Advancing Knowledge of Arctic Lake System Dynamics: A Data-Driven Perspective on Spatiotemporal Patterns

---

Gregor Pfalz

Univ.-Diss.

zur Erlangung des akademischen Grades
“doctor rerum naturalium”
(Dr. rer. nat.)
in der Wissenschaftsdisziplin „Paläoklimatologie“

eingereicht in Form einer kumulativen Arbeit an der
Mathematisch-Naturwissenschaftlichen Fakultät
Institut für Geowissenschaften
der Universität Potsdam
und
angefertigt am
Alfred-Wegener-Institut – Helmholtz Zentrum
für Polar- und Meeresforschung

Potsdam, November 2023

Ort und Tag der Disputation: Potsdam, 29.04.2024

Hauptbetreuer*in: apl. Prof. Dr. Bernhard Diekmann
Betreuer*innen: Prof. Johann-Christoph Freytag, PhD, Dr. Boris K. Biskaborn
Gutachter*innen: Prof. Dr. Elisabeth Dietze, Prof. Dr. John P. Smol

Published online on the
Publication Server of the University of Potsdam:
https://doi.org/10.25932/publishup-63655
https://nbn-resolving.org/urn:nbn:de:kobv:517-opus4-636554

Abstract

Ecosystems play a pivotal role in addressing climate change but are also highly susceptible to drastic environmental changes. Investigating their historical dynamics can enhance our understanding of how they might respond to unprecedented future environmental shifts. With Arctic lakes currently under substantial pressure from climate change, lessons from the past can guide our understanding of potential disruptions to these lakes. However, individual lake systems are multifaceted and complex. Traditional isolated lake studies often fail to provide a global perspective because localized nuances—like individual lake parameters, catchment areas, and lake histories—can overshadow broader conclusions. In light of these complexities, a more nuanced approach is essential to analyze lake systems in a global context.

A key to addressing this challenge lies in the data-driven analysis of sedimentological records from various northern lake systems. This dissertation emphasizes lake systems in the northern Eurasian region, particularly in Russia (n=59). For this doctoral thesis, we collected sedimentological data from various sources, which required a standardized framework for further analysis. Therefore, we designed a conceptual model for integrating and standardizing heterogeneous multi-proxy data into a relational database management system (PostgreSQL). Creating a database from the collected data enabled comparative numerical analyses between spatially separated lakes as well as between different proxies.

When analyzing numerous lakes, establishing a common frame of reference was crucial. We achieved this by converting proxy values from depth dependency to age dependency. This required consistent age calculations across all lakes and proxies using one age-depth modeling software. Recognizing the broader implications and potential pitfalls of this, we developed the LANDO approach ("Linked Age and Depth Modelling"). LANDO is an innovative integration of multiple age-depth modeling software into a singular, cohesive platform (Jupyter Notebook). Beyond its ability to aggregate data from five renowned age-depth modeling software, LANDO uniquely empowers users to filter out implausible model outcomes using robust geoscientific data. Our method is not only novel but also significantly enhances the accuracy and reliability of lake analyses.

Considering the preceding steps, this doctoral thesis further examines the relationship between carbon in sediments and temperature over the last 21,000 years. Initially, we hypothesized a positive correlation between carbon accumulation in lakes and modelled paleotemperature. Our homogenized dataset from heterogeneous lakes confirmed this association, even if the highest temperatures throughout our observation period do not correlate with the highest carbon values. We assume that rapid warming events contribute more to high accumulation, while sustained warming leads to carbon outgassing. Considering the current high concentration of carbon in the atmosphere and rising temperatures, ongoing

climate change could cause northern lake systems to contribute to a further increase in atmospheric carbon (positive feedback loop). While our findings underscore the reliability of both our standardized data and the LANDO method, expanding our dataset might offer even greater assurance in our conclusions.

Zusammenfassung

Ökosysteme spielen eine zentrale Rolle bei der Bewältigung des Klimawandels, gelten jedoch auch als äußerst anfällig für drastische Umweltveränderungen. Die Erforschung ihrer historischen Dynamiken kann unser Verständnis darüber verbessern, wie sich zukünftige Veränderungen angesichts beispielloser Umweltveränderungen auf sie auswirken können. Angesichts des enormen Stresses, dem arktische Seen durch den Klimawandel ausgesetzt sind, können konkrete Fälle aus der Vergangenheit helfen, mögliche Schwankungen im Ökosystem des Sees besser zu verstehen und zu deuten. Einzelne Seesysteme unterliegen jedoch einer inhärenten Komplexität und vielschichtigen Beschaffenheit. Klassische Einzelanalysen von Seen liefern oft keine globale Perspektive, da lokale Besonderheiten – wie individuelle Seeparameter, Einzugsgebiete und Seehistorien – allgemeinere Schlussfolgerungen überlagern können. Eine differenzierte Herangehensweise ist hierbei erforderlich, um Seesysteme im globalen Kontext angemessen zu analysieren.

Ein Schlüssel zur Bewältigung dieser Herausforderung ist die datenwissenschaftliche Analyse von sedimentologischen Daten aus mehreren nördlichen Seesystemen. Diese Dissertation fokussiert sich dabei auf das Gebiet des nördlichen Eurasiens mit einem besonderen Fokus auf Seesystem in Russland (n=59). Die gesammelten sedimentologischen Daten für diese Doktorarbeit mussten hierfür zunächst standardisiert und homogenisiert werden. Hierfür wurde ein konzeptuelles Modell für die Integration und Standardisierung von heterogenen Multi-Proxy-Daten in ein relationales Datenbankverwaltungssystem (PostgreSQL) entworfen. Die Erstellung einer Datenbank aus der gesammelten Datenkollektion ermöglichte die numerische, vergleichende Analyse zwischen räumlich getrennten Seen als auch zwischen verschiedenen Proxys.

Eine Analyse von mehreren Seen erforderte zudem eine gemeinsame Analyseebene, welche wir durch die Umwandlung von einer Tiefenabhängigkeit zu Altersabhängigkeit der Proxywerte erreichten. Diese bedurfte aber, dass die zugehörigen Alter von Proxywerte von allen Seen mit demselben Verfahren einer Alterstiefenmodellsoftware berechnet werden müssen. Angesichts der weitreichenden Implikationen und potenziellen Fallstricke entwickelten wir den LANDO-Ansatz („Linked Age and Depth Modelling“). LANDO stellt eine innovative Integration mehrerer Alters-Tiefen-Modellierungssoftware in eine einheitliche, kohärente Plattform (Jupyter Notebook) dar. Neben seiner Fähigkeit, Daten von fünf renommierten Alters-Tiefen-Modellierungssoftware zu aggregieren, ermöglicht LANDO es den Nutzern auf einzigartige Weise, unbegründete Modellergebnisse anhand robuster geowissenschaftlicher Daten herauszufiltern. Unsere Methode ist nicht nur neuartig, sondern steigert auch signifikant die Genauigkeit und Zuverlässigkeit von Seeanalysen.

Schlussendlich unter Berücksichtigung der vorangegangenen Schritte betrachtet die Doktorarbeit den Zusammenhang zwischen Kohlenstoff in Sedimenten und Temperatur über die letzten

21 000 Jahre. Zunächst nehmen wir an, dass es eine positive Korrelation zwischen Kohlenstoffakkumulation in Seen und modellierter Paläo-Temperatur gibt. Diese kann dank des homogenisierten Datensatzes von heterogenen Seen bestätigt werden, wenn auch die höchsten Temperaturen über unseren Betrachtungszeitraum nicht korrelieren mit den höchsten Kohlenstoffwerten. Wir gehen davon aus, dass schnelle Erwärmungsereignisse eher zu einer hohen Akkumulation beitragen, während bestehende Erwärmung eher zu einer Ausgasung von Kohlenstoff führt. In Anbetracht der aktuellen hohen Konzentration von Kohlenstoff in der Atmosphäre und der steigenden Temperaturen, können bei einem weiterführenden Klimawandel nördliche Seesysteme zu einem weiteren Anstieg von atmosphärischem Kohlenstoff führen (positive Feedbackschleife). Obwohl die bemerkenswerten Ergebnisse zeigen, dass unser Ansatz aus standardisierten Daten und LANDO zuverlässig ist, könnte eine größere Datenmenge das Vertrauen in die Ergebnisse noch weiter stärken.

Table of contents

List of figures

List of tables

1

Introduction

1.1 Scientific background and motivation

1.1.1 Geoscientific perspective

Ecosystems worldwide suffer from the imminent threat of progressive, anthropogenic climate change. According to the sixth assessment report (AR6) of the Intergovernmental Panel on Climate Change (IPCC), it is likely that the global mean surface temperature will reach 1.5 °C above the 1850-1900 pre-industrial baseline by 2030. This projection draws from both significant modelling efforts and the observation that the global mean surface temperature was 1.1 °C above the pre-industrial baseline between 2011 and 2020 (IPCC 2023). The severity of ecosystem failure depends on the emissions pathway we choose in the coming years. However, it is crucial to align our efforts with the goals of the Paris Agreement, which aims to limit global warming well below 2 °C above pre-industrial levels (IPCC 2023). By adhering to this target, we can mitigate the impacts and reduce the likelihood of ecosystem failure.

But in a rapidly warming world, the likelihood of ecosystem collapse becomes highly probable, as evidenced by the current impact on northern lake systems (60° to 90° N) in the Arctic (Dakos et al. 2019). In a recent collaborative study, we have observed that even remote lakes, once considered pristine, like Bolshoe Toko, are showing signs of anthropogenic influence (Biskaborn et al. 2021). This emphasizes the crucial role lakes play as sentinels of environmental change. Around the world, lake systems help regulate local climates, provide habitat for various plant and animal species, and support people's livelihoods, well-being, and local economies (Zhang et al. 2017; Grant et al. 2021). However, rising temperatures and associated changes in precipitation patterns are changing established systems. This includes the loss of habitats and biodiversity resulting from shifts in species distribution and abundances within aquatic ecosystems (Weiskopf et al. 2020; Habibullah et al. 2022), as well as the exacerbation of water quality and quantity issues (Michalak 2016; Konapala et al. 2020; Melaku Melese 2016). Gaining insights into potential trajectories of lake dynamics is crucial for anticipating tipping points that could trigger irreversible changes in the lake ecosystem.

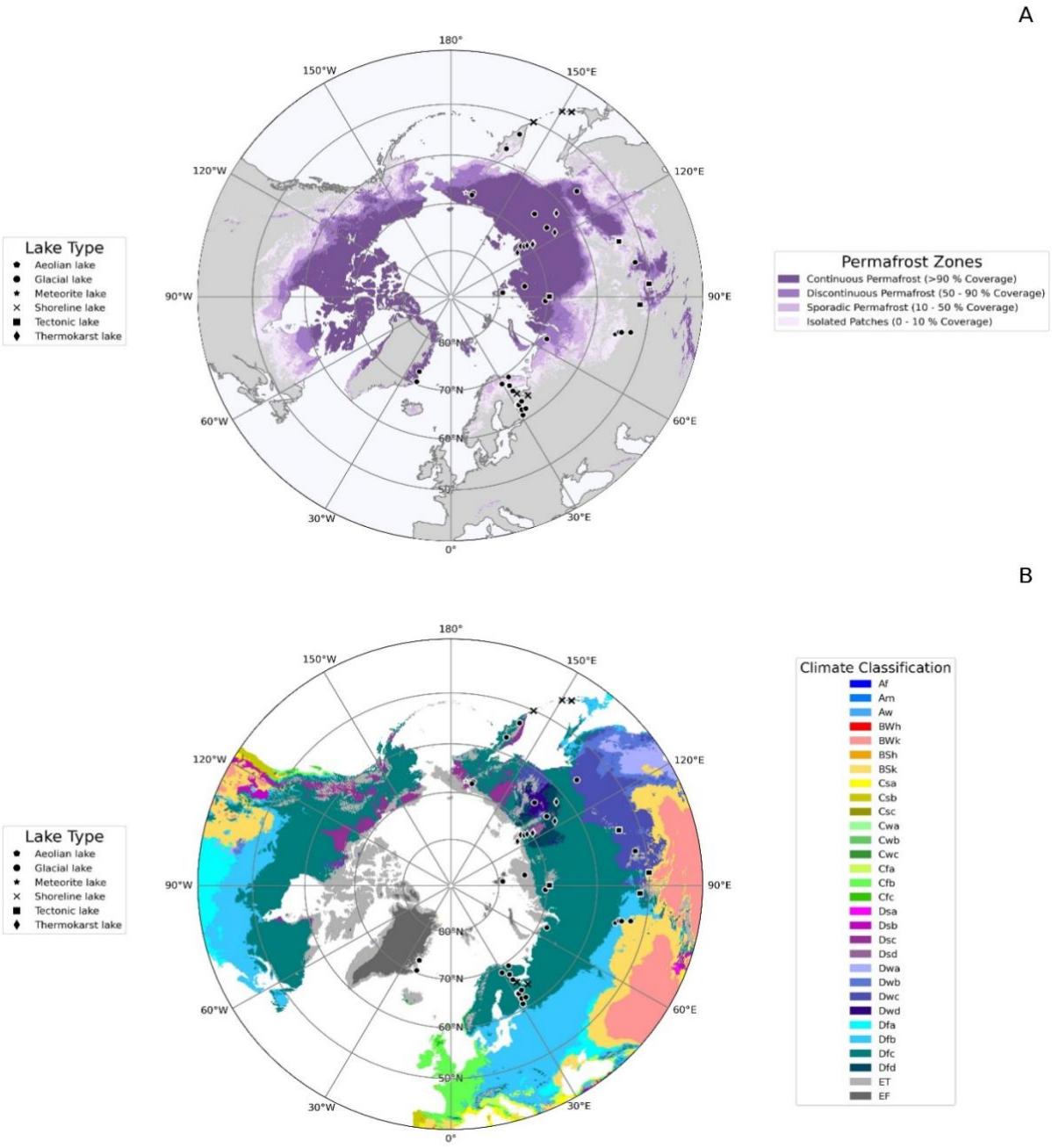

Figure 1.1: A) Permafrost zones according to Obu et al. (2019), including color scheme for each zone (four different shades of purple) to match the original publication. B) Köppen-Geiger climate classification according to Beck et al. (2018) for the study area. Both subplots include the spatial distribution of sediment cores used for data analysis in this dissertation labeled by their lake type (black symbols, n = 59). The outer ring of A and B corresponds to 40° N.

Furthermore, thawing permafrost serves as an additional tipping point, posing substantial threats to these vulnerable lake systems (Miner et al. 2022). Permafrost is hereby defined as perennially frozen ground material, i.e. soil, rock, or unconsolidated sediment (which includes organic material and ice), that remains at or below 0°C for at least two consecutive years (French 2007). This permafrost covers approximately 25% of the Earth's land surface and affects the majority of lakes in our data collection (Figure 1.1) (Olefeldt et al. 2016; French 2007).

Estimates of carbon stored within permafrost range from around 1460 to 1600 GtC (gigatonnes of carbon) (Schuur et al. 2022; Meredith et al. 2019; Hugelius et al. 2014; Miner et al. 2022), which makes it one of the biggest terrestrial carbon pools in the global carbon cycle (Friedlingstein, Jones, et al. 2022; Friedlingstein, O'sullivan, et al. 2022). Thawing permafrost releases carbon into the atmosphere either in the form of carbon dioxide (CO_2) or methane (CH_4), causing an additional warming effect ('permafrost carbon feedback') (Schuur et al. 2015; Miner et al. 2022).

Based on current estimations, a business-as-usual case could result in the release of 5 to 15 percent of the permafrost carbon pool, equivalent to around 67 to 237 GtC by 2100 or approximate 0.5 to 2 GtC yr^{-1} (Schuur et al. 2015, 2022). In comparison, between the time periods of 1970-1979 and 2012-2021, global annual greenhouse gases (GHG) emission increased from an average of 30 ± 4.0 to 54 ± 5.3 $GtCO_2e\ yr^{-1}$ (gigatonnes CO_2-equivalent emissions per year) (Forster et al. 2023). In 2021, total global GHG emissions reached 55 ± 5.2 $GtCO_2e\ yr^{-1}$ with global CO_2 emissions contributing 40.2 $GtCO_2\ yr^{-1}$, which corresponds to 11.0 GtC yr^{-1} (Friedlingstein, O'sullivan, et al. 2022; Forster et al. 2023). As the permafrost is already warming at an accelerated rate compared to the global surface air temperature (Biskaborn, Smith, et al. 2019), the direct impact of CO_2 and CH_4 released from permafrost on aquatic and terrestrial lifeforms is and will remain staggering (Vonk et al. 2015; Baltzer et al. 2014; Iturrate-Garcia et al. 2020).

Despite being under threat, lake systems offer us an invaluable record of past climate, preserved in their sediment. These sediment records serve as a foundation for limnological studies, enabling the reconstruction of a lake's unique history and its catchment by using a wide range of proxies, i.e. measurable characteristics of the sediment that serve as indirect evidence for a specific variable of interest (Zolitschka et al. 2015). However, interpretations from paleolimnological data, while rich and diverse, are also complex. The different proxies obtained from lake sediments can often convey conflicting stories, influenced by a multitude of drivers and processes on varying timescales. Some might indicate decadal changes, while others highlight millennial shifts, making the interpretation both richly detailed and complex.

Lakes often reveal a more localized and variable climate signal due to the significant influence of local factors, including catchment vegetation, lake morphologies, hydrological and hydrobiochemical dynamics. The use of multiple proxies enables the creation of multiple lines of evidence, which in turn allows for the distinction between local and global patterns. (H. H. Birks and Birks 2006). However, other factors can affect the reliability of climate interpretations in lake sediments, such as:

- sediment mixing and disturbance through wave action, bioturbation, and erosion (Zolitschka et al. 2015; Maher, Heiri, and Lotter 2012),
- proxy data interpretations in lake sediments can be ambiguous, as different environmental drivers may produce similar effects, e.g., high organic matter in sediments could result from increased productivity in the lake or from reduced decomposition rates (B. B. Wolfe, Edwards, and Aravena 1999; Meyers and Ishiwatari 1993)

- changes in availability and resolution of specific proxies, like pollen, diatoms, charcoal, stable isotopes, or trace elements (Telford 2019; Huntley 2012; Bradley 2015), and
- radiometric dating uncertainties affecting chronologies and subsequently influence the precision of climate interpretations (Blaauw and Heegaard 2012; Fritz 2008).

To enhance reliability, limnologists can address some of these challenges by using a range of modeling techniques, such as forward modeling of climate proxies (Dolman and Laepple 2018) or using resampling techniques (Reschke, Kunz, and Laepple 2019). The development of advanced geochronological software systems has significantly improved the ability to address dating uncertainties in non-layered sediments. However, Trachsel and Telford (2017) found that while geochronological software systems are essential for creating accurate age-depth models, they differ significantly among themselves. This further presents a conundrum where interpretation relies on models that may not accurately reflect the true chronology. However, as technology advances, these model limitations are becoming smaller through the application of more sophisticated approaches, requiring enhanced interdisciplinary collaboration between the fields of geoscience and computer science (Bergen et al. 2019; Pennington et al. 2020).

1.1.2 Data scientific perspective

Measurement data are the key to understand the world from a geoscience perspective. Temperature has been continuously measured in certain locations for over 350 years, and advancements in present-day sensors enable highly accurate direct measurements, e.g., for humidity or outgassing of CO_2 from lakes (Parker, Legg, and Folland 1992; Manley 1953). The amount of data generated by modern sensors has experienced exponential growth in the past decade, approaching the realm of big data for continuous measurements (Tolle, Tansley, and Hey 2009). Big data, as defined by De Mauro, Greco, and Grimaldi (2016), represents an information asset characterized by such a high volume, velocity, and variety that its conversion into values requires specific technology and analytical methods.

In contrast, palaeoenvironmental reconstructions still rely on indirect measurements through proxies, e.g., using chironomids for temperature reconstruction (Eggermont and Heiri 2012; PAGES 2k Consortium 2017). Although new laboratory equipment is becoming easier to use, the resulting datasets of proxy measurement data range in kilobytes to megabytes, leading to their classification as ‘long-tail research data’ (Ferguson et al. 2014) (Figure 1.2).

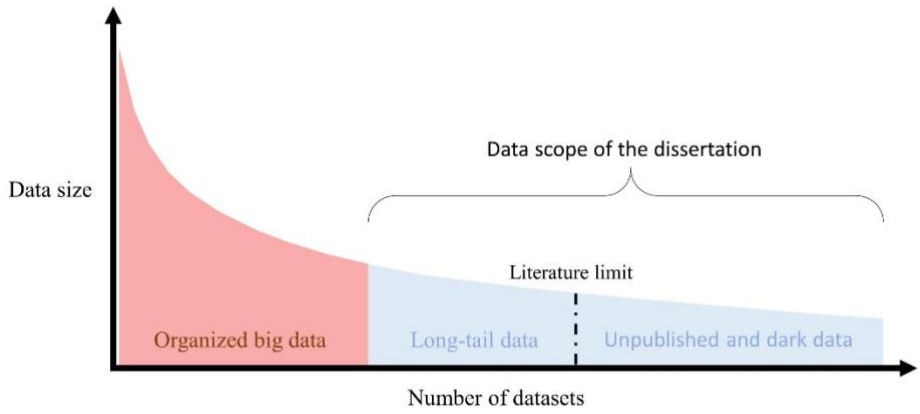

Figure 1.2: Scope of number of datasets and data size used in this dissertation in the context of big data and long-tail research data. Figure adapted from Ferguson et al. (2014). Ferguson et al. (2014) define dark data as "scientific findings [which] do not appear in the published literature and instead reside in file drawers and personal hard drives".

A key feature of long-tail research data is the significant number of individual files, which currently do not require a large amount of storage space. However, these files can vary in format, syntax, and semantics. Due to the inherent complexity of long-tail research data, automation of data preprocessing becomes difficult. Standardization partially allows overcoming this challenge by introducing structure, but still requires human intervention to fit the data into the proposed framework. However, standardization and standardized procedure have proven to be successful in global efforts, such as the Paleoclimate Modelling Intercomparison Project (PMIP) or Coupled Model Intercomparison Project (CMIP), which actively contribute to the IPCC reports (Touzé-Peiffer, Barberousse, and Le Treut 2020; Kageyama et al. 2021). Newly introduced standards such as Paleoclimate Community Reporting Standard (PaCTS) or International Generic Sample Number (IGSN) will help future limnological expeditions to produce robust datasets under the FAIR (Findability, Accessibility, Interoperability, and Reusability) principle (Conze et al. 2017; Khider et al. 2019; Wilkinson et al. 2016).

The downside of proposing a standard after data generation or collection is the risk of missing data and incomplete datasets. While one dataset may exclusively concentrate on a millennial scale with samples collected at one-meter intervals (e.g., biological proxies requiring additional treatment and labor), other datasets may have a higher sampling density, such as 1 mm intervals for x-ray fluorescence (XRF). The primary reasons for this are initially different geoscientific research questions and hypotheses, which lead to distinct data collection objectives, consequently producing contrasting datasets. Constraints in financial resources and inadequate laboratory staffing can also contribute to the disparities between datasets.

Unifying such datasets requires the development of simple data-driven solutions, as any large modeling interventions have the potential to introduce bias or obscure the climate signal (S. G. Dee et al. 2016; Ammann, Genton, and Li 2010). Ensuring high data quality control becomes particularly crucial when dealing with long-tail research data, because of the limited number of available data points. This stands in stark contrast to big data, where techniques such as machine learning and deep learning

can provide substantial advantages in data analysis and data quality control. However, even replacement algorithms or fill-in algorithms used to substitute missing values often require a substantial amount of data points to optimize their performance based on training data (Enders 2022). While certain physics-based models can operate effectively with a small number of samples (W.-C. Chen, Tareen, and Kinney 2018; Sadeghi Eshkevari et al. 2021; Raissi and Karniadakis 2018), models often still rely on additional independent data to ensure high-quality outcomes.

However, researchers can derive more comprehensive and generalizable conclusions from multi-sites investigations with the availability of sufficient long-tail data from multiple sources and the successful assimilation of such data. Global databases such as the PAGES[1] Arctic 2k database (McKay and Kaufman 2014), global multi-proxy database for the Common Era (PAGES 2k Consortium 2017), or the Temperature 12k proxy database (Kaufman, McKay, Routson, Erb, Davis, et al. 2020), provide the basis for the reconstruction of past climate conditions (Erb et al. 2022; Kaufman and Broadman 2023; Nicolle et al. 2018). Nevertheless, the strength of a global effort relies heavily on the collective contributions and collaborations of individual scientists within the research landscape. As researchers invest more resources in addressing the climate crisis, there is a growing willingness to share data and develop open-source applications, facilitating greater collaboration across national borders. Moreover, the synergy between geoscience and data science through interdisciplinary research enables a comprehensive understanding of complex systems, driving impactful change in addressing global challenges.

1.2 Methodological overview

1.2.1 Regional setting

Considering the extensive spatial scope of research conducted in the section "Polar Terrestrial Environmental Systems" at Alfred Wegener Institute, as well as the potential access to valuable external data from partnering institutions, this dissertation specifically narrows its spatial focus to lake systems in northern Eurasia, with a particular emphasis on those within Russia (Figure 1.1). We define northern Eurasia as the landmass extending from the Scandinavian Peninsula across Asia, encompassing regions such as Siberia, the Far Eastern Region, and extending all the way to the Kamchatka Peninsula and the Pacific Ocean. The area borders the Arctic Ocean to the north, while to the south, it encompasses Russian lake systems that extend as far as latitude 40° N.

The region features prominent mountain ranges, including the Scandinavian Mountains in the west, the Ural Mountains that serve as a natural boundary between Europe and Asia, the Altai and Sayan Mountains in southern Siberia, and the Kamchatka Range along the Kamchatka Peninsula. One of the most renowned lakes in the region is Lake Baikal, the world's deepest and oldest freshwater lake located in southern Siberia. Other notable lakes include Lake Teletskoye in the Altai Mountains, Lake

[1] Past Global Changes 2k Consortium (Ahmed et al. 2013)

El'gygytgyn in northeastern Russia, as well as Lake Ladoga and Lake Onega in northwestern Russia. Additionally, the northern Eurasian landscape features remarkable river systems such as the Ob and Yenisei Rivers in western Siberia, the Lena River flowing through eastern Siberia into the Laptev Sea, and the Volga River, Europe's longest river, flowing through western Russia.

Table 1.1: Selection of Köppen-Geiger climate classification criteria for study area based on Beck et al. (2018) with original work from Peel, Finlayson, and McMahon (2007).

Variable explanation for criteria:

MAT = mean annual air temperature (°C); T_{cold} = the air temperature of the coldest month (°C);

T_{hot} = the air temperature of the warmest month (°C); T_{mon10} = the number of months with air temperature >10 °C (unitless);

MAP = mean annual precipitation (mm y^{-1}); P_{dry} = precipitation in the driest month (mm $month^{-1}$);

P_{sdry} = precipitation in the driest month in summer (mm $month^{-1}$);

P_{wdry} = precipitation in the driest month in winter (mm $month^{-1}$);

P_{swet} = precipitation in the wettest month in summer (mm $month^{-1}$);

P_{wwet} = precipitation in the wettest month in winter (mm $month^{-1}$);

$P_{threshold}$=2×MAT if >70% of precipitation falls in winter,

$P_{threshold}$=2×MAT+28 if >70% of precipitation falls in summer, otherwise $P_{threshold}$=2×MAT+14.

During the six-month period from April to September, temperatures are warmer, indicating the summer season. Conversely, from October to March, temperatures become colder, signifying the winter season.

1st	2nd	3rd	Description	Criterion
B			Arid	$MAP<10\times P_{threshold}$
D			Cold	Not (B) & $T_{hot}>10$ & $T_{cold}\leq 0$
	s		- Dry summer	$P_{sdry}<40$ & $P_{sdry}<P_{wwet}/3$
	w		- Dry winter	$P_{wdry}<P_{swet}/10$
	f		- Without dry season	Not (Ds) or (Dw)
		a	- Hot summer	$T_{hot}\geq 22$
		b	- Warm summer	Not (a) & $T_{mon10}\geq 4$
		c	- Cold summer	Not (a, b, or d)
		d	- Very cold winter	Not (a or b) & $T_{cold}<-38$
E			Polar	Not (B) & $T_{hot}\leq 10$
	T		- Tundra	$T_{hot}>0$
	F		- Frost	$T_{hot}\leq 0$

The climate in northern Eurasia is highly diverse due to the immense geographical extent of the region. According to the Köppen-Geiger climate classification, the following eight climate subtypes are most commonly found in the study area (Beck et al. 2018; Rubel et al. 2018; Kottek et al. 2006):

- Dsc
- Dwc
- Dwd
- Dfb
- Dfc
- Dfd
- ET
- EF

Figure 1.1 illustrates the spatial distribution of the climate subtypes, while Table 1.1 outlines the criteria for each climate subtype based on Beck et al. (2018).

Based on data collected from 529 weather stations in Russia, Chernokulsky et al. (2018) reported that the spatially averaged annual precipitation in northern Eurasia between 1966 and 2014 was approximately 253.8 mm for showery precipitation[2] and 186.9 mm for non-showery precipitation[3], with total precipitation amounting to around 444 mm per year. According to the more recent ECMWF[4] ERA5 reanalysis precipitation data (Hersbach et al. 2020), where we focused on the period from 2018 to 2022, the average annual total precipitation for the region spanning 40° to 80° N and 10° to 180° E ranged from 615 mm to 628 mm per year (Figure 1.3). Both highest and lowest mean values for total precipitation are within the same longitudinal sector (120° to 180° E), with the lowest mean occurring between latitude 60° to 80 ° N (369 mm per year) and the highest mean between 40° to 60° N (1074 mm per year).

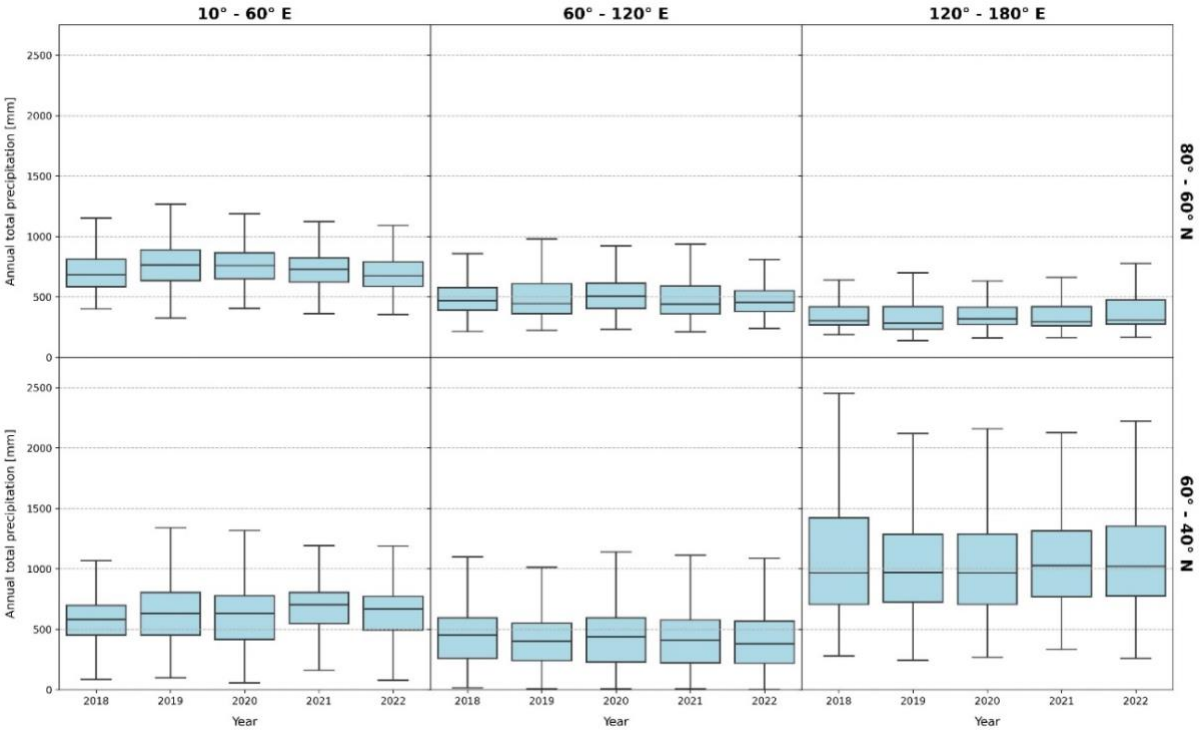

Figure 1.3: Boxplot showing the annual total precipitation in millimeter derived from ERA5 reanalysis for the region spanning 10° to 180° E and 40° to 80° N divided into sextants. ERA5 monthly average data on single levels from 1940 to present[5] including total precipitation data are available on https://cds.climate.copernicus.eu/. To calculate annual precipitation from the ERA5 total precipitation data (in meter per day), multiply the data by 1000 and by the number of days in each month, before summing for each year.

