2019) and therefore affect the results derived from subsequent analyses. Taphonomy is related to the depositional environment and is a known source of palaeoecological data heterogeneity and uncertainty (Cleal et al., 2021; Jackson, 2012; Nieto-Lugilde et al., 2021). For example, the degree of 'redeposition' (Box 1) and preservation can influence the taxonomic composition detected in a record causing an over- or under-representation of taxa ('taxonomic biases', Birks & Birks, 1980; Cushing, 1967; Davis, 1968). Also, some depositional environments, such as fluvial systems, have naturally high variations in water energy and provenance source, which can cause pollen assemblages to change abruptly over short periods of time. Such changes could be incorrectly interpreted as high ecosystem turnover. For quiet-water lakes and mires, where most deposited pollen is the outcome of airborne transport and deposition, another key concept that affects the pollen assemblage recorded from the surrounding vegetation is the source area. Pollen source area is a function of the lake area, with bigger lakes capturing vegetation signals across a broader area. A

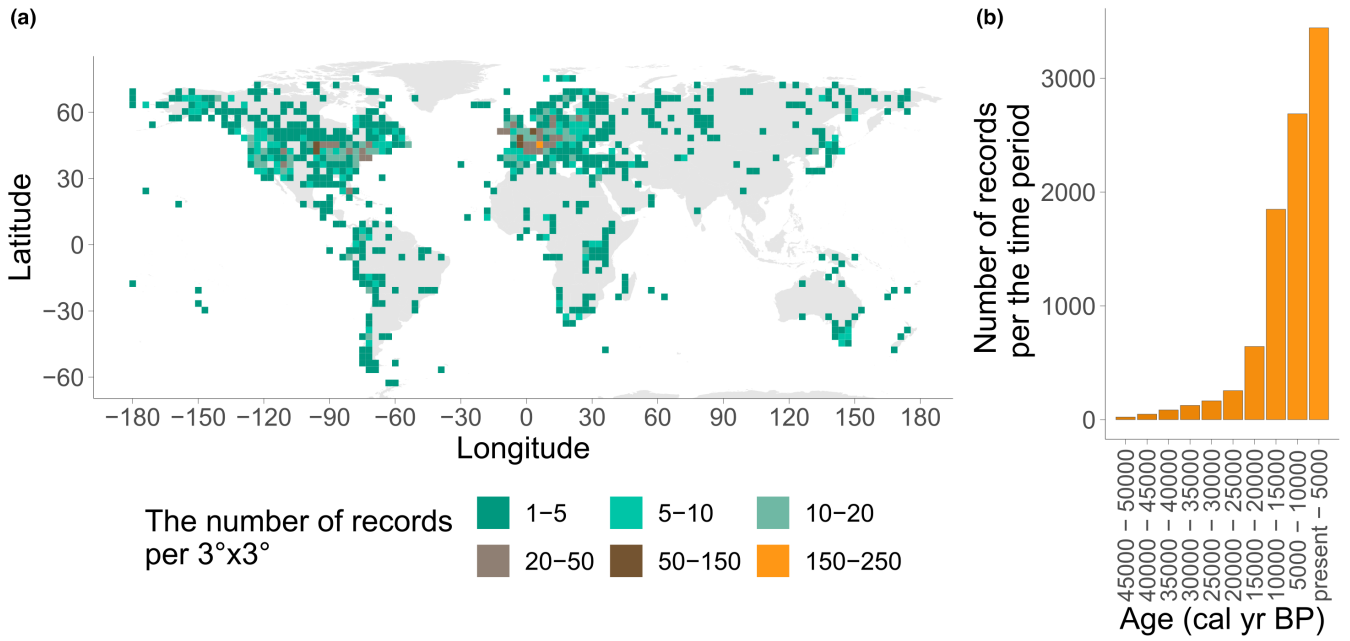

FIGURE 2 Spatial (a) and temporal (b) overview of fossil pollen data of Neotoma. The data were obtained from Neotoma on 20th January 2023. Records without a geographical location and age information were filtered out. In addition, only records spanning at least 100 years between −75 and 50,000 BP years were kept.

small pollen source area (e.g. tens or hundreds of square metres), on the other hand, provides a highly local signal of the vegetation (from, e.g. pollen in a forest hollow). Alternatively, the source area of some environments can be extensive (e.g. colluvial fan, large lake, marine environment) and the signal recorded in the pollen record can reflect plant taxa originating from ecosystems hundreds of kilometres away (Jacobson & Bradshaw, 1981; Prentice, 1985).

The effects of mixed depositional environments in large databases have been flagged as a concern in the palaeoecological literature for several decades (e.g. Goring et al., 2010; Jacobson & Bradshaw, 1981; Wilmshurst & McGlone, 2005). The representation of a pollen taxon can differ substantially among depositional environments (e.g. Zhao et al., 2009). To address this issue, some studies have chosen to exclude certain depositional environments from their analyses (Goring et al., 2010; Mottl et al., 2021). Therefore, it is important to consider how including different depositional environments, each with their unique taphonomic processes, could affect the analysis and results (Birks, 1995). In particular, for continental-scale syntheses of spatial changes in taxa or ecosystems, we recommend avoiding the use of records from local depositional environments such as small ponds, caves or forest hollows. Additionally, records from marine or river depositional environments that capture mixed signals from a wide range of vegetation types should be avoided as well. By following these recommendations, researchers can avoid or minimize the biases in their data analyses potentially incorporated due to the inclusion of varied depositional environments. Depending on the research question, a simple rule of thumb is to use depositional environments from a single broad category, namely archaeological, biological, 'estuarine', 'lacustrine', 'palustrine' or terrestrial (Box 1), or to select only those that correspond to the scale of research interest (Jacobson & Bradshaw, 1981).

2.4 | Estimating chronologies

2.4.1 | Background

To estimate the age of individual levels based on their depth, a chronology or age-depth model needs to be constructed (Figure 1). An age-depth model provides age estimates of each individual level and the full age range of the record. Most datasets in Neotoma have chronologies available and the ages of individual levels are given, but these chronologies often need updating to match current best practices in age-depth modelling (e.g. Herzschuh et al., 2022; Wang et al., 2019). Generally, age-depth models are constructed using 'chronology control points' (Box 1) with known depth, estimated age and associated age uncertainties. The chronology control points for each record are saved in the 'chronology control table' (Box 1).

Each control point in the chronology control table has the following properties: (i) depth, (ii) estimated age, (iii) error of the estimated age and (iv) type of the chronology control point (e.g. radiocarbon, uranium/thorium, biostratigraphic, annual laminations). Each type of chronology control point has different kinds of age uncertainties. Radiocarbon dating (^{14}C) is the most widely used dating technique in palaeoecology, but chronological problems with ^{14}C-dated records are manifold. A detailed discussion on sources of ^{14}C-uncertainties is beyond the scope of this paper, but our recommended reading in Appendix S1 summarizes the most important aspects. As there are over 50 types of control points in Neotoma alone, the data analyst can opt to define criteria that include only records with certain control point types. For instance, some control points relied on indirect dating techniques such as biostratigraphic layers and ages inferred from other records. Such approaches assume synchronicity of events across multiple records but consequently can have age uncertainties of thousands of years (Blaauw et al., 2010; Blois et al., 2011; Flantua et al., 2016; Giesecke et al., 2014). Also, a record with a small number of chronology control points within the focal time period will usually have high uncertainties of predicted ages around this time period (Flantua et al., 2016; Giesecke et al., 2014). Hence, information about the quality of chronologies that considers the types and levels of chronology control points can be a criterion used in the selection of records (Figure 1; Blois et al., 2011).

2.4.2 | Re-estimation of age-depth models

Because radiocarbon ages are not true calendar ages, a 'calibration curve' (Box 1) is needed to convert the raw dates in the chronology control table to calendar ages. These are IntCal (Reimer et al., 2020), SHCal (Hogg et al., 2020) or mixed calibration curves (see 'Radiocarbon calibration' in Box 1). Specific calibration curves are applied to the Northern and Southern Hemispheres and to terrestrial and marine environments (all are available in *clam*, *rbacon* and *Bchron* software). Calibration curves continue to be updated, so one should re-estimate older chronologies based on the latest calibration curves or re-estimate chronologies for all records in the dataset compilation. Nonradiocarbon dates such as defined core-top dates and annual laminations (varves) are considered as calendar ages.

For multi-record data syntheses, we recommend re-estimating age-depth models for all records, so that all age estimates are based upon a common methodology and use the best available statistical methods for age calibration and inference (Figure 1). Broadly, two kinds of age-modelling approaches exist, namely 'classical' and 'Bayesian' (Box 1, see also Appendix S1), with the latter being the most popular approach in the last decade (Blaauw et al., 2018). The popularity of Bayesian age-depth models is mainly due to the probabilistic nature of the models, with better handling of outliers, and the fact that their precision increases with increasing dating density (Blaauw et al., 2018). Software programs for age-depth modelling include *clam* (classical; Blaauw, 2010), *rbacon* (Bayesian; Blaauw & Christen, 2011), *Bchron* (Bayesian; Haslett & Parnell, 2008) and *OxCal* (Bayesian; Bronk Ramsey, 1995, 2001). Independently of the approach selected, all age-depth models rely upon a chronology control table.

