Tropical South America Diatom Database: a tool for studying the macroecology of microorganisms

XAVIER BENITO[1], MELINA FEITL[2,3], MARIA L. CARREVEDO[4], MARIA I VELEZ[5], JAIME ESCOBAR[6,7], PEDRO M. TAPIA[8], MIRIAM STEINITZ-KANNAN[9], SHERILYN C. FRITZ[2]

[1] Marine and Continental Waters Programme, Institute of Agrifood Research and Technology (IRTA), Spain

[2] Department of Earth and Atmospheric Sciences, University of Nebraska-Lincoln, NE, USA

[3] Department of Geography & Geology, University of Nebraska at Omaha, Omaha, NE, USA

[4] Instituto de Ecología y Biodiversidad (IEB), Universidad de Chile- Las Palmeras 3425, Ñuñoa, 7750000, Santiago, Chile

[5] Department of Geology, University of Regina, Canada

[6] Smithsonian Tropical Research Institute, Ciudad de Panamá, Panamá

[7] Universidad del Norte, Barranquilla, Colombia

[8] Instituto Nacional de Glaciares y Ecosistemas de Montaña (INAIGEM), Ancash, Perú

[9] Department of Biological Sciences, Northern Kentucky University, Highland Heights, KY, USA

Correspondence to: Xavier Benito, e-mail: xavier.benito@irta.cat . ORCID: https://orcid.org/0000-0003-0792-2625

Abstract

Determining the mechanisms of community assembly forms the foundation of biogeography and community ecology. Studies of the biodiversity and distribution of Neotropical macro-organisms have revealed the roles of environmental, spatial, and historical factors in structuring communities at different spatial and temporal scales. The role of these factors for species and communities of microorganisms are still poorly understood. Diatoms are a very species-rich group of algae, widely distributed, and sensitive to environmental variation because of their position at the base of aquatic food webs. Here, we present the Tropical South American Diatom Database (TSADB), which contains geographical and ecological information on diatom species across lentic and lotic environments, including predictors that describe local (limnological) and regional (geo-climatic) factors. The open access database can be used to ask fundamental questions in macroecology, including testing foundational theories of metacommunity ecology and biogeography, and evaluating the sensitivity of species and communities to the rapid environmental changes characteristic of tropical regions. The TSADB includes diatom taxa from 437 samples from 326 sites distributed across 26 regions (0 to 5,070 m a.s.l, and between 8°N–35°S; 58–90°W). In addition, long-term, diatom-based paleolimnological records are presented as a complementary tool for identifying geographically well-covered regions with modern and palaeo-datasets. We describe the TSADB structure and functionality, together with the R codes for data manipulation and visualization. Each of the 26 study regions is represented by three data matrices: sampling site information, environmental variables (limnology, climate, and landscape), and diatom community data (relative abundance or presence/absence). Access to data and future additions is through publicly available repositories and a guide to contributors, respectively. Thus, it offers

ample opportunities to complement existing databases on diatoms, allowing optimal usage of TSADB by scientists including diatomists, limnologists, and aquatic ecologists.

Keywords: *Diatoms, Database, Neotropics, Metacommunity, Paleolimnology, Biogeography*

Introduction

Successful application of biological indicators for inferring present and past environmental conditions is dependent on knowing how species vary over space and time in relation to environmental drivers. Whereas studies of the distribution of macro-organisms have illuminated the roles of environmental, spatial, and historical factors in regulating community dynamics (Antonelli et al. 2018), how these factors affect microorganisms, that are often used as aquatic bioindicators, is still poorly known. Diatoms are a diverse and productive group of primary producers in freshwaters, such as lakes, rivers, streams and wetlands (Mann 1999). At the base of aquatic food webs, diatoms often provide the first signal of environmental changes that may cause ecosystem-scale shifts (McGowan & Leavitt 2009). While many resources are available for studying the drivers and dynamics of diatom assemblages in temperate ecosystems, far fewer are available for tropical regions, where rates of environmental change are unprecedented (Trisos et al. 2020).

In tropical South America the wide latitudinal, longitudinal, and elevational ranges, coupled with complex physiography and geology, produce notable climatic and limnological variability that structure species composition and diversity of diatom assemblages (Benito et al. 2018a). Hence, lacustrine sediment cores from lakes of varied origins and ages have enabled the use of subfossil diatoms to track recent (10s-100s yr), historic (1000s yr), and geological (100,000s yr) change in ecosystems (Steinitz-Kannan et al. 1993, Ekdahl et al. 2008, Fritz et al. 2012, Giles et al. 2018, Feitl et al. 2019, Patiño et al. 2020. However, large-scale repositories of raw data combining modern and paleolimnological observations of Neotropical bioindicators are still scarce (e.g. Neotoma paleoecological database; Goring et al. 2015). Thus, it is critical to create and implement a

single-themed diatom database for generating and answering relevant questions about the Neotropics, facilitating access to large archives of data.

Regional collections of diatom datasets that integrate taxonomic, geographic, and ecological information facilitate research on a wide range of ecological and biogeographical topics, such as metacommunity structuring (Dong et al. 2016), resilience to widespread limnological change (Bracht-Flyr & Fritz 2012), formation of aquatic ecosystems (Vélez et al. 2018), latitudinal gradients of diversity (Vyverman et al. 2007), micro-refugia (Nascimento et al. 2019), and evolutionary processes (speciation, extinction) (Spanbauer et al. 2018). Distribution patterns of freshwater diatoms are relatively well documented across mid- and high latitude regions from Eurasia and North America due to regional, national, and international efforts, resulting in the Europeans Diatom Database (Battarbee et al. 2001), and LakeCores (Battarbee et al. 2011), the Circumpolar Diatom Database (Pienitz & Cournoyer 2017), the Russian PaleoLake Database (Syrykh et al. 2021), and the US National Lakes Assessment (US EPA 2009) and Diatoms of North America (diatoms.org). However, there has been less coordinated effort in the tropics (Escobar et al. 2020). More specifically, in tropical South America, despite advances in recent decades to collate and publish floras from early expeditions between the 1980s and 2010s (Benito & Fritz 2020), it is still challenging to find aggregated data on diatom taxonomy, ecology, and distribution for ecological synthesis and application in bioassessment and paleoecology. A consensus among the scientific community is that more emphasis should be given to data archiving, reusability, and making data processing and code-sharing platforms (e.g. GitHub) available to facilitate reproducible and transparent methods. This is particularly crucial for poorly documented regions where community-

curated repositories would enable new research directions and collaborations among institutions (Cheruvelil & Soranno 2018).

Aims and scope

The overarching goal of this paper is to present the tropical South American diatom database (TSADB) as a tool to facilitate analyses of different levels of biological organization (species, populations, metacommunities, and biogeography) and to document the ecology of diatoms across the continent. Note that the term "tropical" is applied widely here, also including sites from sub-tropical locations. Descriptions of the physiography and climate characteristics of the studied regions within the TSADB have been published in previous works (Benito et al. 2018b, Benito & Fritz 2020).