[2] Convection, whether free or forced, often causes showery (convective) precipitation (Chernokulsky et al. 2018)
[3] In cases of large-scale air lifting, such as on warm atmospheric fronts, the upward motion of air causes non-showery precipitation (Chernokulsky et al. 2018)
[4] European Centre for Medium-Range Weather Forecasts
[5] At the time of writing this work, it is November 2023

ECMWF ERA5 2-meter air surface temperature data covering the same period from 2018 to 2022 reveals substantial temperature variability across the study area (Figure 1.4). The southeastern European region of the study area (40° to 60° N and 10° to 60° E) stands out as the warmest part, with mean annual temperatures ranging from 9.07 °C (2021) to 9.95 °C (2020). It also experienced the highest temperatures in both JJA[6] (21.77 °C in 2021) and DJF[7] (1.26 °C in 2019/2020) seasons.

In comparison, extremely cold temperatures of below -20 °C to -50 °C prevail in all other regions within the study area during the winter seasons. However, there are notable differences in these extreme cold temperatures across latitudes. For example, in the region spanning 60° to 80° N and 120° to 180° E (northern Far Eastern Russia), mean DJF temperatures vary from -27.48 °C (2021/2022) to -28.50 °C (2018/2019). In contrast, despite extremely low individual temperatures, within the same longitudinal range but at lower latitudes (40° to 60° N - southern Far Eastern Russia), mean DJF temperatures range between -6.96 °C (2021/2022) and -8.66 °C (2020/2021). At the same time, the northern part of Far Eastern Russia also experiences the coldest summer temperatures of all regions, with mean JJA temperatures being between 7.11 °C (2018) and 7.57 °C (2021).

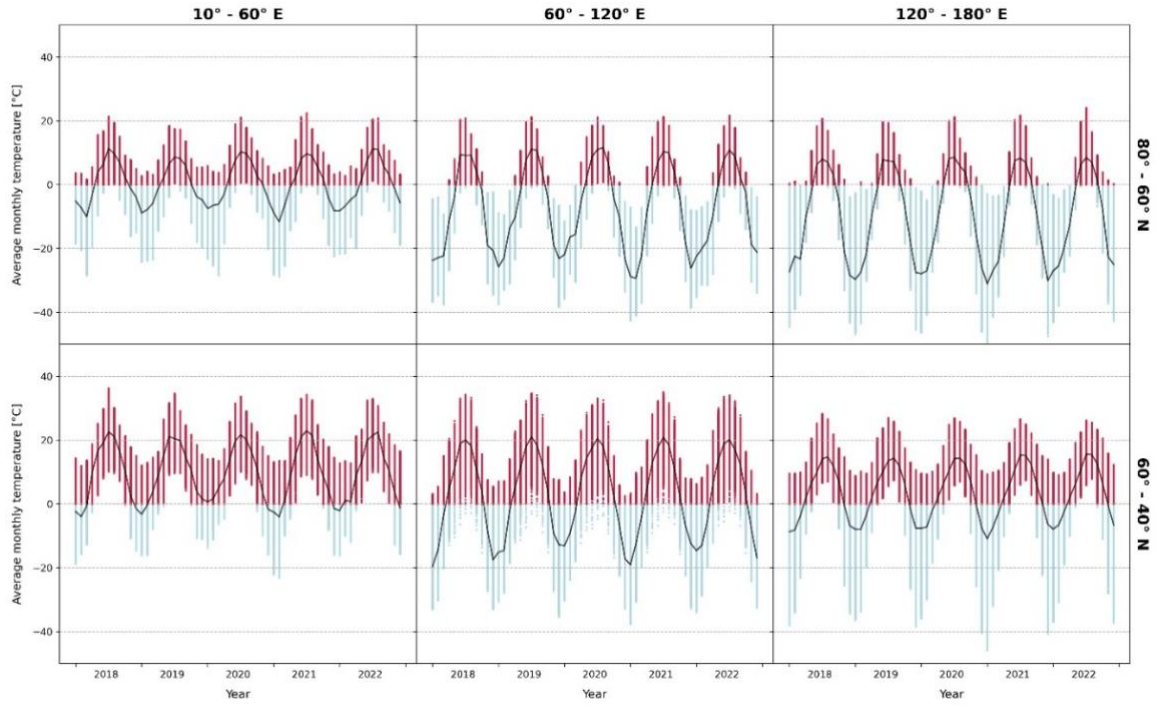

Figure 1.4: Plot showing the average monthly 2-meter air surface temperature in degrees Celsius derived from ERA5 reanalysis for the region ranging from 10° to 180° E and 40° to 80° N divided into sextants. ERA5 monthly average data on single levels from 1940 to present[8] including 2-m air surface temperature data are available on https://cds.climate.copernicus.eu/. To convert the ERA5 2-meter air surface temperature data from Kelvin to degrees Celsius, subtract 273.15 Kelvin from the original values.

[6] June-July-August

[7] December-January-February

[8] At the time of writing this work, it is November 2023

Considering the cold to polar climate condition, the vegetation in the study area has adapted to these often-harsh environments. Lichens, mosses, sedges, small shrubs like Arctic willows (Salix arctica) or dwarf birches (Betula nana), and grass-like plants dominate the landscape in polar tundra climates (Köppen-Geiger classification E). Coniferous (evergreen) trees such as pine (Pinus spp.), spruce (Picea spp.), larches (Larix spp.), and fir (Abies spp.) prevail in colder boreal climates (Köppen-Geiger classification D). In milder climates, including some boreal regions, birch trees (Betula spp.), alders (Alnus spp.), and other deciduous trees are more common. (cf. Glückler et al. 2021)

1.2.2 Data cleaning procedure and data availability

An extensive exploratory analysis yielded 1018 entries on presumed metadata information on lake sediment cores. The exploratory analysis included publicly available repositories, dissertations, journal articles, institute-internal data sources, and personal communication with scientists in the research area. An initial data quality control reduced the total number of unique sediment cores in northern Eurasia (above 40 °N) to 533 unique entries. This was mainly due to an established minimum requirement of essential metadata that we set, i.e., unique CoreID, geographical metadata (longitude, latitude), information on expedition/drilling campaign, drilling water depth and core length.

As this dissertation focused on the analysis of sediment cores on longer time scales, it was important that we had access to the results of the laboratory analysis of each sediment core. However, we often could not access the laboratory data (n=226) or only one data point was available (n=237), which often resembled a short surface sample. A total of 70 sediment cores had either one or multiple sets of available measurement data (Chapter 2). However, age determination data were only present in 55 of these cores (Chapter 3). In our investigation of carbon accumulation rates in lakes (Chapter 4), we found 28 suitable sediment cores within our metadata repository.

There are multiple explanations for this great imbalance between available and unavailable data:

- Scientists recorded both the drilling and its metadata, but the sediment core itself still remains closed and unanalyzed in a cold storage room (“No data available” case);
- Partnering institutions conducted an expedition, but the measurement data remains under an embargo or cannot be made publicly available (“No data available” case);
- Scientist analyzed a core, but the measurement data got lost on old hard drives or was never published (“No data available” case);
- Some investigations focused on the current state of lake systems and therefore only sampled the top centimeter to get an overview for the research question posed (“One data point” case);
- A team pilot drilled for a longer core, measuring only a few parameters at the top (“One data point” case).

While we considered both published and unpublished data when developing our tools, our primary focus during the analyses was on using published data. Our main reason for prioritizing published data during the analyses was to ensure the reproducibility and accessibility of the studies' findings to a wider

audience. However, we often encountered situations where laboratory measurement data for published sediment cores were not readily available in an online repository, requiring us to request the data from the principal investigator of the study. This approach presented a time-consuming challenge, as we had to ensure that we were getting the necessary metadata and raw data from the right source, while strictly adhering to the protocol of not sharing or publishing any data without prior permission. Although the acquisition, standardization, and ETL[9] process of the acquired data into the database took longer than initially anticipated, we are confident that our thorough approach to metadata and data quality control significantly enhanced the comparability of results of this dissertation.

As listed in Appendix A, parts of the metadata and measurement data used in this dissertation are available on PANGAEA (www.pangaea.de). PANGAEA is a renowned open-access data repository, specifically designed for the archiving, publishing, and wide dissemination of georeferenced data (Diepenbroek et al. 2002; Felden et al. 2023; Elger et al. 2016). In cases where data were not publicly available, we have either provided contact details (name of research group, email of working group leader) or referred to the original publication within each chapter.

1.3 Own prior contributions

This thesis includes work that appears in peer-reviewed publications or is currently under review. The following three chapters are the result of the collaboration between the author of this dissertation (GP) and several co-authors. To summarize the contribution of GP and other co-authors (initials) to the research results:

Chapter 2: **G. Pfalz**, B. Diekmann, J.-C. Freytag, and B. K. Biskaborn (2021). Harmonizing heterogeneous multi-proxy data from lake systems. Computers & Geosciences, 153, 104791. https://doi.org/10.1016/j.cageo.2021.104791. Submission: 24 June 2020. Published: 24 April 2021

GP conceptualized the study design, conducted data processing, analysis, and visualization, and wrote the original manuscript. BKB, BD, and JCF supervised the work of GP, contributed to the study concept, and provided input for writing and editing the manuscript. JCF additionally assisted GP in the development of the entity-relationship diagram. BKB and BD shared their unpublished data and metadata for the development of the MAYHEM system and data analysis.

Chapter 3: **G. Pfalz**, B. Diekmann, J.-C. Freytag, L. Syrykh, D. A. Subetto, and B. K. Biskaborn (2022). Improving age-depth relationships by using the LANDO ("Linked age and depth modeling") model ensemble. Geochronology, 4, 269-295. https://doi.org/10.5194/gchron-4-269-2022. Submission: 29 November 2021. Published: 18 May 2022

[9] Extract-Transform-Load

GP conceptualized the study design, developed the LANDO application, implemented the system into a multi-language Jupyter notebook, and conducted testing. GP also conducted data processing, analysis, and visualization, and wrote the original manuscript. All co-authors provided input for writing and editing the manuscript. BKB, BD, and JCF provided guidance and supervision throughout GP's work. Additionally, BKB, BD, LS, and DAS contributed published and unpublished age determination data for this publication.

Chapter 4: **G. Pfalz**, B. Diekmann, J.-C. Freytag, and B. K. Biskaborn (2023). Effect of temperature on carbon accumulation in northern lake systems over the past 21 000 years. Frontiers in Earth Science, 11, 1233713. https://doi.org/10.3389/feart.2023.1233713. Submission: 2 June 2023. Published: 29 August 2023

GP conceptualized the study design, conducted data processing, analysis, and visualization, and wrote the original manuscript. BKB, BD, and JCF supervised the work of GP, contributed to the study concept, and provided input for writing and editing the manuscript.

Harmonized datasets and models created during the dissertation were also part of several co-authored publications:

- S. A. Vyse, U. Herzschuh, **G. Pfalz**, L. A. Pestryakova, B. Diekmann, N. Nowaczyk, and B. K. Biskaborn (2021). Sediment and carbon accumulation in a glacial lake in Chukotka (Arctic Siberia) during the Late Pleistocene and Holocene: combining hydroacoustic profiling and down-core analyses. Biogeosciences, 18, 4791–4816. https://doi.org/10.5194/bg-18-4791-2021.

  BKB and UH designed the study and together with LAP organised fieldwork. SAV obtained hydroacoustic profiling data and was responsible for opening, splicing and subsampling the core as well as all biogeochemical and sedimentological analyses. He also wrote the first version of the manuscript. **GP** was responsible for the age-depth model and sedimentation rates. NN provided magnetic susceptibility data. BKB supervised the works of SAV. BD assisted in the interpretation of hydroacoustic data. All co-authors contributed to the interpretation of the results and commented on the text.

- B. K. Biskaborn, L. Nazarova, T. Kröger, L. A. Pestryakova, L. Syrykh, **G. Pfalz**, U. Herzschuh, and B. Diekmann (2021). Late Quaternary Climate Reconstruction and Lead-Lag Relationships of Biotic and Sediment-Geochemical Indicators at Lake Bolshoe Toko, Siberia. Frontiers in Earth Science, 9. https://doi.org/10.3389/feart.2021.737353.

  BB led the sediment-geochemical and diatom analyses and writing of the manuscript. LN and LS led the chironomid analyses including chironomid-based reconstructions. UH led the pollen-based reconstruction. BD and LP led the fieldwork and contributed to the environmental

interpretation of the data. TK and **GP** assisted with data processing and statistical analyses. All authors contributed to analysis of the results and revision of the manuscript.

- B.K. Biskaborn, A. Forster, **G. Pfalz**, L. A. Pestryakova, K. Stoof-Leichsenring, J. Strauss, T. Kröger, and U. Herzschuh (2023). Diatom responses and geochemical feedbacks to environmental changes at Lake Rauchuagytgyn (Far East Russian Arctic). Biogeosciences, 20, 1691–1712, https://doi.org/10.5194/bg-20-1691-2023.

  BKB conceived the study, conducted fieldwork, and statistical analyses, and wrote the paper. AF performed diatom analysis and counting. **GP** performed age–depth modeling. LAP, KSL, and UH coordinated fieldwork and dating of the short core. JS performed mercury analysis. TK performed correlation with *p*-value adjustment. All authors contributed to generating data as well as writing and reviewing the manuscript.

1.4 Research questions and hypotheses

As lake systems face increasing vulnerability to climate change, it is crucial to gain a comprehensive understanding of their past dynamics across different temporal and spatial scales. To achieve this, we collected datasets from various lake systems across northern Eurasia (above 40 °N), as well as data from two lakes in Greenland. The acquired datasets underwent a thorough data cleaning procedure and rigorous quality control, which included steps such as establishing minimum metadata requirements and checking for data inconsistencies, as previously outlined. The overarching objectives of this dissertation were to (1) collect a comprehensive range of measurement data and ensure data homogeneity in format, syntax, and semantics across the datasets; (2) establish consistent and reliable age-depth models to enable temporal comparability; and (3) explore the relationships between climate-related lake sediment parameters, such as carbon accumulation and environmental forcing. By adhering to these objectives, we aimed to establish a robust foundation for our analyses and enhance the reliability of our findings.

The dissertation addresses the following research questions and hypotheses in Chapter 2 to Chapter 4:

Chapter 2: *Research question:* Can the development of a conceptual model for heterogeneous multi-proxy data from multiple sediment cores in the Arctic, add value through the homogenization of these datasets and by enabling data transformation for comparative analyses?

Hypothesis: By developing a comprehensive conceptual model for integrating and standardizing heterogeneous multi-proxy data, it will be possible to transform the data effectively, enabling comparative analyses and enhancing our understanding of past environmental changes in the region.

Chapter 3: *Research question:* Can we reduce the effort required to apply different methods for determining age-depth relationships and achieve comparability of results across multiple lakes by combining the best available age models in an ensemble to estimate sedimentation rates more reliably?

Hypothesis: By providing a tool to link multiple modeling systems in a single multi-language Jupyter Notebook and introducing an ensemble age–depth model, it is possible to reduce the effort required to determine age–depth relationships and make their results comparable. The application of data-driven, semi-informed age–depth relationships will improve the accuracy and efficiency of dating in sediment cores, resulting in more reliable and comparable results across different methods.

Chapter 4: *Research question:* Can the analysis of carbon accumulation rates across different lake types, using a data-driven framework consisting of a homogenized database and ensemble age-depth models, provide insights into the relationship between carbon in lake sediments and paleotemperature?

Hypothesis: By analyzing carbon accumulation rates in multiple lakes and correlating them with temperature data from the TraCE-21ka climate reanalysis dataset, a significant relationship between carbon in lake sediments and paleotemperature occurs. The hypothesis suggests that higher temperatures are associated with increased carbon accumulation rates in lakes, indicating a positive correlation between temperature and carbon balance. Factors such as the presence of permafrost, composition of vegetation, and specific attributes of the lakes likely influence this relationship, which the analysis will explore.

2

Harmonizing heterogeneous multi-proxy data from lake systems

Gregor Pfalz[1,2,3,4], Bernhard Diekmann[1,2], Johann-Christoph Freytag[3,4], and Boris K. Biskaborn[1,2]

[1] Alfred Wegener Institute, Helmholtz Centre for Polar and Marine Research, Research Unit Potsdam, Telegrafenberg A45, 14473 Potsdam, Germany
[2] University of Potsdam, Institute of Geosciences, Karl-Liebknecht-Str. 24-25, 14476 Potsdam-Golm, Germany
[3] Einstein Center Digital Future, Robert-Koch-Forum, Wilhelmstraße 67, 10117 Berlin, Germany
[4] Humboldt-Universität zu Berlin, Unter den Linden 6, 10099 Berlin, Germany

Published: 24 April 2021 in *Computers & Geosciences* – DOI: 10.1016/j.cageo.2021.104791

Abstract

When performing spatial-temporal investigations of multiple lake systems, geoscientists face the challenge of dealing with complex and heterogeneous data of different types, structure, and format. To support comparability, it is necessary to transform such data into a uniform format that ensures syntactic and semantic comparability. This paper presents a data science approach for transforming research data from different lake sediment cores into a coherent framework. For this purpose, we collected published and unpublished data from paleolimnological investigations of Arctic lake systems. Our approach adapted methods from the database field, such as developing entity-relationship (ER) diagrams, to understand the conceptual structure of the data independently of the source. We demonstrated the feasibility of our approach by transforming our ER diagram into a database schema for PostgreSQL, a popular database management system (DBMS). We validated our approach by conducting a comparative analysis on a set of acquired data, hereby focusing on the comparison of total organic carbon and bromine content in eight selected sediment cores. Still, we encountered serious obstacles in the development of the ER model. Heterogeneous structures within collected data made an automatic data integration impossible. Additionally, we realized that missing error information hampers the development of a conceptual model. Despite the strong initial heterogeneity of the original data, our

harmonized dataset leads to comparable datasets, enabling numerical inter-proxy and inter-lake comparison.

2.1 Introduction

On-going global warming impacts Arctic landscapes through the "Arctic amplification" effect, where temperatures in the Arctic exceed the average Northern Hemisphere surface air temperature change (Biskaborn, Smith, et al. 2019; IPCC 2014; Miller et al. 2010). Lake systems are thereby among the most valuable, but at the same time also the most complex climatic archives of the earth as they enshrine various environmental information into their sediment (Bradley 2015; Brauer 2004; Cohen 2003). The regional and global climate as well as non-climatic influencing factors both affect the sedimentation process of lake systems (Fritz 2008; Wilke et al. 2016; Zolitschka et al. 2015). Understanding them helps to improve our perception of the earth system.

Analytical data derived from determining lake sediment properties, also known as proxy data, are essential for reconstructing lake histories, as they indicate change of environmental conditions (Bradley 2015). While scientists continue to collect new data from lake systems each year, thorough data handling of already existing datasets might help to fill remaining knowledge gaps of past changes. The quality of these older datasets varies depending on different factors, such as date of creation, individual project goals, available laboratory resources, and personnel bias (Cai and Zhu 2015; Heidorn 2008; Wang, Ziad, and Lee 2001). When integrating these existing datasets into a coherent framework and reporting standard, we can work with higher reliability and reproducibility thus enabling large-scale synthesis studies.

The number of repositories containing valuable data for paleolimnological studies has increased in recent years (Elger et al. 2016; Latif, Limani, and Tochtermann 2019; Muster 2018). Various studies using these repositories have already shown the effectiveness of multi-proxy, multi-site investigations through the synthesis of data from various sources (e.g., Bouchard et al. (2016), Kaufman et al. (2020), PAGES 2k Consortium (2017), Subetto et al. (2017)). Khider et al. (2019) recently proposed a reporting standard for new and past ("legacy") paleoclimate datasets, which includes the reporting of metadata information and measured data from lake sediments. The positive reception of their proposed model shows the consensus within the paleoclimate research community in favor of uniform standards for reporting measurement data of various kinds.

Still, there are non-digitalized, unprocessed or unpublished data hidden on local storage devices, old field and lab books as well as in hand-written documents (Curry and Moosdorf 2019; Heidorn 2008). Some "legacy" datasets might not meet the requirements of such a standard as proposed by Khider et al. (2019). This prevents older data from being included in any multi-site investigations. However, these datasets are potentially invaluable sources of almost forgotten knowledge (Muster 2018).

In this paper, we present a *conceptual* integration approach to enable a comprehensive comparison of datasets of varying quality from laboratory analysis of lake sediment. The specific

objectives of this paper are (I) to provide a conceptual entity-relationship (ER) model for merging heterogeneous multi-proxy data into a common framework from a database-centric perspective, and (II) to translate the conceptual model into a reference implementation using the PostgreSQL database management system (DBMS) to perform a comparative analysis on acquired and transformed data. Our approach will allow scientists to perform their data analysis on the integrated data with less effort compared to an analysis using the raw (original) data.

2.2 Methods

2.2.1 Data collection

For the cleansing and integration process presented in this paper, we used a collection of published and unpublished laboratory data and corresponding metadata from lake sediment cores. The majority of data came from online data repositories (e.g., Pangaea, GFZ Data Service) or institute-internal data sources of the Alfred Wegener Institute (e.g., expedition reports, personal communication). For the purpose of reproducibility and tracking, we collected and stored (meta) data about the sources of the collected data. This data is accessible in the repositories mentioned in the Appendix A.

We manually curated the laboratory data by using different data validation approaches (Pannekoek, Scholtus, and Van der Loo 2013; Sun et al. 2011). We assessed both laboratory data and metadata hereby on their completeness, consistency, accuracy, and precision (Batini et al. 2009; Batini and Scannapieca 2006; Sebastian-Coleman 2013). In a first step, we performed type checking, as laboratory data is known to be primarily numeric. We substituted unsuitable characters by numeric values using the Python package *pandas* (Reback et al. 2020). A physical range check followed the previous check to ensure that values do not exceed physical ranges (Sun et al. 2011). If values exceed their logical physical range, we then removed them from the dataset. We standardized names of common proxies (e.g., “Aluminum” to elemental symbol “Al”) or associated units (e.g., core lengths from centimeters to meters) to ensure consistent naming across all datasets. We logged all cleaning actions carried out during the entire validation process for provenance reasons. If possible, we examined original files from measuring instruments as well as original publications to avoid any conversion errors. We consulted the corresponding responsible scientist for any clarification when needed.

We selected the following information as minimum requirements for metadata information to be included in our study:

- unique core identifier (“*CoreID”*),
- geographical information (latitude, longitude),
- information about the field campaign (name, year),
- site name,
- lake type,
- water depth at coring location, and
- composite (i.e. cumulative) core length, derived from overlapping core segments.

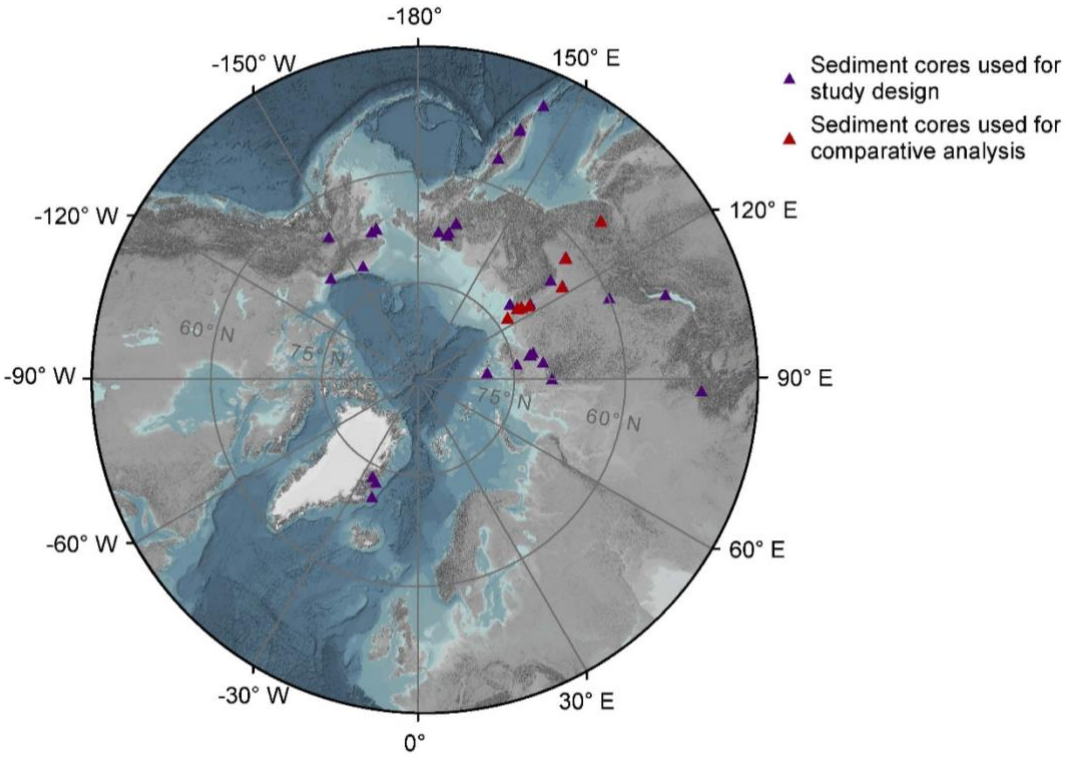

Figure 2.1: Geographical distribution of lake sediment cores used for the study design (triangles, n=70). Red triangles (n=8) indicate lake sediment cores used for the comparative analysis of total organic carbon (TOC) and bromine (Br) content shown in this study. ArcGIS Basemap: GEBCO Grid 2014 modified by AWI. The outer ring in the graphic corresponds to 45° N.

The proposed preselection helped us to distinguish clearly unique entries of lake sediment cores from other entries. If the metadata was not included in the acquired dataset, then we searched in related literature to determine the uniqueness of the core. We generated unique identifiers for the cores using the pattern "*{FirstAuthorLastName} {LakeID} {ExpeditionYear} {CoreNumber}*", when no unique identifier was available. CoreIDs, names of field campaign, and sites are stored in English using the Latin alphabet. The information about latitude and longitude is recorded as decimal degrees (minimum precision: three decimal places), while water depth and core length are given in meters (precision: two decimal places). Besides using the geographical location to validate the uniqueness of a sediment core, we further used the coordinates in ArcGIS, Google Earth, and HydroLAKES database (Messager et al. 2016; Meyer et al. 2020) to check that the scientists correctly placed and assigned site names to cores.

In total, there were 70 metadata information entries for unique core sediments. Figure 2.1 illustrates the geographical distribution of this metadata compilation. It shows that the availability of information with high latitude values (50° N to 90° N) dominates the spread. The dominance of the latitude values in this range is due to the clear research focus by the Alfred Wegener Institute and its partners on the Arctic.

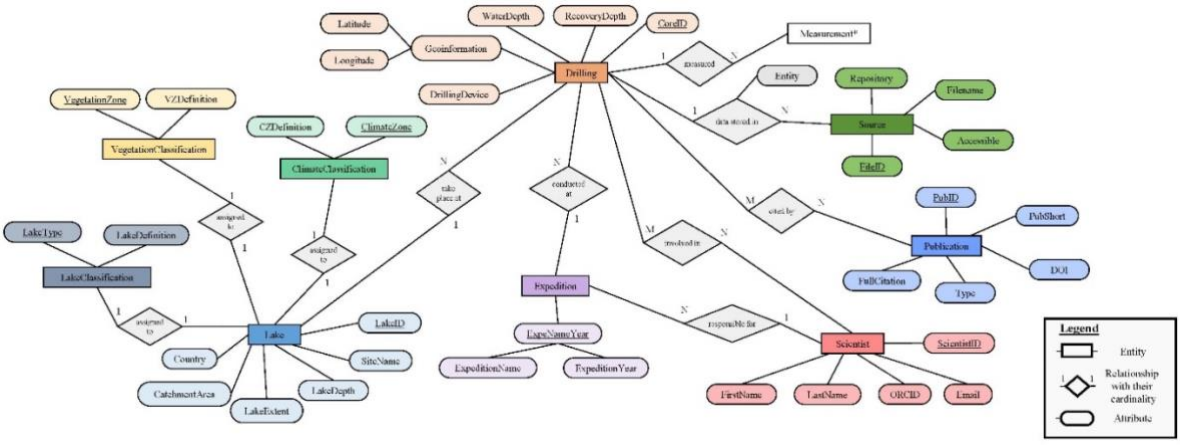

Figure 2.2: Entity-relationship diagram of nine core-specific entities (rectangular boxes) with their attributes (boxes with rounded corners) connected through relationships (diamond-shaped objects). The core-specific entities refer to information about core retrieval and associated information. Entity *Measurement* in the right upper corner establishes the connection between both entity groups.

2.2.2 Conceptual approach

During data collection, we encountered variety of structures and inconsistent data, which makes data of different sources incomparable. In order to facilitate a better integration of those heterogeneous datasets into a database, we decided to design a unified scheme based on the existing structures to store all base values. We defined as base values all values, which cannot be derived from any other data values in any of the data source. Under this assumption, the first step within the (conceptual) schema design approach was to develop generic concepts, which should resemble aspects of limnological studies. We followed the principles of database design for a redundant free data representation. As such, a (conceptual) data model describes all aspects of the real world by identifying relevant entities and the relationships among them in the domain of discourse (Codd 1970; Elmasri and Navathe 2009; Teorey et al. 2008).

Both, entities and relationships are associated with specific attributes, i.e. descriptive properties that are inherent for each of them in the real world. For instance, researchers conduct a core drilling at a specific geographical location (described by latitude and longitude) and water depth with a particular drilling device. The resulting core has an unique core identifier (e.g., *"International GeoSample Number"* (Conze et al. 2017), or simply *"CoreID"*) and a composite core length. Therefore, we reduced this process to one core-specific entity *Drilling* with five attributes (see Figure 2.2), while one attribute is a composite attribute (*"Geoinformation"*) and another is the key attribute (*"CoreID"*) of this entity.

Naturally, entities have an interdependency with other entities. Teorey et al. (2008) provides guidance for how to design entities and their relationships. As an example for paleolimnological studies: It is only possible to conduct a core drilling at one location at any given time; hence, there is a one-to-one relationship between the entities *Drilling* and *Lake*. Other binary data modeling cardinalities are one-to-many or many-to-many relationships (Garcia-Molina, Ullman, and Widom 2002; Teorey et al. 2008). In general, entity-relationship diagrams, also known as "*ER diagrams"* (Figure 2.2, Figure 2.3),

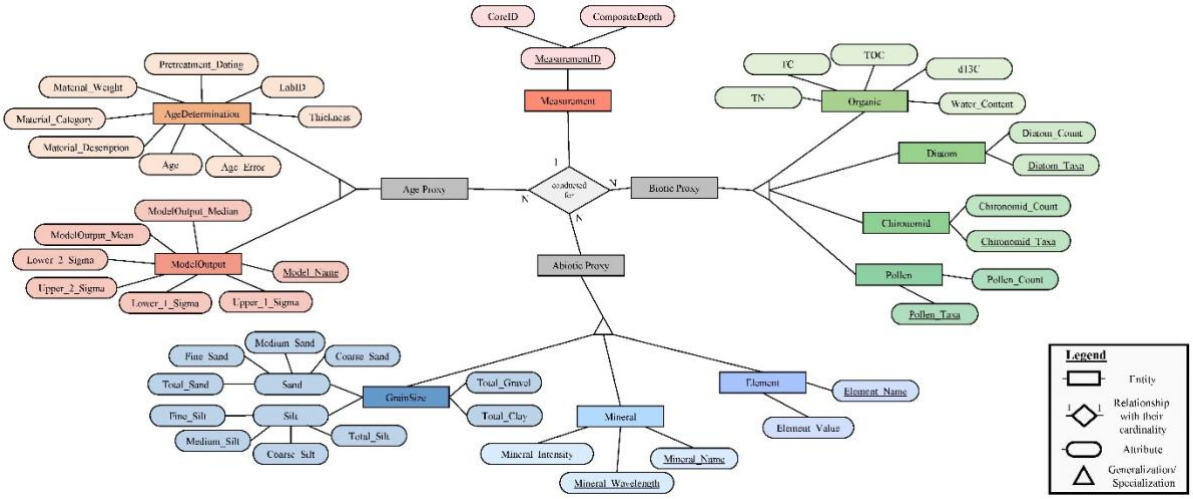

Figure 2.3: Entity-relationship diagram of ten measurement-specific entities (rectangular boxes) with their attributes (boxes with rounded corners) connected through relationships (diamond-shaped object). Measurement-specific entities are consistent with the measured laboratory data of sediment cores.

or alternatively, diagrams using the Unified Modeling Language (UML) (Figure 2.4, Figure 2.5), allow us to visualize entities together with their relationships (P. P. S. Chen 1975; Garcia-Molina, Ullman, and Widom 2002; Teorey et al. 2008). This visual representation accelerates the implementation of the developed concept in a database management system (DBMS).