Generally, the data analyst can choose to use the original chronology control table as provided by the data contributors or to create an adjusted table based on user-defined criteria. The latter

involves: (i) filtering out unwanted control point types and (ii) filtering out control points without known age error or an error that is considered too large (e.g. >2000 yr). Moreover, it may be preferable to define the minimum acceptable number of control points (e.g. at least three). This leads to a trade-off between accepting chronologies with a high number of control points (which will usually provide more robust age-depth models) and the number of records that will be able to fulfil the stricter criteria.

Age-depth models should be evaluated before being incorporated. The first way to do so is by visual inspection of the age-depth curves to look for unrealistically large age estimates or error bars, hiatuses or extreme extrapolations toward the present or the past. Whenever possible, we recommend checking the original publication to assess whether, for example, the authors identified outlier age controls or events such as changes in sedimentation regime that might lead to changes in the depth-age relationship.

2.5 | Implementing taxonomic harmonization

The goal of taxonomic harmonization (Figure 1) is to standardize all site-level names to the same pollen morphotypes (set of pollen and spore morphotypes used for all pollen records) and thus reduce the effect of taxonomic uncertainty and nomenclatural complexity. There are several underlying issues: (i) morphotypes differ in their identifiability; (ii) pollen analysts vary in their choices of names for morphotypes; and (iii) mapping of taxon names to morphotypes can vary among spatial regions. These issues are most acute at finer taxonomic resolutions (e.g. species) and are less so as one aggregate nomenclature to the genus or family level. For this purpose, a 'harmonization table' (Box 1) can be created that groups the morphotypes into the highest taxonomic level that is most likely to be identified by most of the pollen analysts. Skill of analyst, preservation, different plant nomenclature, regional floras, inconsistent naming of identification uncertainty level (type, cf., undifferentiated etc.), spatiotemporal domain and range of modern pollen reference material available ideally should all be considered when deciding which morphotypes to merge.

Harmonization tables are created for a certain region or project to address particular scientific questions and spatiotemporal domains, which can affect decisions about how best to lump or split taxa. For example, if a project specifically aims at detecting human impact in past vegetation, taxa considered as human indicators should not be lumped into a higher taxonomic level, e.g. Cerealia into Poaceae (Deza-Araujo et al., 2022). Harmonization tables for fossil pollen data have already been published (Appendix S1). However, we strongly recommend that data analysts carefully consider whether these harmonization tables are appropriate given the scientific question and geographical region of interest. It is always advisable to work with an expert familiar with the modern and the pollen flora of the spatiotemporal domain of interest in order to create a reliable project-specific table of harmonized taxa.

When attempting to link harmonized palaeoecological and neoecological datasets, it is important to note that discrepancies in taxonomic nomenclatures will exist. For taxonomic consistency, both fields require standardization as a first step (Grenié et al., 2023). To successfully integrate data from these fields, data analysts will need a 'common currency', which can be achieved through a linking table that is created during the data-handling process prior to starting integrative analyses (Rapacciuolo & Blois, 2019). It is worth noting that creating a linking table is not part of the FOSSILPOL workflow presented here, but see, for instance, Blaus et al. (2020).

2.6 | Filtering datasets

To obtain a comprehensive compilation of multiple fossil pollen records, to increase its overall quality and to answer research questions reliably, we recommend the data analyst to further trim down the data selection by filtering out individual levels and/or whole records (Figure 1) by considering the following criteria:

1. *Pollen count*—The number of counted pollen grains at each level is an index of data quality. To obtain a reliable representation of the vegetation, researchers often aim to count more than 300 pollen grains (following Moore et al., 1991), but other recommendations may have been followed (>150; e.g. Djamali & Cilleros, 2020) and will vary with region and by scientific question (Birks & Birks, 1980). For example, to achieve a representative sample of the regional pollen pool, counts in Arctic records may only reach c. 100 grains per level, whereas counts in Mediterranean sites can be as high as 1000 (Birks & Birks, 1980, p. 165), but the main determinant can be the preference of the pollen analyst. Reasons for low numbers (<100) are often mainly due to time constraints of the data contributor but can also be natural depositional phenomena causing poor pollen preservation, such as low local pollen production in arctic or alpine environments. Given that statistical inferential power is proportional to sample size, we recommend defining a *minimum* number of total pollen grains in each level. Subsequently, whole records can be selected on the proportion of levels with a selected minimum number of pollen grains counted per level.
2. *Age criteria*—To reduce processing time, we recommend using predefined age criteria to select records that span the time period related to the research question and filter out levels beyond this focal period.
3. *Age extrapolation*—Extrapolation of age inferences to samples beyond the set of chronology control points is another factor in quality control. Samples older than the oldest chronology control point have no other chronology control point to constrain the age inference and, therefore, have increasingly large uncertainty as extrapolation increases. To limit the use of samples based on large extrapolations, we recommend selecting a *maximum*

extrapolation age, i.e. samples older than *X* years relative to the oldest chronology control point are filtered out.

4. *Number of samples*—The total number of samples in a record is an important quality criterion. Records might have been sampled at low resolution (e.g. depth intervals >30 cm) leaving substantial unassessed gaps—and thus time periods—between samples. In addition, records with few samples will contribute poorly to studies focussed on specific time periods and can result in outlier values. Therefore, we recommend selecting a minimum number of samples within the time period of interest and using this as an additional criterion to filter out unwanted records.

3 | *FOSSILPOL* WORKFLOW

The *FOSSILPOL* workflow is designed to process multiple fossil pollen records to create a comprehensive, standardized dataset compilation, ready for multi-record and multiproxy analyses at macroecological scales (Figure 3). The workflow automatically handles all the previously described analytical steps and flags places where the analyst needs to specify their choices (Appendix S2), while following a simple chain of actions: (i) retrieve data from Neotoma and/or 'other datasets' (Box 1), (ii) re-estimate age-depth models, (iii) taxonomically harmonize the pollen taxa, (iv) filter out records based on user-defined criteria and (v) save the prepared compilation in a standardized format for easy data processing and reviewing.

The workflow is modular: All steps are organized sequentially and guided by one main configuration file (called *Config file*) where all criteria and setup configurations are predefined by the data analyst and saved as a reference file (Figure 3; Appendix S3; see next section). A more detailed description of individual steps of the *FOSSILPOL* workflow, including guidance about possible issues and corresponding solutions for multi-record fossil pollen datasets, can be found in the *A step-by-step guide to data processing* on the *FOSSILPOL* website (https://hope-uib-bio.github.io/FOSSILPOL-website/).