The specific aims of the TSADB are to: *i*) contribute to information on the autoecology, palaeolimnology, and macroecology (metacommunity and biogeography) of tropical South American diatoms, *ii*) identify spatial overlap between contemporary diatom sites and paleolimnological records, and *iii*) develop a ready-to-use workflow in R to facilitate consultation, visualization, and sharing of diatom taxa and associated environmental data. The TSADB is intended to be a live database, integrating new data as they become available. To maintain the database in the future, we have designated a contributor guide on how to share diatom datasets (Supplementary Material). Furthermore, the TSADB is currently a constituent database of Neotoma (Williams et al. 2018), a global community-curated database by regional experts for multiple types of palaeoecological data. TSADB will act as a platform to link diatom data stewards in South America and Neotoma to archive, maintain and manage datasets.

Database compilation

The rationale of TSADB is to study diatom ecology in relation to macroecological patterns of lentic (lake, wetland, pond) and lotic (river, streams) habitats. The TSADB is not the result of a systematic literature review of all diatom studies published on South America. The TSADB began with a compilation of sites mostly sampled from the Andes of Peru and Bolivia as part of diatom ecological and palaeolimnological studies by members of the diatom lab at the University of Nebraska-Lincoln (USA). Recent data incorporated from other regions (Amazon lowlands, Argentina, Brazil, Colombia, Chile, Ecuador, and Perú) included diatom material that was processed and counted by the first author of the manuscript, hence allowing inter-region methodological consistency. A further search was conducted to cover ecoregions of limnological interest, such as the hypersaline shallow lakes in Sud Lipez and the cloud forest streams of Bolivia (Servant-Vildary 1986, Morales & Vis 2007). In total, the TSADB contains 437 samples from 326 sites in 26 regions. The study regions are broadly defined by the geographical extent of the original studies and include both published and unpublished collections (See *Metadata*). Diatoms were identified to species level, where possible, by one taxonomic expert in each region. Species were classified into the following ecological groups: freshwater planktic (e.g., *Cyclostephanos andinus* (Theriot, Carney & Richerson) Tapia, Theriot, Fritz, Cruces & Rivera*, Discostella stelligera* (Cleve & Grunow) Houk & Klee), oligosaline planktic (*Cyclotella meneghiniana* Kützing), obligate epiphytic (*Cocconeis placentula* Ehrenberg), saline (*Amphora carvajaliana* Patrick , *Pseudostaurosira zeilleri* (Héribaud) D.M. Williams & Round), and benthic (*Encyonema* Kützing spp., *Epithemia* Kützing spp., and other benthic taxa). Such classification has proved useful for modelling the dispersal

potential (planktic as "strong dispersers" and benthic as "weak" dispersers) of tropical

diatoms (Wetzel et al. 2012, Benito et al. 2018b).

Structure and functionality

The TSADB is structured as follows. Each region comprises three different data matrices:

diatom community data (relative abundance, presence/absence), site information,

environmental variables (limnology, climate, and landscape). Each sample is identified by a

unique ID ('code'), which links all the data tables to allow access to all information

relevant to each region, taxon, and associated environmental variables.

To explore the database, we created two simple R shiny applications (Winston et al.

2021), allowing the user to search for data on: 1) region site's information, 2)

environmental descriptors, 3) biogeographic distribution of diatom taxa and their relative

abundance, and 4) characterization of diatom taxa based on broad ecological groups

(freshwater planktic, oligosaline planktic, tychoplanktic, benthic, epiphytic, and saline).

The first R shiny app ("*TSADB_shiny_sites.R*") allows the user to obtain site characteristics

for each region, including geographical distribution of sites, sampling information (year of

sampling, sample type, habitat type, substratum), and environmental raw data, as well as

boxplots that summarize a region's limnological, climatic, and landscape context. The

second R shiny app ("*TSADB_shiny_species.R*") makes it possible to search for diatom

taxa. The output search shows the geographical distribution of the taxon, its relative

abundance, and the basic site characteristics (region, year of sampling, habitat type) where

the taxon is found. The TSADB contains an extended metadata file that provides general

information for each region and a list of contributors who contributed data to the respective

dataset, as well as associated references.

The TSADB is hosted in two different data repositories (Dryad for Excel spreadsheets, and Zenodo for R files; see *Data availability statement*) via two different formats to provide users alternative ways to access and use the data. The database contributor guide is included in both repositories.

Excel spreadsheets: 26 excel files (one for each study region). Depending on the original information, each file has three sheets containing the three data matrices (site information, environmental variables, and diatom community data). A metadata sheet describing sources and descriptions of environmental variables is common in all excel files.

R folder: Files consisting of the TSADB's region datasets as a R list object ("TSABD.RData"), three R scripts for data processing, plotting (Figs 1-4) and updating taxon nomenclature, and the R shiny apps. The "TSADB.Rdata" is a list of objects, in which each element refers to a data matrix ($**diatoms,** $**sites**, $**environment**) that itself contains a list of study regions. The R script "*1-cleaning and formatting.R*" allows for data preparation, checking the correct format (i.e., numeric, text, ordered factors, region name spelling), and merging the original datasets by unique IDs into regions. Database manipulation, plotting, and generating metadata is available via the R script "*2-visualization.R*". Finally, the R script "*3-taxonomic-harmonisation.R*" run steps for nomenclature synchronization and authority harmonization (*see below*). The R code routines generate output files and plots that can be turned on or off by commenting or uncommenting lines. All steps were conducted using several R packages, including

tidyverse (Wickham et al. 2019), *ggplot2* (Wilke et al. 2019), *cowplot* (Wilke et al. 2019), and *shiny* (Winston et al. 2021).

Harmonization

Taxonomic harmonization consisted of a two-step process. First, we standardized taxon nomenclature across sites by scanning the names for synonyms and updating nomenclature according to Guiry & Guiry (2017) using the *algaeClassify* R package (Patil et al. 2019). Unfortunately, for most of the regions (except certain Ecuadorean lakes) slide re-counting was not possible, and hence, harmonization was largely the result of merging the resulting synonymy and lumping species complexes (e.g., *Achnanthidium minutissimum* (Kützing) Czarnecki*, Sellaphora pupula* (Kützing) Mereschkowsky*, D. stelligera, Ulnaria ulna* (Nitzsch) Compère) from the database entries. In a second more detailed step, we created a master list of harmonized diatom taxa from TSADB. This was then used to produce a series of R scripts that performed automated steps to synchronize nomenclature for additional datasets entries. Subsequently, we harmonized the authority of diatom names to reflect current taxonomy in BioData (MacCoy 2011) and Omnidia 2015 version (Lecointe et al. 1993) using the previous mentioned R scripts. For names that were not found in these two databases, we assigned diatom names to the most recent relevant taxonomic treatment Finally, limnological data were simply harmonized by ensuring that the variables were expressed in common units of measurement across the datasets.