In our case of paleolimnological studies, we identified two groups of entities: measurement-specific and core-specific entities. Measurement-specific entities represent individual laboratory measurements for proxy determination. Core-specific entities do not only characterize the core retrieval, but also operational metadata describing the drilling. The operational data includes data about the surveyed lake, field campaign/expedition, responsible scientist, and publications. This allocation further supports initiatives for optimal metadata management and investigating for possible systematic errors, i.e. data trustworthiness (Batini et al. 2009; Bertino and Lim 2010).

Besides the superordinate entity *Drilling* in the group of core-specific entities (Figure 2.2), we split the field campaign into the entities *Lake*, *Expedition*, and *Scientist*. *Lake* describes further details about the lake at which the drilling took place. The *Lake* entity also relates to the entities *ClimateClassification*, *VegetationClassification*, and *LakeClassfication* entity, which describe the climate, vegetation and lake origin at time of core retrieval to the lake, respectively. For operational data about the field campaigns in relation to the drillings, we designed both *Expedition* and *Scientist* entities for temporal attribution and contact details. *Publication* consists of the important publications relating to each drilling, while *Source* gives information about the files used to produce the measurement-specific entities and therefore allows us to reproduce them with higher precision.

For the measurement-specific entities (Figure 2.3), the attributes describe measured quantities of the individual laboratory measurements. It was therefore vital for us to gain an understanding of the different laboratory methods used to analyze proxies. Based on the available data, we determined eleven proxies that were analyzed frequently in the datasets (Table 2.1). We examined each related study

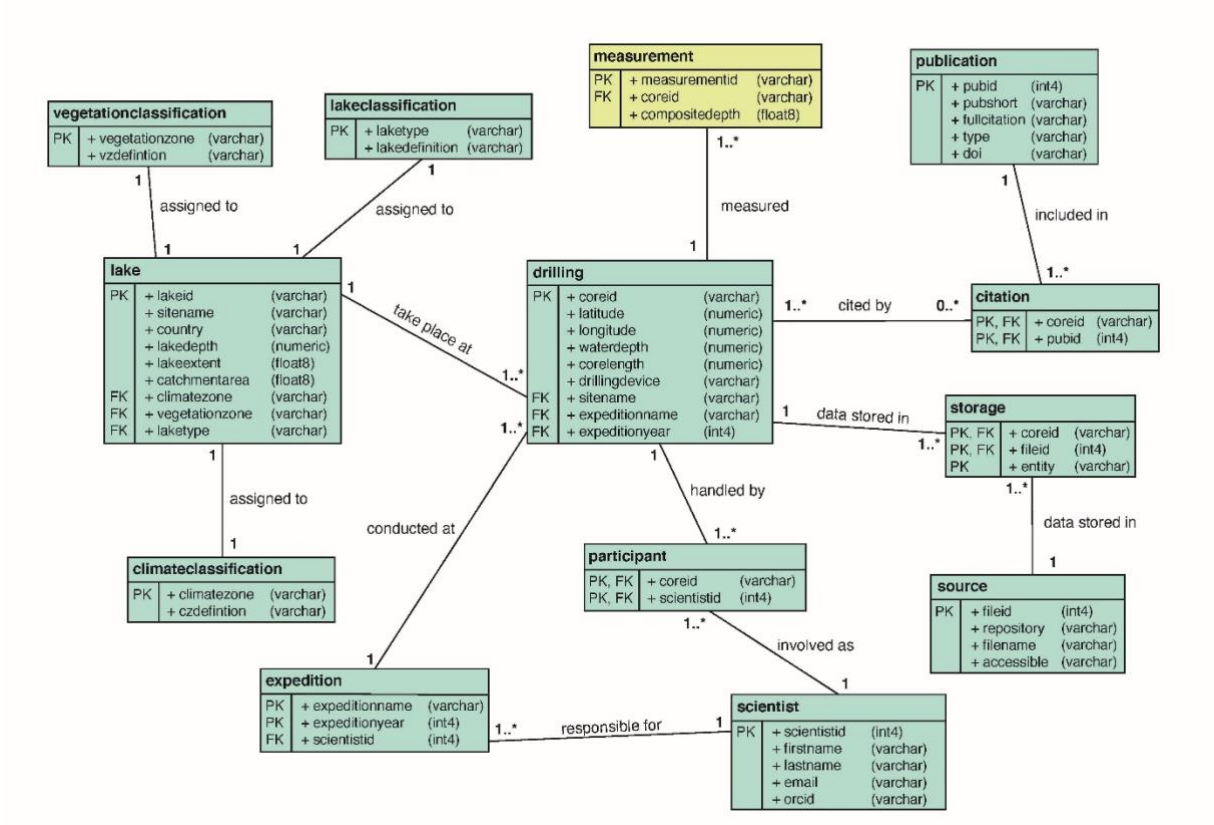

Figure 2.4: Unified Modeling Language (UML) diagram of core-specific entities for the reference implementation. Entities (rectangular boxes) consist of their name, attributes and their PostgreSQL data type in tabular form, and an indication whether the attributes are primary keys (PK), foreign keys (FK), or both. The entities are connected by a relationship (solid line) to another entity. The numbers on the solid line indicate the cardinality of that relationship. This figure includes the entity *Measurement* to show connection to the other derived measurement-specific entity group. List B1 in the Appendix further describes each entity.

closely on its applied methodology to determine said proxies. While most studies followed the same methodological approach, we found variations for three proxies, namely for elements, grain size, and Total Organic Carbon (TOC) (Figure B1 in the Appendix provides a pictorial representation). In our design, we decided to focus on one approach as a representative for both TOC and grain size. For the elemental proxy we kept the method selection optional.

Regarding the attributes for measurement-specific entities, we decided for the lowest common denominator to avoid artificial inflation. For instance, for a measurement independent abstraction of elemental data, the reduction leads to one entity *Element* with two attributes for (i) the name or symbol of the element, and (ii) the associated value at a specific depth inside the core. To place an emphasis on the applied method, we appended the unit of the measurement to the element symbol in the attribute "*Element_Name*", e.g., Aluminum measured in parts-per-million as "*Al_ppm*". This strategy is only possible for elemental data, as the unit of measurement does not change and values from different methods for this proxy are not directly comparable.

Ultimately, we derived nine core-specific (Figure 2.2) and ten measurement-specific (Figure 2.3) essential entities from the acquired data. The entities *Measurement* and *Drilling* establish the

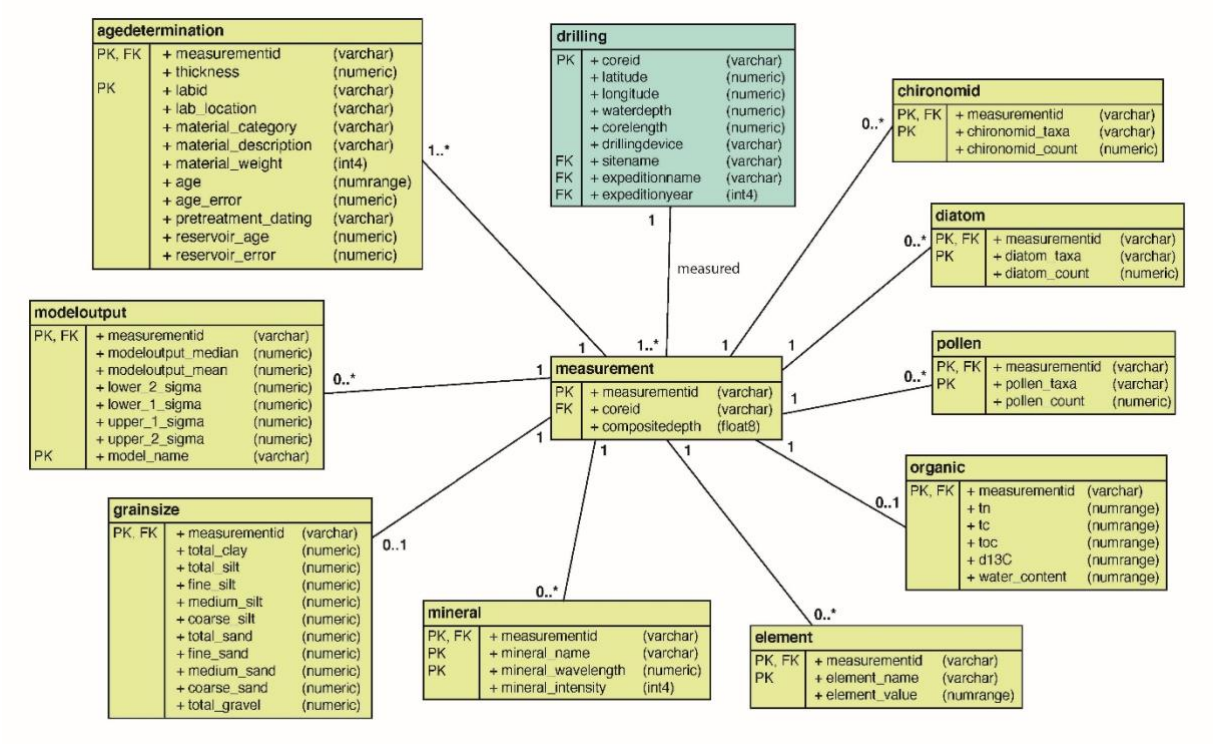

Figure 2.5: Unified Modeling Language (UML) diagram of measurement-specific entities for the reference implementation. Entities (rectangular boxes) consist of their name, attributes and their PostgreSQL data type in tabular form, and an indication whether the attributes are primary keys (PK), foreign keys (FK), or both. The entities are connected by a relationship (solid line) to another entity. The numbers on the solid line indicate the cardinality of that relationship. This figure includes the entity *Drilling* to show connection to the other derived core-specific entity group. List B1 in the Appendix further describes each entity.

linkage between the two entity groups. When analyzing sediment cores, scientists extract multiple proxy measurements from a specific depth along the core's length (one-to-many relationship). Hence, unique identifiers in *Measurement* contain a composite attribute consisting of the composite depth and the corresponding core identifier. As the core identifier "*CoreID"* is also the key attribute of the *Drilling* entity, it enables us to extract additional operational information belonging to the measurement.

In the second step of our conceptual approach, we had to ensure that a comparison between different datasets is feasible. Harmonizing laboratory data from geographically dispersed cores necessitates finding a common anchor point between those datasets. The sampling scheme of the individual proxies strongly depends on the depth within the sediment core, which itself depends on the research questions posed. Laboratory measurements could therefore be taken very frequently (i.e. every one to two millimeters), less frequently (i.e. every five to ten centimeters), or where distinctive changes within the sediment core are visible. The time axis is the only constantly running physical quantity and common denominator on which we can place all measurements. For this conversion, one might use existing age-depth modeling software.

While users can set defined depth resolutions for the age-depth relationship within the modeling software to match varying proxy resolutions, a complete conversion of the proxies from individual

Table 2.1: Selection of proxies, which were frequently determined in the acquired laboratory datasets

Abiotic proxies	Biotic proxies
Elements	Diatoms
Minerals	Chironomids
Grain size	Pollen
Water content	δ13C
	Total organic carbon (TOC)
	Total carbon (TC)
	Total nitrogen (TN)

depth-dependent to joint age-dependent will leave blanks. Logical approximations or interpolations of the individual proxy have to fill those vacant time slices when seeking comparable multi-site investigations (H. H. Birks and Birks 2006). There are various well-established techniques available, such as in-filling techniques, spline interpolation or machine-learning based interpolation (Harry John Betteley Birks 2012).

At first, we define the proxy with the lowest age resolution as the *base proxy* after the conversion from depth-dependent to age-dependent. We then match other proxies with a higher resolution (*desired proxies*) to the base proxy. That is to say, we select values of the desired proxy at the time slices equal to the time slices of the base proxy. If the value of the desired proxy is not available at the exact time slice of the base proxy, we perform an interpolation of the desired proxy. We hereby follow the advice by Blaauw (2012) to avoid the use of overfitted values which could potentially result in misinterpretations. If we were to excessively interpolate values of a proxy with a lower resolution in order to fit the curve of proxies with a higher resolution, we would increase the likelihood of a misinterpretation. Additionally, internal lake dynamics influence biological activity within lake systems, which means that a higher or lower abundance of biological proxies within a sediment core might be an indication for a specific event (Biskaborn, Nazarova, et al. 2019). It is therefore debatable, whether an interpolation of biological proxy is reasonable, or we should use Bayesian modelling approaches instead (Huntley 2012). For simplicity, we assume in our study that interpolation is feasible for the proxies involved in this study.

To synchronize values between sediment cores, we use binning to create equally spaced bins of time. We calculate the optimal bin size using the mean of the maximum proxy age resolution for the base proxy across all sediment cores. We then select the interpolated central values for each proxy measurement at the same interval as the bin size. We use all remaining values within an age bin to calculate the minimum and maximum value range for each proxy.

2.2.3 Comparative Analysis

To set up for our comparative analysis, we implemented the developed ER data model as a database schema for a PostgreSQL database system (Version 11.2; PostgreSQL Global Development Group (2018)). For the implementation, we used the open-source software tool *DBeaver* (Version 7.0; DBeaver Community (2020)). Figure 2.4 and Figure 2.5 provide a visual representation of the reference implementation as UML diagram. List B1 in the Appendix shows the individual entities with their attributes, a short explanation, the PostgreSQL data type, and an example.

After implementing the schema in PostgreSQL, we had to transform the available (raw) data to fit the proposed schema. Therefore, we set up a spreadsheet template with the same schema as the database to support the integration process. We show an example spreadsheet template for our reference implementation in the repository mentioned in the Appendix A. The standardized data was then inserted into the database using a *Jupyter Notebook* (Kluyver et al. 2016) using the package *SQLAlchemy* (Bayer 2012).

We performed the comparative analysis in a separate Jupyter notebook – for more information on the code we refer to Appendix A. We selected total organic carbon (TOC) content and X-ray fluorescence (XRF) measured bromine (Br) content to showcase the basic functionality of our approach. Previous studies conducted by Biskaborn et al. (2016) and Kalugin et al. (2007) showed that bromine is a good indicator for changes of organic content in lake sediment and should therefore agree well with the TOC content (Rothwell and Croudace 2015). Both proxies were measured in eight sediment cores within our database. First, we converted the measurement depths to the median ages from the corresponding age-depth model. To allow a transparent and comprehensible (re-)modeling of already existing age-depth relationships, we gathered all information regarding laboratory age analysis with its associated uncertainty. For a better reproduction of the age-depth relationship, we stored further information regarding the age determination. This information included the description of the dated material, involved laboratory, pretreatment methods, and thickness of the dated sediment layer, if bulk sediment was dated.

For age-depth modeling, we used the existing open-source MATLAB software package *Undatable* (Lougheed and Obrochta 2019) with the improved IntCal20 calibration curve (Reimer et al. 2020). We implemented an additional script to access all eight sediment cores from the database and then computes age-depth models in bulk (Undatable settings: nsim = 10^5, bootpc = 30, xfactor = 0.1). This recalculation reduces potential biases introduced by the authors and possible differences between modeling software output (Trachsel and Telford 2017; Wright et al. 2017). Age-depth relationships produced during our harmonization process are no replacement for the original relationships identified by the contributing authors (cf. McKay and Kaufman 2014). We stored the resulting output created by the modeling software as "*ModelOutput"* in the database using an additional script.

We determined our base proxy for our comparative analysis using the mean proxy age resolution. We also considered the impact the higher resolution proxy had on the data, if we were to use

the proxy with higher resolution as the base proxy. To enable a synchronized comparison of TOC and bromine, we used interpolation to replace missing values by approximated values. We applied a piecewise polynomial interpolation for existing gaps using the Python package *SciPy* (Virtanen et al. 2020). We assessed all cores on their maximum proxy age resolution for the base proxy to determine the optimal bin size. Ultimately, we binned each measurement into its respective age bin and determined from all measurements the minimum and maximum value ranges within each bin.

2.3 Results and Discussion

Using research data from external sources always contains the risk that undocumented transformations (knowingly or unknowingly) changed the data after a laboratory analysis. This might lead to erroneous and incomprehensible data. Therefore, we contacted the responsible scientists for further inquiries regarding the data handling to avoid propagating possible errors. If there were inconsistencies between different approaches, then we documented this circumstance for traceability. Due to the design of the data model, we provide data, which allows a scientist to retrieve the original data files and publication for each proxy dataset. Such reference supports the important concept of lineage thus providing an improved contextualizing of the data, which might be important for further use of the data. We claim that good data cleansing can foster an interoperability amongst geoscientist and the use of automated data integration tools. However, the biggest challenge during the harmonization process of our collected dataset was the handling of varying data qualities. The most noticeable inconsistency was the heterogeneous structure within the data. While almost all data from online repositories followed syntactic rules, data from other sources did not stay within a coherent framework. Therefore, over the course of our investigation, we had to exclude the possibility of automated data integration. If future datasets follow the FAIR principle (Findable, Accessible, Interoperable, Reusable), we are convinced that automated data integration becomes possible (Latif, Limani, and Tochtermann 2019; Stall et al. 2018; Wilkinson et al. 2016).

The use of a database management system (DBMS) for the comparative analysis has clear advantages over loosely connected, personal spreadsheets without the ability for integration. Currently, measurement data exists in different labs on different computer without the ability for a common usage and understanding. Our approach presents first steps towards a data-driven integration of such data. There are multiple reasons for using database techniques in the context of data transformation and integration resulting in set of homogenized data for further analysis. Once the data are in a standardized format, the database provides high availability, high flexibility, synchronization, error recovery, and great efficiency. Integrated datasets further support data integrity within the database. Despite new developments in database research, relational database management systems (RDBMSs) still provide the best fit for laboratory data from paleolimnological studies as most of data generated by measuring instruments can be stored and accessed in a tabular form. Other geoscientific databases, such as

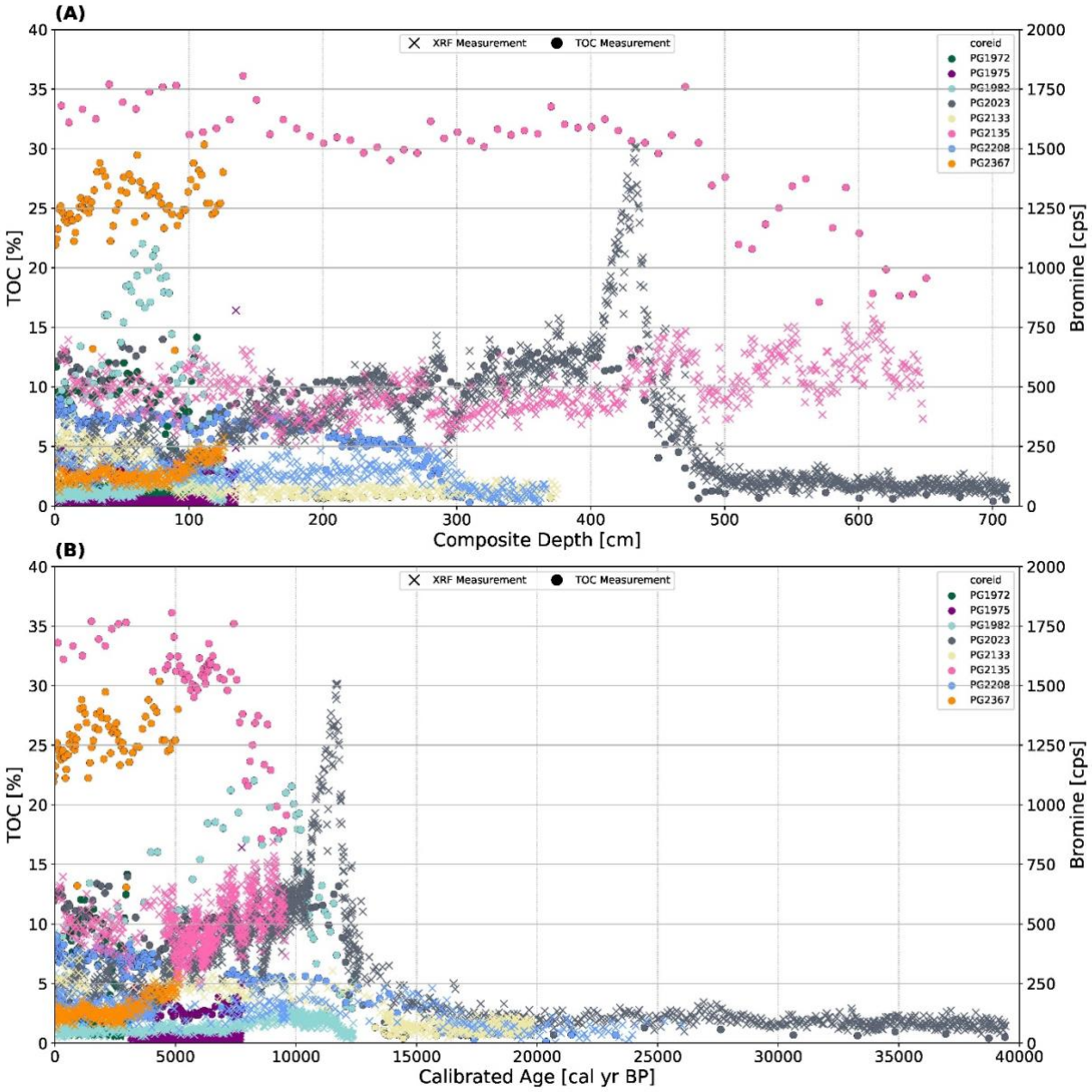

Figure 2.6: Comparison of total organic carbon (TOC) and bromine (Br) content for eight selected sediment cores within the reference implementation and after being age-transformed. Panel A shows TOC and bromine measurements against the individual composite depth in centimeter of each sediment core, as existent in the reference implementation. In panel B, same measurements are transformed from composite depth to the corresponding calibrated median ages. The x axis in panel B is based on calibrated median ages (calibrated years Before Present, cal yr BP) derived from age-depth modeling software *Undatable* (Lougheed and Obrochta 2019) using the IntCal20 calibration curve (Reimer et al. 2020). Color codes are consistent over all measurements and plots for each sediment core. Circles markers represent TOC measurements in percent and cross markers indicate bromine measurements in counts per seconds (cps) using X-ray fluorescence (XRF).

Neotoma, Pangaea, or GTN-P, proved the reliability of RDBMSs (Biskaborn et al. 2015; Diepenbroek et al. 2002; Williams et al. 2018). Additionally, most database management systems provide interfaces (APIs) to a series of programming languages, which makes it easier to retrieve and to analyze the stored data efficiently and effectively (cf. Elmasri and Navathe 2009; Teorey et al. 2008).

Table 2.2. Statistics of proxy sampling and proxy age resolution for each proxy in reference implementation. Q_{25}, Q_{50}, and Q_{75} correspond to the 25% quantile, median, and 75% quantile within each proxy resolution, respectively. Note: Total organic carbon (TOC) and Total nitrogen (TN) were measured together, hence the resolution and number of data points are the same

Proxy	Number of data points	Proxy age resolution [yr]				Proxy sampling resolution [cm]			
		Mean	Q_{25}	Q_{50}	Q_{75}	Mean	Q_{25}	Q_{50}	Q_{75}
Element	8388	33.38	9.00	13.71	30.00	3.97	0.50	1.00	1.02
TOC / TN	2130	187.65	22.67	72.00	158.00	5.99	1.95	4.75	9.65
TC	1169	181.92	37.75	73.75	125.15	3.88	1.95	2.02	5.29
δ13C	1166	163.24	16.67	55.00	152.00	6.86	1.78	6.24	9.94
Pollen	760	217.07	51.00	126.00	185.00	8.84	3.28	8.78	10.93
Grain Size	462	312.61	76.50	134.00	355.00	15.51	2.06	7.93	18.99
Diatom	437	372.99	82.88	173.75	356.75	9.88	4.63	9.67	12.74
Mineral	418	295.37	77.00	119.50	232.50	12.22	4.77	5.87	9.48
Chironomid	152	472.87	126.87	236.50	566.75	14.08	12.41	15.01	16.68

Figure 2.6 (A) shows the untransformed output from the reference implementation, where TOC and bromine content are dependent on the depth within each sediment core. We then transformed all values from depth-dependent to age-dependent (Figure 2.6 B). Figure 2.6 (B) illustrates the TOC and bromine values against their corresponding median age derived from the age-depth modeling software *Undatable* (Lougheed and Obrochta 2019). What stands out in Figure 2.6 is the variability of consecutive measurements along the x axis. Still, we determined elemental and TOC measurements to have the highest and third highest age resolution in our reference implementation, respectively (see Table 2.2). The resolution depends highly on the level of automation and treatment processes for each proxy. We can measure elemental data from non-destructive X-ray fluorescence (XRF) core scanning without any pretreatment at a depth resolution of two to five millimeter. Other proxy groups such as diatoms, chironomids, or pollen require time-consuming pretreatment and microscope-based analyses performed by individual scientists. Scientists accommodate the additional preparation by commonly taking fewer samples.

Based on these results, we selected TOC as our base proxy for the comparison of TOC and bromine. Figure 2.7 compares the results from using (A) TOC and (B) bromine as base proxy for the matching process. Panel A is following our approach of choosing the proxy with the lower resolution as the appropriate base proxy and generating corresponding interpolated bromine values for each TOC value, if needed. In panel B we show that using the higher resolution proxy as base proxy instead, leads to an overestimation of specific events due to necessity of excessive interpolation. With the results from panel A we started to calculate optimal bin size and minimum and maximum value ranges for each bin. We determined 700-year bins to be the optimal bin size for our comparative analysis. Figure 2.8 illustrates bromine and TOC values for all sediment cores binned into 700-year bins with their minimum

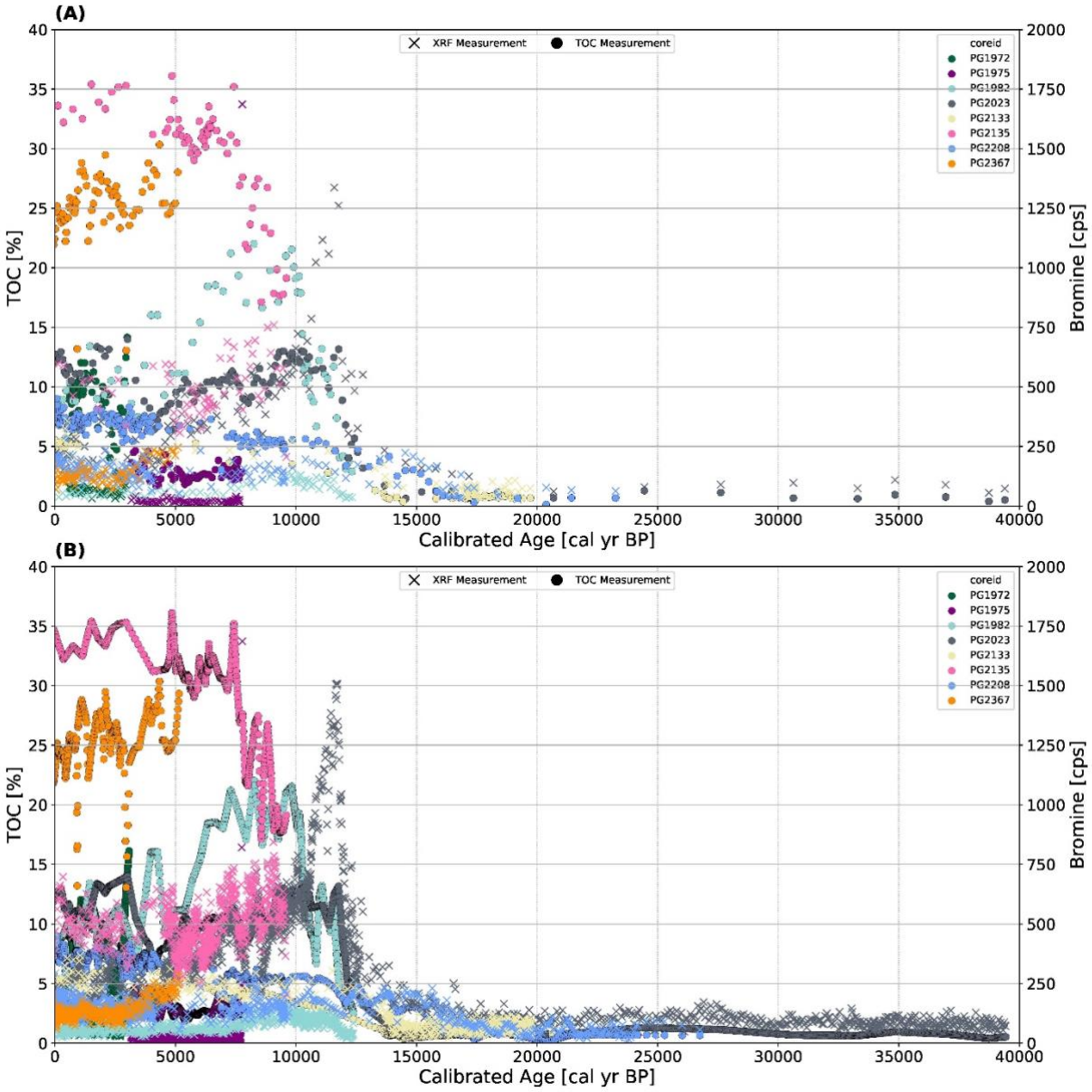

Figure 2.7: Difference in interpolation approaches to synchronize measurements from total organic carbon (TOC) and bromine (Br). Panel A shows bromine values being synchronized to measurements of base proxy TOC. Panel B displays TOC values being resampled to match measurements in higher resolution of the base proxy bromine. In the case that the exact value of the desired proxy was not present at the specific age of the base proxy, we applied a piecewise polynomial interpolation to the desired proxy. Panel B therefore demonstrates a case were harmonization would result in a strong bias of resulting interpolated values. Color codes are consistent over all measurements and plots for each sediment core. Circles markers represent TOC measurements in percent and cross markers indicate bromine measurements in counts per seconds (cps) using X-ray fluorescence (XRF).

and maximum value ranges. Through this approach we are able to transform both high and low resolution data into one single matrix containing a minimum amount of null values. By defining value ranges for each bin, we quantify the uncertainty of interpolated values at the center of each bin. We claim that through our approach, sediment cores are now comparable on both a temporal scale and an inter-proxy level.