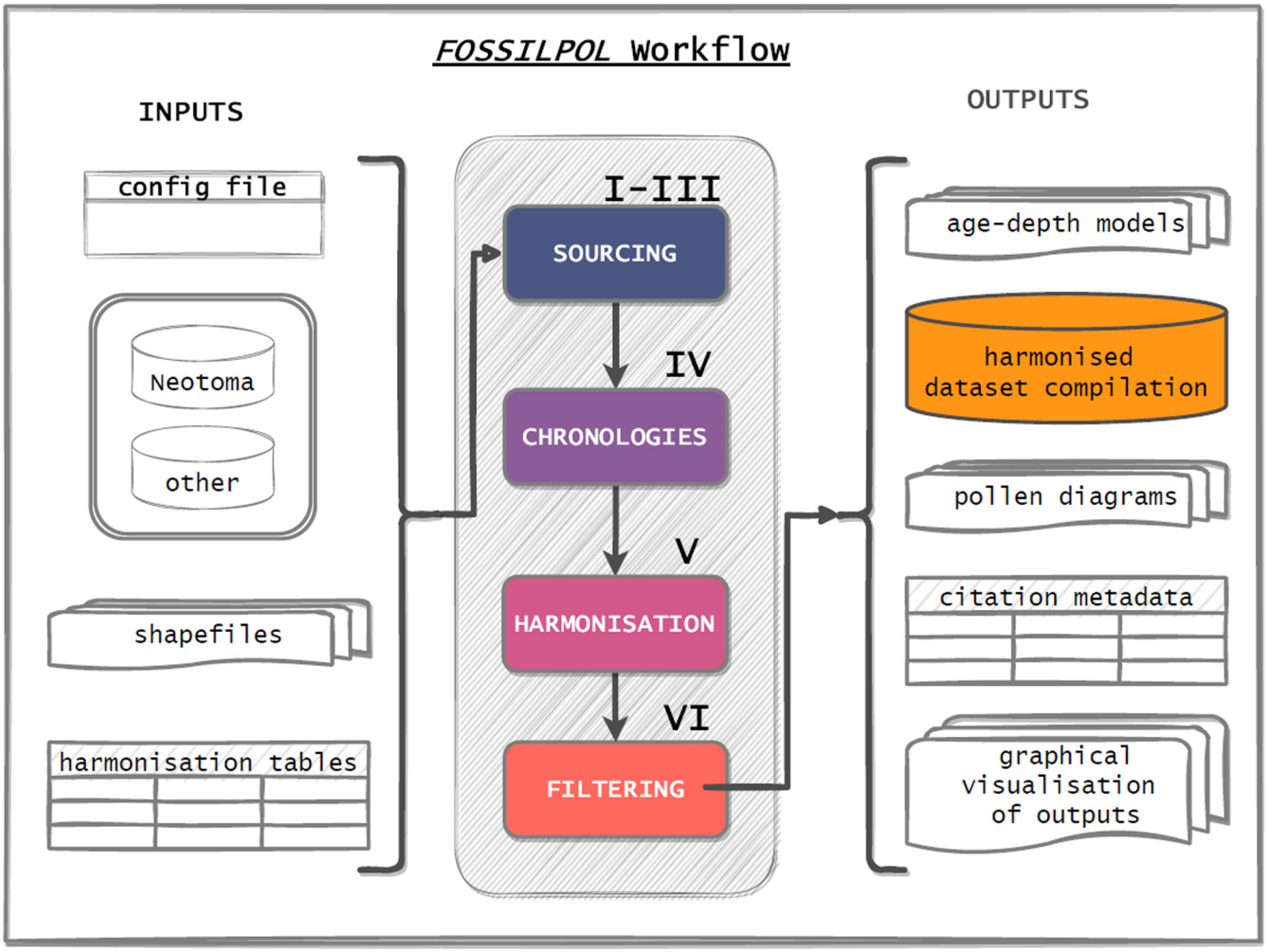

FIGURE 3 Summary figure of *FOSSILPOL* workflow providing an overview of the inputs, main workflow steps and outputs. Major workflow steps are I–III. Sourcing. Note that this section consists of three parts merged into one to simplify the figure, namely downloading/accessing/merging raw datasets, IV. Creating chronologies, V. Performing taxonomic harmonization, VI. Applying user-based criteria for filtering. A detailed version is presented in Appendix S2 and the step-by-step guide on the *FOSSILPOL* website (https://hope-uib-bio.github.io/FOSSILPOL-website/). See Box 1 for term definitions.

The *FOSSILPOL* workflow is coded in the R environment (R Core Team, 2021) and accessible through GitHub as a template repository for data analysts to copy and adjust to their project-based interests (https://github.com/HOPE-UIB-BIO/FOSSILPOL-workflow). Several R-packages are used throughout the workflow, but most processes are accessible through the *R-Fossilpol* package, a new R-package developed specifically for this workflow (currently hosted by GitHub).

3.1 | Data inputs

The *FOSSILPOL* workflow is set up in a way that data from Neotoma are the primary data input, but other data sources can be used in parallel by using our predefined format (Figure 3). The data analyst thus has the flexibility to source data from either Neotoma or from another data source, as long as our predefined format is used. This includes: (a) metadata (geographical location, author of the data, etc.), (b) depositional information (Section 2.3), (c) chronology information (Section 2.4), (d) level information and (e) pollen count information (Section 2.6). See more on the usage of other datasets in *A step-by-step guide to data processing* on the *FOSSILPOL* website.

Three additional data inputs are required for the initial setup of the *FOSSILPOL* workflow: (i) configuration file, (ii) geographical 'shapefile(s)' (Box 1) and (iii) harmonization table(s) (Figure 3):

1 *Configuration file*—The configuration file contains all the user-selected settings, which will be applied throughout the workflow. These range from technical settings (e.g. location of data storage) to specific criteria for records inclusion. The configuration file template is provided with the workflow guide on how to apply criteria (see Appendix S3).
2 *Geographical shapefiles*—The workflow is internally set up so that data are processed by geographical regions and shapefiles are used to assign relevant geographical information to the records to process. First, the workflow is conceptualized for a global project, so the default structure of data processing is done per continent (i.e. region = continent), but the data analyst can use any other regionalization of interest. The workflow comes with a default shapefile roughly delimiting continents, but it can be adjusted or replaced to fit project needs. Second, the taxonomic harmonization of records is structured by 'harmonization regions' (Box 1), provided by a 'harmonization region shapefile' (Box 1). By default, this shapefile is a copy of the continental shapefile, but as harmonization tables are region-specific (see next data input item) this shapefile can be adjusted to represent the geographical boundaries of the harmonization regions used. Finally, if the analyst is interested in other biogeographical, climatic or ecological units of interest to be linked to each record (e.g. ecozones, biome type, climate zones), then additional shapefiles (or TIF files) can be added to the workflow (see *FOSSILPOL* website for further details).
3 *Harmonization tables*—In each project, one harmonization table must be provided for each harmonization region (delimitated by the corresponding harmonization region shapefile). A harmonization table always comes with two columns: (i) original taxa with taxonomic names as originally provided by Neotoma or other data sources and (ii) harmonized taxa with the final taxonomic names. The workflow will detect if a harmonization table has been provided by the data analyst, or if not, create a new table with all detected original taxa names for each harmonization region. The latter can then serve as the template for the data analyst to create the harmonized taxa column.

3.2 | Data outputs

The final outputs from the *FOSSILPOL* workflow (Figure 3) include (i) a ready-to-use standardized compilation of taxonomically harmonized fossil pollen data in an .rds format (R language data storing format), ready for the analytical stage, (ii) plots of modelled age-depth curves for each record in PDF format, (iii) pollen diagrams for each record in PDF format, (iv) a metadata table listing the main data contributor, contact information and corresponding publications of the used datasets and (v) overview figures of the spatial and temporal distribution of the dataset compilation, namely a map and a graph of the record lengths, respectively.

3.3 | *FOSSILPOL* example project for Scandinavia

To illustrate the application of *FOSSILPOL* and the overview figures that can be readily achieved after running the workflow, we present an example project from Scandinavia. We used Neotoma as input data and our *Config file* containing all our user-defined, level-filtering criteria (Appendix S4). The selected area of interest is part of the European Pollen Database (http://www.europeanpollendatabase.net/) as a constituent database of Neotoma. Our goal was to assess the availability of records fulfilling certain quality criteria (spatial extent, pollen sums, levels and chronology), within the last 8500 cal yr BP. The latter was set as the period of interest in the regional age limits file, where young_age = 1000; old_age = 5000, end_of_interest_period = 8500. We also added to the workflow a global biome-shapefile (Olson et al., 2001) and assigned these ecological units of interest to the records. Following the sequential steps in *FOSSILPOL*, we obtained 246 datasets in the initial download (Figure 4). However, the subsequent steps filtered out substantial numbers of records (filtered out records: $n = 149$). Once the data compilation was finalized (in this case after 4 h, most of the time being re-estimation of age-depth models), *R-Fossilpol* provided an overview as shown in Figure 5. The overview of the spatial (Figure 5a) and temporal (Figure 5b) distribution of records can now be checked visually before proceeding to further analysis. The GitHub repository of the project including the figures and

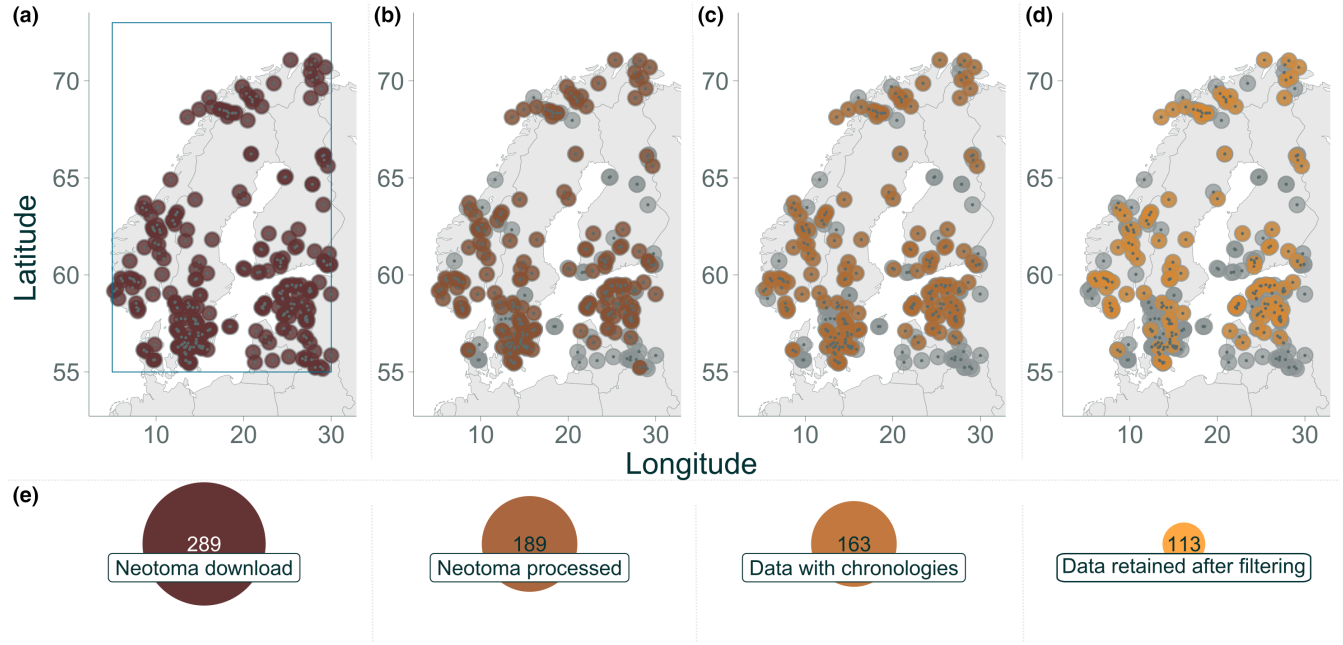

FIGURE 4 Overview of the effect of user-defined criteria in *R-Fossilpol* on the number of available records between the initial download and the final data compilation. At each step, certain user-defined criteria have been applied, which reduces the number of suitable records available for data analysis. In each map, the coloured points represent the records present in the compilation, while grey points represent those filtered out from the initial download. The number of records decreases at each step, but data quality increases when criteria are applied that are aimed at decreasing temporal and taxonomic uncertainties. (a) The distribution of fossil pollen records downloaded from Neotoma limited by the choice of spatial domain (visualized by dark blue rectangle). These records are part of the European Pollen Database (http://www.europeanpollendatabase.net/) as a constituent database of Neotoma. (b) The distribution of fossil pollen records after initial data processing. This includes, e.g. selection of depositional environments or selection of ecological groups (e.g. trees and shrubs, herbs). (c) The distribution of fossil pollen records with age-depth models reconstructed under user-defined criteria. Only records with robust age-depth models were kept. (d) The distribution of fossil pollen records after applying criteria aimed at filtering individual levels and whole records. For example, selection based on pollen sums, age limitation of records and setting the minimum number of levels for each record within the period of interest.

the Reproducibility bundle link can be found at https://github.com/HOPE-UIB-BIO/FOSSILPOL-example-Scandinavia.