Database overview

The geographical range of the database spans 8°N–35°S, 58–90°W, and altitudes from 0 to 5,070 m a.s.l (Fig. 1). Data are in site-by-taxon matrices containing relative abundance

(~80% of samples) or presence-absence data (remaining samples, ~20%). Lakes are by far the most represented ecosystem in the database (80%; Fig. 2), followed by rivers and streams. Samples are mostly from surface sediment and modern collections (periphyton, plankton), followed by core tops (top 1cm). Composite samples (i.e., counts included surface sediment plus modern presence/absence data) only applied to a subset of samples from Ecuador. Diatom samples were collected between 1977 to 2017, with many samples from the early 2000s (Fig. 2c).

The database contains highly heterogeneous physical-chemical observations (see *Metadata*). Limnological variables include parameters that have been identified as important for diatoms in the tropical Andes and adjacent Amazon lowlands in previous studies (Steinitz-Kannan 1979, Sylvestre et al. 2001, Tapia et al. 2006, Benito et al. 2018a) that were available for a large number of sites (i.e. water temperature, pH, conductivity, cations [Ca^{2+}, Na^{2+}], anions [Cl^-]). Additional regional-scale information are also provided, including: latitude, longitude, elevation (m), mean annual air temperature (MAT, °C), mean annual precipitation (MAP, mm), temperature seasonality (standard deviation; °C), precipitation seasonality (coefficient of variation; mm), % aquatic habitat, connectivity, Terrain Ruggedness Index (TRI), lake depth (m), lake surface area (km^2), and Human Footprint (Venter et al. 2016). The area of the waterbody was either calculated by digitalization using the ESRI World Imagery (lakes), or extracted using the lowest basin-level classification from USGS's HydroSHEDS database (Lehner et al. 2008). Lake depth was extracted from the HydroLAKES database (Messager et al. 2016) if not available in the original dataset. The extended dataset includes the sources and methods used to extract environmental predictors and contains additional variables related to land-cover and geology (Table S1).

A total of 1890 diatom taxa representing 143 genera are present in the TSADB. *Navicula* sensu stricto and *Nitzschia* are the richest diatom genera, with 166 and 165 species respectively (9% each of the total number of taxa), followed by *Eunotia* (8.2%), *Pinnularia* (7.3%), and *Gomphonema* (6.9%) (Fig. 3). Remarkably, 355 diatom taxa (18%) could not be assigned to described species based on current literature (Frenguelli 1939, Manguin 1964, Servant-Vildary 1986, Oliveira & Steinitz-Kannan 1992, Metzeltin & Lange-Bertalot 1998, 2007, Rumrich et al. 2000, Metzeltin et al. 2005), hence there is great potential for new taxa to be described. By ecological group, benthic and planktic diatoms dominate the freshwater lakes (Fig. 4), with stream and riverine epiphytic diatoms the second most abundant group, followed by lacustrine tychoplanktonic and saline-tolerant species (i.e., water conductivity > 18,000 uS/cm), and benthic taxa inhabiting lentic habitats (wetlands, ponds). Among the freshwater benthic taxa, *C. placentula, Gomphonema parvulum* (Kützing) Kützing, *Nitzschia palea* (Kützing) W. Smith, *Staurosirella pinnata* (Ehrenberg) D.M. Williams & Round 1988, and *A. minutissimum* made up most of the occurrences in the samples. Freshwater planktic diatoms were dominated by *C. meneghiniana* ., *Aulacoseira granulata* (Ehrenberg) Simonsen, *D. stelligera*, and the endemic *C. andinus*. Saline-tolerant diatoms include the endemic *A. carvajaliana*., *Entomoneis paludosa* var *subsalina* (Cleve) Krammer, and *Navicula salinicola* Hustedt. Abundant tychoplanktonic diatom taxa occurring in lentic habitats are *Ulnaria delicatissima* (W. Smith) Aboal & Silva, whereas *Diatoma mesodon* (Ehrenberg) Kützing, *Navicula trivialis* Lange-Bertalot and *Nitzschia pusilla* Grunow dominate in lotic environments.

Regions are unequally represented in terms of elevation and latitude in the TSADB (Fig. 5). Sites at latitudes between 20 S and 10 N above 2000 m a.s.l. are more common in

the database, whereas freshwaters sites below 1000m a.s.l., except for coastal (Ecuador mainland, Galapagos, and Chile) and Andean foreland plains (Colombia, Ecuador, Peru, Bolivia, and Brazil), are less common. Regions also differed in diatom richness and composition. Colombian and Ecuadorean lowlands had the highest taxon richness, while equatorial sites across the elevation gradient had relatively high species richness (>150 average species richness per site) (Fig. 5). However, major differences exist between highlands among Andean regions, for instance, between the Colombian Andes, the central and southern Altiplano of Bolivia and Perú, and the Peruvian Andes. This suggests the relevance of physiography and micro-altitudinal gradients that influence both regional and local climate (Valencia et al. 2016), and ultimately species richness. The study regions of TSADB, while interesting as a generalized classification for visualization and analyses, might be less suited for analyzing large-scale diatom community assessments targeting transition regions that are not presently covered (i.e. páramos and upper Andean forest in the Northern Andes).

Applications and future developments

The database serves as a repository that collates modern diatom studies and compares these with paleolimnological observations to inform ecological synthesis on species-environment relationships at varied spatial and temporal scales. There are several research questions that could be explored with the database:

What insights do different levels of biological organization provide to the study of diatom macroecology?

The TSADB promotes the study of microbial assemblage processes by considering niche (i.e. species filtering) and dispersal factors (Ricklefs & Jenkins 2011). For instance, metacommunity ecology, defined as a set of local populations connected through dispersal, may benefit from a georeferenced database of diatom community data that can be used to calculate ecologically meaningful spatial distances, such as landscape resistance to genetic flow (i.e., dispersal), in combination with water chemistry conditions (i.e., niche), to shed light on the local and regional processes structuring communities (Dong et al. 2016, Soininen & Teittinen 2019).

Improved knowledge of the relative effects of local and regional factors will help to inform management practices. For instance, biomonitoring tools have historically assumed species sorting only and hence spatial factors, such as dispersal limitation, have been underestimated in ecological assessment (Vilmi et al. 2016). Combining metacommunity approaches with traditional locally based ecological assessments from different habitats (e.g. lentic and lotic) is timely, because habitat fragmentation and water degradation (local processes) will likely scale-up to the catchment level, while local human impacts are exacerbated by the impact of global change and the increased interaction of multiple stressors (e.g. global warming, water abstraction, land-use change) (Grimm & Ambrizzi 2009).

Including dispersal-based metrics, which have rarely been considered due to insufficient spatial/environmental gradients, could also improve the accuracy of species distribution models (Jamoneau et al. 2018, Nieto- Lugilde et al. 2018). In this context, TSADB is especially relevant, because it provides information on ecological guilds at the genus level, which is generally considered sufficient taxonomic resolution for inferring the dispersal abilities of the taxa (Rimet & Bouchez 2011).