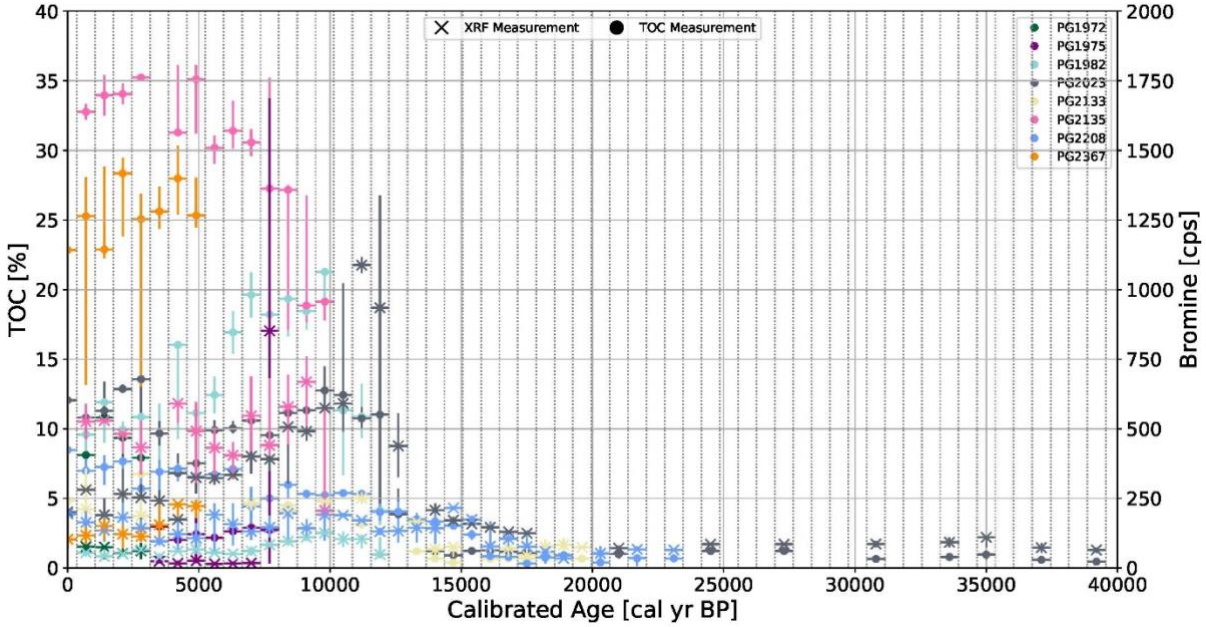

Figure 2.8: Synchronized values for both total organic carbon (TOC) and X-ray fluorescence (XRF) derived bromine (Br) measurements using 700-year binning. We calculated the bins using the mean of the maximum resolution for the base proxy of each involved sediment core. Markers represent the interpolated value at the center of the bin at a continuous 700-year interval. We included all measurements within each 700-year bin to determine minimum and maximum values within each bin, represented by connected vertical lines. Circles markers represent TOC measurements in percent and cross markers indicate bromine measurements in counts per seconds (cps) using X-ray fluorescence (XRF). Dashed vertical lines display boundaries between 700-year bins. Horizontal lines display the age range for which measurement are applicable. Color codes are consistent over all measurements for each sediment core.

However, natural systems hold some degree of uncertainty, as the analysis of proxies itself inhibits inevitable uncertainties (Amrhein 2019; Goswami 2014; Reschke, Rehfeld, and Laepple 2019). Therefore, it is the responsibility of scientists to handle and to report the uncertainties with their data measurements. By default, manufacturers of measuring devices usually report error intervals for their devices. The laboratory staff on-site refine the accuracy of these devices through the implementation of improved calibration methods. In many cases, however, the inaccuracy/deviation of results is not reported, visible in publications, or stated in any supplementary material. This situation is a serious obstacle in a multi-site investigation, especially when minor alterations in the data can determine distinct points of change. As a result, we omit static uncertainties and error information from our conceptual model and reference implementation in favor of dynamical error adjustment in the comparative analysis and appraisal of the comparable multi-proxy results.

While we designed our approach with the clear research focus on Arctic lake systems, we believe that our conceptual approach could be applicable to other research areas, such as long-term based in-situ data record (Su et al. 2018; Zeng et al. 2019). The practical implementation depends on the deliberate selection of the reference frame, i.e. universe of discourse (Elmasri and Navathe 2009), and choosing the appropriate entities and relationships for the abstracted aspects of the real world. Further

on, it is crucial to consider the specific domain knowledge and long-term scientific goals of the harmonization, when converting our conceptual approach into another domain.

2.4 Conclusions

The goal of this study was to provide paleolimnologists with a conceptual framework to integrated heterogeneous multi-proxy data from lake systems. The conceptual data model allows scientists to integrate heterogeneous data into a common database for further comparative analyses. We presented additional steps to prepare datasets for multi-site statistical investigation. We found that heterogeneous structures within the data, differing methods for determining proxy values, and missing error information still pose major challenges in developing a comprehensive data model. However, we concluded that despite strong initial heterogeneity our harmonized dataset still leads to comparable values, enabling numerical inter-proxy and inter-lake comparison.

3

Improving age-depth relationships by using the LANDO ("Linked age and depth modeling") model ensemble

Gregor Pfalz[1,2,3,4], Bernhard Diekmann[1,2], Johann-Christoph Freytag[3,4], Liudmila Syrykh[5], Dmitry A. Subetto[5,6], Boris K. Biskaborn[1,2]

[1]Alfred Wegener Institute, Helmholtz Centre for Polar and Marine Research, Research Unit Potsdam, Telegrafenberg A45, 14473 Potsdam, Germany

[2]Institute of Geosciences, University of Potsdam, Karl-Liebknecht-Str. 24–25, 14476 Potsdam-Golm, Germany

[3]Einstein Center Digital Future, Robert-Koch-Forum, Wilhelmstraße 67, 10117 Berlin, Germany

[4]Department of Computer Science, Humboldt-Universität zu Berlin, Unter den Linden 6, 10099 Berlin, Germany

[5]Department of Physical Geography and Environment, Herzen State Pedagogical University of Russia, Moyka Emb. 48, St. Petersburg 191186, Russia

[6]Institute for Water and Environmental Problems of the Siberian Branch of the Russian Academy of Sciences, Molodezhnayastr.1, Barnaul 656038, Russia

Published: 18 May 2022 in *Geochronology* – DOI: 10.5194/gchron-4-269-2022

Abstract

Age–depth relationships are the key elements in palaeoenvironmental studies to place proxy measurements into a temporal context. However, potential influencing factors of the available radiocarbon data and the associated modeling process can cause serious divergences of age–depth relationships from true chronologies, which is particularly challenging for paleolimnological studies in Arctic regions. This paper provides geoscientists with a tool-assisted approach to compare outputs from age–depth modeling systems and to strengthen the robustness of age–depth relationships. We primarily focused on the development of age determination data from a data collection of high-latitude lake systems (50° to 90° N, 55 sediment cores, and a total of 602 dating points). Our approach used five age–

depth modeling systems (*Bacon, Bchron, clam, hamstr, Undatable*) that we linked through a multi-language *Jupyter Notebook* called LANDO ("Linked age and depth modeling"). Within LANDO we implemented a pipeline from data integration to model comparison to allow users to investigate the outputs of the modeling systems. In this paper, we focused on highlighting three different case studies: comparing multiple modeling systems for one sediment core with a continuously deposited succession of dating points (CS1), for one sediment core with scattered dating points (CS2), and for multiple sediment cores (CS3). For the first case study (CS1), we showed how we facilitate the output data from all modeling systems to create an ensemble age–depth model. In the special case of scattered dating points (CS2), we introduced an adapted method that uses independent proxy data to assess the performance of each modeling system in representing lithological changes. Based on this evaluation, we reproduced the characteristics of an existing age–depth model (Lake Ilirney, EN18208) without removing age determination data. For multiple sediment cores (CS3) we found that when considering the Pleistocene–Holocene transition, the main regime changes in sedimentation rates do not occur synchronously for all lakes. We linked this behavior to the uncertainty within the dating and modeling process, as well as the local variability in catchment settings affecting the accumulation rates of the sediment cores within the collection near the glacial–interglacial transition.

3.1 Introduction

Lake sediments are important terrestrial archives for recording climate variability in the high latitudes of the Northern Hemisphere (Biskaborn et al. 2016a; Smol 2016; Lehnherr et al. 2018; Subetto et al. 2017; Syrykh, Subetto, and Nazarova 2021; Diekmann et al. 2017). The identification of age–depth relationships in those lake sediments helps us to put their measured sediment properties in a temporal context (Bradley 2015; Lowe and Walker 2014; Blaauw and Heegaard 2012). We can determine these relationships by directly counting the annual laminated layers (varves) (Brauer 2004; Zolitschka et al. 2015), or by using indirect age determination methods such as radiocarbon, optically stimulated luminescence (OSL), or lead–cesium (lead-210/cesium-137) dating (Lowe and Walker 2014; Bradley 2015; Appleby 2008; Hajdas et al. 2021). Defining a reliable age–depth relationship for palaeoenvironmental studies in cold regions is particularly challenging, as varves only exist in rare cases and the determination of ages mostly depends on radiocarbon dating (Strunk et al. 2020, and references therein). Because of primarily financial restrictions, however, only a few selected samples are taken from sediment core sections to determine the corresponding ages of certain depths (Blaauw et al. 2018; Ciarletta et al. 2019; Olsen et al. 2017). We therefore rely on model calculations to define the ages between the samples. In addition to the mathematical challenges that arise when establishing age–depth relationships, the selection of appropriate dating material has an impact on the modeling process.

In the special case of Arctic lake systems, the amount of material for radiocarbon dating, i.e. aquatic/terrestrial macrofossils and organic remains, is extremely low (Abbott and Stafford 1996; Colman et al. 1996; Strunk et al. 2020). Radiocarbon dating is therefore often based on the organic

carbon content in bulk sediment samples, which can be relatively small due to the lower bioproductivity in those lakes (Strunk et al. 2020, and references therein). However, the use of bulk sediments is problematic, as some portions of contributing carbon are not occurring at the same time as the deposition but may reveal inherited ages from reworked older materials (Rudaya et al. 2016; Biskaborn, Herzschuh, Bolshiyanov, Schwamborn, et al. 2013b; Biskaborn, Nazarova, et al. 2019; Schleusner et al. 2015; Palagushkina et al. 2017). Several methods are available for pre-treating bulk sediment samples to address sample-based dating uncertainties (Brock et al. 2010; Strunk et al. 2020; Rethemeyer et al. 2019; Bao et al. 2019; M. W. Dee et al. 2020). Each pre-treatment method may yield a different result for the same material due to the influence of humic acids, fulvic acids, and humins (Brock et al. 2010; Strunk et al. 2020; Abbott and Stafford 1996). Similarly, older, inert material incorporated by living organism, known as "reservoir effect" or "hard-water effect", distorts the actual radiocarbon age by up to ±10 000 years (Ascough, Cook, and Dugmore 2005; Austin et al. 1995; Lougheed, Van Der Lubbe, and Davies 2016). Such a distortion creates methodological and mathematical errors in the development of age–depth relationships, which possibly leads to a misinterpretation of these relationships.

There are numerous geochronological software systems (from now on simply called modeling systems) available to the geoscientific community, which try to solve the challenges stated above (Trachsel and Telford 2017; Wright et al. 2017; Lacourse and Gajewski 2020). Methods have been implemented for detecting outliers, accounting for varying sedimentation rates, or using bootstrapping processes to support the construction of an age–depth model (Parnell, Buck, and Doan 2011; Lougheed and Obrochta 2019; Bronk Ramsey 2009, 2008). However, the correct usage of those systems requires a high degree of understanding of the underlying mathematical methods and models. Trachsel and Telford (2017) noted that, despite the users' impact on the outcome of the model by setting priors and parameters, most users do not have any prior objective insights into appropriately choosing the right parameters. Wright et al. (2017), Trachsel and Telford (2017), and Lacourse and Gajewski (2020) even showed that the results produced by modeling systems could diverge from the true chronology. An in-depth comparison of the results is therefore extremely error-prone. Due to time constraints, users usually only select and apply one modeling system for palaeoenvironmental interpretation. However, comparing multiple modeling systems, despite their inherent differences, offers the benefit of reducing biases towards interpreting of age–depth relationships.

The objective of this paper is to reduce the effort involved in applying different methods for determining age–depth relationships and to make their results comparable. We provide a tool to link five selected modeling systems in a single multi-language *Jupyter Notebook*. We introduce an ensemble age–depth model that uses uninformed models to create data-driven, semi-informed age–depth relationships. We demonstrate the power of our tool by highlighting three case studies in which we examine our application for individual sediment cores and a collection of multiple sediment cores. Throughout this paper, the term "LANDO" refers to our implementation, which stands for "Linked age

and depth modeling". The current development version of LANDO is accessible via GitHub (https://github.com/GPawi/LANDO, last access: 20 April 2022).

In this paper, we use published age determination data from 55 sediment cores from high-latitude lake systems (50° to 90° N). This unique collection of age determination data allows us to thoroughly test LANDO by examining changes in sedimentation rates over time for various modeling and lake systems. The harmonization of the acquired data follows the conceptual framework described in Pfalz et al. (2021).

3.2 Methods

A key element in our data-science based approach for developing comparable age–depth relationships was to facilitate the use of modeling systems independent from their original proprietary development environment. A multi-language data analysis environment, such as *SoS Notebook* (Peng et al. 2018) or *GraalVM* (Niephaus, Felgentreff, and Hirschfeld 2019), provides an interface that enables the comparison of modeling systems without being limited to one programming language or environment. Our implementation used SoS Notebook as its backbone. *SoS Notebook* is a native Python- and JavaScript-based *Jupyter Notebook* (Kluyver et al. 2016), which extends to other languages through so-called "Jupyter kernels". We developed our implementation with the focus on four languages and their respective kernels: Python, R, Octave, and MATLAB. This selection allowed us to use the most common modeling systems.

According to Lacourse and Gajewski (2020), the most commonly used modeling systems are *Bacon* (Blaauw and Christen 2011), *Bchron* (Haslett and Parnell 2008; Parnell et al. 2008), *OxCal* (Bronk Ramsey 1995; Bronk Ramsey and Lee 2013), and *clam* (Blaauw 2010). We additionally considered the MATLAB/Octave software *Undatable* (Lougheed and Obrochta 2019), as an alternative to the classical Bayesian approach, and the R package *hamstr* (Dolman 2022).

In our study, we were able to connect five of the abovementioned modeling systems in SoS Notebook, namely *Bacon*, *Bchron*, *clam*, *hamstr*, and *Undatable*. All modeling systems assume a monotonic deposition process, i.e. a positive accumulation rate over the entire core length (Trachsel and Telford 2017; Lougheed and Obrochta 2019). The modeling system clam uses five different regression-based techniques in combination with a Monte Carlo procedure to repeatedly interpolate between calibrated dates. Because *clam* tries to fit the regression curves to the data, in some cases this can lead to age inversions, which *clam* automatically filters out (cf. Trachsel and Telford 2017; Blaauw 2010).

The modeling procedure of *Undatable* involves a weighted random sampling from both calibrated age and depth uncertainties (expressed as a probability density functions) for all dating points and an advanced bootstrapping process over a user-defined number of simulations. The advanced bootstrapping procedure includes removing age inversions from the simulation runs as well as inserting connection points between calibrated dates to account for uncertainties in sediment accumulation rates between the dating points (cf. Lougheed and Obrochta 2019).

The Bayesian modeling systems *Bacon, Bchron,* and *hamstr* subdivide the sediment core into smaller increments for the modeling process but differ in their division technique. *Bacon* separates the core into equal segments, while hamstr extends *Bacon's* algorithm by adding additional hierarchical accumulation structures to each segment (Trachsel and Telford 2017; Dolman 2022; Blaauw and Christen 2011). *Bchron* estimates the number of increments between calibrated dates by a compound Poisson-gamma distribution (Trachsel and Telford 2017; Parnell, Buck, and Doan 2011). For age–depth calculations, *Bacon* uses prior distributions for the accumulation rate (gamma distribution) and autocorrelation memory (beta distribution) between segments, which users can fit with values for the mean and shape of these distributions (Blaauw and Christen 2011). Similarly, *hamstr* relies on user input for the shape of the gamma distribution and values for the memory but estimates the mean value for the accumulation rate from the available age determination data by using a robust linear regression (Dolman 2022). *Bchron* does not require any specific hyperparameter selection due to its fully automated numerical best-fit approach (Wright et al. 2017; Haslett and Parnell 2008). All three Bayesian modeling systems use iterations of the Markov chain Monte Carlo (MCMC) algorithm to estimate the calibrated ages and confidence intervals at each depth within the sediment core (Dolman 2022; Blaauw and Christen 2011; Haslett and Parnell 2008).

The workflow of LANDO consists of five major components: input – preparation – execution – result aggregation – evaluation of model performance.

3.2.1 Input

To work with LANDO users need to provide age determination data, e.g., data from radiocarbon or OSL dating, and associated metadata as listed in Table 3.1. We developed two import options for the users: through a single spreadsheet or a connection to a database. For this study, we used a connection to a PostgreSQL database, which we developed after the conceptual framework as described in Pfalz et al. (2021), via the Python package *SQLAlchemy* (Bayer 2012).We divided age determination input data into two attribute categories: necessary and recommended. The category "necessary" focused on the prerequisites of the individual modeling systems as well as project-related attributes, such as unique identifiers, i.e., "*measurementid*" or "*labid*". However, a larger comprehensive set of descriptive metadata helps a better understanding of the data (Cadena-Vela, Mazón, and Fuster-Guilló 2020; Thanos 2017). We added four additional attributes from the category "recommended" to facilitate the interpretation of age–depth models regarding their age determination data.

If users decide to use a spreadsheet as an input option, then the spreadsheet should follow the same attribution as the database. In addition, we implemented an input prompt for further information, such as the year of core drilling and core length, to ensure comparability to our database implementation. We provide an example spreadsheet with all attributes in the expected format in the repository mentioned in the Appendix A.

Table 3.1: Necessary and recommended attributes for age determination input data, when used with LANDO. Attributes apply for both input methods through either a database or a spreadsheet.

Attribute	Description	Data type	Necessary/ recommended
measurementid	Composite key composed of a unique CoreID, a blank space, and the depth below sediment surface (mid-point cm) with a maximum of two decimal digits of corresponding analytical age measurement – example: "CoreA1 100.5" when users obtained sample of CoreA1 between 100 and 101 cm depth	String	Necessary
thickness	Thickness of the sample slice used for age determination in cm	Float	Necessary
labid	Unique sample identifier that was provided by the laboratory for age determination	String	Necessary
lab_location	Name of city, where laboratory that conducted the analysis resides	String	Recommended
material_category	One of the eight categories that describes the material best, based on the categories from age-depth modeling system Undatable (Lougheed and Obrochta 2019) ^{14}C marine fossil – ^{14}C terrestrial fossil – ^{14}C sediment – tephra – tie point – paleomag. – U/Th – other	String	Necessary
material_description	Short description of the used material	String	Recommended
material_weight	Weight of analyzed carbon used in radiocarbon dating in µgC	Float	Recommended
age	Uncalibrated radiocarbon age in uncal. yr BP or non-radiocarbon ages as values in yr BP (BP: before present (before 1950 CE))	Float	Necessary
age_error	Error of the uncalibrated radiocarbon age or non-radiocarbon age in yr	Float	Necessary
pretreatment_dating	Concise description or abbreviation of sample pre-treatment – example: "ABA" when radiocarbon pre-treatment is comprised of an acid–base–acid sequence	String	Recommended
reservoir_age	Additional reservoir effect (also known as hard-water effect or age offset) identified by the user in yr; if unknown, then insert 0	Float	Necessary
reservoir_error	Error of reservoir age known to the user in yr; if unknown, then insert 0	Float	Necessary

3.2.2 Preparation

The preparation component consisted of two separate steps. First, we checked each age determination dataset to find out whether a reservoir effect was influencing the radiocarbon data. In the absence of a known reservoir age or recent surface sample, we used available radiocarbon data points and a fast-calculating modeling system to predict the age of the uppermost layer within a sediment core. In our approach, we used the hamstr package with a default value of 6000 iterations. We then compared the predicted value for the uppermost layer with the year of the core retrieval, i.e., our target age. We accounted for an uncertainty in the estimate by allowing an extra 10% error between predicted age and target age. If a gap between predicted and target age is observable, then we assumed a reservoir effect is present. We approximated the reservoir effect by subtracting the target age from the mean predicted

age, whereas we based the associated error on the 2σ uncertainty ranges of the prediction. LANDO allows users to add the calculated reservoir age and its uncertainty range to the corresponding attributes ("*reservoir_age*" and "*reservoir_error*"). Depending on the choice of the user, this addition affects either all radiocarbon samples or only bulk sediment samples, or users completely discard the output for the subsequent modeling process.

As the second step in the preparation component, we built a module that automatically changes the format of the available data to the individually desired input of each of the five modeling systems implemented in LANDO. We primarily used the Python package *pandas* (Reback et al. 2020) for the transformation within the module. We transferred the newly transformed age determination data to the corresponding programming language for age–depth modeling using the built-in %get function of *SoS Notebook*.

3.2.3 Execution

We developed LANDO with the specific ability of creating multiple age–depth models for multiple dating series from spatially distributed lake systems. Hence, reducing overall computing time was one of our highest priorities. We achieved this reduction by applying existing parallelization back ends for both R and Python, such as *doParallel* (Microsoft Corporation and Weston 2020b) and *Dask* (Dask Development Team 2016), respectively. For each modeling system in R, we wrote a separate script that takes advantage of the parallelization back end *doParallel*. Besides the individual modeling system packages, we made use of different R libraries, such as *tidyverse* (Wickham et al. 2019), *parallel* (R Core Team 2019), *foreach* (Microsoft Corporation and Weston 2020c), *doRNG* (Gaujoux 2020), and *doSNOW* (Microsoft Corporation and Weston 2020a). We neglected the use of parallelization for the *Undatable* software in MATLAB, since even the sequential execution for several sediment cores in our test setup was on the order of a few minutes. However, we achieved comparable results with *Undatable* in Octave using the parallelization package *parallel* (Fujiwara, Hajek, and Till 2021).

As mentioned before, the selection of model priors and parameters has an impact on the modeling outcome. This is challenging if no objective prior knowledge exists. To lower our impact and to avoid introducing biases in the modeling process, we used the default values from each modeling system as our own default values (Blaauw, Christen, and Aquino-López 2021; Blaauw 2021; Parnell et al. 2008; Dolman 2022; Lougheed and Obrochta 2019). In our adaptation of *clam*, the parameter "*poly_degree*" controls the polynomial degree of models for type 2, while the parameter "*smoothing*" controls the degree of smoothing for types 4 and 5. In the original version of *clam*, users adjust both parameters with the single option "*smooth*" (Blaauw 2021). Furthermore, the default value for "*ssize*" within the original version of *Bacon* is 2000. We increased this value to 8000 to ensure good MCMC mixing for problematic cores, as recommended by Blaauw, Christen, and Aquino Lopez (2021). In the case of the user having in-depth knowledge about their sediment core and wanting to change certain values, we opted for making crucial parameters accessible within *SoS Notebook* outside of the executing

Table 3.2: Default values for each modeling system, which users can access and change within LANDO.

Modeling system	Parameter	Default value
Bacon	acc.shape	1.5
	acc.mean	20
	mem.strength	10
	mem.mean	0.5
	ssize	8000
Bchron	*not applicable*	-
clam	types	1 to 5
	poly_degree	1 to 4
	smoothing	0.1 to 1.0
hamstr	K	c(10,10)
Undatable	xfactor	0.1
	bootpc	30

scripts. Table 3.2 provides an overview of all values which users can access and change for the individual systems. However, we limited the access to some parameters for operational purposes, such as the number of iterations or the resolution of the output.

3.2.4 Result aggregation

After every model run, we received 10 000 age estimates (also known as "iterations" or "realizations") per centimeter from each modeling system for every sediment core. We transferred these results back to Python using the built-in %put function of *SoS Notebook*, where in the next module, we calculated the median and mean age values per centimeter as well as 1σ and 2σ age ranges. For the summarizing statistics, we used standard Python libraries such as *pandas* (Reback et al. 2020) and *numpy* (Harris et al. 2020). We appended the model name as an attribute to the statistics to allocate each result to its modeling system. In addition, we implemented a module, which helped us to push the aggregated result to our initial database to reuse in follow-up research projects. In a similar approach to the input component, we established the connection to our designed PostgreSQL database via the package *SQLAlchemy* (Bayer 2012).

Similarly, we used the 10 000 age estimates per centimeter for calculating the sedimentation rates. Our calculation used three different approaches to calculate sedimentation rates: "*naïve*", "*moving average over three depths*", and "*moving average over five depths*". Table 3.3 lists the appropriate equations for each approach. The user can decide which one of the three approaches best applies to the individual sediment record. We summarized the output into the basic summarizing statistics (mean, median, 1σ ranges, and 2σ ranges) accessible to the users but added the model name and employed approach as additional attributes. If users use more than one sediment core for sedimentation rate calculation, then LANDO will automatically execute the sedimentation rate calculation in parallel using the *Dask* back end (Dask Development Team 2016) and the *joblib* Python package (Joblib Development Team 2020).

Table 3.3: Approaches to calculate sedimentation rates within LANDO. The value represents the layer of interest within a sediment core for which the calculation is necessary. Both x_{i+1} and x_{i+2} are the following layers, while x_{i-1} and x_{i-2} are the previous layers. The unit for the resulting sedimentation rate is centimeter per year (cm yr^{-1}).

Approach	Equation
Naïve (default)	$\text{sedimentation rate}\,(x_i) = \frac{\text{depth}(x_i) - \text{depth}\,(x_{i-1})}{\text{age}(x_i) - \text{age}(x_{i-1})}$
Moving average over three depths	$\text{sedimentation rate}\,(x_i) = \frac{\text{depth}(x_{i+1}) - \text{depth}\,(x_{i-1})}{\text{age}(x_{i+1}) - \text{age}(x_{i-1})}$
Moving average over five depths	$\text{sedimentation rate}\,(x_i) = \frac{\text{depth}(x_{i+2}) - \text{depth}\,(x_{i-2})}{\text{age}(x_{i+2}) - \text{age}(x_{i-2})}$

3.2.5 Evaluation of model performance

To evaluate the performance of each modeling system, we looked at three different case studies:

- Case Study no. 1 – Comparison of multiple modeling systems for one sediment core with a continuously deposited sequence of dating points ("Continuously deposited sequence" – CS1)
- Case Study no. 2 – Comparison of multiple modeling systems for one sediment core with a disturbed sequence (including inversions) of dating points ("Inconsistent sequence" – CS2)
- Case Study no. 3 – Comparison of sedimentation rate changes for multiple sediment cores ("Multiple cores" – CS3).

We examined both sedimentation rate and age–depth modeling results in each of the three case studies. For the first case study, we selected the sediment core EN18218 (Vyse et al. 2021) to showcase the generated output of LANDO. The 6.53m long sediment record obtained from Lake Rauchuvagytgyn, Chukotka (67.78938° N, 168.73352° E; core location water depth: 29.5 m) during an expedition in 2018 consisted of 23 bulk sediment samples used for radiocarbon sampling. The authors determined an existing age offset of 785±31 years BP (years before present, i.e., before 1950 CE), which we used in our modeling process as well.

As a counterexample, for the second case study we have chosen the sediment core EN18208 (Vyse et al. 2020b). During the same expedition to Russia's Far East in 2018, scientists recovered this EN18208 core from Lake Ilirney, Chukotka (67.34030° N, 168.29567° E; core length: 10.76 m; core location water depth: 19.0 m). The authors based their age–depth model on 4 OSL dates and 17 radiocarbon dates from bulk sediment samples as well as an age offset of 1721±28 years BP. However, in addition to the age offset, we included all 7 available OSL and 25 radiocarbon dates for this core in our study.

Both cores are also part of the "Multiple cores" case study with a total of 55 sediment cores (Figure 3.1). More details on each sediment cores are accessible in the corresponding references, which we list in Table 3.4.

Table 3.4: List of all datasets used in this study. Main data source or repository are either the Pangaea database, PaleoLake database, or tables within the main body or supplementary material of publications. Data accessible links to the main data source. Paper reference includes citation to the latest version of the corresponding dataset.