4 | DISCUSSION

Global analyses of palaeoecological data are essential for answering macroecological research questions related to understanding ecological patterns, processes and dynamics in time dimensions that cannot be observed directly (Fordham et al., 2020; Jackson & Blois, 2015; Woodbridge et al., 2020). Multi-dataset studies using palaeoecological proxies have been used for many years (e.g. Cao et al., 2013; Giesecke et al., 2014; Huntley & Birks, 1983; Kuneš et al., 2019). So far, no guide to data preparation for quantitative analyses has been published. The workflow presented here is a modular tool specifically designed to be flexible for the needs and wishes of a variety of users, while at the same time guiding the data analyst through procedures that help to create a standardized and high-quality compilation of palaeoecological data, especially fossil pollen datasets.

There are many benefits to following such a stepwise procedure. First, the reproducibility of the selection and filtering of high-quality palaeoecological datasets is guaranteed, and the criteria used to do so are transparent. Second, data analysts outside the field of palaeoecology are made aware of selection criteria that palaeoecologists would routinely use to filter out low-quality datasets or levels within a record. This straightforward workflow facilitates the use of palaeoecological data across disciplines. Third, for the comparison or integration of fossil data with datasets of present-day vegetation, for instance with the Botanical Information and Ecology Network (BIEN; http://bien.nceas.ucsb.edu/bien/) or TRY plant trait database (Kattge et al., 2020), project-based taxonomic harmonizations are imperative. Given that the degree of harmonization can significantly influence the outcome of analyses further down the line, this step is carefully integrated and documented in the workflow. Fourth, these procedures allow a step in the direction of uniform presentation of processed palaeoecological data, where there is a common understanding among researchers from different fields about how the compiled data were derived and processed to form ultimately the basis for further quantitative analyses at macroecological scales. Finally, the modular structure of the workflow reduces processing time and increases rapid access to large dataset compilations, but it does not replace the need for joint decision-making and discussions with expert colleagues about the datasets and their

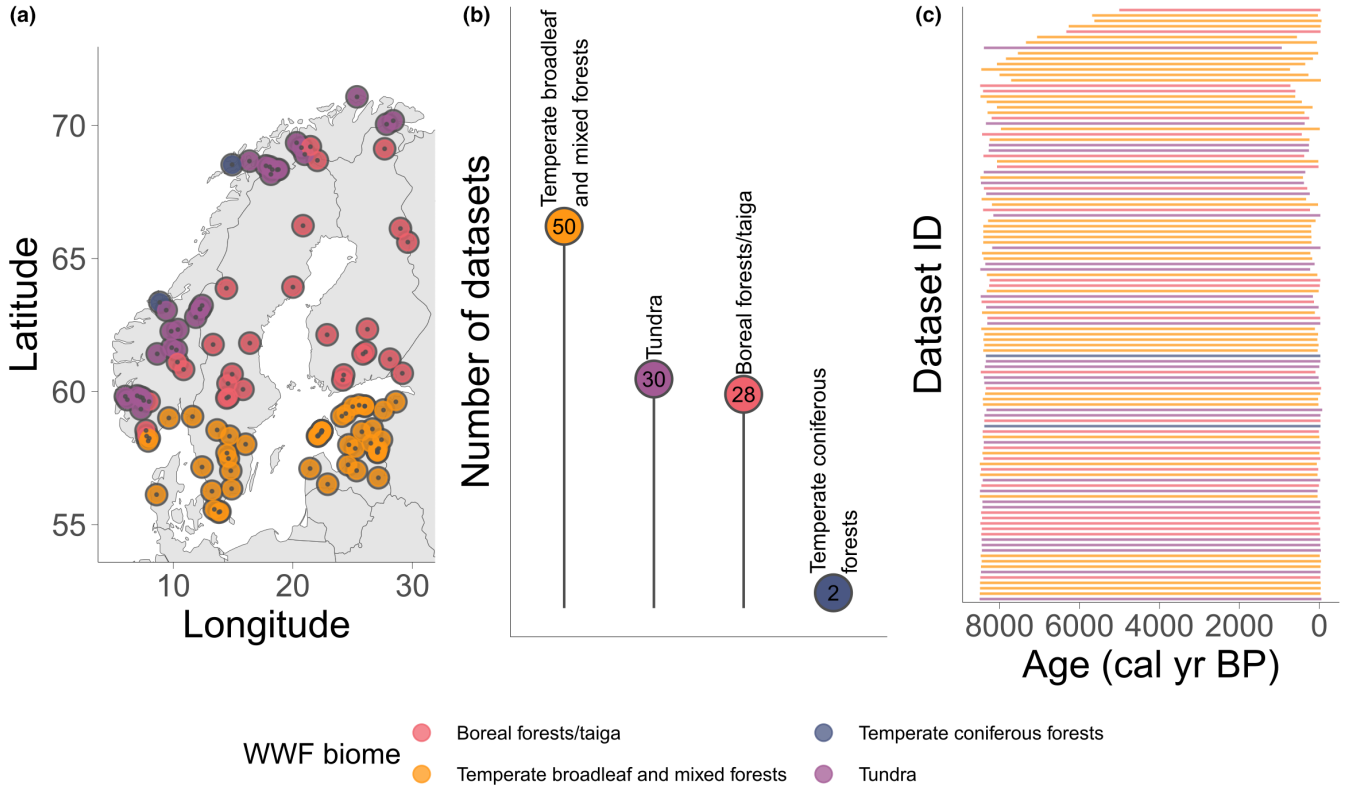

FIGURE 5 Data overview of an example project from Scandinavia using the *FOSSILPOL* workflow. (a) Geographical position of all records in this data compilation, where one point represents one record. (b) Number of records in each biome (Terrestrial Ecoregions of the World, Olson et al., 2001). (c) Temporal distribution of all records in this data compilation, where one line represents the temporal span of one record. Records and background regions are coloured based on biomes ('WWF biomes'). Records were derived from the European Pollen Database (http://www.europeanpollendatabase.net/) in Neotoma.

complexity. The input from experts in palaeoecological and fossil pollen research is essential. We provide relevant anchor points for discussion with our workflow, glossary and Appendices, but, if necessary, urge users to gain more in-depth understanding in conjunction with expert palaeoecologists.

Using Scandinavia as an example (Figures 4 and 5), we created a workflow that is easily reproducible at any spatial and temporal scale and combines data processing criteria with inputs and outputs designed for easy storage on Figshare and Dryad (e.g. https://github.com/HOPE-UIB-BIO/FOSSILPOL-example-Scandinavia). With the *FOSSILPOL* workflow, we promote the improvement of the documentation during the processing of palaeoecological data, consisting of a protocol for best practice and sets of inferences for appropriate analyses of fossil pollen data. Such successful management and preservation of processed data and the corresponding analytical steps will help in FAIR data practices and re-usability in science (Wilkinson et al., 2016). Finally, given the open-source character of the workflow itself, the current version is a starting point for further improvements (see user guidelines for commenting under *Get in touch* on the *FOSSILPOL* website).

The *FOSSILPOL* workflow is designed to work efficiently with Neotoma as an external data source and use the provided metadata for filtering purposes. These metadata fields are, however, not universal across the palaeoecological research infrastructure. Structures vary and recommendations for required metadata are not standardized (Emile-Geay et al., 2018). Although Neotoma will probably continue to maintain its leading role in global-change research, frameworks to connect more easily among different databases are being developed (e.g. between LinkedEarth and pSESYNTH project; Nieto-Lugilde et al., 2021). Overall, more work is still needed to increase interoperability (Emile-Geay et al., 2018). In our workflow, we facilitate the use of data not derived from Neotoma by providing a cross-archive datafile as a metadata standard. Ideally, similar workflows should be developed for the efficient sourcing and processing of data from parallel research infrastructures (see table 3 in Nieto-Lugilde et al., 2021).