Can diatoms be used to indicate and reconstruct the multivariate environment of tropical aquatic ecosystems?

The TSADB provides modern training-sets that encompass wide gradients of water chemistry, climate, and landscape conditions across tropical and sub-tropical South America. Specifically, the TSADB has a large number of sediment surface samples with consistent coverage of in-lake variables (e.g. pH, conductivity, water temperature), which would made it possible to obtain reliable reconstructions based on the sensitivity of a large number of species to such variables (Benito et al. 2018a, 2018b). The large multivariate environmental context provided by TSADB could increase the model accuracy of transfer functions by factoring out, as much as possible, the influence of secondary variables in the training set or of autocorrelation between spatial factors and the proximal variable of interest (e.g., water temperature and elevation) (Rivera-Rondón & Catalan 2020). However, the intricate statistical correlative relationships between variables are often very difficult to tease apart in some calibration datasets, hence complementary methodological approaches are required (Juggins 2013). In combination with individual species' optima and the range of tolerances to several environmental variables, the qualitative approach of indicator species analysis (Dufrene & Legendre 1997) could better reflect the specific habitat conditions between and among groups of lakes in the TSADB.

Which regions have the best potential to act as a long-term observatory of limnological and biodiversity change?

A current challenge in aquatic ecology is to properly appraise limnological and biodiversity change at temporal scales before the onset of major human impacts. Escobar et al. (2020), show that palaeolimnology is strongly biased towards temperate and high-latitude regions.

Of 1278 articles reviewed in *Journal of Paleolimnology* between 1997 and 2015 with geographical coordinates, 118 (9.2%) were focused on tropical regions, and of these, only 10 included diatom paleolimnological time-series from tropical South America (Fig. 1). A query in the Neotoma paleoecological database yielded an additional nine records, and diatom palaeolimnology papers in *Diatom Research* are restricted to three studies for the period 1997-2021 (Vázquez & Caballero 2013, Weide et al. 2017, Ruwer et al. 2021). Despite these significant gaps in coverage from the South American continent, the TSADB shows that the Peruvian-Bolivian Altiplano is the region with a larger overlap between modern and palaeolimnological diatom observations (Fig. 1). This region therefore has the potential to act as a long-term observatory of tropical limnological and biodiversity change. It remains a challenge to merge datasets in a single time series with a common temporal scale and harmonized taxonomy.

The database could be expanded to identify sites and provide background information on the extent of the interaction of climate change and anthropogenic stressors (e.g. fish stocking, land use changes, extractive mining) affecting this portion of the South American continent (Aguilera et al. 2006, Heine-Fuster et al. 2021). The central Andean Altiplano also is associated with the onset and legacy of pre-European occupation affecting beta diversity patterns of diatoms (Goldberg et al. 2016, Benito et al. 2020). Although few diatom sedimentary sequences are publicly available in the above analyzed sources (Fig. 1; Table S2), many palaeolimnological studies could be available in other journals and the "grey literature", particularly in Ecuador, which has historically been better studied than most adjoining tropical regions (Seddon et al. 2014, Michelutti et al. 2015a, 2015b, Van Colen et al. 2017, Benito et al. 2019, Steinitz-Kannan et al. 2020).

How can the TSADB facilitate synthesis research with other biological databases?

The TSADB is a live database: it will be maintained and upgraded to enlarge its geographical coverage with new diatom sites across physical and chemical gradients of aquatic ecosystems (*see* database contributor guide in the Supplementary Material). Importantly, the TSADB is a current constituent database of the Neotoma paleoecology database (Goring et al. 2015). The level of regional data sharing is key to guide large-scale projects before archiving datasets in online platforms (Flantua et al. 2015). Anyone interested in sharing diatom data is welcome to use the contributor guide as a platform before archiving data in Neotoma to make sure their diatom counts, and taxonomy meet the standards of data reusability and reproducibility for the larger diatom community. The TSADB will not only reduce data loss through time, it also will provide a shared environment for comparing and displaying data for diatom researchers, without the need to understand the functionality of database cyberinfrastructure. In the longer term, the TSADB has the potential to contribute to long-term biodiversity time-series through spatially replicated diatom community time-series, such as BioTIME (Dornelas et al. 2018), metacommunity databases, such as CESTES (Jeliazkov et al. 2020) or DISPERSE (Sarremejane et al. 2020), or palaeoclimate databases (e.g. Linked PaleoData; McKay & Emile-Geay 2016). Finally, the TSADB aims to be a future Pan-tropical reference database and hence extend its potential to address synthetic work on ecology, macroecology, taxonomy and biodiversity.

Conclusions

The TSADB facilitates exploration, visualization, and analysis of diatom community data and associated local and regional environmental factors. It synthesizes diatom inventory

data of varied ecosystems, elevations, climates and physiography. The database reflects the spatial distribution of diatom sites by providing publicly available modern diatom samples and reproducible tools for easy visualization and manipulation. As freshwater biodiversity declines in tropical regions, collating, harmonizing, and sharing existing datasets is fundamental to foster new interdisciplinary collaborations. In the case of diatoms, the TSADB can provide an additional freshwater biodiversity component (microorganisms) for generating new, interdisciplinary, large-scale hypotheses on latitude-altitude diversity gradients, functional biogeography, and long-term ecological change across tropical regions.

Data availability statement

All diatom and environmental datasets, and R scripts to perform analyses and visualization of the TSADB are available from GitHub (https://github.com/xbenitogranell/diatoms-biogeography-southamerica). In addition, the TSADB has been stored in a zip folder at Dryad (https://doi.org/10.5061/dryad.jm63xsjbs) and Zenodo (https://doi.org/10.5281/zenodo.5415446). *Note*: GitHub repository represents the dynamic portal to access the TSADB, while Dryad and Zenodo are permanent repositories.

References

AGUILERA X., DECLERCK S., DE MEESTER L., MALDONADO M. & OLLEVIER F. 2006. Tropical high Andes lakes: A limnological survey and an assessment of exotic rainbow trout (*Oncorhynchus mykiss*). *Limnologica-Ecology and Management of Inland Waters* 36: 258–268.

ANTONELLI A., ARIZA M., ALBERT J., ANDERMANN T., AZEVEDO J., BACON C., FAURBY S., GUEDES T., HOORN C., LOHMANN L. G., MATOS-MARAVÍ P., RITTER C. D., SANMARTÍN I., SILVESTRO D., TEJEDOR M., STEEGE H. TER, TUOMISTO H., WERNECK F. P., ZIZKA A. & EDWARDS S. V. 2018. Conceptual and empirical advances in Neotropical biodiversity research. *PeerJ* 6: e5644.