CoreID	PaleoLake database ID	Age-depth model available	Main data source / repository	Data accessible	Paper reference
16-KP-04-L19		Yes	Publication	https://doi.org/10.1111/bor.12521	(Andreev et al. 2021)
2008-3		Yes	Publication	https://doi.org/10.1016/j.quascirev.2012.06.002	(Rudaya et al. 2012)
BC2008		No	Publication	https://doi.org/10.1016/j.rgg.2016.07.005	(Zhdanova et al. 2017)
BL02-2007		No	Publication	https://doi.org/10.1016/j.rgg.2015.05.012	(Khazin et al. 2016)
BN2016-1		Yes	Publication	https://doi.org/10.1177/09596836211019093	(Rudaya et al. 2021)
Chupa-8	295	No	PaleoLake DB	https://clck.ru/N5ksZ – PALEOLAKE DATABASE ID 295	(Kolka et al. 2015)
Co1309	76	Yes	Publication	https://doi.org/10.1111/bor.12379	(Gromig et al. 2019)
Co1412		Yes	Publication	https://doi.org/10.1111/bor.12476	(Baumer et al. 2021)
CON01-603-5		Yes	PANGAEA	https://doi.pangaea.de/10.1594/PANGAEA.856103	(Piotrowska et al. 2004, 2005)
Dolgoe2012	335	No	Publication	https://doi.org/10.7868/S0435428118020049	(Kolka et al. 2018)
EN18208		Yes	PANGAEA	https://doi.pangaea.de/10.1594/PANGAEA.921228	(Vyse et al. 2020b, 2020a)
EN18218		Yes	Publication	https://doi.org/10.5194/bg-18-4791-2021	(Vyse et al. 2021)
ESM-1		Yes	Publication	https://doi.org/10.1016/j.quascirev.2012.03.004	(Mackay et al. 2012)
KAS-1		No	Publication	https://doi.org/10.1017/qua.2017.21	(Anatoly Lozhkin et al. 2017)
Korzhino2010	336	No	PaleoLake DB	https://clck.ru/N5ksZ - PALEOLAKE DATABASE ID 336	(Syrykh, Subetto, and Nazarova 2021)
LENDERY180-4	342	No	PaleoLake DB	https://clck.ru/N5ksZ - PALEOLAKE DATABASE ID 342	(Shelekhova et al. 2021)
LENDERY192	343	No	PaleoLake DB	https://clck.ru/N5ksZ - PALEOLAKE DATABASE ID 343	(Shelekhova et al. 2021)
LENDERY200-1	344	No	PaleoLake DB	https://clck.ru/N5ksZ - PALEOLAKE DATABASE ID 344	(Shelekhova et al. 2021)
LENDERY203-3	345	No	PaleoLake DB	https://clck.ru/N5ksZ - PALEOLAKE DATABASE ID 345	(Shelekhova et al. 2021)
LOT83-7	321	No	PaleoLake DB	https://clck.ru/N5ksZ - PALEOLAKE DATABASE ID 321	(Syrykh, Subetto, and Nazarova 2021)
LS-9		Yes	Publication	https://doi.org/10.1016/S0277-3791(00)00120-7	(Pisaric et al. 2001)
Maloye-1		No	Publication	https://doi.org/10.1017/qua.2017.21	(Anatoly Lozhkin et al. 2017)
MC2006		No	Publication	https://doi.org/10.1016/j.rgg.2015.05.012	(Khazin et al. 2016)
Muan2018	339	No	PaleoLake DB	https://clck.ru/N5ksZ - PALEOLAKE DATABASE ID 339	(Shelekhova and Lavrova 2020)
Okun2018	338	No	Publication	https://doi.org/10.17076/lim1319	(Shelekhova, Tikhonova, and Lazareva 2021)
OSIN	110	No	Publication	https://doi.org/10.17076/lim305	(Tolstobrova et al. 2016)
PER3		Yes	Publication	https://doi.org/10.1007/s10933-015-9858-y	(P. M. Anderson et al. 2015)
PG1111		Yes	Publication	https://doi.org/10.1016/j.quaint.2004.01.032	(Andreev et al. 2004)
PG1205		Yes	PANGAEA	https://doi.pangaea.de/10.1594/PANGAEA.734962	(Wagner et al. 2000b, 2000a)
PG1214		Yes	PANGAEA	https://doi.pangaea.de/10.1594/PANGAEA.734137	(Cremer et al. 2001b, 2001a)
PG1228		Yes	PANGAEA	https://doi.pangaea.de/10.1594/PANGAEA.726591	(Andreev et al. 2003b, 2003a)
PG1238		Yes	Publication	https://doi.org/10.1016/S0277-3791(03)00139-2	(Raab et al. 2003)
PG1341		Yes	Publication	https://doi.org/10.1101/2021.11.05.465756	(von Hippel et al. 2021)
PG1351		Yes	Publication	https://doi.org/10.1046/j.1365-246X.2002.01625.x	(Nowaczyk et al. 2002)
PG1437		Yes	PANGAEA	https://doi.pangaea.de/10.1594/PANGAEA.728450	(Andreev et al. 2005b, 2005a)
PG1746		Yes	PANGAEA	https://doi.pangaea.de/10.1594/PANGAEA.802677	(Nazarova et al. 2013b, 2013a)
PG1755		Yes	Publication	https://doi.org/10.1016/j.quascirev.2010.04.024	(S. Müller et al. 2010)
PG1756		Yes	PANGAEA	https://doi.pangaea.de/10.1594/PANGAEA.708169	(S. Müller et al. 2009, 2008)
PG1856		Yes	Publication	https://doi.org/10.1016/j.gloplacha.2015.07.011	(Hoff et al. 2015)
PG1857		Yes	Publication	https://doi.org/10.1016/j.gloplacha.2015.07.011	(Hoff et al. 2015)
PG1858		Yes	Publication	https://doi.org/10.1007/s10933-012-9580-y	(Hoff et al. 2012)
PG1890		Yes	Publication	https://doi.org/10.1016/j.gloplacha.2015.07.010	(Dirksen et al. 2015)
PG1972		No	PANGAEA	https://doi.pangaea.de/10.1594/PANGAEA.780526	(Biskaborn, Herzschuh, Bolshiyanov, Savelieva, et al. 2013b, 2013a)
PG1975		No	PANGAEA	https://doi.pangaea.de/10.1594/PANGAEA.780385	(Biskaborn, Herzschuh, Bolshiyanov, Schwamborn, et al. 2013b, 2013a)
PG1984		Yes	PANGAEA	https://doi.pangaea.de/10.1594/PANGAEA.776407	(Biskaborn et al. 2012a, 2012b)
PG2023		Yes	PANGAEA	https://doi.pangaea.de/10.1594/PANGAEA.848897	(Biskaborn et al. 2016a, 2016b)
PG2133		Yes	Publication	https://doi.org/10.3389/fevo.2021.625096	(Courtin et al. 2021)
PG2201		Yes	Publication	https://doi.org/10.3389/feart.2021.710257	(Hughes-Allen et al. 2021)
PG2208		Yes	Publication	https://doi.org/10.3389/feart.2021.737353	(Biskaborn et al. 2021)
Tel2006		Yes	PANGAEA	https://doi.pangaea.de/10.1594/PANGAEA.914417	(Rudaya et al. 2016; Rudaya 2020)
Teriberka17	341	No	Publication	https://doi.org/10.17076/lim865	(Tolstobrov et al. 2018)
TKT-3		Yes	Publication	https://doi.org/10.1016/j.quaint.2020.05.023	(Anatoly Lozhkin et al. 2020)
TL-1-1		No	Publication	https://doi.org/10.1191/095968399669823431	(B. B. Wolfe, Edwards, and Aravena 1999)
TULOMA27	23	No	Publication	https://doi.org/10.1016/S0921-8181(01)00118-7	(Corner et al. 2001)
UKhau2015	337	No	Publication	https://doi.org/10.31857/S0869607121060070	(Shelekhova, Lavrova, and Subetto 2021)

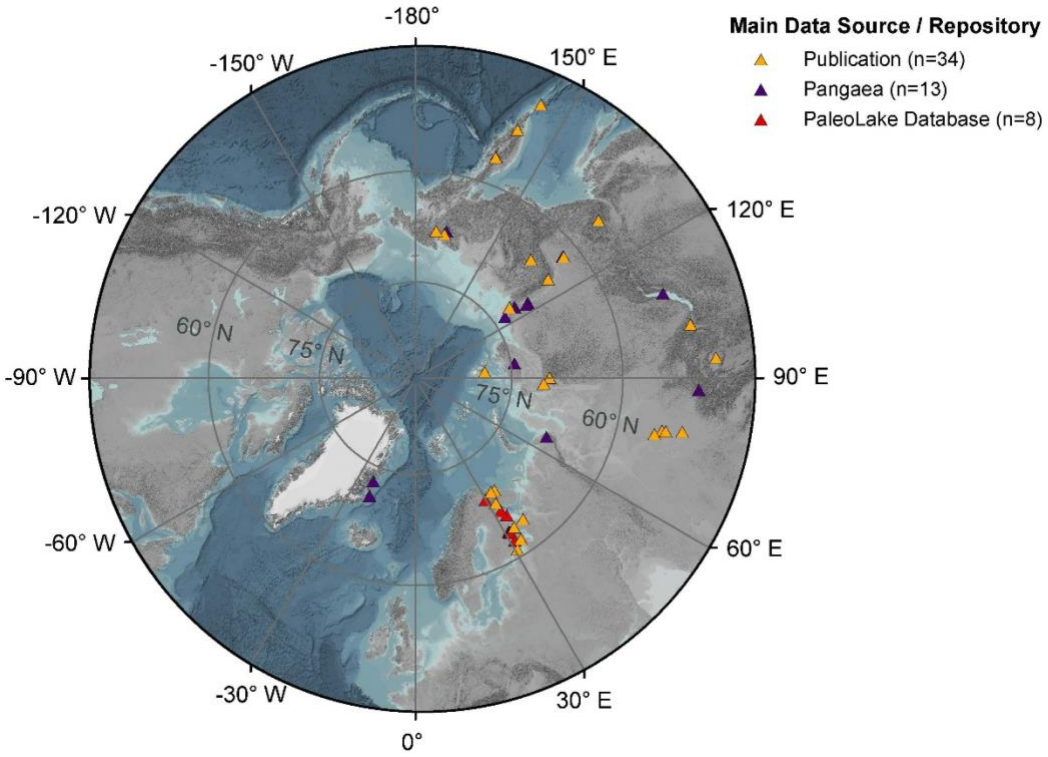

Figure 3.1: Map of the geographical distribution of lake sediment cores used for our study (triangles, n=55). Orange triangles (n=34) represent sediment cores for which we obtained age determination data from a related publication. Purple triangles (n=13) show datasets we collected from the publicly accessible PANGAEA database (Diepenbroek et al. 2002). Red triangles (n=8) indicate referenced datasets provided by the PaleoLake Database (Syrykh, Subetto, and Nazarova 2021). ArcGIS Basemap: GEBCO Grid 2014 modified by AWI. The outer ring in the graphic corresponds to 45° N.

3.2.5.1 Numerical combination of model outputs

To introduce the ensemble model in LANDO, we combined the outputs from all five modeling systems into one composite model. We considered the outermost limits (min and max values) of all confidence intervals (1σ or 2σ) as our boundary for the ensemble model. By taking these outermost limits into account, we artificially increased the area of uncertainty covered by the ensemble model, but we made sure that we were representing all possible outcomes and maximizing the likelihood of including the true chronology. We also included a weighted average ($\bar{x}$) of the age estimates and sedimentation rates, which we calculated using the following equations:

$$\bar{x} = \sum_{k=1}^{m} \frac{n_k}{n} \times \bar{x}_k \qquad \text{Eq. (1)}$$

$$n = \sum_{k=1}^{m} n_k \qquad \text{Eq. (2)}$$

with *m* being the number of participating modeling systems, *n* the total number of iterations, and $\bar{x}_k$ and n_k the median value (either for age estimate or sedimentation rate) and the associated number of

iterations from each modeling system, respectively. In some cases, the weights from each modeling system are equal, as they produce the same number of iterations. Then we can simplify Eq. (1) to represent the arithmetic mean:

$$\bar{x} = \frac{1}{m}\sum_{k=1}^{m} \bar{x}_k \qquad \text{Eq. (3)}$$

For our "Multiple cores" case study (CS3), we additionally had to ensure the comparability of sedimentation rates between sediment cores, since each model assigns a different age value to its sedimentation rate value per centimeter. Therefore, we binned sedimentation rate results into 1000- year bins for each age–depth model as well as the ensemble model and calculated the weighted averages and their confidence intervals within these bins. Inside LANDO, users can change the initial bin size of 1000 years to the desired resolution.

3.2.5.2 Detection and filtering of unreasonable models

For cases in which age–depth models do not agree with each other, e.g., "Inconsistent sequence" case study (CS2), we have built in the option of importing data from measured sediment properties, also known as proxies. Because of compositional and density variations in deposits, changes in sedimentation rates imply changes in the deposition of proxies (Baud et al. 2021; Biskaborn et al. 2021; Vyse et al. 2021). By including appropriate, independent proxy data on lithological changes within the sediment core, we can weight each model based on its performance to represent these variations in sedimentation rate. Users should provide the independent sediment proxy data as a file with two columns, namely "*compositedepth*", which should be the measurement depths (as mid-point centimeter below sediment surface), and "*value*", representing the values of the proxy. This simplification makes it possible to import different available proxies or statistical representations of proxy data, i.e., results from ordination techniques (PCA, MDS, etc.), into the optimization process and to visualize the behavior of the age–depth models in comparison to these proxies.

In order to evaluate the performance, we adapted the fuzzy change point approach by Hollaway et al. (2021) to work with our input data and desired outcome on a depth-dependent scale instead of a time series. Similarly to Hollaway et al. (2021), our approach firstly detected change points within the proxy data and each modeling system output by fitting an ARIMA (autoregressive integrated moving average) model to the data and then extracted change points by using the *changepoint* R package (Killick and Eckley 2014; Killick, Haynes, and Eckley 2016) on the residuals of the ARIMA model. If we found no change points in the proxy data via this approach, we applied the *changepoint* R package on the raw independent sediment proxy data instead. Through the additional bootstrapping process introduced by Hollaway et al. (2021), we were able to set up confidence intervals for the extracted change points. Subsequently, we searched for the intersection between the change points plus their confidence interval for each age–depth model with the independent proxy data. After converting the change points for both age–depth model and independent proxy data into triangular fuzzy numbers, we obtained similarity

scores using the Jaccard similarity score of the fuzzy number pairs as described in Hollaway et al. (2021). The similarity score can reach numbers between 0 (no match) and 1 (perfect match). However, the threshold of excluding an age–depth model from the generated combined model depends on the imported proxy data and number of detected change points. Therefore, the user can set the threshold accordingly to their proxy within LANDO, but we have implemented the default value for this threshold as 0.1, which corresponds to an overlap of 10% of the change points between model and proxy data.

In addition to the criterion of preparing the proxy data in the format of depth vs. value in a separate file, we suggest using a proxy with a high resolution. As a high-resolution proxy, we define a proxy with more than 50 measurements per meter of core length. For our "Inconsistent sequence" case study (CS2), we used high-resolution elemental proxy data from XRF (X-ray fluorescence) measurement as our independent proxy data. As our evaluation element to optimize the age–depth models, we selected zircon (Zr), which itself is an indicator for minerogenic/detrital input (Vyse et al. 2020b, and references therein). The zircon proxy data of EN18208 have a resolution of 200 measurements per meter of core length.

To achieve a realistic comparison between sediment cores in the "Multiple cores" case study (CS3), we looked at the individual age–depth model outputs for each sediment core to determine whether an optimization step was required. We have only selected sediment cores with a published age–depth model (n=33) so that we can refer to lithological boundaries from the original publication. During the analysis, we saw that nine sediment cores needed to be optimized due to strong inconsistencies between models over the entire length of each core. In 12 cases, where models within the lower section of the cores did not match, we considered proxy-based optimization to improve the model outcome when high-resolution data were available.

3.2.5.3 Display of models

To display the results from age–depth modeling and sedimentation rate calculation, we decided to create our own plots, instead of reusing the plots from each individual modeling system. Our plot header contains the unique CoreID; additionally, the header indicates whether the user decided to apply a reservoir correction to the radiocarbon data or not. Our single core plots consist of two main panels: on the left-hand side, the panel shows the results from the age–depth modeling process with the calibrated ages (in calibrated years BP) on the x axis and the composite depth of the sediment core (in centimeters) on the inverted y axis. On the right-hand side, the panel displays the result from the sedimentation rate calculation (in cm yr^{-1}, centimeter per year) on the x axis plotted against the same composite depth on the inverted y axis. For better readability of the strong variability of sedimentation rate, we used the log scale for the x axis of the right panel. Generally, LANDO draws the ensemble age–depth model and sedimentation rate in gray with the weighted average as a dashed line.

For all models, LANDO will display the median values for age and sedimentation rate as solid lines. Both panels further display the corresponding 1σ range and 2σ range per centimeter for each

model. Depending on the user's selection, users can plot both sigma ranges, only one of the two sigma ranges, or just the median ages. To include age determination data within the plots, LANDO internally calibrates the radiocarbon data with the BchronCalibrate function of the *Bchron* package (Haslett and Parnell 2008; Parnell et al. 2008) with either the IntCal20 (Reimer et al. 2020), Marine20 (Heaton et al. 2020), or SHCal20 (Hogg et al. 2020) calibration curve. This allows users to analyze samples from locations other than the terrestrial Northern Hemisphere. By default, the left panel contains each age data point as a predefined symbol with its 1σ uncertainty as an error bar. The symbol used by LANDO depends on the material category defined in the input file for each dating point.

If users decide to filter out unreasonable age–depth models, similar to the "Inconsistent sequence" case study (CS2), we added the option to plot the independent proxy data and therefrom derived lithology as an additional panel on the left-hand side for a better interpretability. Further, LANDO highlights the boundaries of lithological change and its confidence interval in both sedimentation rate and age–depth model plots. The optimized plot includes a goodness of fit for each involved modeling system to represent the change points at the bottom of the plot.

When using LANDO for multiple sediment cores, for each sediment core, the overall plot holds the results from the binned weighted average sedimentation rate calculation (as median sedimentation rate in cm yr^{-1}, centimeter per year) against the selected age bins (in calibrated years BP) for each modeling system. This visual illustration allows user to compare multiple sediment cores based on the time axis.

For people with color vision deficiency, we incorporated the extra option to plot the resulting age–depth plots with different line styles and textures to support the visual differentiation between each model. Figure C4 in the Appendix shows the color-blind friendly output created by LANDO. With LANDO we want to support inclusivity in science, but we look forward to feedback from the community on how we can improve LANDO in this regard.

3.2.6 Further analysis – sedimentation rate development over time

To identify similar temporal shifts in sedimentation regimes in our case study "Multiple cores" (CS3), we examined our data collection of 55 sediment cores regarding a general tendency in sedimentation rate shifts. First, we considered the 11 700 years BP boundary as our marker for the change between Holocene and Late Pleistocene to separate the datasets (S. Olander Rasmussen et al. 2006; Lowe and Walker 2014; Walker et al. 2008). We selected this marker because numerous studies suggest a general difference in sedimentation regimes between these periods (e.g., Baumer et al. 2021; Bjune et al. 2021; Kublitskiy et al. 2020; Müller et al. 2009; A. P. Wolfe 1996; Vyse et al. 2021). As some of the models were below the 11 700 years BP marker, the calculation of the mean sedimentation rate for the Late Pleistocene featured only a subset of sediment cores (total number of sediment cores with measurement in Late Pleistocene: 20). Then, for each age model of the sediment cores in the subset, we used the 2σ ranges around 11 700 years BP to determine whether the maximum absolute change occurred exactly at

11 700 years BP or around our set marker. For this investigation, we changed the bin size to 100-year bins to allow comparison between each modeling system and the combined models. Using the maximum from the interquartile ranges of the 2σ ranges for each model (see Figure C3 in the Appendix), we defined the observation period from 8700 to 14 700 years BP (corresponds to a range of ±3000 years). We then checked the data within the time span to see where the maximum change in sedimentation rate occurred. If the calculated age for the new marker was at the edge of our time span, we iteratively increased the outer limit by 100 years (up to a maximum of 18 000 years BP) to see if the calculated age still reflected the maximum absolute change. We then used the newly defined marker to calculate the mean sedimentation rate for before and after the marker.

3.3 Results

3.3.1 "Continuously deposited sequence" – Case Study no. 1

All five age–depth models were able to produce an age–depth relationship for sediment core EN18218 (Lake Rauchuvagytgyn) with only small diversions in between some of the calibrated ages. Figure 3.2 depicts the two visual outputs produced by LANDO. Figure 3.2a displays all models side by side, while Figure 3.2b shows the combined output from all models.

All models revealed the highest sedimentation rates for the interval between 108 and 133 cm. Mean values ranged from 0.242 cm yr^{-1} (*hamstr*) to 0.764 cm yr^{-1} (*clam*) within this interval, whereas the median sedimentation rate varied between 0.107 cm yr^{-1} (*Bacon*) and 0.314 cm yr^{-1} (*clam*). In the lower segment of EN18218 (653 to 504 cm), the models showed a stronger disagreement among each other with larger varying mean and median values for sedimentation rate. In three instances, the majority of models noticeably dropped to lower sedimentation rate values. We found the first two declines in sedimentation rate between 366 and 339 cm and between 249 and 222 cm with median sedimentation rates from 0.012 cm yr^{-1} (*hamstr*) to 0.027 cm yr^{-1} (*Bacon*) and from 0.013 cm yr^{-1} (*hamstr*) to 0.025 cm yr^{-1} (*Bacon*), respectively. The last significant downward shift occurred between 66 and 57 cm, where *hamstr* decreased the median sedimentation rate 10-fold from 0.15 to 0.015 cm yr^{-1} between 66 and 64 cm. In our ensemble model, we found the highest value for weighted average sedimentation rate at 128 cm with 0.4483 cm yr^{-1} (2σ range: 0.032–2.338 cm yr^{-1}), which corresponded to a weighted average age estimate of 4846 cal yr BP (2σ range: 4301–5384 cal yr BP). Throughout the core, the cumulative 2σ uncertainty of the ensemble model ranged from 0.002 to 2.486 cm yr^{-1}.

3.3.2 "Inconsistent sequence" – Case Study no. 2

For the second case study, we considered an example where the underlying age determination data within the core are very contradictory to each other (see Figure 3.3). Before considering modeling such an age–depth relationship with conflicting data, users need to investigate and try to understand the reasons for any outliers. Fitting any age–depth model, including the LANDO ensemble, to such

divergent data should be done with extreme caution, and we do not recommend doing so without further deliberate investigation. Here we primarily aim to illustrate the range of age–depth models obtained within the ensemble as well as the results of the optimization with our proxy-based lithology.

During the standard modeling procedure with LANDO, four out of five modeling systems produced an output for sediment core EN18208 (Lake Ilirney). The modeling system *clam* was unable to produce an age–depth model for this core. Figure 3.3 shows the visual outputs with all models in panel (a) and the combined model in panel (b). Figure 3.4 consists of three panels showing the results from the proxy-based optimization process using zircon (Zr). Figure 3.4a shows the visual output from the optimization process, while Fig. 3.4b and c illustrate the optimized age–depth model with the highest matching score and the resulting ensemble model, respectively.

While *Undatable* was the only modeling system that considered the dating point at 1066 cm before following the next dating point at 966 cm, all remaining three modeling systems assumed a steady accumulation (mean sedimentation rate: 0.0575 cm yr^{-1}) from 1076 cm before their paths overlapped with *Undatable*. At the depth of 795 cm, we found the next divergence between the age–depth models. *Undatable* followed the younger OSL dates and the young radiocarbon date at 666 cm. *Bacon*, *Bchron*, and *hamstr* continued with the radiocarbon date at 561 cm before taking different paths until the age determination point at 184 cm. All modeling systems's paths again overlapped from 184 cm to the sediment surface with a mean sedimentation rate of 0.0277 cm yr^{-1}.

During the optimization process, our adapted algorithm located four lithological boundaries with their uncertainty ranges from the independent proxy data: 189.5 cm (182– 192.5 cm), 646 cm (638–657 cm), 890.5 cm (874–912 cm), and 1051.5 cm (1043–1061.5 cm). We found the highest matching score from the optimization for hamstr (Score: 0.0237). Table 5 shows the average sedimentation rate for each proxy-derived lithological unit (PLU) of the ensemble model of EN18208.

3.3.3 "Multiple cores" – Case Study no. 3

In contrast to the previous case studies, this case study focused on understanding the development of sedimentation rates over time, with the emphasis on the transition from the Holocene to the Pleistocene. We used age determination data from 33 sediment cores with a published age–depth model to show the standard output of LANDO for multiple sediment cores, while using all datasets for the subsequent analyses. Figure 3.5 shows the ensemble models with weighted average sedimentation rates binned into 1000-year bins from our multi-core investigation with 33 published sediment cores (see Fig. C1 for the individual models in the Appendix). We set the boundaries from 0 to 21 000 cal BP within these figures to cover the time span from the present to the Last Glacial Maximum (LGM) (Clark et al. 2009). Below the number for each core in Figure 3.5 are the proxies used for their optimization. In 17 out of 55 cases within our entire collection, the ensemble model was based on four out of five models, as neither *clam* or *Undatable* was able to find a suitable age–depth model (for more details, please see Table C1 in the Appendix).

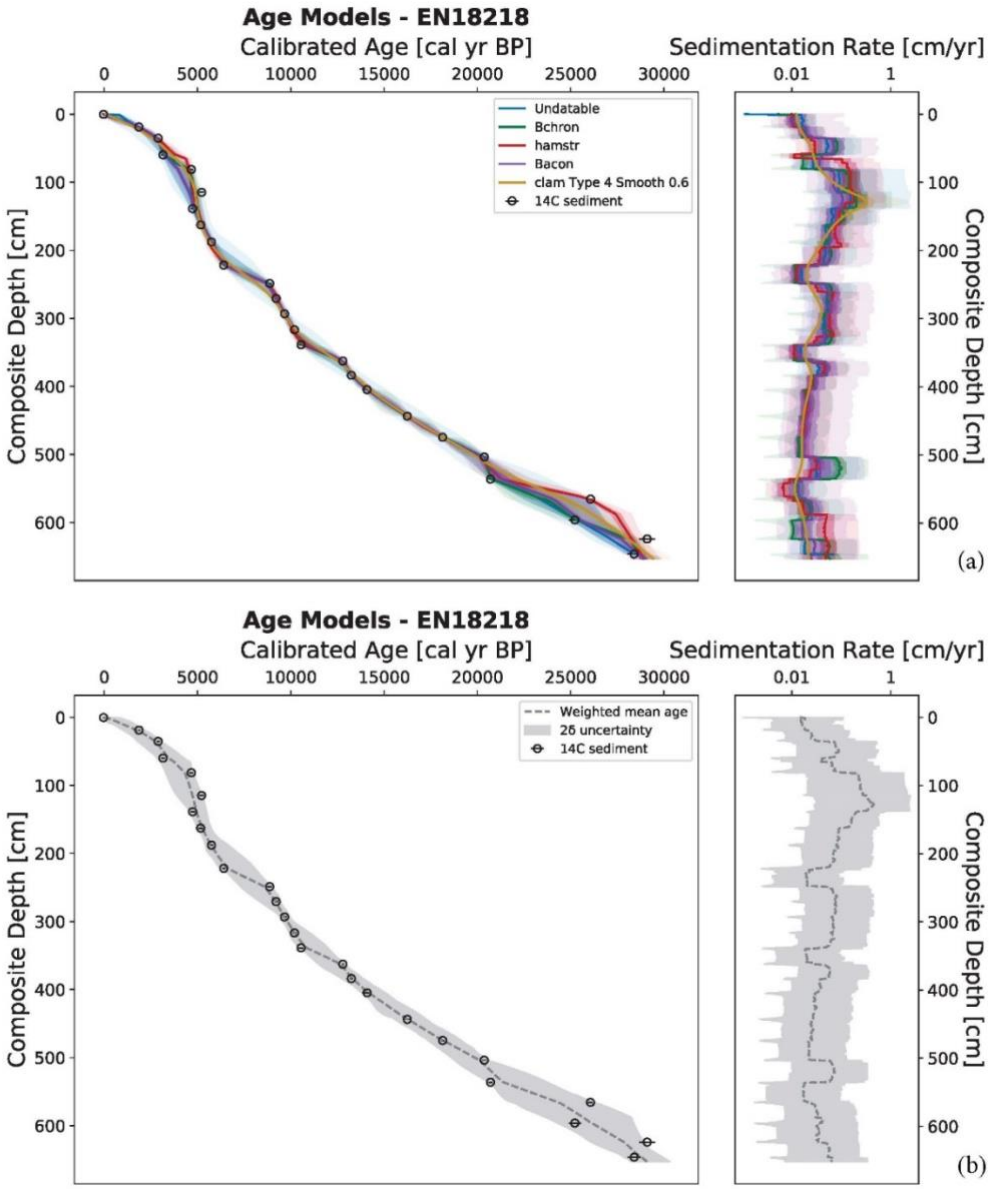

Figure 3.2: Generated output from LANDO for sediment core EN18218 (^{14}C data from Vyse et al. (2021)) as an example of continuous lacustrine sedimentation over time. Panel (a) consists of a comparison between age-depth models from all five implemented modeling systems (left plot) and their calculated sedimentation rate (right plot). Colored solid lines indicate both the median age and median sedimentation rate for all models, while shaded areas represent their respective 1σ and 2σ ranges in the same colors with decreasing opacities. Panel (b) shows the ensemble age-depth model (left plot) and its sedimentation rate (right plot). The dashed line in panel (b) represents the weighted average age estimates (left plot) and the weighted average sedimentation rates (right plot) for the ensemble model, while the grey area represents the 2σ uncertainty, i.e., the outermost limits of 2σranges from all models. Both plots on the left of (a) and (b) show the depth below sediment surface on the inverted y axis as composite depth of the sediment core in centimeter (cm) and the calibrated ages on the x axis in calibrated years Before Present (cal. yr BP, i.e., before 1950 CE). Black circles within (a) and (b) indicate the calibrated ^{14}C bulk sediment samples with their mean calibrated age using the IntCal20 calibration curve (Reimer et al. 2020) and their 1σ uncertainty as error bars. The plots on the right display the sedimentation rate in centimeter per year (cm yr^{-1}, x axis as log-scale) against the depth below sediment surface as the composite depth of the sediment core in centimeter (cm, inverted y axis).

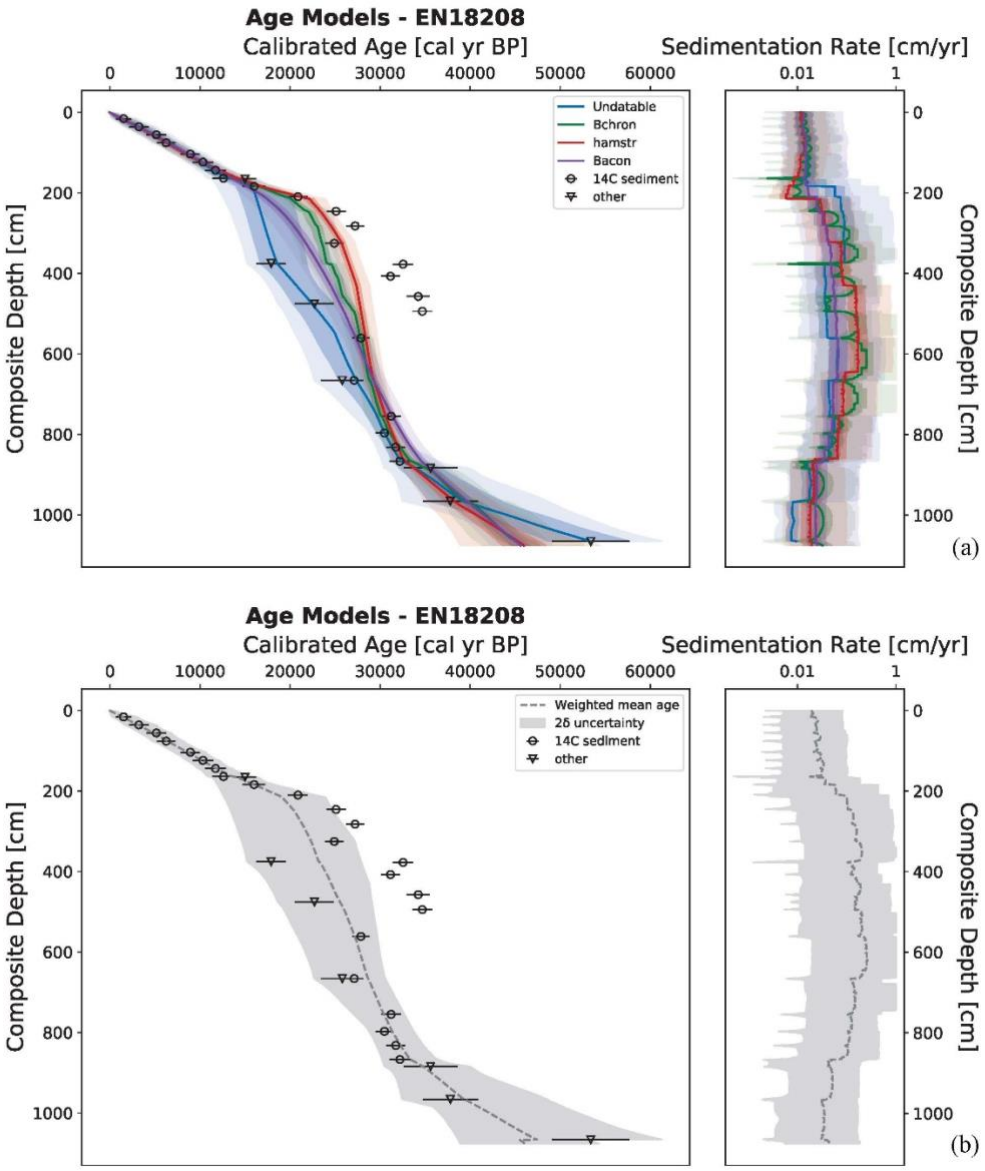

Figure 3.3: Generated output from LANDO for sediment core EN18218 (OSL and ^{14}C data from Vyse et al. (2020b)) as an example of discontinuous lacustrine sedimentation. Panel (a) consists of a comparison between age-depth models from four out of five implemented modeling systems (left plot) and their calculated sedimentation rate (right plot). The modeling system clam was unable to produce an age-depth model for this core. Colored solid lines indicate both the median age and median sedimentation rate for all four models, while shaded areas represent their respective 1σ and 2σ ranges in the same colors with decreasing opacities. Panel (b) shows the ensemble age-depth model (left plot) and its sedimentation rate (right plot). The dashed line in panel (b) represents the weighted average age estimates (left plot) and the weighted average sedimentation rates (right plot) for the ensemble model, while the grey area represents the 2σ uncertainty, i.e., the outermost limit of 2σ ranges from all four models. Both plots on the left of (a) and (b) show the depth below sediment surface on the inverted y axis as composite depth of the sediment core in centimeter (cm) and the calibrated ages on the x axis in calibrated years Before Present (cal. yr BP, i.e., before 1950 CE). Black circles within (a) and (b) indicate the calibrated ^{14}C bulk sediment samples with their mean calibrated age using the IntCal20 calibration curve (Reimer et al. 2020) and their 1σ uncertainty as error bars. Black down-pointing triangles show mean ages from OSL analysis and their 1σ uncertainty as error bars. The plots on the right display the sedimentation rate in centimeter per year (cm yr^{-1}, x axis as log-scale) against the depth below sediment surface as the composite depth of the sediment core in centimeter (cm, inverted y axis).