We emphasize that the use of *FOSSILPOL* and the vast opportunities ahead for fossil pollen synthesis work rely heavily on the availability of open research and data as guided by FAIR practices (Wilkinson et al., 2016). We acknowledge the work of the data contributors, data stewards and the Neotoma and EPD community for their extensive work and dedication and encourage the users of *FOSSILPOL* to use the metadata overview table (Appendix S2) to acknowledge correctly data contributors and corresponding publications of datasets. Finally, we hope that our workflow will motivate more researchers to contribute their data for open science while also becoming data analysts themselves by using our guide and workflow.

Though our workflow is currently tailor-made for fossil pollen records, we foresee a future potential to adjust the stepwise approach

to other palaeoecological proxies. Fossil pollen records are often accompanied by reconstructions using other proxies. For example, nonpollen palynomorphs (NPPs) and charcoal can be used as indicators of human impact, eutrophication and changes in erosion, and may be complementary to the analyses of fossil pollen. In the future, the workflow could be adjusted to accommodate other taxonomic groups, such as diatoms and foraminifera, obtained via similar archive collection procedures as fossil pollen with corresponding chronologies (e.g. Benito et al., 2022; Yasuhara et al., 2020).

Easier access to palaeoecological data processing will stimulate inter- and trans-disciplinary use of information on the history of biota for complex issues linked to global-change drivers and biodiversity crises. Using palaeoecological data with a solid knowledge of the data's strengths and weaknesses—and how to overcome them—enables researchers in any field to address research questions that can ultimately help decision-makers, conservation organizations and governments understand biodiversity and ecological processes. With our guide and workflow, we provide a tool that the scientific community can use to address an array of research questions of broad interest in macroecology, biogeography, palaeoecology and conservation.

AUTHOR CONTRIBUTIONS

Suzette G. A. Flantua, Ondrej Mottl, Vivian A. Felde and Kuber P. Bhatta (HOPE's postdocs) conceived the idea with Alistair W. R. Seddon and jointly developed this paper over the course of HOPE's project (2018–2022). Suzette G. A. Flantua led the writing, and Ondřej Mottl led the development of the *FOSSILPOL* workflow, *R-Fossilpol* package, website and the European example with critical contributions by all the postdocs. H. John B. Birks and Alistair W. R. Seddon led the theoretical content of the workflow with conceptual contributions by Hilary H. Birks and John-Arvid Grytnes. All authors contributed to the drafts and gave the final approval for publication.

ACKNOWLEDGEMENTS

We thank Manuel Steinbauer, Franka Gaiser, Gregor Mathes, David R. Early, Sarah J. Ivory and Hannah Wauchope for testing the package during its development. Funding was provided by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No. 741413) to the project called 'HOPE Humans On Planet Earth—Long-term impacts on biosphere dynamics'. S.G.A.F. additionally acknowledges support from Trond Mohn Stiftelse (TMS) and the University of Bergen for the startup grant 'TMS2022STG03'. We thank Cathy Jenks for help with text edits.

CONFLICT OF INTEREST STATEMENT

The authors have no conflicts of interest to declare.

DATA AVAILABILITY STATEMENT

All relevant code and data can be found at GitHub: The *FOSSILPOL* workflow https://github.com/HOPE-UIB-BIO/FOSSILPOL-workflow, the *R-Fossilpol* package https://github.com/HOPE-UIB-BIO/R-Fossilpol-package and the *FOSSILPOL* example for Scandinavia https://github.com/HOPE-UIB-BIO/FOSSILPOL-example-Scandinavia.

ORCID

Suzette G. A. Flantua https://orcid.org/0000-0001-6526-3037
Ondrej Mottl https://orcid.org/0000-0002-9796-5081
Vivian A. Felde https://orcid.org/0000-0002-2855-0894
Kuber P. Bhatta https://orcid.org/0000-0001-7837-1395
Hilary H. Birks https://orcid.org/0000-0001-6881-9133
John-Arvid Grytnes https://orcid.org/0000-0002-6365-9676
Alistair W. R. Seddon https://orcid.org/0000-0002-8266-0947
H. John B. Birks https://orcid.org/0000-0002-5891-9859

REFERENCES

Antonelli, A., Kissling, W. D., Flantua, S. G. A., Bermudez, M., Mulch, A., Muellner-Riehl, A. N., Kreft, H., Linder, P., Badgley, C., Fjeldså, J., Fritz, S. A., Rahbek, C., Herman, F., Hooghiemstra, H., & Hoorn, C. (2018). Geological and climatic influences on mountain biodiversity. *Nature Geoscience*, *11*(10), 718–725. https://doi.org/10.1038/s41561-018-0236-z

Barnosky, A. D., Hadly, E. A., Gonzalez, P., Head, J., Polly, P. D., Lawing, A. M., Eronen, J. T., Ackerly, D. D., Alex, K., Biber, E., Blois, J., Brashares, J., Ceballos, G., Davis, E., Dietl, G. P., Dirzo, R., Doremus, H., Fortelius, M., Greene, H. W., … Zhang, Z. (2017). Merging paleobiology with conservation biology to guide the future of terrestrial ecosystems. *Science*, *355*(6325), eaah4787. https://doi.org/10.1126/science.aah4787

Benito, X., Feitl, M., Carrevedo, M. L., Vélez, M. I., Escobar, J., Tapia, P., Kannan, M. S., & Fritz, S. C. (2022). Tropical South America diatom database: A tool for studying the macroecology of microorganisms. *Diatom Research*, 1–13. https://doi.org/10.1080/0269249X.2022.2078429

Birks, H. J. B. (1995). Quantitative palaeoenvironmental reconstructions. In D. Maddy & J. S. Brew (Eds.), *Statistical modelling of quaternary science data technical guide 5* (pp. 161–254). Quaternary Research Association.

Birks, H. J. B., & Birks, H. H. (1980). *Quaternary palaeoecology*. Edward Arnold.

Blaauw, M. (2010). Methods and code for 'classical' age-modelling of radiocarbon records. *Quaternary Geochronology*, *5*(5), 512–518. https://doi.org/10.1016/j.quageo.2010.01.002

Blaauw, M., & Christen, J. A. (2011). Flexible paleoclimate age-depth models using an autoregressive gamma process. *Bayesian Analysis*, *6*(3), 457–474. https://doi.org/10.1214/11-BA618

Blaauw, M., Christen, J. A., Bennett, K. D., & Reimer, P. J. (2018). Double the dates and go for Bayes—Impacts of model choice, dating density and quality on chronologies. *Quaternary Science Reviews*, *188*, 58–66. https://doi.org/10.1016/j.quascirev.2018.03.032

Blaauw, M., Wohlfarth, B., Christen, J. A., Ampel, L., Veres, D., Hughen, K. A., Preusser, F., & Svensson, A. (2010). Were last glacial climate events simultaneous between Greenland and France? A quantitative comparison using non-tuned chronologies. *Journal of Quaternary Science*, *25*, 387–394. https://doi.org/10.1002/jqs.1330

Blaus, A., Reitalu, T., Gerhold, P., Hiiesalu, I., Massante, J. C., & Veski, S. (2020). Modern pollen–plant diversity relationships inform palaeoecological reconstructions of functional and phylogenetic diversity in calcareous fens. *Frontiers in Ecology and Evolution*, *8*, 207. https://doi.org/10.3389/fevo.2020.00207

Blois, J. L. (2012). Update: Stemming "ignorance creep" in paleoecology and biogeography. *Frontiers of Biogeography*, *4*(3). http://escholarship.org/uc/item/00c905kp

Blois, J. L., Williams, J. W., Grimm, E. C., Jackson, S. T., & Graham, R. W. (2011). A methodological framework for assessing and reducing temporal uncertainty in paleovegetation mapping from late-quaternary pollen records. *Quaternary Science Reviews*, *30*(15), 1926–1939. https://doi.org/10.1016/j.quascirev.2011.04.017

Bronk Ramsey, C. (1995). Radiocarbon calibration and analysis of stratigraphy: The OxCal program. *Radiocarbon*, *37*(2), 425–430. https://doi.org/10.1017/S0033822200030903

Bronk Ramsey, C. (2001). Development of the radiocarbon calibration program. *Radiocarbon*, *43*(2A), 355–363. https://doi.org/10.1017/S0033822200038212

Brovkin, V., Brook, E., Williams, J. W., Bathiany, S., Lenton, T. M., Barton, M., DeConto, R. M., Donges, J. F., Ganopolski, A., McManus, J., Praetorius, S., Vernal, A., Abe-Ouchi, A., Cheng, H., Claussen, M., Crucifix, M., Gallopín, G., Iglesias, V., Kaufman, D. S., … Yu, Z. (2021). Past abrupt changes, tipping points and cascading impacts in the earth system. *Nature Geoscience*, *14*(8), 550–558. https://doi.org/10.1038/s41561-021-00790-5

Buma, B., Harvey, B. J., Gavin, D. G., Kelly, R., Loboda, T., McNeil, B. E., Marlon, J. R., Meddens, A. J. H., Morris, J. L., Raffa, K. F., Shuman, B., Smithwick, E. A. H., & McLauchlan, K. K. (2019). The value of linking paleoecological and neoecological perspectives to understand spatially-explicit ecosystem resilience. *Landscape Ecology*, *34*(1), 17–33. https://doi.org/10.1007/s10980-018-0754-5