BATTARBEE R. W., JUGGINS S., GASSE F., ANDERSON N. J., BENNION H., CAMERON N. G., RYVES D. B., PAILLES C., CHALIE F. & TELFORD R. 2001. European Diatom Database (EDDI). An information system for palaeoenvironmental reconstruction. ECRC Report 81. Available at: http://craticula.ncl.ac.uk/Eddi/jsp/index.jsp (accessed November 2021).

BATTARBEE R. W., MORLEY D., BENNION H., SIMPSON G. L., HUGHES M. & BAUERE V. 2011. A palaeolimnological meta-database for assessing the ecological status of lakes. *Journal of Paleolimnology* 45: 405–414.

BENITO X., FEITL M., FRITZ S. C., MOSQUERA P. V., SCHNEIDER T., HAMPEL H., QUEVEDO L. & STEINITZ-KANNAN M. 2019. Identifying temporal and spatial patterns of diatom community change in the tropical Andes over the last ~150 years. *Journal of Biogeography* 46: 1889–1900.

BENITO X. & FRITZ S. C. 2020. Diatom Diversity and Biogeography Across Tropical South America. In: *Neotropical Diversification: Patterns and Processes* (Ed. by V. Rull & A. C. Carnaval), pp. 121–143. Springer International Publishing, Cham.

BENITO X., FRITZ S. C., STEINITZ-KANNAN M., TAPIA P., KELLY M. A. & LOWELL T. V. 2018a. Geo-climatic factors drive diatom community distribution in tropical South American freshwaters. *Journal of Ecology* 106: 1660–1672.

BENITO X., FRITZ S. C., STEINITZ- KANNAN M., VÉLEZ M. I. & MCGLUE M. M. 2018b. Lake regionalization and diatom metacommunity structuring in tropical South America. *Ecology and Evolution* 8: 7865–7878.

BENITO X., VILMI A., LUETHJE M., CARREVEDO M. L., LINDHOLM M. & FRITZ S. C. 2020. Spatial and Temporal Ecological Uniqueness of Andean Diatom Communities Are Correlated With Climate, Geodiversity and Long-Term Limnological Change. *Frontiers in Ecology and Evolution* 8: 1-14.

BRACHT-FLYR B. & FRITZ S. C. 2012. Synchronous climatic change inferred from diatom records in four western Montana lakes in the US Rocky Mountains. *Quaternary research* 77: 456–467.

CHERUVELIL K. S. & SORANNO P. A. 2018. Data-intensive ecological research is catalyzed by open science and team science. *BioScience* 68: 813–822.

DONG X., LI B., HE F., GU Y., SUN M., ZHANG H. & TAN L. 2016. Flow directionality, mountain barriers and functional traits determine diatom metacommunity structuring of high mountain streams. *Sci. Rep.* 6: 24711.

DORNELAS M., ANTÃO L. H., MOYES F., BATES A. E., MAGURRAN A. E., ADAM D., AKHMETZHANOVA A. A., APPELTANS W., ARCOS J. M., ARNOLD H., AYYAPPAN N., BADIHI G., BAIRD A. H., BARBOSA M., BARRETO T. E., BÄSSLER C., BELLGROVE A., BELMAKER J., BENEDETTI- CECCHI L., BETT B. J., BJORKMAN A. D., BŁAŻEWICZ M., BLOWES S. A., BLOCH C. P., BONEBRAKE T. C., BOYD S., BRADFORD M., BROOKS A. J., BROWN J. H., BRUELHEIDE H., BUDY P., CARVALHO F., CASTAÑEDA-MOYA E., CHEN C. A., CHAMBLEE J. F., CHASE T. J., COLLIER L. S., COLLINGE S. K., CONDIT R., COOPER E. J., CORNELISSEN J. H. C., COTANO U., CROW S. K., DAMASCENO G., DAVIES C. H., DAVIS R. A., DAY F. P., DEGRAER S., DOHERTY T. S., DUNN T. E., DURIGAN G., DUFFY J. E., EDELIST D., EDGAR G. J., ELAHI R., ELMENDORF S. C., ENEMAR A., ERNEST S. K. M., ESCRIBANO R., ESTIARTE M., EVANS B. S., FAN T.-Y., FARAH F. T., FERNANDES L. L., FARNEDA F. Z., FIDELIS A., FITT R., FOSAA A. M., FRANCO G. A. D. C., FRANK G. E., FRASER W. R., GARCÍA H., GATTI R. C., GIVAN O., GORGONE- BARBOSA E., GOULD W. A., GRIES C., GROSSMAN G. D., GUTIERRÉZ J. R., HALE S., HARMON M. E., HARTE J., HASKINS G., HENSHAW D. L., HERMANUTZ L., HIDALGO P., HIGUCHI P., HOEY A., HOEY G. V., HOFGAARD A., HOLECK K., HOLLISTER R. D., HOLMES R., HOOGENBOOM M., HSIEH C., HUBBELL S. P., HUETTMANN F., HUFFARD C. L., HURLBERT A. H., IVANAUSKAS N. M., JANÍK D., JANDT U., JAŻDŻEWSKA A., JOHANNESSEN T., JOHNSTONE J., JONES J., JONES F. A. M., KANG J., KARTAWIJAYA T., KEELEY E. C., KELT D. A., KINNEAR R., KLANDERUD K., KNUTSEN H., KOENIG C. C., KORTZ A. R., KRÁL K., KUHNZ L. A., KUO C.-Y., KUSHNER D. J., LAGUIONIE- MARCHAIS C., LANCASTER L. T., LEE C. M., LEFCHECK J. S., LÉVESQUE E., LIGHTFOOT D., LLORET F., LLOYD J. D., LÓPEZ- BAUCELLS A., LOUZAO M., MADIN J. S., MAGNÚSSON B., MALAMUD S., MATTHEWS I., MCFARLAND K. P., MCGILL B., MCKNIGHT D., MCLARNEY W. O., MEADOR J., MESERVE P. L., METCALFE D. J., MEYER C. F. J., MICHELSEN A., MILCHAKOVA N., MOENS T., MOLAND E., MOORE J., MOREIRA C. M., MÜLLER J., MURPHY G., MYERS- SMITH I. H., MYSTER R. W., NAUMOV A., NEAT F., NELSON J. A., NELSON M. P., NEWTON S. F., NORDEN N., OLIVER J. C., OLSEN E. M., ONIPCHENKO V. G., PABIS K., PABST R. J., PAQUETTE A., PARDEDE S., PATERSON D. M., PÉLISSIER R., PEÑUELAS J., PÉREZ- MATUS A., PIZARRO O., POMATI F., POST E., PRINS H. H. T., PRISCU J. C., PROVOOST P., PRUDIC K. L., PULLIAINEN E., RAMESH B. R., RAMOS O. M., RASSWEILER A., REBELO J. E., REED D. C., REICH P. B., REMILLARD S. M., RICHARDSON A. J., RICHARDSON J. P., RIJN I. VAN, ROCHA R., RIVERA- MONROY V. H., RIXEN C., ROBINSON K. P., RODRIGUES R. R., ROSSA- FERES D. DE C., RUDSTAM L., RUHL H., RUZ C. S., SAMPAIO E. M., RYBICKI N., RYPEL A., SAL S., SALGADO B., SANTOS F. A. M., SAVASSI- COUTINHO A. P., SCANGA S., SCHMIDT J., SCHOOLEY R., SETIAWAN F., SHAO K.-T., SHAVER G. R., SHERMAN S., SHERRY T. W., SICIŃSKI J., SIEVERS C., SILVA A. C. DA, SILVA F. R. DA, SILVEIRA F. L., SLINGSBY J., SMART T., SNELL S. J., SOUDZILOVSKAIA N. A., SOUZA G. B. G., SOUZA F. M., SOUZA V. C., STALLINGS C. D., STANFORTH R., STANLEY E. H., STERZA J. M., STEVENS M., STUART- SMITH R., SUAREZ Y. R., SUPP S., TAMASHIRO J. Y., TARIGAN S., THIEDE G. P., THORN S., TOLVANEN A., TONIATO M. T. Z., TOTLAND Ø., TWILLEY R. R., VAITKUS G., VALDIVIA N., VALLEJO M. I., VALONE T. J., COLEN C. V., VANAVERBEKE J., VENTUROLI F., VERHEYE H. M., VIANNA M., VIEIRA R. P., VRŠKA T., VU C. Q., VU L. V., WAIDE R. B., WALDOCK C., WATTS D., WEBB S., WESOŁOWSKI T., WHITE E. P., WIDDICOMBE C. E., WILGERS D., WILLIAMS R., WILLIAMS S. B., WILLIAMSON M., WILLIG M. R., WILLIS T. J., WIPF S., WOODS K. D., WOEHLER E. J., ZAWADA K. & ZETTLER M. L. 2018. BioTIME: A database of biodiversity time series for the Anthropocene. *Global Ecology and Biogeography* 27: 760–786.