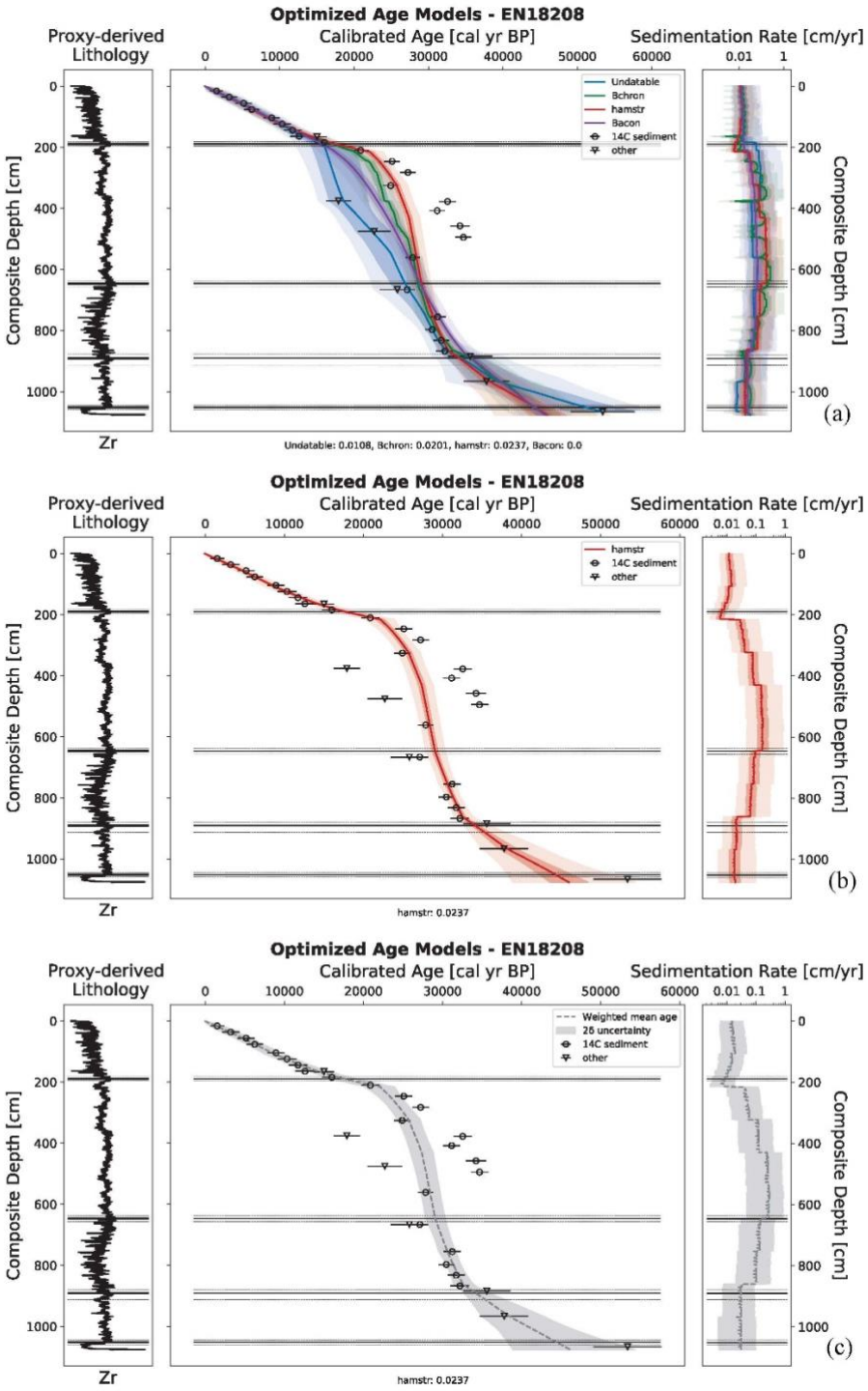

Figure 3.4: Optimized visual output for EN18208 (OSL and ^{14}C data from Vyse et al. (2020b)). We used high-resolution X-ray fluorescence (XRF) measurements of zircon (Zr) as independent proxy to evaluate model performance to represent lithological changes. Panel (a) extends the existing panel (a) of Figure 3.3 by adding a plot on the left to show the proxy-derived lithology used to filter unreasonable models. This added plot consists of the proxy measurements of Zr (in counts per second) along the depth below sediment surface as the composite depth of the sediment core in centimeter (cm) and the derived lithological boundaries (solid horizontal lines) plus their uncertainty range (dashed horizontal lines). Both age-depth model and sedimentation rate plot contain the same lithological boundaries as visual aid. The text box in the bottom middle lists the models with their matching score related to the proxy-derived lithology. Panel (b) shows the model (*hamstr*) with the highest matching score (0.0237). Panel (c) depicts our ensemble model based on this model. The age-depth models displayed in panel (b) and (c) show strong similarities with the age-depth model developed by Vyse et al. (2020b).

Table 3.5: Average sedimentation rate of EN18208 divided into proxy-derived lithological units. The calibrated mean model range indicates the mean age estimates of the ensemble model for the corresponding depths of the proxy-derived lithological unit (PLU).

Proxy-derived lithological unit	Corresponding depths below sediment surface [cm]	Calibrated mean model range [cal yr BP]	Average sedimentation rate [cm yr^{-1}]
PLU1	0 – 190	-67 – 17752	0.0152
PLU2	190 – 646	17752 – 29073	0.1664
PLU3	646 – 891	29073 – 34244	0.1073
PLU4	891 – 1052	34244 – 44499	0.0307

The maximum time span covered by the sediment cores varied between 2000 years BP (CoreID: PG1972) and 320 000 years BP (CoreID: PG1351). The average non-optimized sedimentation rate ranged between 0.004 cm yr^{-1} (CoreID: LOT83-7) and 1.142 cm yr^{-1} (CoreID: PG1228). In total, we optimized seven sediment cores, as in most cases high-resolution data were not available nor did the provided proxy data represent a lithological proxy when crosschecked with the original publication. From these seven sediment cores, we reconstructed the proxy-based lithology twice with TOC (total organic carbon) as a low-resolution proxy (CoreID: PG1228 & PG1437).

To visualize the difference in sedimentation rates between two neighboring and fundamentally different environmental settings, i.e. Pleistocene glacial and Holocene interglacial, we used the datasets that were split at the Holocene–Pleistocene boundary at 11 700 years BP. Figure 3.6 shows the mean sedimentation rate for the Holocene and Late Pleistocene for each model with its 1σ uncertainty. Figure C2 in the Appendix gives an overview over the overall uncertainty for all models. Among all models, *clam* models have the lowest range on average for both Holocene (0.0135 cm yr^{-1}) and Late Pleistocene (0.0011 cm yr^{-1}), while the combined models show the greatest uncertainty on average in the Holocene (0.0942 cm yr^{-1}) and for the Late Pleistocene (0.0711 cm yr^{-1}). The sediment core PG1228 (latitude: 74.473° N) showed the highest individual sedimentation rate for the Holocene in *Undatable* (median sedimentation rate: 1.1013 cm yr^{-1}). We observed a significant reduction of about 77% for the optimized model of the same core (0.1264 cm yr^{-1}), compared to its combined model (0.5615 cm yr^{-1}).

For our data compilation, we found the largest absolute change in sedimentation rates within the modeling systems on average between 9600 and 11 900 years BP (Figure 3.7). For our combined and optimized models, however, the largest change averaged between 10 500 and 10 700 years BP. Still, all sediment cores covered the entire range of our initial time span from 8700 to 14 700 years BP within the models. Using the results of the largest change in sedimentation rate for each sediment core and model as new markers, we again split the datasets into two separate datasets. One dataset contained mostly Holocene sedimentation rate values (Holocene dataset), while the other contained mostly Late Pleistocene values (Late Pleistocene dataset). Therefore, the initial display (Figure 3.6) changed slightly to Figure 3.8. The increase in total number of sediment cores in the Late Pleistocene dataset with an individual separation (n=38) compared to the Late Pleistocene dataset with the separation at 11 700 years BP (n=19) was most notable.

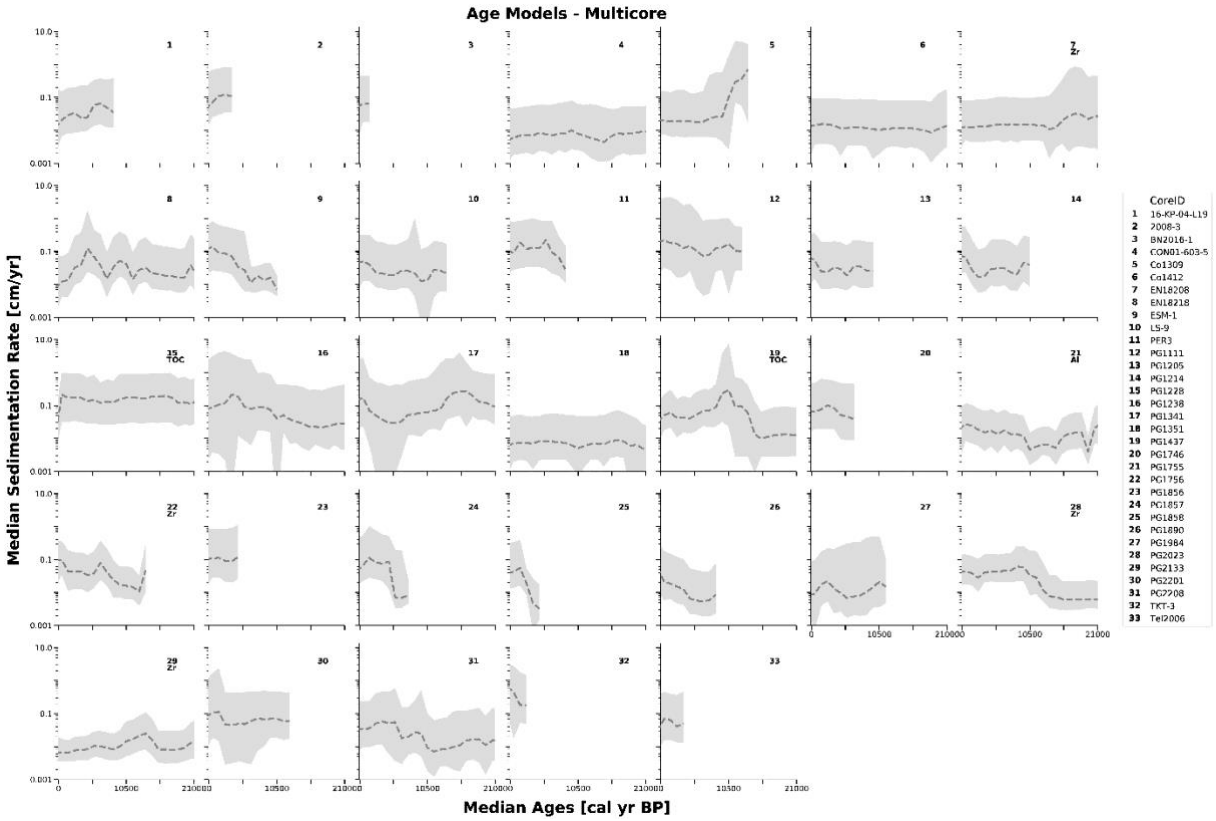

Figure 3.5: Optimized combined models for 33 sediment cores with a published age-depth model displayed as weighted average sedimentation rate (in centimeter per year, cm yr^{-1} – y axis) binned into 1000-year bins (in calibrated years Before Present, cal. yr BP, i.e. before 1950 CE – x axis) for the last 21 000 years. Dashed line represents the weighted average sedimentation rate, whereas the grey areas are the respective two-sigma ranges. Each grid cell contains the unique core identifier of each involved sediment core. In seven cases, the letters below each number give the name of the independent proxy used for optimization process.

3.4 Discussion

3.4.1 Assessment of different case studies

By comparing the cases for the two single-sediment cores, it becomes clear how age–depth relationships may diverge depending on the individual modeling system and its treatment of available dating points (cf. Wright et al. 2017; Trachsel and Telford 2017; Lacourse and Gajewski 2020). In the case of EN18218 ("Continuously deposited sequence" – CS1), all five implemented modeling systems yield an agreeing and continuous chronology. However, the two radiocarbon dates at 81.25 and 114.75 cm have a significant impact on the model's interpretation for these depths. Vyse et al. (2021) argued that these two dates are outliers resulting from reworking and mixing effects within the sediment column. According to the authors, no additional proxy data from EN18218 would support the immediate increase in sedimentation rate for these depths, and, hence, they excluded both dates from the modeling process. Because we are not considering any additional proxy data to evaluate age–depth models in their geoscientific context but rather include all provided age determination data in the modeling process, the

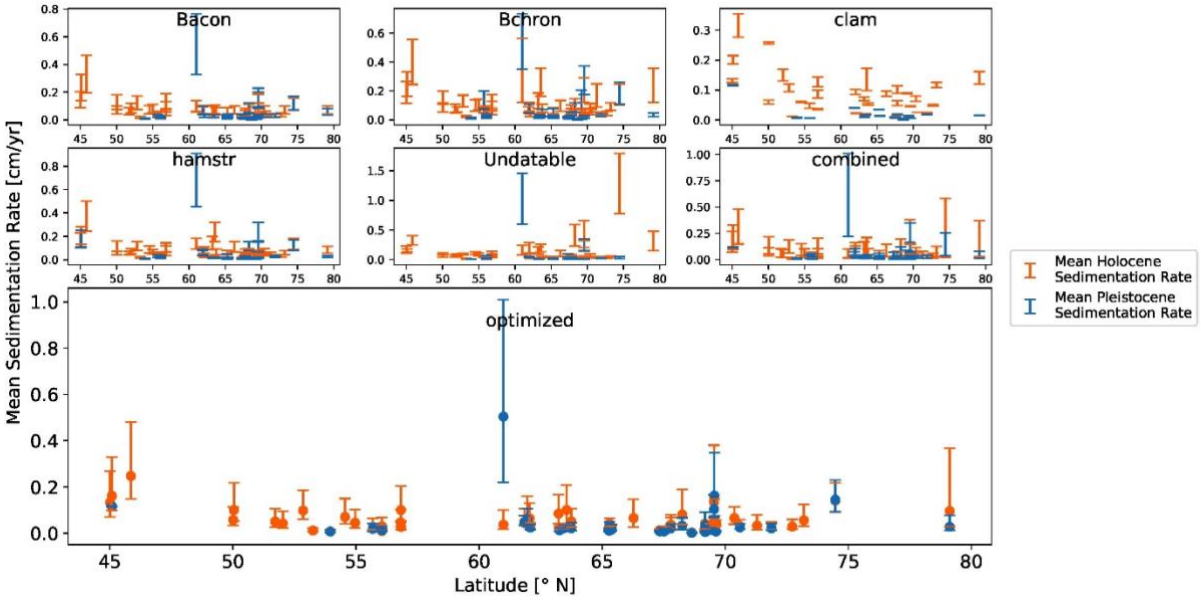

Figure 3.6: Average sedimentation rate in centimeter per year (cm yr^{-1}) for each sediment core in our data collection of 55 sediment cores divided into Holocene dataset (from present to 11 700 yr BP, orange lines) and Late Pleistocene dataset (from 11 700 yr BP to 21 000 yr BP, blue lines). Each plot displays the one-sigma range of sedimentation rate within each dataset for each model and sediment core. In addition, filled circles represent the mean value for the optimized models.

consideration of these two radiocarbon dates on the basis of all available models leads to a higher sedimentation rate. Nonetheless, the example here shows how the comprehensive application of the different modeling systems may help to identify doubtful dating points.

We saw a disagreement between the modeling systems in the case of sediment record EN18208 ("Inconsistent sequence" – CS2), which we expected prior to the execution of our application, due to the scattered dating points in the original data. Vyse et al. (2020b) linked this scatter of age data points observed in the interval between 282 and 755 cm of EN18208 to the redeposition of older carbon. They implied that to produce a reliable age–depth model they had to exclude both OSL and radiocarbon dating points for these depths. However, our optimized combined model agrees with their established age–depth model and can reproduce the characteristics of the existing model by Vyse et al. (2020b), without removing dating points. In addition, in three out of four cases, our proxy-derived lithology with its uncertainty matches the lithological boundaries set by the authors of the EN18208 study, according to criteria based on acoustic sub-bottom profiling. Only the first original boundary (196 cm) is outside our confidence interval from 182 to 192 cm. We still showed that our approach could set logical boundaries for sediment cores by solely relying on high-resolution proxy data.

Despite a strong similarity between our optimized model and the existing model developed by Vyse et al. (2020b), the highest score showed a low similarity value (0.0237) using our similarity scale from 0 (no match) to 1 (perfect match). Although we chose the highest matching score to demonstrate LANDO's ability of filtering out disagreeing models, we do not support the strategy of choosing a single age–depth model with such a low matching score. Rather, users should investigate the cause of the

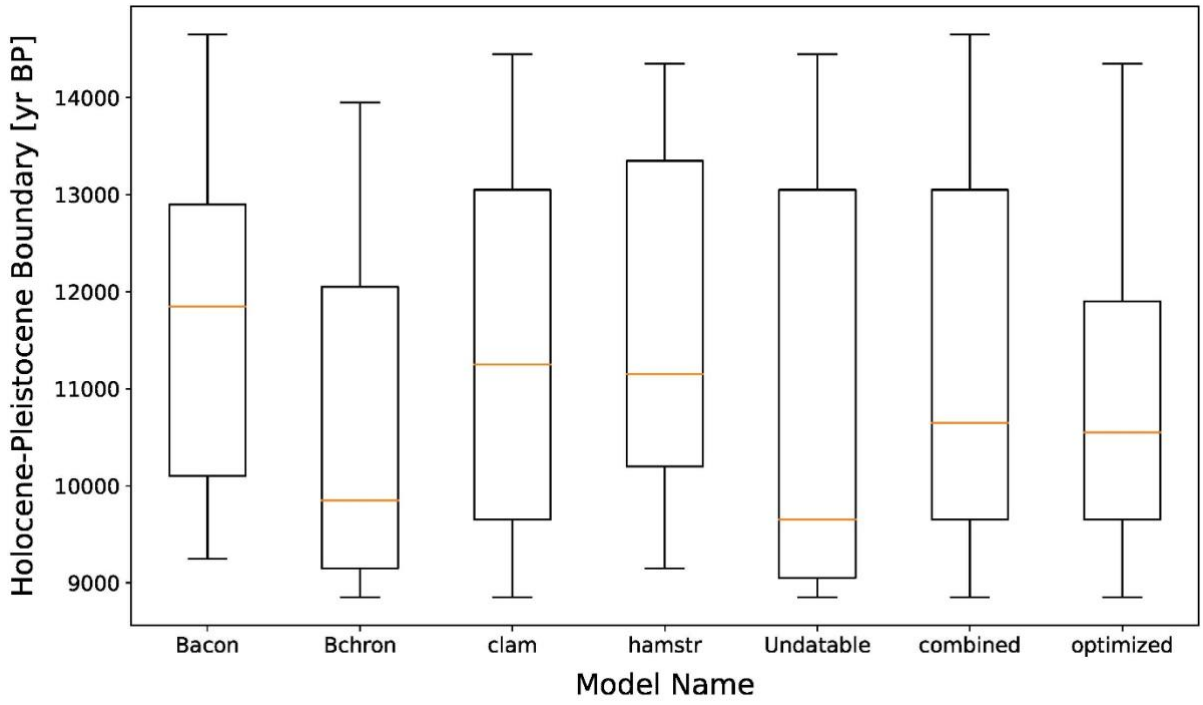

Figure 3.7: Boxplot representing the years with the biggest absolute change in sedimentation rate for our data collection of 55 sediment cores. Sedimentation rate results from each model binned into 100-year bins to allow comparisons between the modeling systems. The initial observation time span covers 8700 to 14 700 yr BP. The orange line corresponds to the median value for each model.

scatter in the age determination data and/or change the default values within LANDO. For example, to deal with the scatter in the data, users can increase the *Undatable* parameter "*bootpc*" to a higher value – as suggested by Lougheed and Obrochta (2019) – to account for a higher uncertainty in the given data. For palaeoenvironmental reconstruction, users should also propagate these increased uncertainties into their proxy interpretation, which is often underrepresented (Lacourse and Gajewski 2020; McKay, Emile-Geay, and Khider 2021).

Even though LANDO can produce age–depth models for multiple sediment cores ("Multiple cores" – CS3), we must assume limitations in the geoscientific validity for some of the results. In a few cases, an optimization of age–depth models with independent proxy data would have been necessary, but such independent data were inaccessible or did not exist. As for these cases age–depth relationships between implemented modeling systems seem to disagree (see Figure C1 in the Appendix), the results from our combined model might over- or underestimate the true sedimentation rate. On the other hand, optimization using proxy data can reduce these biases.

For instance, during the examination of the Holocene and the Pleistocene sedimentation rates (Figure 3.6), we noticed that one sediment core (PG1228) had an extremely high mean sedimentation rate for the Holocene dataset in *Undatable*. Similar to the second case study ("Inconsistent sequence" – CS2), we found scattered age data points for this sediment core, which influenced the modeling process of *Undatable*. Further, the result then affected our combined model by increasing the overall

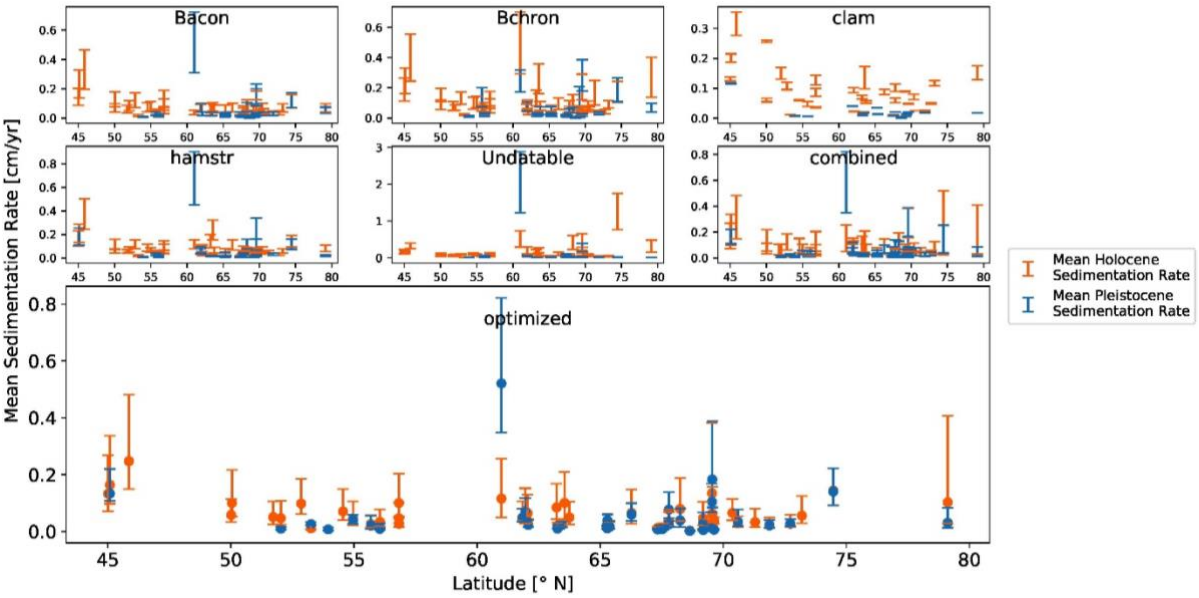

Figure 3.8: Average sedimentation rate in centimeter per year (cm/yr) for each sediment core in our data collection of 55 sediment cores divided into Holocene dataset (orange lines) and Late Pleistocene dataset (blue lines). The exact value for the split of the datasets for each individual core and each model depends on the results of the maximum change in sedimentation rate within the observation period 8700 to 14 700 yr BP. Each plot displays the one-sigma range of sedimentation rate within each dataset for each model and sediment core. In addition, filled circles represent the mean value for the optimized models.

sedimentation rate for the Holocene in this core. However, LANDO identified the *Undatable* model as an outlier based on the lithology established through independent TOC proxy data. The optimized model then agreed well with the original publication by Andreev et al. (2003b), which further increased the validity of our approach. Our findings suggest that high-resolution proxy data should accompany geochronological studies to enable a more concise and realistic assessment of the development of sedimentation rates over time in high-latitude lake systems.

We further improved the validity of some results of our multi-core study by comparing our LANDO output with the available age–depth models from publications. In four cases (CoreID: 2008-3, Co1309, LS-9, PG1205), we adjusted our initial output to the previously published age–depth models (Rudaya et al. 2012; Gromig et al. 2019; Pisaric et al. 2001; Wagner et al. 2000b). One reason for the discrepancy was that the age determination data were not available for the entire length of sediment cores and LANDO extrapolated beyond these dating points to match the core length. In the case of PG1205 (Wagner et al. 2000b) with a core length of 9.85 m, dating points were available for the upper 2.5 m (Table 3.4), and therefore LANDO extrapolated the remaining 7m to cover the entire sediment core. However, the extrapolated results in accumulation rates do not reflect the geological history of the lake record provided by Wagner et al. (2000b). We have therefore changed the length of the sediment core to the last dating point to avoid strong extrapolation. In the case of Co1309 (Gromig et al. 2019), the age–depth model required the introduction of a hiatus that would span from 14 to 80 cal BP (Andreev et al. 2019; Savelieva et al. 2019). However, while a specific customization (such as a hiatus) is possible for single core cases, this is not possible in the current version of LANDO for multi-core investigation.

To overcome this, we reduced the length of the record used in our study for core Co1309 to the depth of the last available dating point (Table 3.4), such that the LANDO output matches the age–depth relationship reported by Gromig et al. (2019).

The detection of sedimentation rate change as an indicator of the Holocene–Pleistocene boundary yielded contrasting results. While the results from hamstr were closest to the 11 700-year boundary, all other modeling systems place the largest change in sedimentation rate either before or after 11 700 years BP. We hypothesize that three factors may have influenced all model results. (1) The age uncertainty (1σ range) within each individual model varied on average between 1000 and 3000 years for the period of 11 600 to 11 800 years BP (Figure C3 in the Appendix). This wide range of uncertainty does not provide confidence in pinpointing the boundary to an exact time slice. We expect that a higher amount of dating points close to the Holocene– Pleistocene boundary could constrain the models (Blaauw et al. 2018; Lacourse and Gajewski 2020; Trachsel and Telford 2017), which would lead to a better estimate of the boundary. (2) The age output for each model is not evenly distributed, which means that in the period from 11 600 to 11 800 years BP there are different numbers of observations for each core and each modeling system. We took this behavior into account by using binning (Alasadi and Bhaya 2017). Otherwise, an interpolation between both age and sedimentation rate values could lead to potential biases in the interpretation. (3) While we assumed in our first setup that the main sedimentation rate change would occur at 11 700 years BP consistently for all sediment cores (Figure 3.6), we cannot rule out the possibility that the sedimentation rate has changed significantly at different times for different lake systems. As our data collection covers a large area both in latitude and longitude (Figure 3.1), the variability between the models indicates the local variability between the climate and lithological preferences of the lake catchment for the involved sediment cores (e.g., Lozhkin et al. 2018; Finkenbinder et al. 2015; Anderson and Lozhkin 2015; Kokorowski et al. 2008; Biskaborn et al. 2016b; Courtin et al. 2021).

3.4.2 Design of LANDO

From the beginning of the development of LANDO, we decided to integrate most of the default settings for each modeling system as default values (Table 3.2). Regional studies, such as the one performed by Goring et al. (2012), have shown that specific prior information for the Bayesian modeling systems is needed to best fit the models to lakes within a geographical area. Without this regional information, changing settings within the modeling system to an arbitrary higher or lower value without considering the regional diversity could lead to under- or overfitting if the constraints are too loose or too strict (Trachsel and Telford 2017). For the special case that users have in-depth knowledge of one lake or multiple lake system, users can easily adapt these parameters within LANDO, as we have made these settings accessible in the Jupyter Notebook itself.

Part of the reason we made this decision was that we acquired external age determination datasets where we may not necessarily have all the essential information to specify each model. But we

also wanted to simplify the process for users who do not have in-depth modeling knowledge. By using the default values, we can compare models based on their ability to work with the available data. On the other hand, we are sure that the developers have set their default values based on systematic testing. Since we did not tune the age–depth models to the existing core, i.e. changing the parameters within each modeling system, we generated "uninformed" models that solely work with the available age determination data. By combining these uninformed models into one model, we have created an ensemble model that we consider to be data-driven and "semi-informed".

The advantage of this data-driven, semi-informed model approach is that we are reducing the risk of overfitting by considering the uncertainty of all modeling systems. This allows us to reevaluate existing geoscientific interpretations with larger uncertainty by taking advantage of the ensemble outcome. Additionally, we found that the more information is accessible to generate age–depth models, the more accurate and less uncertain these models become. A higher density of age determination along the depth of the sediment core is desirable for future drilling campaigns (cf. Blaauw et al. 2018).

The disadvantage arises in our second case study ("Inconsistent sequence" – CS2) and the multi-core investigation ("Multiple cores" – CS3). For both cases we needed the optimization step to narrow down the most suitable age–depth models for each sediment core, since the unoptimized uncertainty band was otherwise too wide for a clear interpretation. The optimization requires additional and independent proxy data, which are not available for some of our cores, especially for sediment cores obtained some decades ago. Our optimizing step is therefore mainly suitable for recently retrieved and analyzed sediment cores.

In addition to the assessment of age-modeling quality, we also checked the time and effort to conduct dating routines (Table C2 in the Appendix). We saw that *Bacon* had the highest runtime overall in all three case studies of our study design, which we link to our adjustment of the *ssize* parameter from 2000 (per default) to 8000 within the application. We increased this value to ensure good MCMC mixing for problematic cores, as suggested by Blaauw, Christen, and Aquino Lopez (2021), as well as to guarantee we had enough iterations for our summarizing statistics to compare with other modeling systems. If users decide to reduce the value of *ssize*, we implemented an iterative process, which checks whether *Bacon* produced enough iterations. If this is not the case, then LANDO will iteratively rerun the same sediment core with a higher *ssize* to produce 10 000 iterations.

One unique feature of our application is the predominant use of parallelization within the age–depth modeling of multiple sediment cores. For instance, we used the *Dask* back end for our sedimentation rate calculation. The advantage over the popular Scala-based *Apache Spark* and its Python interface *PySpark* (Zaharia et al. 2016) is that the *Dask* back end is Python-based and well integrated into the Python ecosystem (cf. Dask Development Team 2016). Therefore, *Dask* natively works with Python packages already implemented in LANDO. The key difference is that *Dask* neither provides a query optimizer nor relies on Map–Shuffle–Reduce, a data-processing technique for distributed computing, but instead uses generic task scheduling (cf. Dask Development Team 2016). Still,

parallelization libraries and back ends provide LANDO with additional speed-up that can promote future multi-core studies.

Within the ensemble model, we faced the challenge that the combination of all age distributions from the underlying age–depth models per centimeter represents a multi-modal distribution, especially in cases such as the "Inconsistent sequence" case study (CS2). It also means that the output of the ensemble model in these cases is susceptible to inclusion/ exclusion of any model. However, we consider using the weighted average median age to be a suitable solution for the multi-model distribution problem, as it is a good indicator on the most probable age within each centimeter based on all modeling systems. But we advise users to use the age confidence intervals per centimeter in subsequent analyses, instead of relying solely on the weighted average median age (cf. Telford, Heegaard, and Birks 2004). By optimizing the ensemble model with the ability to include independent proxy data, users can increase the likelihood of a more probable mean age for their sediment core.

3.4.3 Technical specifications of LANDO

In the further course of development, we decided to limit the resolution of the age–depth relationships. Using a resolution of one-centimeter increments allows us to match most proxy measurements from each sediment core with our age–depth models, apart from high-resolution measurement, such as XRF measurements. To allow a matching with high-resolution proxy data, we tested for a higher resolution of 0.25 cm for our application. In the single-sediment-core cases (CS1 and CS2), this change did not affect the workflow of LANDO. By contrast, the "Multiple cores" case (CS3) ran into memory issues. Since *SoS Notebook* and our parallel back ends store the resulting data frames in memory, expanding the resulting data frames to a 0.25 cm resolution causes a 4-fold increase in memory use, which limits our capability to run our application on a single laptop. As an intermediate solution, we stored the results from each parallelization worker on disk to free the memory and performed combining operations later. Based on this experience, we recommend working with data centers or increasing the available main memory (RAM) of the operating computer for multi-core studies with expected high-resolution output.

Another advantage of parallelization is that most modeling systems only run on one CPU/thread. Nowadays, however, both personal computers and data centers are made up of multiple CPUs/threads. Especially for larger multi-site studies, our application has the advantage of cutting the overall computing time by running each modeling system on multiple CPUs/threads simultaneously, even for personal computers. In comparison to serial execution of multiple models on one CPU/thread, which would take several hours, our parallel execution reduced the computing time per modeling system by a factor up to 4. When considering that our setup consisted of 6 CPUs (12 threads) and 16GB RAM, user can increase this factor even further by using larger computing facilities.