Cao, X., Ni, J., Herzschuh, U., Wang, Y., & Zhao, Y. (2013). A late quaternary pollen dataset from eastern continental Asia for vegetation and climate reconstructions: Set up and evaluation. *Review of Palaeobotany and Palynology*, *194*, 21–37. https://doi.org/10.1016/j.revpalbo.2013.02.003

Chaudhary, C., Richardson, A. J., Schoeman, D. S., & Costello, M. J. (2021). Global warming is causing a more pronounced dip in marine species richness around the equator. *Proceedings of the National Academy of Sciences of the United States of America*, *118*(15), e2015094118. https://doi.org/10.1073/pnas.2015094118

Chen, I.-C., Hill, J. K., Ohlemüller, R., Roy, D. B., & Thomas, C. D. (2011). Rapid range shifts of species associated with high levels of climate warming. *Science*, *333*(6045), 1024–1026. https://doi.org/10.1126/science.1206432

Chevalier, M., Davis, B. A. S., Heiri, O., Seppä, H., Chase, B. M. K., Gajewski, K., Lacourse, T., Telford, R. J., Finsinger, W., Joel, G., Kühl, N., Maezumi, S. Y., Tipton, J., Carter, V. A., Brussel, T., Phelps, L. N., Dawson, A., Zanon, M., Vallé, F., … Kupriyanov, D. (2020). Pollen-based climate reconstruction techniques for late quaternary studies. *Earth-Science Reviews*, *210*, 103384. https://doi.org/10.1016/j.earscirev.2020.103384

Cleal, C., Pardoe, H. S., Berry, C. M., Cascales-Miñana, B., Davis, B. A. S., Diez, J. B., Filipova-Marinova, M. V., Giesecke, T., Hilton, J., Ivanov, D., Kustatscher, E., Leroy, S., Mcelwain, J. C., Opluštil, S., Popa, M. E., Seyfullah, L., Stolle, E., Thomas, B. A., & Uhl, D. (2021). Palaeobotanical experiences of plant diversity in deep time. 1: How well can we identify past plant diversity in the fossil record? *Palaeogeography, Palaeoclimatology, Palaeoecology*, *576*, 110481. https://doi.org/10.1016/j.palaeo.2021.110481

Cushing, E. J. (1967). Evidence for differential pollen preservation in late quaternary sediments in Minnesota. *Review of Palaeobotany and Palynology*, *4*(1), 87–101. https://doi.org/10.1016/0034-6667(67)90175-3

Daniau, A.-L., Desprat, S., Aleman, J. C., Bremond, L., Davis, B., Fletcher, W. J., Marlon, J. R., Marquer, L., Montade, V., Morales-Molino, C., Naughton, F., Rius, D., & Urrego, D. H. (2019). Terrestrial plant microfossils in palaeoenvironmental studies, pollen, microcharcoal and phytolith. Towards a comprehensive understanding of vegetation, fire and climate changes over the past one million years. *Revue de Micropaleontologie*, *63*, 1–35. https://doi.org/10.1016/j.revmic.2019.02.001

Davies, A. L., Streeter, R., Lawson, I. T., Roucoux, K. H., & Hiles, W. (2018). The application of resilience concepts in palaeoecology. *The Holocene*, *28*(9), 1523–1534. https://doi.org/10.1177/0959683618777077

Davis, M. B. (1968). Pollen grains in lake sediments: Redeposition caused by seasonal water circulation. *Science*, *162*(3855), 796–799. https://doi.org/10.1126/science.162.3855.796

Dawson, T. P., Jackson, S. T., House, J. I., Prentice, I. C., & Mace, G. M. (2011). Beyond predictions: Biodiversity conservation in a changing climate. *Science*, *332*(6025), 53–58. https://doi.org/10.1126/science.1200303

Deza-Araujo, M., Morales-Molino, C., Conedera, M., Pezzatti, G. B., Pasta, S., & Tinner, W. (2022). Influence of taxonomic resolution on the value of anthropogenic pollen indicators. *Vegetation History and Archaeobotany*, *31*(1), 67–84. https://doi.org/10.1007/s00334-021-00838-x

Dietl, G. P., Kidwell, S. M., Brenner, M., Burney, D. A., Flessa, K. W., Jackson, S. T., & Koch, P. L. (2015). Conservation paleobiology: Leveraging knowledge of the past to inform conservation and restoration. *Annual Review of Earth and Planetary Sciences*, *43*(1), 79–103. https://doi.org/10.1146/annurev-earth-040610-133349

Dillon, E., Dunne, E., Womack, T., Kouvari, M., Larina, E., Claytor, J., Ivkić, A., Juhn, M., Milla Carmona, P. S., Robson, S. V., Saha, A., Villafaña, J. A., & Zill, M. (2023). Challenges and directions in analytical paleobiology. *Paleobiology*, *1*, 1–17. https://doi.org/10.1017/pab.2023.3

Divíšek, J., Hájek, M., Jamrichová, E., Petr, L., Večeřa, M., Tichý, L., Willner, W., & Horsák, M. (2020). Holocene matters: Landscape history accounts for current species richness of vascular plants in forests and grasslands of eastern Central Europe. *Journal of Biogeography*, *47*(3), 721–735. https://doi.org/10.1111/jbi.13787

Djamali, M., & Cilleros, K. (2020). Statistically significant minimum pollen count in quaternary pollen analysis; the case of pollen-rich lake sediments. *Review of Palaeobotany and Palynology*, *275*, 104156. https://doi.org/10.1016/j.revpalbo.2019.104156

Ellis, E. C., Gauthier, N., Goldewijk, K. K., Bird, R. B., Boivin, N., Díaz, S., Fuller, D. Q., Gill, J. L., Kaplan, J. O., Kingston, N., Locke, H., McMichael, C. N. H., Ranco, D., Rick, T. C., Shaw, M. R., Stephens, L., Svenning, J.-C., & Watson, J. E. M. (2021). People have shaped most of terrestrial nature for at least 12,000 years. *Proceedings of the National Academy of Sciences of the United States of America*, *118*(17), e2023483118. https://doi.org/10.1073/pnas.2023483118

Ellis, E. C., & Ramankutty, N. (2008). Putting people in the map: Anthropogenic biomes of the world. *Frontiers in Ecology and the Environment*, *6*(8), 439–447. https://doi.org/10.1890/070062

Emile-Geay, J., Khider, D., McKay, N., Gil, Y., Garijo, D., & Ratnakar, V. (2018). LinkedEarth: Supporting paleoclimate data standards and crowd curation. *Past Global Change Magazine*, *26*(2), 62–63. https://doi.org/10.22498/pages.26.2.62

Fægri, K., & Iversen, J. (1964). *Text book of pollen analysis* (2nd ed.). Blackwell.

Fægri, K., Iversen, J., Kaland, P. E., & Krzywinski, K. (1989). *Textbook of pollen analysis* (4th ed.). John Wiley and Sons.

Finsinger, W., Giesecke, T., Brewer, S., & Leydet, M. (2017). Emergence patterns of novelty in European vegetation assemblages over the past 15,000 years. *Ecology Letters*, *20*(3), 336–346. https://doi.org/10.1111/ele.12731

Flantua, S. G. A., Blaauw, M., & Hooghiemstra, H. (2016). Geochronological database and classification system for age uncertainties in neotropical pollen records. *Climate of the Past*, *12*(2), 387–414. https://doi.org/10.5194/cp-12-387-2016

Flantua, S.G.A., Hooghiemstra, H., Grimm, E. C., Behling, H., Bush, M. B., Gonzalez, C., Gosling, W. D., Ledru, M.-P., Lozano, S., Maldonado, A., Prieto, A. R., Rull, V., & Van Boxel, J. (2015). Updated site compilation of the Latin American pollen database. *Review of Palaeobotany*

and Palynology, *223*, 104–115. https://doi.org/10.1016/j.revpalbo.2015.09.008

Fordham, D. A., Jackson, S. T., Brown, S. C., Huntley, B., Brook, B. W., Dahl-Jensen, D., Gilbert, M. T. P., Otto-Bliesner, B. L., Svensson, A., Theodoridis, S., Wilmshurst, J. M., Buettel, J. C., Canteri, E., McDowell, M., Orlando, L., Pilowsky, J., Rahbek, C., & Nogues-Bravo, D. (2020). Using paleo-archives to safeguard biodiversity under climate change. *Science*, *369*(6507), 12. https://doi.org/10.1126/science.abc5654

Giesecke, T., Davis, B., Brewer, S., Finsinger, W., Wolters, S., Blaauw, M., de Beaulieu, J.-L., Binney, H., Fyfe, R. M., Gaillard, M.-J., Gil-Romera, G., van der Knaap, W. O., Kuneš, P., Kühl, N., van Leeuwen, J. F. N., Leydet, M., Lotter, A. F., Ortu, E., Semmler, M., & Bradshaw, R. H. W. (2014). Towards mapping the late quaternary vegetation change of Europe. *Vegetation History and Archaeobotany*, *23*(1), 75–86. https://doi.org/10.1007/s00334-012-0390-y

Giesecke, T., Wolters, S., van Leeuwen, J. F. N., van der Knaap, W. O., Leydet, M., & Brewer, S. (2019). Postglacial change of the floristic diversity gradient in Europe. *Nature Communications*, *10*(1), 5422. https://doi.org/10.1038/s41467-019-13233-y

Goring, S., Lacourse, T., Pellatt, M. G., Walker, I. R., & Mathewes, R. W. (2010). Are pollen-based climate models improved by combining surface samples from soil and lacustrine substrates? *Review of Palaeobotany and Palynology*, *162*(2), 203–212. https://doi.org/10.1016/j.revpalbo.2010.06.014

Grenié, M., Berti, E., Carvajal-Quintero, J., Dädlow, G. M. L., Sagouis, A., & Winter, M. (2023). Harmonizing taxon names in biodiversity data: A review of tools, databases and best practices. *Methods in Ecology and Evolution*, *14*(1), 12–25. https://doi.org/10.1111/2041-210X.13802

Grimm, E. C., Blaauw, M., Buck, C., & Williams, J. W. (2014). Age models, chronologies, and databases workshop. *Past Global Changes Magazine*, *22*(2), 104. https://doi.org/10.22498/pages.22.2.104

Haslett, J., & Parnell, A. (2008). A simple monotone process with application to radiocarbon-dated depth chronologies. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *57*(4), 399–418.