DUFRENE M. & LEGENDRE P. 1997. Species assemblages and indicator species: the need for a flexible asymmetrical approach. *Ecological monographs* 67: 345–366.

EKDAHL E. J., FRITZ S. C., BAKER P. A., RIGSBY C. A. & COLEY K. 2008. Holocene multidecadal-to millennial-scale hydrologic variability on the South American Altiplano. *The Holocene* 18: 867–876.

ESCOBAR J., SERNA Y., HOYOS N., VELEZ M. I. & CORREA-METRIO A. 2020. Why we need more paleolimnology studies in the tropics. *Journal of Paleolimnology* 64: 47–53.

FEITL M., KERN A. K., JONES A., FRITZ S. C., BAKER P. A., JOECKEL R. M., SALENBIEN W. & WILLARD D. 2019. Paleoclimate of the subtropical Andes during the latest Miocene, Lauca Basin, Chile. *Palaeogeography, Palaeoclimatology, Palaeoecology* 534: 109336.

FLANTUA S. G., HOOGHIEMSTRA H., GRIMM E. C., BEHLING H., BUSH M. B., GONZÁLEZ-ARANGO C., GOSLING W. D., LEDRU M.-P., LOZANO-GARCÍA S. & MALDONADO A. 2015. Updated site compilation of the Latin American Pollen Database. *Review of Palaeobotany and Palynology* 223: 104–115.

FRENGUELLI J. 1939. Diatomeas del lago Titicaca. *Notas del Museo de la Plata* 4: 175–198.

FRITZ S. C., BAKER P. A., TAPIA P., SPANBAUER T. & WESTOVER K. 2012. Evolution of the Lake Titicaca basin and its diatom flora over the last ~370,000 years. *Palaeogeography, Palaeoclimatology, Palaeoecology* 317–318: 93–103.

GILES M. P., MICHELUTTI N., GROOMS C. & SMOL J. P. 2018. Long-term limnological changes in the Ecuadorian páramo: Comparing the ecological responses to climate warming of shallow waterbodies versus deep lakes. *Freshwater Biology* 63: 1316–1325.

GOLDBERG A., MYCHAJLIW A. M. & HADLY E. A. 2016. Post-invasion demography of prehistoric humans in South America. *Nature* 532: 232-235.

GORING S., DAWSON A., SIMPSON G., RAM K., GRAHAM R., GRIMM E. & WILLIAMS J. 2015. neotoma: A Programmatic Interface to the Neotoma Paleoecological Database. *Open Quaternary* 1(1): 1-17.

GRIMM A. M. & AMBRIZZI T. 2009. Teleconnections into South America from the tropics and extratropics on interannual and intraseasonal timescales. In*: Past Climate Variability in South America and Surrounding Regions: From the Last Glacial Maximum to the Holocene, Developments in Paleoenvironmental research* (Ed. by F Vimeux, F Sylvestre & M Khodri), pp 159-191. Springer: Netherlands.

GUIRY M. D. & GUIRY G. M. 2017. *AlgaeBase. World-wide electronic publication, National University of Ireland, Galway*.

HEINE-FUSTER I., LÓPEZ-ALLENDES C., ARÁNGUIZ-ACUŇA A. & VÉLIZ D. 2021. Differentiation of Diatom Guilds in Extreme Environments in the Andean Altiplano. *Frontiers in Environmental Science* 9: 1-12..

JAMONEAU A., PASSY S. I., SOININEN J., LEBOUCHER T. & TISON-ROSEBERY J. 2018. Beta diversity of diatom species and ecological guilds: Response to environmental and spatial mechanisms along the stream watercourse. *Freshwater Biology* 63: 62–73.

JELIAZKOV A., MIJATOVIC D., CHANTEPIE S., ANDREW N., ARLETTAZ R., BARBARO L., BARSOUM N., BARTONOVA A., BELSKAYA E., BONADA N., BRIND'AMOUR A., CARVALHO R., CASTRO H., CHMURA D., CHOLER P., CHONG-SENG K., CLEARY D., CORMONT A., CORNWELL W., CAMPOS R. DE, VOOGD N. DE, DOLEDEC S., DREW J., DZIOCK F., EALLONARDO A., EDGAR M. J., FARNEDA F., HERNANDEZ D. F., FRENETTE-DUSSAULT C., FRIED G., GALLARDO B., GIBB H., GONÇALVES-SOUZA T., HIGUTI J., HUMBERT J.-Y., KRASNOV B. R., SAUX E. L., LINDO Z., LOPEZ-BAUCELLS A., LOWE E., MARTEINSDOTTIR B., MARTENS K., MEFFERT P., MELLADO-DÍAZ A., MENZ M. H. M., MEYER C. F. J., MIRANDA J. R., MOUILLOT D., OSSOLA A., PAKEMAN R., PAVOINE S., PEKIN B., PINO J., POCHEVILLE A., POMATI F., POSCHLOD P., PRENTICE H. C., PURSCHKE O., RAEVEL V., REITALU T., RENEMA W., RIBERA I., ROBINSON N., ROBROEK B., ROCHA R., SHIEH S.-H., SPAKE R., STANIASZEK-KIK M., STANKO M., TEJERINA-GARRO F. L., BRAAK C. TER, URBAN M. C., KLINK R. VAN, VILLÉGER S., WEGMAN R., WESTGATE M. J., WOLFF J., ŻARNOWIEC J., ZOLOTAREV M. & CHASE J. M. 2020. A global database for metacommunity ecology, integrating species, traits, environment and space. *Scientific Data* 7: 1–15.