Sediment core length is the most limiting factor that determines the overall computing time in our application. However, we want to ensure that users can model each sediment core over its entire length to match proxy data with the correct age–depth relationships. Within our LANDO system, we

faced this problem by using extrapolation to calculate ages beyond available dating points. The exception here is the modeling system *Undatable*, which models only between the first and last dating point, as these two dating points act as anchors for the bootstrapping process (Lougheed and Obrochta 2019). As a result, we saw the sedimentation rate dropping twice to zero at the end of the sedimentation rate calculations. We link this behavior to the end of the individual modeling processes of *Undatable* as well as the other implemented systems.

Extrapolating the age–depth models beyond age determination points always bears the risk that the extrapolated dates do not reflect the actual age. The implemented modeling systems account for this circumstance by increasing the uncertainty for these undated regions (Blaauw 2010). While we are aware of this potential issue, we wanted to allow users to take advantage of the full age–depth coverage for their sediment core. Blaauw et al. (2018) pointed out in their findings that "most existing late-Quaternary studies contain fewer than one date per millennium" and recommended to increase the number of dating points to "a minimum of 2 dates per millennium". This recommendation would further decrease the need for extrapolation and reduce the overall uncertainty of age–depth models. We agree that more age control can improve the age–depth modeling results, but until the associated costs of analyzing organic material for radiocarbon dating decrease significantly (Hajdas et al. 2021; Zander et al. 2020), we recommend LANDO as tool to improve age–depth modeling.

3.4.4 Current and future model implementation in LANDO

During the development of our approach, we realized that some programs were not executable or parallelizable under the current circumstances. For instance, we tested OxCal 4.4 as stand-alone version on Windows with *NodeJS* (version 12.13.1.0) and the R package *oxcAAR* (Martin et al. 2021) within our application. In the case of EN18208, execution duration was above 3 h until the notebook lost connection to the *OxCal* interface. Furthermore, some cores never fully reached convergence within *OxCal*. We tried adapting our setups including changing the internal constraints, i.e. placement and number of boundaries, or using different depositions models, i.e. alternating between sequential model (*Sequence*) and Poisson-process deposition model (*P_Sequence*). According to Bronk Ramsey and Lee (2013), the long-term plan of *OxCal* is to make the entire source code openly accessible, which we fully support. An open source code would allow us to identify the current bottleneck so that we could implement *OxCal* in a future release.

To determine the best-fitting age–depth model through the *clam* modeling software, we added the "best-fit" option to LANDO by default. The best-fit option utilizes the negative log fit results from all *clam* outputs and identifies the fit with the lowest result as best fit. We included two further exclusion criteria for *clam* models within LANDO: if (a) there are too many age reversals within the models or (b) the fit reaches infinity. Under specific circumstances, some sediment cores will not have a fitting model, as is the case, for instance, in the "Inconsistent sequence" case study (CS2). Including models that do not fit the data would lead to erroneous estimations of the age–depth relationship. This comes with the

cost of losing an established model in the combined model if no fitting *clam* model is available. However, we think that the benefit of having a better-fitting model outweighs this cost.

Although *Undatable* is open source and the fastest modeling system within LANDO, its original development environment (MATLAB) is not free of charge. That is why we implemented *Undatable* in the open-source MATLAB-equivalent Octave. Since the Octave version of *Undatable* was slower than the original MATLAB version, we used the parallelization package *parallel* (Fujiwara, Hajek, and Till 2021) to provide comparable results in terms of computing time. To use *Undatable* with MATLAB within our application, users must acquire a license for MATLAB and link the MATLAB kernel to their license. Unfortunately, we do not have the capacity to provide individual licenses with LANDO. For users with an active MATLAB license, we provide the appropriate code to run the MATLAB version of *Undatable* in LANDO in the repository mentioned in the Appendix A.

We highly appreciate all the work that went into developing the stand-alone versions of each modeling system. Because LANDO relies on the work of these modeling systems, we encourage users of LANDO to cite the original modeling software alongside the LANDO publication in their work. Additionally, users should try the stand-alone versions for each modeling system to provide feedback to both LANDO and modeling system maintainers.

A potential expansion option of LANDO within the multi-language environment is to extend the application and allow future data analysis to use powerful tools, such as Python's machine learning libraries, e.g., *keras* (Chollet and others 2015) and *tensorflow* (Abadi et al. 2016). We anticipate that other developers can use LANDO as their starting point in building a larger limnological data analysis application.

3.5 Conclusion

This paper introduced our application LANDO – a linked age–depth modeling notebook approach. We presented an improved age–depth modeling procedure for sediment cores from high-latitude lake systems by linking five established systems: *Bacon, Bchron, clam, hamstr,* and *Undatable*. The added value of our application is the reduced effort to use established modeling systems in a single *Jupyter Notebook* for both single and multiple dating series and at the same time make the results comparable. In addition, we introduced an ensemble model that uses the output from all models to create a more robust age–depth relationship. In the case of scattered age determination data, we further implemented an adapted version of the fuzzy change point approach that allows users to integrate independent proxy data as indicators of lithological changes. This option helps evaluate the performance of modeling systems across lithological boundaries while providing a more reliable ensemble age–depth model by filtering inappropriate model runs for problematic datasets. Our application also allows users to run large datasets with multiple sediment cores in parallel to reduce the overall computation time. In our data collection of 55 sediment cores from northern lake systems at high latitudes, we found that the main regime changes in sedimentation rates do not occur synchronously for all lakes at the Pleistocene–

Holocene boundary. However, we linked this behavior to the uncertainty within the modeling process as well as the local variability of the sediment cores within the collection.

4

Effect of temperature on carbon accumulation in northern lake systems over the past 21 000 years

Gregor Pfalz[1,2,3,4], Bernhard Diekmann[1,2], Johann-Christoph Freytag[3,4], and Boris K. Biskaborn[1,2]

[1] Alfred Wegener Institute, Helmholtz Centre for Polar and Marine Research, Research Unit Potsdam, Telegrafenberg A45, 14473 Potsdam, Germany
[2] University of Potsdam, Institute of Geosciences, Karl-Liebknecht-Str. 24-25, 14476 Potsdam-Golm, Germany
[3] Einstein Center Digital Future, Robert-Koch-Forum, Wilhelmstraße 67, 10117 Berlin, Germany
[4] Humboldt-Universität zu Berlin, Unter den Linden 6, 10099 Berlin, Germany

Published: 29 August 2023 in *Frontiers in Earth Science* – DOI: 10.3389/feart.2023.1233713

Abstract

Rising industrial emissions of carbon dioxide and methane highlight the important role of carbon sinks and sources in fast-changing northern landscapes. Northern lake systems play a key role in regulating organic carbon input by accumulating carbon in their sediment. Here we look at the lake history of 28 lakes (between 50° N to 80° N) over the past 21 000 years to explore the relationship between carbon accumulation in lakes and temperature changes. For this study, we calculated organic carbon accumulation rates (OCAR) using measured and newly generated organic carbon and dry bulk density data. To estimate new data, we used and evaluated seven different regression techniques in addition to a log-linear model as our base model. We also used combined age-depth modeling to derive sedimentation rates and the TraCE-21ka climate reanalysis dataset to understand temperature development since the Last Glacial Maximum. We determined correlation between temperature and OCAR by using four different correlation coefficients. In our data collection, we found a slightly positive association between OCAR and temperature. OCAR values peaked during warm periods Bølling Allerød (38.07 g m^{-2} yr^{-1}) and the Early Holocene (40.68 g m^{-2} yr^{-1}), while lowest values occurred during the cold phases of Last Glacial Maximum (9.47 g m^{-2} yr^{-1}) and Last Deglaciation (10.53 g m^{-2} yr^{-1}). However, high temperatures did not directly lead to high OCAR values. We assume

that rapid warming events lead to high carbon accumulation in lakes, but as warming progresses, this effect appears to change as increased microbial activity triggers greater outgassing. Despite the complexity of environmental forcing mechanisms affecting individual lake systems, our study showed statistical significance between measured OCAR and modelled paleotemperature for 11 out of 28 lakes. We concluded that air temperature alone appears to drive the carbon accumulation in lakes. We expected that other factors (catchment vegetation, permafrost, and lake characteristics) would influence accumulation rates, but could not discover a conclusive factor that had a statistical significant impact. More data available on long-term records from northern lake systems could lead to more confidence and accuracy on the matter.

4.1 Introduction

Northern lake systems (50° N to 80° N) have been subject to an increase in mean annual surface air temperature up to 2.7 ºC over the last few decades (Box et al. 2019; Meredith et al. 2019; Ballinger et al. 2020). Temperature is one of the key control variable for the mineralization and burial of carbon in lakes, regardless of the origin of carbon (i.e., autochthonous or allochthonous) (Gudasz et al. 2010, 2015). Not only is an increase in temperature associated with higher carbon mineralization and burial, but also favors higher turnover of carbon through more in-lake primary production by macrophytes / aquatic plants (Z. Li et al. 2017; Velthuis et al. 2018) and algae (Biskaborn et al. 2023). As a consequence, lake systems can shift from being a net carbon sink to net carbon source and vice versa (Sobek et al. 2014; Heathcote et al. 2015; Denfeld et al. 2018).

Dean and Gorham (1998) estimated that lakes on a global scale accumulate in their sediment about 42 TgC (teragrams of carbon, i.e., one million metric tons of carbon) per year. Based on a new modeling approach, N. J. Anderson et al. (2020) approximated that accumulation rates have almost tripled over the past 100 years by about 72 TgC from 0.05 PgC to 0.12 PgC per year. In this model, the authors estimated that lakes in boreal biome contribute the highest (24%) to the global carbon burial rate, while tundra lakes are the lowest at only 2% due to their low carbon burial rate (Anderson et al., 2020). Despite potentially lower carbon burial rates in northern lakes due to current lower temperatures (Gudasz et al. 2010), Sobek et al. (2014) found that Arctic lakes show similar burial efficiencies as other lakes at lower latitudes. In addition, climate change-induced shifts in vegetation (Cramer et al. 2001; Pearson et al. 2013), lake aquatic biomass production (Biskaborn et al. 2023), and increased carbon release from permafrost thawing (Meredith et al. 2019; Schuur et al. 2022) may raise carbon burial rates in Arctic lakes due to the growing availability of carbon within the lakes (N. J. Anderson et al. 2020).

Comprehending the complex burial process requires a thorough understanding of how the carbon cycle in a lake responds to temperature fluctuations. Temperature plays a crucial role in shaping the interactions between dissolved organic carbon (DOC) and dissolved inorganic carbon (DIC) in lake ecosystems (Gudasz et al. 2010). DIC comprises carbon in the form of inorganic carbon species, primarily bicarbonate (HCO_3^-), carbonate (CO_3^{-2}), and dissolved carbon dioxide (CO_2), while DOC

refers to the fraction of organic carbon compounds dissolved in water. Higher temperatures can enhance microbial activity, leading to increased breakdown of organic matter and subsequent release of DOC into the lake water (Middelboe and Lundsgaard 2003; Adams, Crump, and Kling 2010). This process elevates the concentration of DOC in the lake, which influences organic carbon burial rates in lake sediments. Additionally, raised temperatures promote primary production by aquatic plants and algae, which enhances photosynthesis and the uptake of DIC from the water column (Hein 1997; Hammer, Kragh, and Sand-Jensen 2019). As a result, this process can either increase the outgassing of CO_2 from the lake or promote more carbonate precipitation of carbonated minerals within the lake.

In-lake bioproductivity and carbon accumulation also depend on catchment vegetation and the availability of allochthonous carbon (Roiha et al. 2016). During the Last Glacial Maximum, sparse vegetation and a reduced flux of allochthonous carbon to the lakes prevailed the Arctic due to the severe climatic conditions (Melles et al. 2012). In most areas the lack of nutrients in the underlying permafrost soil prevented further advances of boreal forests (M. K. Sundqvist et al. 2014). However, as the climate warmed and glaciers retreated, vegetation types shifted from tundra to boreal forest, which substantially increased the availability of organic carbon (A. V. Lozhkin et al. 2007; Anatoly Lozhkin et al. 2018; Biskaborn et al. 2016a; Diekmann et al. 2017).

Nutrient fertilization and atmospheric deposition played a crucial role in the Holocene in enhancing the productivity of the Arctic vegetation (Galloway et al. 2004; Choudhary et al. 2016). A prolonged growing season due to a warmer climate and shorter ice coverage further contributes to an upsurge in carbon turnover within lakes (Walther et al. 2002; Vuglinsky and Valatin 2018; Sharma et al. 2019, 2020). However, eutrophication and browning can in turn negate these effects, leading to stable water stratification with anoxic conditions at the bottom of the lake (Bartosiewicz et al. 2019).

In addition to in-lake primary productivity, other factors can affect the overall carbon balance within a lake, such as sediment resuspension/re-mineralization (Guillemette et al. 2017; Klump et al. 2020), or lake characteristics (e.g., morphology, catchment characteristics, or geographical location) (Ferland et al. 2014; Clow et al. 2015; Denfeld et al. 2018; Zwart et al. 2019). Nevertheless, changes in land use and changing precipitation patterns will in turn affect the distribution and storage of carbon in the Arctic in the future (Tchebakova, Parfenova, and Soja 2009; Bartsch et al. 2016; Windirsch et al. 2022).

While studies have focused on the carbon balance of lakes in the Holocene (e.g., N. J. Anderson, D'Andrea, and Fritz 2009; Sobek et al. 2014; Heathcote et al. 2015), investigations into past carbon accumulation rates back to the Late Pleistocene are lacking. Since the burial of organic carbon can react sensitively to temperature changes (Gudasz et al. 2010, 2015), a longer observation period with larger temperature differences can reveal new perspectives. To test whether temperature is the key driver in northern high-latitude lakes, we need to consider other influencing factors in our analysis, such as catchment vegetation, underlying permafrost, and lake-specific properties.

The main objective of this paper is to investigate the relationship between temperature and carbon in northern lakes over the past 21 000 years. We estimate the amount of carbon accumulated in 28 lakes since the Late Pleistocene using a combination of measured and newly generated organic carbon and dry bulk density data. To generate new data, we test seven different regression techniques as prediction models and evaluate them against common assessment metrics. We then correlate the obtained accumulation rates with temperature from re-analysis data (TraCE-21ka climate reanalysis dataset) to understand the relationship between these rates and changing temperature. Given the large time span covered by the datasets and the geographic spread of the sediment cores, we further create relationships to permafrost, vegetation, and lake-specific attributes.

4.2 Methods

To determine the amount of carbon that accumulated over the past 21 000 years, we need to calculate the "organic carbon accumulation rate" (OCAR, in g m^{-2} yr^{-1}) using the following equation (Eq. 1):

$$OCAR = \left(DBD \times \left(\frac{TOC}{100}\right)\right) \times SR \qquad \text{(Eq. 1)}$$

where DBD is dry bulk density (in g cm^{-3}), TOC is the total organic carbon content (in weight %), and SR is the age-depth-model-derived sedimentation rate (in cm yr^{-1}). We divided the resulting unit (g cm^{-2} yr^{-1}) by 0.0001 to get the desired OCAR unit (g m^{-2} yr^{-1}).

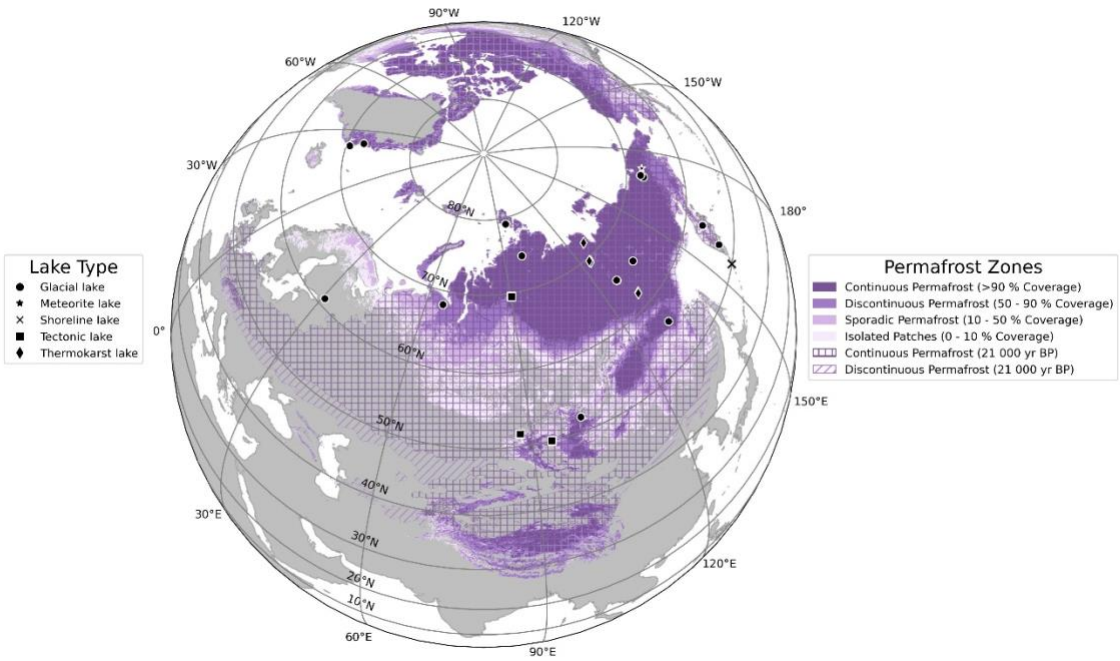

Figure 4.1: Spatial distribution of sediment cores from northern lakes (50° N to 90° N) used in this study labeled by their lake type (black symbols, n = 28). Underlying permafrost zones for the present time (solid colored areas) are from Obu et al. (2019), while permafrost distribution of the last 21000 years (shaded areas) originates from Lindgren et al. (2016). We adapted the color scheme for permafrost zones (four different shades of purple) from Obu et al. (2019) to be consistent with the original publication.

Table 4.1: Summary table containing individual datasets used for this study. Note that some CoreIDs appear multiple times, as some studies did not measure all proxies at the same core depth. For the dataset used to predict DBD, water content was only partially available, which we supplemented with modeled water content data.

Group	Subset of proxies	CoreID	Data points	Paper reference
Complete datasets	***C1 - Full dataset*** *(DBD, composite depth, TOC, silt, clay, WC)*		***Total: 264***	
		EN18208	26	Vyse et al. (2020)
		EN18218	63	Vyse et al. (2021)
		PG1205	159	Wagner et al. (2000)
		PG2201	16	Hughes-Allen et al. (2021)
	C2 - Wet bulk density dataset *(WBD, composite depth, TOC, WC)*		***Total: 260***	
		PG1214	56	Wagner and Melles (2008)
		PG1228	96	Ebel et al. (1999)
		PG1238	108	Raab et al. (2003)
	C3- Remaining dataset *(DBD, composite depth, TOC)*		***Total: 96***	
		EN18208	44	Vyse et al. (2020)
		PG2201	52	Hughes-Allen et al. (2021)
Augmented datasets	***A1 - Predict DBD with models dataset*** *(Composite depth, TOC, silt, clay, WC)*		***Total: 446***	
		Co1309	95	Gromig et al. (2019)
		Co1412	148	Baumer et al. (2021)
		PG1755	47	Diekmann et al. (2017)
		PG1756	28	Diekmann et al. (2017)
		PG1984	54	Biskaborn et al. (2012)
		PG2133	57	Courtin et al. (2021)
		PG2208	17	Biskaborn et al. (2021)
	A2 - Estimate DBD from beta distribution dataset *(Composite depth, TOC)*		***Total: 1790***	
		BN2016-1	39	Rudaya et al. (2021)
		ESM-1	36	Mackay et al. (2012)
		LS-9	59	Wolfe et al. (2000)
		PER3	68	Anderson et al. (2015)
		PG1111	127	Harwart et al. (1999)
		PG1341	141	von Hippel et al. (2021)
		PG1351	335	Asikainen et al. (2007) & Melles et al. (2007)
		PG1437	211	Andreev et al. (2005)
		PG1746	80	Nazarova et al. (2013)
		PG1857	34	Hoff et al. (2015)
		PG1858	37	Hoff et al. (2012)
		PG1890	75	Hoff et al. (2014)
		PG2023	113	Biskaborn et al. (2016)
		PG2208	172	Biskaborn et al. (2021)
		Tel2006	263	Rudaya et al. (2016)

To acquire the necessary data for this project, we conducted a comprehensive data collection process that focused on TOC and DBD measurements. In total, we collected 28 datasets from high latitude lake systems (50° N to 80° N – Figure 4.1) containing TOC, which we standardized following the procedure

introduced by Pfalz et al. (2021). In addition to DBD and TOC, our data collection focused on two additional data series: (1) sediment water content (WC) data, and (2) grain size measurements divided into the three subgroups of sand, silt, and clay (in weight-%).

Given the variable data availability of sediment cores with DBD, we divided the sediment cores into two subgroups: “Complete datasets” and “Augmented datasets” (Table 4.1). "Complete datasets" consist of subsets of sediment cores that contain (C1) DBD, TOC, sand, silt, and WC data, (C2) Wet bulk density, TOC, and WC data, and (C3) both DBD and TOC. On the other hand, "Augmented datasets" refer to datasets that were lacking DBD information but had (A1) grain size and partially WC data available or (A2) neither grain size nor WC data available.

While both C1 and C3 datasets were directly usable for OCAR calculation, in three instances of our data collection (subset C2 – “Wet bulk density dataset” – Table 4.1), we collected values for wet bulk density instead of dry bulk density. Because these datasets also provided data on the water content, we were able to calculate dry bulk density with the following equation (Eq. 2):

$$DBD = \left(1 - \left(\frac{WC}{100}\right)\right) \times WBD \qquad \text{(Eq. 2)}$$

with DBD being dry bulk density (in g cm^{-3}), WC being the water content (in weight-%), and WBD being the wet bulk density (in g cm^{-3}).

As both “augmented datasets” A1 and A2 were lacking DBD measurements, we considered predicting DBD from existing data. A large number of empirically derived pedotransfer functions and techniques for predicting bulk density exist in the literature (e.g., Hollis, Hannam, and Bellamy 2012; Martín, Reyes, and Taguas 2017; Lu et al. 2021; Palladino et al. 2022; Qin et al. 2022)). The majority of these prediction techniques use variations of linear models to predict bulk density. To enable comparison with the existing literature, we decided to use a log-linear model as our base model, which we built in Python using “scikit-learn” and its “LinearRegression” function (Pedregosa et al. 2012). In contrast to other pedotransfer functions, we included both the depth of a given sample within the sediment core (in cm) and its water content, which gave us the following equation (Eq. 3):

$$\ln(DBD) = \alpha + \beta_1 \times (CDepth) + \beta_2 \times (TOC) + \beta_3 \times (Silt) + \beta_4 \times (Clay) + \beta_5 \times (WC) \qquad \text{(Eq. 3)}$$

where DBD is dry bulk density (in g cm^{-3}), CDepth is the composite depth below sediment surface (as mid-point cm), TOC is total organic carbon content (in mg g^{-1}), Silt and Clay are the silt and clay content from grain size measurements (in weight ratios), WC is the water content (in weight-%), α is the intercept, and β_1 to β_5 are the individual coefficients. We obtained the unit “mg g^{-1}” for the TOC measurements by multiplying weight percent by factor 10, and unit “weight ratios” for clay and silt data by dividing the weight percent by factor 100.

Considering the significant impact of sediment water content on sediment compaction, we recognized its importance in predicting DBD. However, 11 sediment core datasets (39% of the collected datasets) lacked water content data. To address this limitation and to test whether other regression methods can

outperform linear models, we decided to predict WC alongside DBD in several multiple output regression methods in addition to the linear model. We opted for non-linear machine learning techniques to allow for a better comparison with the (non-linear) log-linear model. This includes the following regression methods from the "scikit-learn" and "xgboost" package in Python (Pedregosa et al. 2012; T. Chen and Guestrin 2016):

- Random Forest Regression
- Extreme Gradient Boosting (XGBoost)
- Gradient Boosting
- K-nearest Neighbor
- Support Vector Regression
- AdaBoost Regression.

For training and evaluation purposes, we split the "Full dataset" C1 (Table 4.1) into a training (80%) and test set (20%), but also used fivefold cross-validation to alleviate potential biases in the splitting process. We scored the individual models by using the following metrics: mean absolute error (MAE), relative absolute error (RAE), mean squared error (MSE), root mean squared error (RMSE), root relative squared error (RRSE), and R^2 score. The Appendix D contains the equations used for these metrics. We further checked if hyperparameter tuning would improve our results by adding an additional pipeline with the "GridSearchCV" and "RandomizedSearchCV" optimization algorithms from the "scikit-learn" package (Pedregosa et al. 2012).

For subset A1, we used the log-linear model and the regression methods to predict DBD and, where necessary, WC. However, for subset A2 ("Estimate DBD from beta distribution dataset" – Table 4.1), we only had TOC measurements for 15 sediment cores available for bulk density prediction. We therefore used existing grain size data from eleven sediment cores in the data collection (710 data points) to generate beta distributions for clay and silt. These beta distributions rely on the two parameters α_{beta} and β_{beta}, which we individually calculated using the following two equations (Eq. 4 and Eq. 5):

$$\alpha_{beta} = \left(\frac{1-\mu}{\sigma^2} - \frac{1}{\mu}\right) \times \mu^2 \qquad \text{(Eq. 4)}$$

$$\beta_{beta} = \alpha_{beta} \times \left(\frac{1}{\mu} - 1\right) \qquad \text{(Eq. 5)}$$

with μ and σ^2 being the mean and the variance of the existing clay or silt data, respectively. After obtaining α_{beta} and β_{beta} values for both clay and silt content, we drew 10,000 silt and clay values for each TOC measurement from the newly constructed beta distributions using "random sampling" of the Python package "numpy" (Harris et al. 2020). To reduce overall computing time, we performed random sampling and subsequent prediction of dry bulk density (Eq. 3) in parallel using the "Dask" back-end (Dask Development Team 2016) and the "joblib" Python package (Joblib Development Team 2020).
However, we constrained the possible values for clay and silt in two ways before using them in our models to predict dry bulk density ranges. Given that grain size data is compositional data, i.e., the sum of its components should add up to 1 or 100% (Greenacre 2021), we first removed sums of clay and silt

weight ratios that were greater than one. Since grain size data consists of a third component, which is the grain size range for sand, we also considered a lower bound for the sums to account for sand occurrence in the sediment cores. From the given data, we estimated that a maximum of 20% sand in the sediment column would be possible for our data collection. Therefore, we also removed sums of clay and silt weight ratios that were smaller than 0.8.

In any case, working with modeled values can introduce potential errors that could affect the interpretation of results. The quality and accuracy of the input data play a crucial role in determining the model's performance and output (Rebba, Mahadevan, and Huang 2006; Huang and Laffan 2009). Errors during data collection or sample measurement can propagate into the model outputs. We mitigated the risk by relying on original raw data as much as possible and used a database to ensure the values fell within physically plausible ranges (Pfalz et al. 2021). While both Eq. 1 and Eq. 3 represent simplifications of complex systems, and overly simplistic models can overlook important processes that lead to incorrect results (Andersson, Rantzer, and Beck 1999; Mathews and Vial 2017). However, we struck a balance between simplification and computational feasibility, using only publicly available data and ensuring the reproducibility of our results.

We used the LANDO age-depth modeling result from Pfalz et al. (2022) to derive sedimentation rates (SR) for each sediment core. LANDO links five age-depth modeling systems (Bacon, Bchron, clam, hamstr, Undatable) in one multi-language Jupyter Notebook (Haslett and Parnell 2008; Parnell et al. 2008; Blaauw 2010; Blaauw and Christen 2011; Kluyver et al. 2016; Peng et al. 2018; Lougheed and Obrochta 2019; Dolman 2022). For all sediment cores, we combined the results from the five modeling systems with standard settings into an ensemble model with two-sigma uncertainty (as described in Pfalz et al. (2022)). We propagated these sedimentation rate uncertainties into the OCAR calculations (Eq. 1) to obtain OCAR uncertainty ranges.

To understand the temperature development since the Last Glacial Maximum (Clark et al. 2009), we used the TraCE-21ka climate reanalysis dataset (He 2011), which we will refer to simply as TraCE dataset hereinafter. We divided the timespan covered by the TraCE dataset into the following periods (Walker et al. 2019; Head et al. 2021; Kuang et al. 2021):

- Last Glacial Maximum - 22 000 to 18 000 years BP (years before present, i.e., before 1950 CE)
- Last Deglaciation – 18 000 to 14 300 years BP
- Bølling Allerød – 14 300 to 12 700 years BP
- Younger Dryas – 12 700 to 11 700 years BP
- Early Holocene – 11 700 to 8200 years BP
- Mid-Holocene – 8200 to 4200 years BP
- Late Holocene – 4200 years BP to present.

For each core location, we extracted the surface air temperature at reference height (TREFHT) from the nearest grid cell of the TraCE dataset (grid cell resolution: 2.5° × 2.5° (He 2011; S. C. Brown et al.

2020)). We then converted the temperature from Kelvin (K) to Celsius (°C) by subtracting 273.15 K from each value and then averaging values for the summer months June-July-August (JJA). Following the procedure introduced by Kaufman et al. (2020), we converted OCAR values to z-score. The z-score measures how many standard deviations each point is away from the mean, and thus normalizes the data. To comprehend how vegetation affects carbon accumulation, we used the vegetation reconstruction by Dallmeyer et al. (2022) (depicted in Figure 4.2), which incorporates the TraCE dataset into its remodeling.

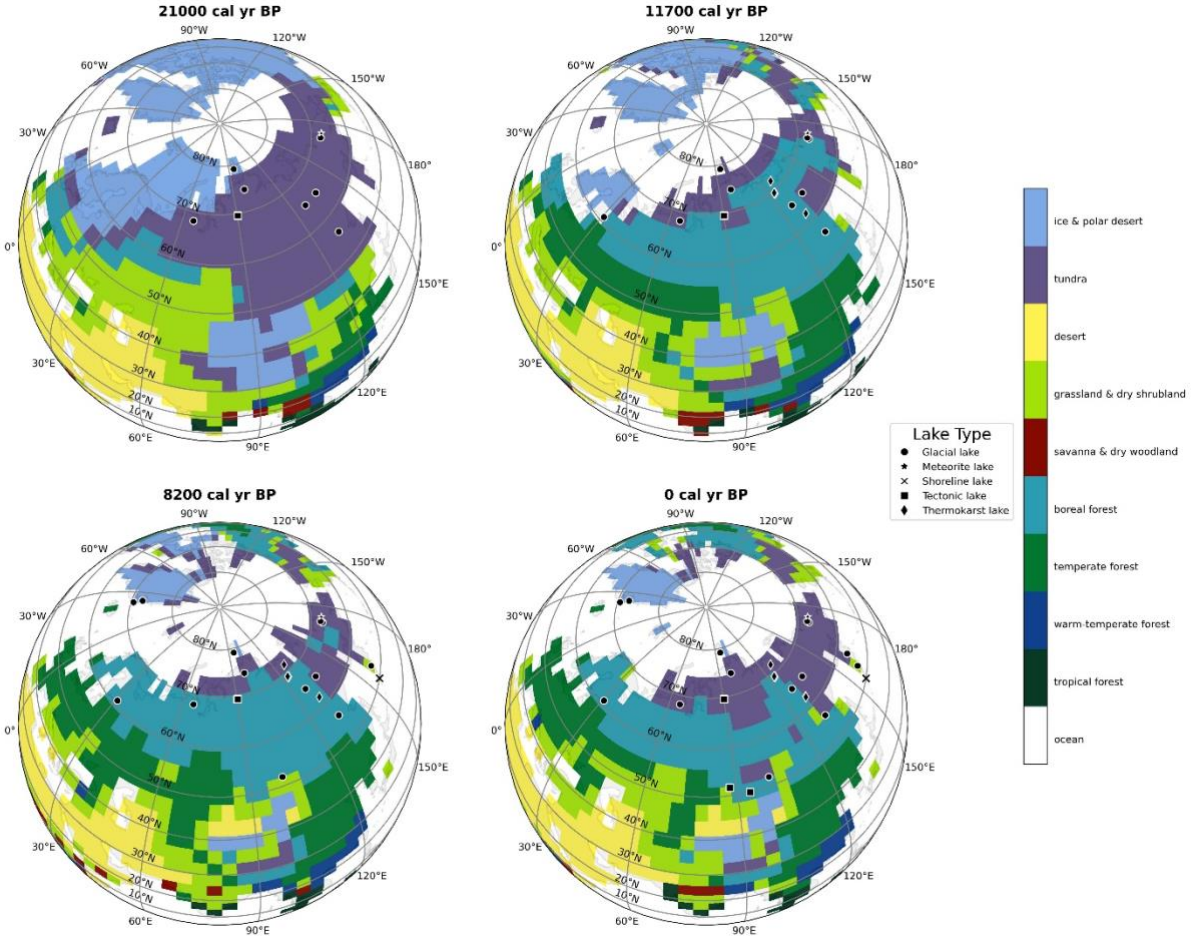

Figure 4.2: Biome distribution based on Dallmeyer et al. (2022) vegetation reconstruction for the past 21 000 years represented in four snapshots (21 000, 11 700, 8200, and 0 calibrated years Before Present, i.e. before 1950 Common Era). We include sediment cores with their respective lake type in the snapshot if there is TOC data available for them (black symbols).