Hébert, R., Herzschuh, U., & Laepple, T. (2022). Millennial-scale climate variability over land overprinted by ocean temperature fluctuations. *Nature Geoscience*, *15*, 899–905. https://doi.org/10.1038/s41561-022-01056-4

Herzschuh, U. (2020). Legacy of the last glacial on the present-day distribution of deciduous versus evergreen boreal forests. *Global Ecology and Biogeography*, *29*, 198–206. https://doi.org/10.1111/geb.13018

Herzschuh, U., Li, C., Böhmer, T., Postl, A. K., Heim, B., Andreev, A. A., Cao, X., Wieczorek, M., & Ni, J. (2022). LegacyPollen 1.0: A taxonomically harmonized global late quaternary pollen dataset of 2831 records with standardized chronologies. *Earth System Scientific Data*, *14*, 3213–3227. https://doi.org/10.5194/essd-14-3213-2022

Hogg, A. G., Heaton, T. J., Hua, Q., Palmer, J. G., Turney, C. S., Southon, J., Bayliss, A., Blackwell, P. G., Boswijk, G., Ramsey, C. B., Pearson, C. L., Petchey, F., Reimer, P. J., Reimer, R., & Wacker, L. (2020). SHCal20 southern hemisphere calibration, 0–55,000 years cal BP. *Radiocarbon*, *62*(4), 759–778. https://doi.org/10.1017/RDC.2020.59

Huntley, B., & Birks, H. J. B. (1983). *An atlas of past and present pollen maps for Europe: 0–13,000 years ago*. Cambridge University Press.

IPCC. (2021). Summary for policymakers. In V. Masson-Delmotte, P. Zhai, A. Pirani, S. L. Connors, C. Péan, S. Berger, N. Caud, Y. Chen, L. Goldfarb, M. I. Gomis, M. Huang, K. Leitzell, E. Lonnoy, J. B. R. Matthews, T. K. Maycock, T. Waterfield, O. Yelekçi, R. Yu, & B. Zhou (Eds.), *Climate change 2021: The physical science basis. Contribution of working group I to the sixth assessment report of the intergovernmental panel on climate change* (pp. 3–32). Cambridge University Press.

Ivory, S., Lézine, A.-M., Grimm, E. C., & Williams, J. W. (2020). Relaunching the African pollen database: Abrupt change in climate and ecosystems. *Past Global Changes Magazine*, *28*(1), 26. https://doi.org/10.22498/pages.28.1.26

Jackson, S. T. (2007). Looking forward from the past: History, ecology, and conservation. *Frontiers in Ecology and the Environment*, *5*(9), 455. https://doi.org/10.1890/1540-9295(2007)5[455:LFFTPH]2.0.CO;2

Jackson, S. T. (2012). Representation of flora and vegetation in quaternary fossil assemblages: Known and unknown knowns and unknowns. *Quaternary Science Reviews*, *49*, 1–15. https://doi.org/10.1016/j.quascirev.2012.05.020

Jackson, S. T., & Blois, J. L. (2015). Community ecology in a changing environment: Perspectives from the quaternary. *Proceedings of the National Academy of Sciences of the United States of America*, *112*(16), 4915–4921. https://doi.org/10.1073/pnas.1403664111

Jacobson, G. L., & Bradshaw, R. H. W. (1981). The selection of sites for paleovegetational studies. *Quaternary Research*, *16*(1), 80–96. https://doi.org/10.1016/0033-5894(81)90129-0

Jeffers, E. S., Bonsall, M. B., & Willis, K. J. (2011). Stability in ecosystem functioning across a climatic threshold and contrasting forest regimes. *PLoS ONE*, *6*(1), e16134. https://doi.org/10.1371/journal.pone.0016134

Jeffers, E. S., Nogué, S., & Willis, K. J. (2015). The role of palaeoecological records in assessing ecosystem services. *Quaternary Science Reviews*, *112*, 17–32. https://doi.org/10.1016/j.quascirev.2014.12.018

Kattge, J., Bönisch, G., Díaz, S., Lavorel, S., Prentice, I. C., Leadley, P., Tautenhahn, S., Werner, G. D. A., Aakala, T., Abedi, M., Acosta, A. T. R., Adamidis, G. C., Adamson, K., Aiba, M., Albert, C. H., Alcántara, J. M., Alcázar, C. C., Aleixo, I., Ali, H., … Wirth, C. (2020). TRY plant trait database—Enhanced coverage and open access. *Global Change Biology*, *26*(1), 119–188. https://doi.org/10.1111/gcb.14904

Kuneš, P., Abraham, V., & Herben, T. (2019). Changing disturbance-diversity relationships in temperate ecosystems over the past 12,000 years. *Journal of Ecology*, *107*, 1678–1688. https://doi.org/10.1111/1365-2745.13136

Lenoir, J., Bertrand, R., Comte, L., Bourgeaud, L., Hattab, T., Murienne, J., & Grenouillet, G. (2020). Species better track climate warming in the oceans than on land. *Nature Ecology & Evolution*, *4*(8), 1044–1059. https://doi.org/10.1038/s41559-020-1198-2

Lenoir, J., & Svenning, J.-C. (2013). Latitudinal and elevational range shifts under contemporary climate change. In *Encyclopedia of biodiversity* (pp. 599–611). Elsevier. http://linkinghub.elsevier.com/retrieve/pii/B9780123847195003750

Michener, W. K., & Jones, M. B. (2012). Ecoinformatics: Supporting ecology as a data-intensive science. *Trends in Ecology & Evolution*, *27*(2), 85–93. https://doi.org/10.1016/j.tree.2011.11.016

Moore, P. D., Webb, J. A., & Collinson, M. E. (1991). *Pollen analysis*. Blackwell Science.

Mottl, O., Flantua, S. G. A., Bhatta, K. P., Felde, V. A., Giesecke, T., Goring, S., Grimm, E. C., Haberle, S., Hooghiemstra, H., Ivory, S., Kuneš, P., Wolters, S., Seddon, A. W. R., & Williams, J. W. (2021). Global acceleration in rates of vegetation change over the past 18,000 years. *Science*, *372*(6544), 860–864. https://doi.org/10.1126/science.abg1685

Nieto-Lugilde, D., Blois, J. L., Bonet-García, F. J., Giesecke, T., Gil-Romera, G., & Seddon, A. (2021). Time to better integrate paleoecological research infrastructures with neoecology to improve understanding of biodiversity long-term dynamics and to inform future conservation. *Environmental Research Letters*, *16*(9), 095005. https://doi.org/10.1088/1748-9326/ac1b59

Nogué, S., Santos, A. M. C., Birks, H. J. B., Björck, S., Castilla-Beltrán, A., Connor, S., de Boer, E. J., de Nascimento, L., Felde, V. A., Fernández-Palacios, J. M., Froyd, C. A., Haberle, S. G., Hooghiemstra, H., Ljung, K., Norder, S. J., Peñuelas, J., Prebble, M., Stevenson, J., Whittaker, R. J., … Steinbauer, M. J. (2021). The human dimension of biodiversity changes on islands. *Science*, *372*(6541), 488–491. https://doi.org/10.1126/science.abd6706

Nolan, C., Overpeck, J. T., Allen, J. R. M., Anderson, P. M., Betancourt, J. L., Binney, H. A., Brewer, S., Bush, M. B., Chase, B. M., Cheddadi,

R., Djamali, M., Dodson, J., Edwards, M. E., Gosling, W. D., Haberle, S., Hotchkiss, S. C., Huntley, B., Ivory, S. J., Kershaw, A. P., … Jackson, S. T. (2018). Past and future global transformation of terrestrial ecosystems under climate change. *Science*, *361*(6405), 920–923. https://doi.org/10.1126/science.aan5360

Olson, D. M., Dinerstein, E., Wikramanayake, E. D., Burgess, N. D., Powell, G. V. N., Underwood, E. C., D'amico, J. A., Itoua, I., Strand, H. E., Morrison, J., Loucks, C. J., Allnutt, T. F., Ricketts, T., Kura, Y., Lamoreux, J., Wettengel, W. W., Hedao, P., & Kassem, K. (2001). Terrestrial ecoregions of the world: A new map of life on earth: A new global map of terrestrial ecoregions provides an innovative tool for conserving biodiversity. *Bioscience*, *51*(11), 933–938. https://doi.org/10.1641/0006-3568(2001)051[0933:TEOTWA]2.0.CO;2