JUGGINS S. 2013. Quantitative reconstructions in palaeolimnology: new paradigm or sick science?. *Quaternary Science Reviews* 64: 20–32.

LECOINTE C., COSTE M. & PRYGIEL J. 1993. "Omnidia": software for taxonomy, calculation of diatom indices and inventories management. *Hydrobiologia* 269: 509–513.

LEHNER B., VERDIN K. & JARVIS A. 2008. New global hydrography derived from spaceborne elevation data. *Eos, Transactions American Geophysical Union* 89: 93–94.

MACCOY D. 2011. BioData: A National Aquatic Bioassessment Database. US Geological Survey.

MANGUIN E. 1964. Contribution à la connaissance des diatomées des Andes du Pérou. *Mémoires du Muséum national d'Histoire naturelle, Sér. B – Botanique* 12: 41–98.

MANN D. G. 1999. The species concept in diatoms. *Phycologia* 38: 437–495.

MCGOWAN S. & LEAVITT P. R. 2009. The role of paleoecology in whole-ecosystem science. In: *Real World Ecology* (Ed. by M Shili, S Carstenn & M Nungesser), pp. 161–208. Springer: Netherlands.

MESSAGER M. L., LEHNER B., GRILL G., NEDEVA I. & SCHMITT O. 2016. Estimating the volume and age of water stored in global lakes using a geo-statistical approach. *Nature Communications* 7: 13603.

METZELTIN D. & LANGE-BERTALOT H. 1998. *Tropical diatoms of South America I.* Iconographia Diatomologica 5, A.R.G. Gantner Verlag K.G., Königstein.

METZELTIN D. & LANGE-BERTALOT H. 2007. *Tropical diatoms of South America II*. In: H. Lange-Bertalot (ed) Iconographia Diatomologica 18, A.R.G. Gantner Verlag K.G., Königstein.

METZELTIN D., LANGE-BERTALOT H. & GARCIA-RODRIGUEZ F. 2005. Diatoms of Uruguay. *Iconographia diatomologica* 15: 1–736.

MICHELUTTI N., COOKE C. A., HOBBS W. O. & SMOL J. P. 2015a. Climate-driven changes in lakes from the Peruvian Andes. *Journal of Paleolimnology* 54: 153–160.

MICHELUTTI N., WOLFE A., COOKE C., HOBBS W. & VUILLE M. 2015b. Climate change forces new ecological states in tropical Andean lakes. *PloS one* 10: e0115338.

MORALES E. & VIS M. 2007. Epilithic diatoms (Bacillariophyceae) from cloud forest and alpine streams in Bolivia, South America. *Proceedings of the Academy of Natural Sciences of* 156: 123–155.

NASCIMENTO M. N., LAURENZI A. G., VALENCIA B. G., VAN R. & BUSH M. 2019. A 12,700-year history of paleolimnological change from an Andean microrefugium. *The Holocene* 29: 231–243.

NIETO- LUGILDE D., MAGUIRE K. C., BLOIS J. L., WILLIAMS J. W. & FITZPATRICK M. C. 2018. Multiresponse algorithms for community-level modelling: Review of theory, applications, and comparison to species distribution models. *Methods in Ecology and Evolution* 9: 834–848.

OLIVEIRA P. D. & STEINITZ-KANNAN M. 1992. The diatom flora (Bacillariophyceae) of the Cuyabeno Faunistic Reserve, Ecuadorian Amazonia. *Nova Hedwigia* 54: 515–552.

PATIL V., SELTMANN T., SALMASO N., ANNEVILLE O., LAJEUNESSE M. & STRAILE D. 2019. *algaeClassify: Determine Phytoplankton Functional Groups Based on Functional Traits*. R package version 1.2.0. https://CRAN.R-project.org/package=algaeClassify

PATIÑO L., VELEZ M. I., WEBER M., VELÁSQUEZ- R C. A., DAVID S., RUEDA M., CASTAÑEDA I. & ARBOLEDA D. 2020. Late Pleistocene–Holocene environmental and climatic history of a freshwater paramo ecosystem in the northern Andes. *Journal of Quaternary Science* 35: 1046–1056.

PIENITZ R. & COURNOYER L. 2017. Circumpolar Diatom Database (CDD): a new database for use in paleolimnology and limnology. *Journal of Paleolimnology* 57: 213–219.

RICKLEFS R. E. & JENKINS D. G. 2011. Biogeography and ecology: towards the integration of two disciplines. *Philosophical Transactions of the Royal Society of London B: Biological Sciences* 366: 2438–2448.

RIMET F. & BOUCHEZ A. 2011. Use of diatom life-forms and ecological guilds to assess pesticide contamination in rivers: lotic mesocosm approaches. *Ecological Indicators* 11: 489–499.

RIVERA-RONDÓN C. A. & CATALAN J. 2020. Diatoms as indicators of the multivariate environment of mountain lakes. *Science of The Total Environment* 703: 135517.

RUMRICH U., LANGE-BERTALOT H. & RUMRICH M. 2000. *Diatoms of the Andes, from Venezuela to Patagonia/Tierra del Fuego, and two additional contributions*. Iconographia diatomologica 9 ARG Gartner Verlag KG, Königstein.

RUWER D. T., OSÓRIO N. C., COSTA-BÖDDEKER S. & RODRIGUES L. 2021. Influence of temporal changes on a diatom species in a floodplain lake: paleolimnological and biomonitoring approach. *Diatom Research* 36: 1–9. Taylor & Francis.

SARREMEJANE R., CID N., STUBBINGTON R., DATRY T., ALP M., CAÑEDO-ARGÜELLES M., CORDERO-RIVERA A., CSABAI Z., GUTIÉRREZ-CÁNOVAS C., HEINO J., FORCELLINI M., MILLÁN A., PAILLEX A., PAŘIL P., POLÁŠEK M., TIERNO DE FIGUEROA J. M., USSEGLIO-POLATERA P., ZAMORA-MUÑOZ C. & BONADA N. 2020. DISPERSE, a trait database to assess the dispersal potential of European aquatic macroinvertebrates. *Scientific Data* 7: 386.

SEDDON A. W., FROYD C. A., WITKOWSKI A. & WILLIS K. J. 2014. A quantitative framework for analysis of regime shifts in a Galápagos coastal lagoon. *Ecology* 95: 3046–3055.