Parmesan, C., & Yohe, G. (2003). A globally coherent fingerprint of climate change impacts across natural systems. *Nature*, *421*(6918), 37–42. https://doi.org/10.1038/nature01286

Poloczanska, E.S., Brown, C. J., Sydeman, W. J., Kiessling, W., Schoeman, D. S., Moore, P. J., Brander, K., Bruno, J. F., Buckley, L. B., Burrows, M. T., Duarte, C. M., Halpern, B. S., Holding, J., Kappel, C. V., O'Connor, M. I., Pandolfi, J. M., Parmesan, C., Schwing, F., Thompson, S. A., & Richardson, A. J. (2013). Global imprint of climate change on marine life. *Nature Climate Change*, *3*(10), 919–925. https://doi.org/10.1038/nclimate1958

Prentice, I. C. (1985). Pollen representation, source area, and basin size—Toward a unified theory of pollen analysis. *Quaternary Research*, *23*(1), 76–86. https://doi.org/10.1016/0033-5894(85)90073-0

Prentice, I. C. (1988). Records of vegetation in time and space: The principles of pollen analysis. In B. Huntley & T. Webb (Eds.), *Vegetation history* (pp. 17–42). Kluwer Academic Publishers.

R Core Team. (2021). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Rapacciuolo, G., & Blois, J. L. (2019). Understanding ecological change across large spatial, temporal and taxonomic scales: Integrating data and methods in light of theory. *Ecography*, *42*(7), 1247–1266. https://doi.org/10.1111/ecog.04616

Reimer, P. J., Austin, W. E. N., Bard, E., Bayliss, A., Blackwell, P. G., Ramsey, C. B., Butzin, M., Cheng, H., Edwards, R. L., Friedrich, M., Grootes, P. M., Guilderson, T. P., Hajdas, I., Heaton, T. J., Hogg, A. G., Hughen, K. A., Kromer, B., Manning, S. W., Muscheler, R., … Talamo, S. (2020). The IntCal20 northern hemisphere radiocarbon age calibration curve (0–55 cal kBP). *Radiocarbon*, *62*(4), 725–757. https://doi.org/10.1017/RDC.2020.41

Rull, V. (2010). Ecology and palaeoecology: Two approaches, one objective. *The Open Ecology Journal*, *3*(2), 1–5. https://doi.org/10.2174/1874213001003020001

Rull, V. (2012). Community ecology: Diversity and dynamics over time. *Community Ecology*, *13*(1), 102–116. https://doi.org/10.1556/ComEc.13.2012.1.13

Rull, V. (2020). *Quaternary ecology, evolution, and biogeography* (1st ed.). Elsevier.

Runge, J., Gosling, W. D., Lézine, A.-M., & Scott, L. (Eds.). (2021). *Quaternary vegetation dynamics—The African pollen database* (1st ed.). CRC Press.

Sevillano, V., Holt, K., & Aznarte, J. L. (2020). Precise automatic classification of 46 different pollen types with convolutional neural networks. *PLoS ONE*, *15*(6), e0229751. https://doi.org/10.1371/journal.pone.0229751

Staples, T. L., Kiessling, W., & Pandolfi, J. M. (2022). Emergence patterns of locally novel plant communities driven by past climate change and modern anthropogenic impacts. *Ecology Letters*, *25*, 1497–1509. https://doi.org/10.1111/ele.14016

Stephens, L., Fuller, D., Boivin, N., Rick, T., Gauthier, N., Kay, A., Marwick, B., Armstrong, C. G., Barton, C. M., Denham, T., Douglass, K., Driver, J., Janz, L., Roberts, P., Rogers, J. D., Thakar, H., Altaweel, M., Johnson, A. L., Vattuone, M. M. S., … Ellis, E. (2019). Archaeological assessment reveals Earth's early transformation through land use. *Science*, *365*(6456), 897–902. https://doi.org/10.1126/science.aax1192

Svenning, J.-C., Fløjgaard, C., Marske, K. A., Nógues-Bravo, D., & Normand, S. (2011). Applications of species distribution modeling to paleobiology. *Quaternary Science Reviews*, *30*(21–22), 2930–2947. https://doi.org/10.1016/j.quascirev.2011.06.012

Wang, Y., Goring, S. J., & McGuire, J. L. (2019). Bayesian ages for pollen records since the last glaciation in North America. *Scientific Data*, *6*, 176.

Webb, T., III. (1993). Constructing the past from late-quaternary pollen data: Temporal resolution and a zoom lens space-time perspective. In S. M. Kidwell & A. K. Behrensmeyer (Eds.), *Taphonomic approaches to time resolution in fossil assemblages* (Vol. *6*, pp. 79–101). Paleontological Society.

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., … Mons, B. (2016). Comment: The FAIR guiding principles for scientific data management and stewardship. *Scientific Data*, *3*, 1–9. https://doi.org/10.1038/sdata.2016.18

Williams, J. W., Grimm, E. C., Blois, J. L., Charles, D. F., Davis, E. B., Goring, S. J., Graham, R. W., Smith, A. J., Anderson, M., Arroyo-Cabrales, J., Ashworth, A. C., Betancourt, J. L., Bills, B. W., Booth, R. K., Buckland, P. I., Curry, B. B., Giesecke, T., Jackson, S. T., Latorre, C., … Takahara, H. (2018). The Neotoma Paleoecology Database, a multiproxy, international, community-curated data resource. *Quaternary Research*, *89*(1), 156–177. https://doi.org/10.1017/qua.2017.105

Williams, J. W., Jackson, S. T., & Kutzbach, J. E. (2007). Projected distributions of novel and disappearing climates by 2100 AD. *Proceedings of the National Academy of Sciences of the United States of America*, *104*(14), 5738–5742. https://doi.org/10.1073/pnas.0606292104

Williams, J. W., Shuman, B. N., Webb, T., Bartlein, P. J., & Leduc, P. L. (2004). Late-quaternary vegetation dynamics in North America: Scaling from taxa to biomes. *Ecological Monographs*, *74*(2), 309–334.

Willis, K. J., Bailey, R. M., Bhagwat, S. A., & Birks, H. J. B. (2010). Biodiversity baselines, thresholds and resilience: Testing predictions and assumptions using palaeoecological data. *Trends in Ecology & Evolution*, *25*(10), 583–591. https://doi.org/10.1016/j.tree.2010.07.006

Willis, K. J., & MacDonald, G. M. (2011). Long-term ecological records and their relevance to climate change predictions for a warmer world. *Annual Review of Ecology, Evolution, and Systematics*, *42*(1), 267–287. https://doi.org/10.1146/annurev-ecolsys-102209-144704

Wilmshurst, J. M., & McGlone, M. S. (2005). Origin of pollen and spores in surface lake sediments: Comparison of modern palynomorph assemblages in moss cushions, surface soils and surface lake sediments. *Review of Palaeobotany and Palynology*, *136*(1–2), 1–15. https://doi.org/10.1016/j.revpalbo.2005.03.007

Woodbridge, J., Fyfe, R., Smith, D., Pelling, R., de Vareilles, A., Batchelor, R., Bevan, A., & Davies, A. L. (2020). What drives biodiversity patterns? Using long-term multidisciplinary data to discern centennial-scale change. *Journal of Ecology*, *109*(3), 1396–1410. https://doi.org/10.1111/1365-2745.13565

Yasuhara, M., Huang, H.-H., Hull, P., Rillo, M., Condamine, F., Tittensor, D. P., Kucera, M., Costello, M. J., Finnegan, S., O'Dea, A., Hong, Y., Bonebrake, T., McKenzie, N. R., Doi, H., Wei, C.-L., Kubota, Y., & Saupe, E. (2020). Time machine biology: Cross-timescale integration of ecology, evolution, and oceanography. *Oceanography*, *33*(2), 16–28. https://doi.org/10.5670/oceanog.2020.225

Zhao, Y., Xu, Q., Huang, X., Guo, X., & Tao, S. (2009). Differences of modern pollen assemblages from lake sediments and surface soils in arid and semi-arid China and their significance for pollen-based quantitative climate reconstruction. *Review of Palaeobotany and Palynology*, *156*, 519–524. https://doi.org/10.1016/j.revpalbo.2009.05.001

BIOSKETCH

All authors are members of the Humans on Planet Earth (HOPE) project (ERC 2018–2022) dedicated to understanding the impact of prehistoric people on the biosphere and its dynamics. The authors' research interests cover palaeoecology and macroecology, and their overlapping fields. They have a special interest in using palaeoecological data to address research questions with palaeoecological, macroecological and biogeographical perspectives.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Flantua, S. G. A., Mottl, O., Felde, V. A., Bhatta, K. P., Birks, H. H., Grytnes, J.-A., Seddon, A. W. R., & Birks, H. J. B. (2023). A guide to the processing and standardization of global palaeoecological data for large-scale syntheses using fossil pollen. *Global Ecology and Biogeography*, *00*, 1–18. https://doi.org/10.1111/geb.13693