SERVANT-VILDARY S. 1986. Les diatomées actuelles des Andes de Bolivie (Taxonomie, écologie). *Cahiers de Micropaléontologie* 1: 99–124.

SOININEN J. & TEITTINEN A. 2019. Fifteen important questions in the spatial ecology of diatoms. *Freshwater Biology* 64: 2071–2083.

SPANBAUER T. L., FRITZ S. C. & BAKER P. A. 2018. Punctuated changes in the morphology of an endemic diatom from Lake Titicaca. *Paleobiology* 44: 89–100.

STEINITZ-KANNAN M. 1979. Comparative Limnology of Ecuadorian Lakes: A study of species number and composition of plankton communities of the Galapagos Islands and Equatorial Andes. PhD Dissertation. The Ohio State University.

STEINITZ-KANNAN M., LÓPEZ C., JACOBSEN D. & GUERRA M. DE L. 2020. History of limnology in Ecuador: a foundation for a growing field in the country. *Hydrobiologia* 847: 4191–4206.

STEINITZ-KANNAN M., NIENABER M. A. & RIEDINGER-WHITMORE M. A. 1993. The fossil diatoms of Lake Yambo, Ecuador: a 2500 year record of intense El Niño events.. *Bull. Inst. fr. etudes andines* 22: 227–241.

SYLVESTRE F., SERVANT-VILDARY S. & ROUX M. 2001. Diatom-based ionic concentration and salinity models from the south Bolivian Altiplano (15-23S). *Journal of Paleolimnology* 25: 279–295.

SYRYKH L., SUBETTO D. & NAZAROVA L. 2021. Paleolimnological studies on the East European Plain and nearby regions: the PaleoLake Database. *Journal of Paleolimnology* 65: 369–375.

TAPIA P. M., FRITZ S. C., SELTZER G. O., RODBELL D. T. & METIEVER S. 2006. Contemporary distribution and late-quaternary stratigraphy of diatoms in the Junin Plain, central Andes, Peru. *Bol. Soc. Geol. Peru* 101: 19–42.

TRISOS C. H., MEROW C. & PIGOT A. L. 2020. The projected timing of abrupt ecological disruption from climate change. *Nature* 580: 496–501.

US EPA 2009. National Lakes Assessment: a collaborative survey of the Nation's lakes. EPA 841-R-09-001. Office of Water and Office of Research and Development, US Environmental Protection Agency.

VALENCIA B. G., MATTHEWS-BIRD F., URREGO D. H., WILLIAMS J. J., GOSLING W. D. & BUSH M. 2016. Andean microrefugia: Testing the Holocene to predict the Anthropocene. *New Phytologist* 212: 510–522.

VAN COLEN W. R., MOSQUERA P., VANDERSTUKKEN M., GOIRIS K., CARRASCO M.-C., DECAESTECKER E., ALONSO M., LEÓN-TAMARIZ F. & MUYLAERT K. 2017. Limnology and trophic status of glacial lakes in the tropical Andes (Cajas National Park, Ecuador). *Freshwater biology* 62: 458–473.

VÁZQUEZ G. & CABALLERO M. 2013. The structure and species composition of the diatom communities in tropical volcanic lakes of eastern Mexico. *Diatom Research* 28: 77–91.

VÉLEZ M. I., JARAMILLO C., SALAZAR-RÍOS A., BENITO X., FRITZ S., TAPIA P., LUBINIECKI D., KAR N. & ESCOBAR J. 2018. Aquatic ecosystems in a newly formed ecospace: Early Pliocene lakes in the Central Andean Altiplano. *Palaeogeography, Palaeoclimatology, Palaeoecology* 490: 218–226.

VENTER O., SANDERSON E. W., MAGRACH A., ALLAN J. R., BEHER J., JONES K. R., POSSINGHAM H. P., LAURANCE W. F., WOOD P. & FEKETE B. M. 2016. Global terrestrial Human Footprint maps for 1993 and 2009. *Scientific data* 3: 160067.

VILMI A., KARJALAINEN S. M., HELLSTEN S. & HEINO J. 2016. Bioassessment in a metacommunity context: Are diatom communities structured solely by species sorting? *Ecological Indicators* 62: 86–94.

VYVERMAN W., VERLEYEN E., SABBE K., VANHOUTTE K., STERKEN M., HODGSON D. A., MANN D. G., JUGGINS S., VAN DE VIJVER B., JONES V., FLOWER R., ROBERTS D., CHEPURNOV V. A., KILROY C., VANORMELINGEN P. & WEVER A. D. 2007. Historical processes constrain patterns in global diatom diversity. *Ecology* 88: 1924–1931.

WEIDE D. M., FRITZ S. C., BRINSON B. E., THOMPSON L. G. & BILLUPS W. E. 2017. Freshwater diatoms in the Sajama, Quelccaya, and Coropuna glaciers of the South American Andes. *Diatom Research* 32: 153–162.

WETZEL C. E., DE BICUDO D. C., ECTOR L., LOBO E. A., SOININEN J., LANDEIRO V. L. & BINI L. M. 2012. Distance Decay of Similarity in Neotropical Diatom Communities. *PLoS One* 7: 10–11.

WICKHAM H., AVERICK M., BRYAN J., CHANG W., MCGOWAN L. D., FRANÇOIS R., GROLEMUND G., HAYES A., HENRY L. & HESTER J. 2019. Welcome to the Tidyverse. *Journal of open source software* 4: 1686.

WILKE C. O., WICKHAM H. & WILKE M. C. O. 2019. cowplot: Streamlined Plot Theme and Plot Annotations for 'ggplot2'. R package version 1.1.1. https://CRAN.R-project.org/package=cowplot

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., NICHOLS J., PURDUM T., ROTH R. E., STRYKER M. & TAKAHARA H. 2018. The Neotoma Paleoecology Database, a multiproxy, international, community-curated data resource. *Quaternary Research* 89: 156–177.

WINSTON C., CHENG J., ALLAIRE J., SIEVERT C., SCHLOERKE B., XIE Y., ALLEN J. & MCPHERSON J. 2021. shiny: Web Application Framework for R. package version 1.6.0. https://CRAN.R-project.org/package=shiny

Figure legends

Fig. 1. Diatom sites coloured by habitat in the Tropical South America Diatom Database (TSADB). Diatom palaeolimnological sites from Escobar et al. (2020) and Neotoma database are shown in red. See Table S2 for the complete list of diatom-based palaeolimnological records.

Fig. 2. Distribution of diatom sample types, habitats, and year of sampling in the TSADB.

Fig. 3. Species number of diatom genus (having more than 2 species) identified in the TSADB. Non-identified diatoms are taxa tentatively designated as *cf. aff.* or *sp.* in the database.

Fig. 4. Most common (>40% relative abundance in at least one sample) diatom taxa across regions, classified by ecological groups. Note that regions with presence/absence data only (e.g. Galapagos) are not shown.

Fig. 5. Distributional patterns of diatom species richness by latitude and elevation of the study regions in the TSADB.