To determine the correlation between temperature and OCAR, we first had to check for normality of the two variables. For this reason, we visually inspected the data by plotting quantile-quantile plots (Q-Q plots) using the package "statsmodels" (Seabold and Perktold 2010). We then used both the Shapiro-Wilk as well as the D'Agostino and Pearson's test from the Python package "scikit-learn" (Pedregosa et al. 2012) for our statistical tests. As both temperature and OCAR did not display normality, we used Spearman's and Chatterjee's rank correlation coefficient to check for the correlation between the two

variables. In contrast to Chatterjee's coefficient, Spearman's coefficient is a well-established, robust correlation metric often used for variables from non-normal distributions (Sadeghi 2022). However, the Chatterjee's coefficient showed promising results for testing the non-linear functional correlation between two variables (Chatterjee 2021; Sadeghi 2022). In addition to the Spearman and Chatterjee correlation coefficient, we checked the Pearson and Kendall-Tau correlation coefficient on both the untransformed and z-transformed variables. The methods for the more common correlation coefficients came from the Python package "scipy" (Virtanen et al. 2020), while to calculate the Chatterjee coefficient we used the script provided by Chatterjee (2021).

4.3 Results

While our data collection yielded 620 data points (Table 4.1 – Complete datasets) containing DBD and TOC which were directly usable for OCAR calculation (Eq. 1), the majority of data points from the augmented datasets (n = 2236, Table 4.1 – Augmented datasets) required further calculations. In preparation for both A1 and A2, we fitted the log-linear model with training dataset of our subset C1 (Table 4.1 – "Full dataset") to obtain the following equation to predict dry bulk density using a linear regression:

$$\ln(DBD) = 1.3337 + 0.0001 \times (CDepth) - 0.0016 \times (TOC) - 0.3986 \times (Silt) + 0.353 \times (Clay) - 0.0267 \times (WC) \quad \text{(Eq.6)}$$

where DBD is dry bulk density (in g cm^{-3}), CDepth is the composite depth below sediment surface (as mid-point cm), TOC is total organic carbon content (in mg g^{-1}), Silt and Clay are the silt and clay content from grain size measurements (in weight ratios), WC is the water content (in weight-%).

Table 4.2: Summary of evaluation of regression methods for predicting dry bulk density and water content with metrics such as mean absolute error (MAE), relative absolute error (RAE), mean squared error (MSE), root mean squared error (RMSE), root relative squared error (RRSE), and R2 score.

	Linear Regression	**Random Forest Regression**	**Extreme Gradient Boosting**	**Gradient Boosting**	**K-nearest Neighbor**	**Support Vector Regression**	**AdaBoost Regression**
Dry Bulk Density							
MAE	0.1486	0.0946	0.0956	0.0941	0.1354	0.1310	0.0954
RAE	0.2327	0.1527	0.1501	0.1516	0.2121	0.2045	0.1564
MSE	0.0416	0.0179	0.0169	0.0173	0.0342	0.0317	0.0183
RMSE	0.2027	0.1324	0.1297	0.1313	0.1842	0.1776	0.1348
RRSE	0.5704	0.3754	0.3690	0.3717	0.5192	0.5005	0.3842
R^2 Score	0.6728	0.8551	0.8611	0.8602	0.7288	0.7491	0.8483
Water Content							
MAE	6.6296	3.9758	4.4254	4.2335	6.3042	6.8274	4.1054
RAE	0.1529	0.1028	0.1106	0.1062	0.1594	0.1578	0.1127
MSE	66.7129	30.4543	35.6869	32.6959	72.5261	71.1432	37.3450
RMSE	8.1604	5.4941	5.9163	5.6851	8.4945	8.4289	6.0397
RRSE	0.5323	0.3610	0.3887	0.3724	0.5535	0.5520	0.3971
R^2 Score	0.7153	0.8662	0.8433	0.8580	0.6920	0.6915	0.8350

Since we wanted to test whether other regression methods could outperform the log-linear model, we had to ensure that there were no calibration and validation issues. Figure 4.3 and Table 4.2 provide an overview of the train and test performance with five-fold cross-validation of each regression methods we used to predict DBD (first column) and WC (second column). Ensemble methods such as AdaBoost, Random Forest, Gradient Boosting, and Extreme Gradient Boosting achieved the highest R^2 score and lowest error across all error scores. The Random Forest regression performed best for both DBD and WC with an R^2 score of 0.8551 and 0.8662, respectively. However, all regression methods yielded a high mean absolute error for WC between 3.9758 (Random Forest Regression) and 6.6296 (Linear Regression), which translated to a deviation from true water content percentages, i.e. 3.98% to 6.63%. Hyperparameter tuning using grid search and randomized search did not yield any improvement of these results.

Table 4.3: Mean prediction error of water content across the seven prediction methods.

Prediction method	Deviation from true water content [%]
Linear Regression	9.70
Random Forest Regression	12.13
Extreme Gradient Boosting	12.08
Gradient Boosting	11.90
K-nearest Neighbor Regression	10.32
Support Vector Regression	11.67
AdaBoost Regression	11.54

During the first predictions for DBD with the subset A1, we used the mean error between predicted WC versus measured WC as an additional measure of quality for all prediction methods. Table 4.3 summarizes the results across all prediction methods. In contrast to the previous test and training performance, Linear Regression performed with the smallest error (9.70%), while Random Forest Regression had the highest error (12.13%) amongst the methods.

Table 4.4: Occurrence statistics within dataset of clay and silt and their calculated parameters α_{beta} and β_{beta} for the beta distribution.

Grain size category	Mean	Variance	α_{beta}	β_{beta}
Clay	0.1919	0.0162	1.6459	6.9295
Silt	0.7431	0.0209	6.0322	2.086

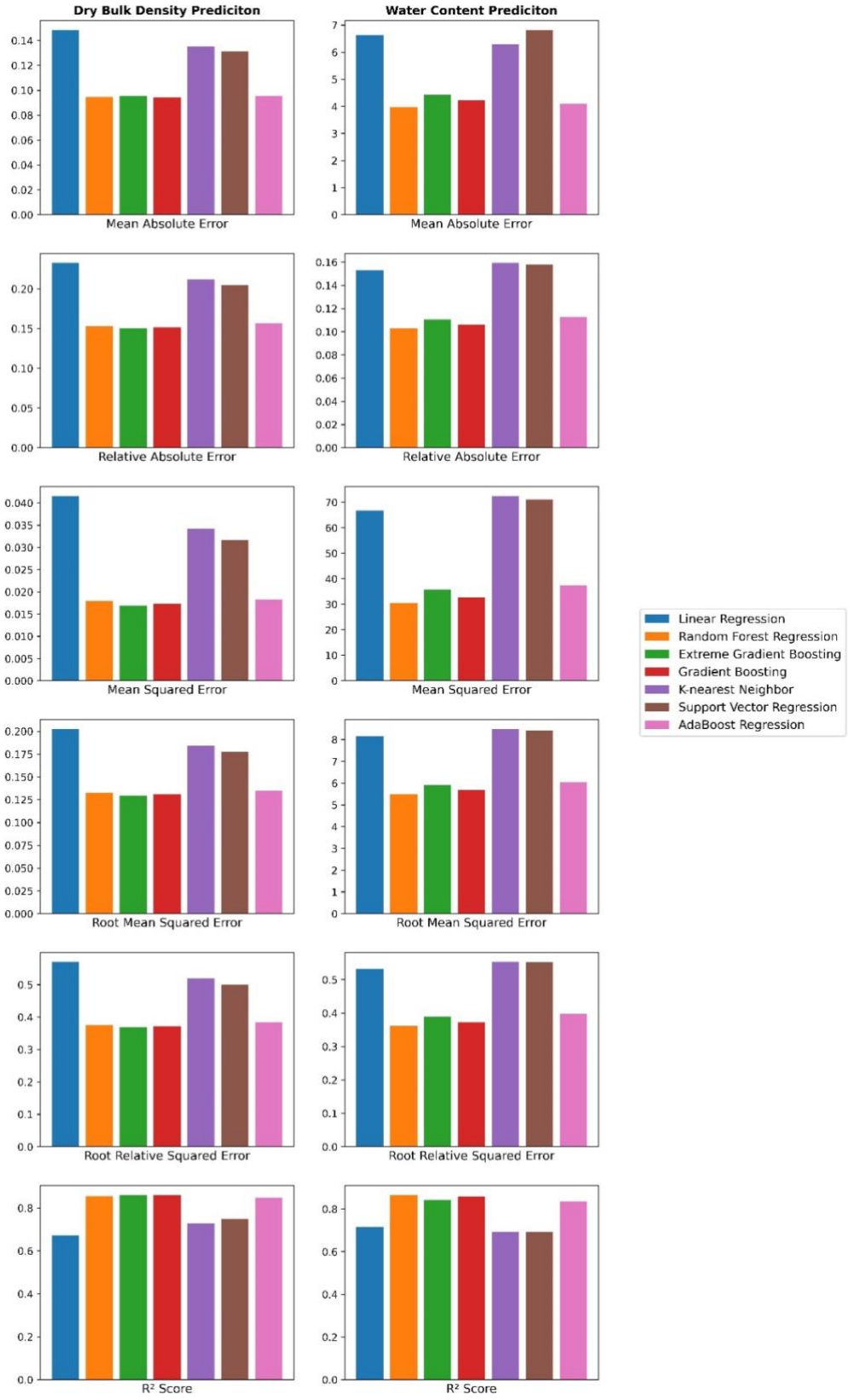

Figure 4.3: Visual representation of evaluation of seven regression methods for predicting dry bulk density and water content with metrics such as mean absolute error (MAE), relative absolute error (RAE), mean squared error (MSE), root mean squared error (RMSE), root relative squared error (RRSE), and R^2 score.

Before we could predict DBD for A2, we required a clay-silt pair for each of the 1790 data points in A2. Figure 4.4 shows the beta distributions for clay and silt based on the existing data across all datasets (Table 4.4). Clay content peaks between 10 to 20 percent, while silt content has its highpoint between 70 to 80 percent. We used random samples from these beta distributions with the subset A2 (Table 4.1) to estimate dry bulk density for all regression methods. To verify that model results were within a reasonable range and to allow comparison with the literature, we compared TOC values with measured and predicted DBD values. In Figure 4.5 we summarize the results obtained from the model predictions for Random Forest regression, Support Vector regression, and Linear Regression as well as measured values of TOC and dry bulk density. For completion, we show the results of the remaining four methods in Appendix D (see Figure D1).

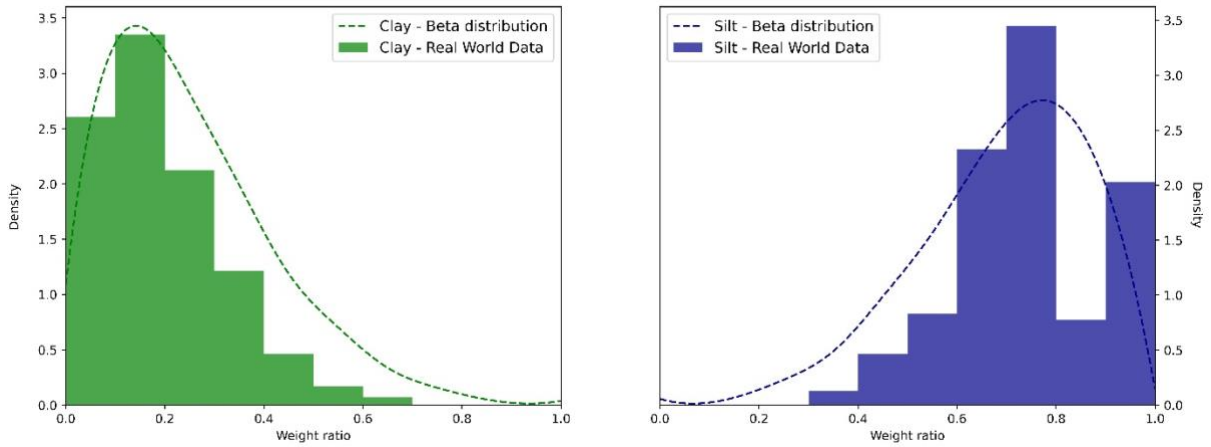

Figure 4.4: Derived beta distribution for clay (left, green dotted line) and silt (right, blue dotted line) from available grain size data

In Figure 4.5A, the random forest regression displays a step-like curve with a plateau between about 20 and 150 mg g^{-1} TOC. Predicted values of dry bulk density ranged from 0.18 to 1.21 g cm^{-3}. We associated this pattern with potential overfitting, as the method ignores other measured values. Support Vector regression (Figure 4.5B) predicted values follow the measured values with a shallow increase from 300 to 500 mg g^{-1} TOC. DBD values for this method varied between 0.18 and 1.25 g cm^{-3}. The curve within Figure 4.5C (Linear regression) shows a similar log curve as presented in the literature, however, it exceeds the level of possible values (linear regression method maximum value: 4.06 g cm^{-3}; maximum physically possible value: 2.65 g cm^{-3} (Avnimelech et al. 2001)). The minimum value of this method was 0.07 g cm^{-3}. The exceeding values (n = 27) corresponded to deep level samples of core PG1341 (deeper than 15 meters) where compactions plays a greater role than the method can reflect. To allow comparison with the literature, we excluded the values above 2.65 g cm^{-3} from the linear model and continued with linear model further.

Figure 4.6 contains the comparison between mean OCAR and the temperature data from TraCE dataset for the 28 sediment cores. Some temperature data showed similarities as the core-drilling locations were

in the same TraCE grid cell, e.g., EN18208 (Core No. 4 – Lake Ilirney) and EN18218 (Core No. 5 – Lake Rauchuvagytgyn), or PG2133 (Core No. 25 – Lake Bolshoe Toko) and PG2208 (Core No. 27 – Lake Bolshoe Toko). However, temperature ranges strongly varied within the dataset depending on the core location. For instance, for sediment core Tel2006 (Core No. 28 – Lake Teletskoye) we saw temperatures ranged from a minimum of 22.61 °C to a maximum of 17.88 °C, while temperatures for EN18218 (Core No. 5 – Lake Rauchuvagytgyn) only spanned from 8.06 °C to 2.60 °C. Regarding OCAR values, we obtained the lowest overall values for PG1351 (Core No. 15 – Lake El'gygytgyn) with a mean of 0.286 g m^{-2} yr^{-1} (uncertainty range max: 6.131 g m^{-2} yr^{-1}, uncertainty range min: 0.008 g m^{-2} yr^{-1}). We saw the highest OCAR value in ESM-1 (Core No. 6 – East Sayan Mountains Lake) with 278 g m^{-2} yr^{-1}, however, this value came with a large uncertainty range (min: 8.78 g m^{-2} yr^{-1}, max: 3018.625 g m^{-2} yr^{-1}). We calculated a mean OCAR of 24.615 g m^{-2} yr^{-1} for all collected sediment cores in our dataset.

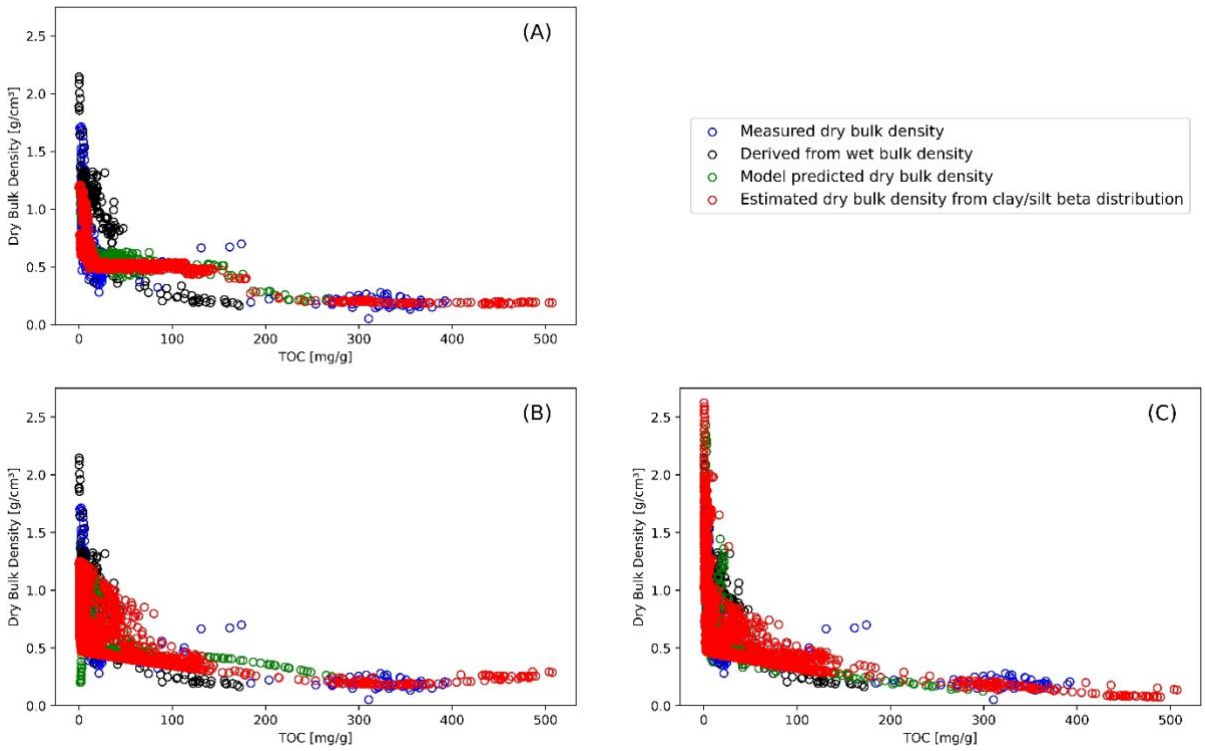

Figure 4.5: Model prediction results for dry bulk density against total organic carbon (TOC) for Random Forest regression (A), Support Vector regression (B), and Linear Regression (C). Directly measured dry bulk density (blue circles) and dry bulk density derived from measured wet bulk density (black circles) data are the same across all subplots. Green circles represents predicted dry bulk density values for each prediction method, where all input values were available. Red circles are estimated mean dry bulk density values for each prediction method with grain size data based on beta distribution for clay and silt.

Figure 4.7 contains two comparisons: on the left, a direct comparison between mean OCAR values and JJA temperature data across all sediment cores, while on the right is a comparison of z-transformed OCAR values over time. We obtained mean OCAR values for Last Glacial Maximum (9.47 g m^{-2} yr^{-1}) and Last Deglaciation (10.53 g m^{-2} yr^{-1}) at the lowest mean temperatures 12.75 °C (range:

24.14 to -0.51 °C) and -10.62 °C (range: 20.58 to 2.12 °C), respectively. We observed the highest temperature ranges in the Late Holocene with a mean temperature of 3.37 °C (range: 12.69 to 14.76 °C), but only with a mean OCAR of 21.8 g m^{-2} yr^{-1}. The highest OCAR values occurred in Bølling Allerød (38.07 g m^{-2} yr^{-1}) and Early Holocene (40.68 g m^{-2} yr^{-1}), where temperature ranged from 14.51 to 1.69 °C (mean value: 4.28 °C) and -12.02 to 8.20 °C (mean value: 1.65 °C), respectively.

Table 4.5: Shapiro-Wilk and D'Agostino's distribution results for OCAR and temperature as untransformed and z-transformed values.

	OCAR	**Temperature**	**OCAR (z-transformed)**	**Temperature (z-transformed)**
Shapiro-Wilk test	0.61966	0.96647	0.96743	0.97967
p-value	0	$3.786 * 10^{-22}$	$7.777 * 10^{-22}$	$4.112 * 10^{-17}$
Distributed	Non-normal	Non-normal	Non-normal	Non-normal
D'Agostino's K^2 test	1411.31	33.09	164.68	144.82
p-value	$3.459 * 10^{-307}$	$6.504 * 10^{-8}$	$1.736 * 10^{-36}$	$3.564 * 10^{-32}$
Distributed	Non-normal	Non-normal	Non-normal	Non-normal

However, when comparing the normalized data over time, we found that Mid-Holocene (mean z-score: 0.126, median z-score: 0.015) and Late Holocene (mean z-score: 0.089, median z-score: 0.055) were among the higher z-transformed OCAR values. Both Mid-Holocene and Late Holocene showed the highest temperature ranges, as shown on the left side of Figure 4.7. Periods that displayed a lower temperature range, i.e., Last Glacial Maximum, Last Deglaciation, and Younger Dryas, also revealed lower z-transformed OCAR values. Mean z-scores were 0.643, 0.402, and 0.616, while their median z-scores were 0.735, 0.429, and 0.577 for Last Glacial Maximum, Last Deglaciation, and Younger Dryas, respectively.

To check for correlation between temperature and OCAR, we first had to inspect visually Q-Q plots of those variables to check for normality (see Figure D2). The visual inspection of the Q-Q plots, however, showed that variables were non-normal distributed. D'Agostino's K^2 and Shapiro-Wilk test (Table 4.5) confirmed this numerically. We then determined the appropriate correlation coefficients for both variables untransformed, both variables z-transformed, and one where only OCAR was z-transformed while temperature was untransformed. Table 4.6 shows the correlation coefficients for Pearson, Spearman, Kendall-Tau, and Chatterjee, their p-value, and their statistical significance for the above cases. Except for Pearson correlation coefficient for both untransformed variables, all coefficient showed a statistical significance for the relationship between temperature and OCAR. Chatterjee coefficient displayed in all three cases a positive significance. Spearman's rho value denoted a weak negative relationship (-0.06367) for untransformed values, while for z-transformed values it saw a weak positive relationship between temperature and OCAR (0.2394 and 0.0863).

Table 4.6: Correlation statistic between OCAR and temperature using four different correlation techniques and an alpha value of 0.05 for the p-value.

OCAR vs temperature	Both untransformed	Both z-transformed	Only OCAR z-transformed
Pearson r value	-0.01705	0.2356	0.1075
p-value	0.4254	6.025 * 10-29	4.733 * 10-7
Statistically significant	No	Yes	Yes
Spearman rho value	-0.06367	0.2394	0.0863
p-value	0.002899	7.084 * 10-30	5.374 * 10-5
Statistically significant	Yes	Yes	Yes
Kendall tau value	-0.03879	0.1646	0.0577
p-value	0.006566	8.687 * 10-31	5.266 * 10-5
Statistically significant	Yes	Yes	Yes
Chatterjee xi value	0.0682	0.0477	0.0311
p-value	2.317 * 10-7	0.0002	0.0108
Statistically significant	Yes	Yes	Yes

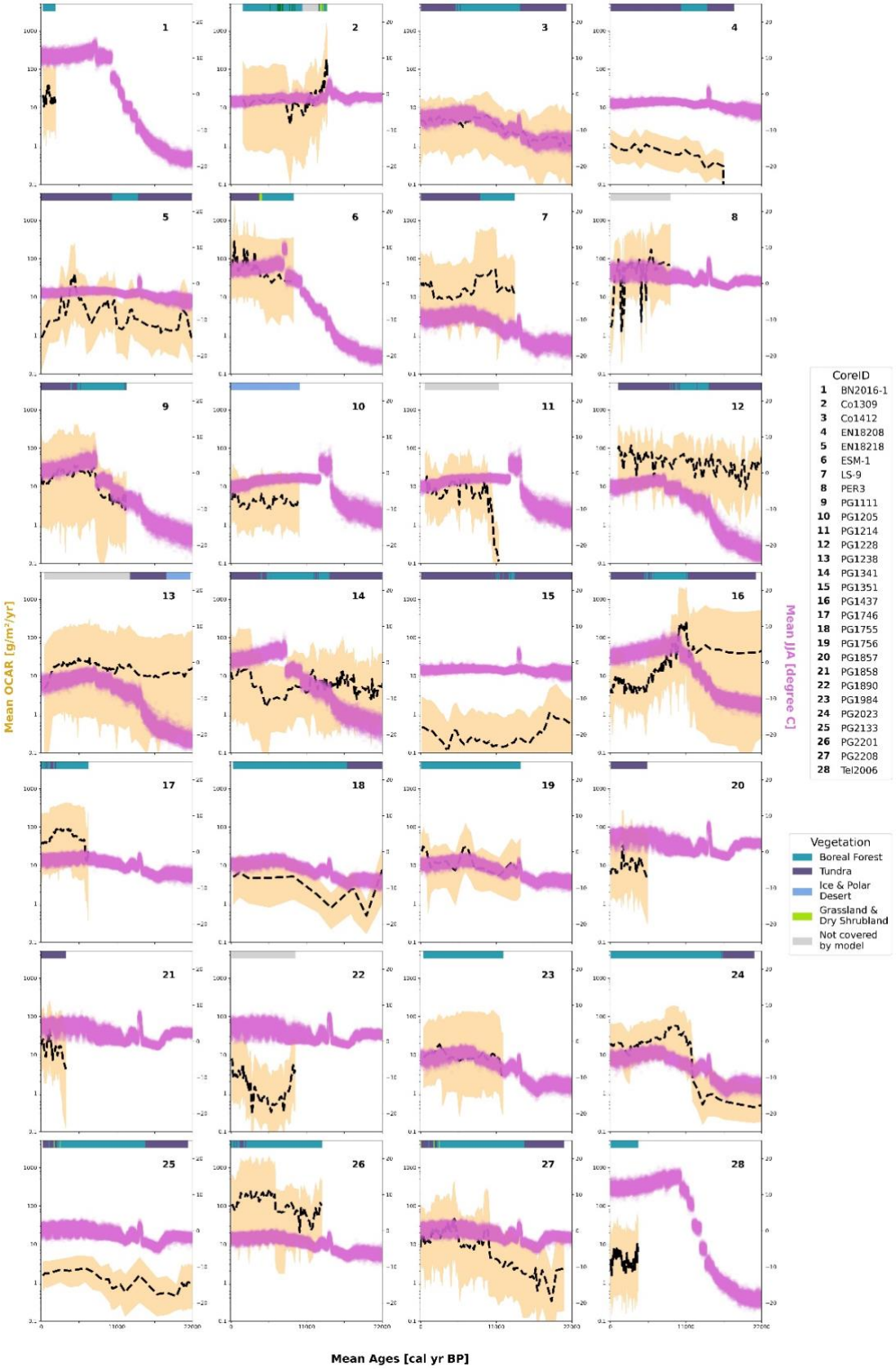

Figure 4.6: Organic carbon accumulation rate (mean values – black dashed line, 2σ uncertainty – golden shaded area) and June-July-August temperature from TraCE-21k temperature reconstruction (violet line) for the sediment cores used in this study (n=28). Vegetation for each sediment core based on Dallmeyer et al. (2022) biome reconstruction.

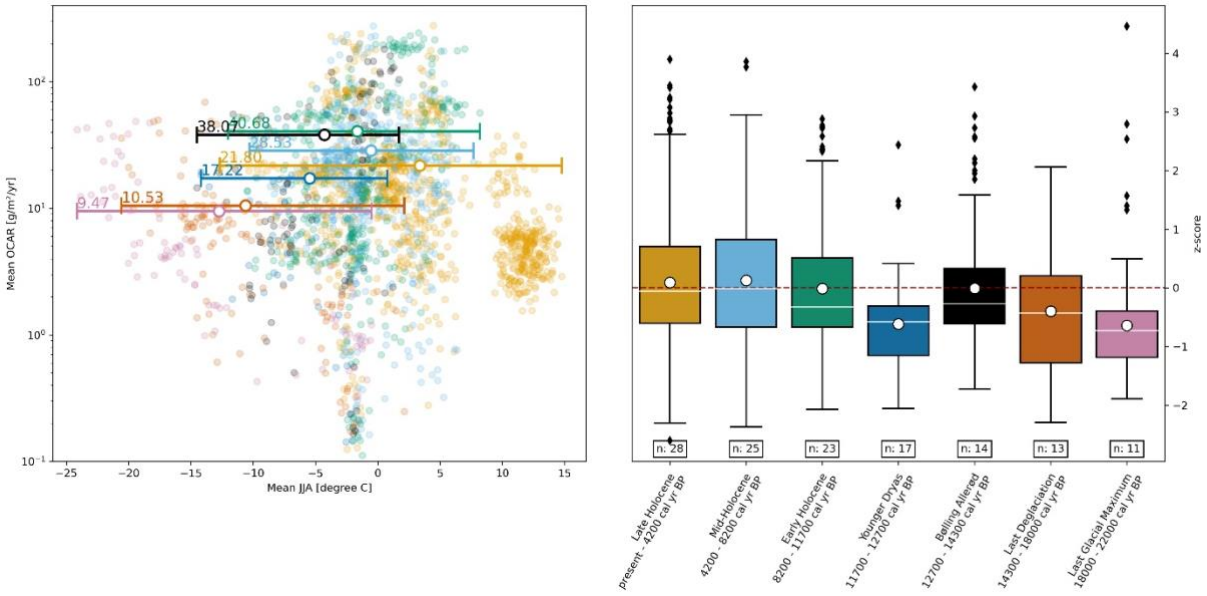

Figure 4.7: Left plot: Scatter plot showing the relationship between mean June-July-August (JJA) temperature and mean organic carbon accumulation rate (OCAR). For coloring, we have chosen the same color code as for the periods on the right. Vertical lines indicate the temperature range, while white dots show the mean temperature. Numbers on the left side of vertical lines are the overall mean OCAR for this period. Right plot: The OCAR z-scores grouped by the individual periods. The number below each box indicates the number of sediment cores contributing to this specific period. White dots are showing the overall mean z-scores, while the white lines are the overall median values.

4.4 Discussion

4.4.1 Lake carbon-temperature relationship across millennia

The Last Glacial Maximum (22 000 to 18 000 years BP) marks the lowest carbon accumulation in our observation period (9.47 g m^{-2} yr^{-1}), followed by the Last Deglaciation (18 000 to 14 300 years BP – 10.53 g m^{-2} yr^{-1}) and Younger Dryas (12 700 to 11 700 years BP – 17.22 g m^{-2} yr^{-1}). Our finding suggest that lower OCAR values tend to occur in lower temperature ranges (Figure 7). Conversely, however, the highest temperatures did not directly result in the highest OCAR values, with the mean OCAR above 10°C being 5.97 g m^{-2} yr^{-1} (range 1.55 to 37.33 g m^{-2} yr^{-1}). Even removing the cluster between 10 °C and 15 °C by excluding measurements from BN2016 1 (Lake Bayan Nuur) and Tel2006 (Lake Teletskoye) only slightly raises the mean OCAR values for the Late Holocene (present to 4200 years BP) to 28.41 g m^{-2} yr^{-1}.

The statistical analysis further supports this trend with only slight positive statistical significance between OCAR and temperature for our data collection (Table 4.6 – see z transformed values). However, both Bølling Allerød (14 300 to 12 700 years BP – 38.07 g m^{-2} yr^{-1}) and Early Holocene (11 700 to 8200 years BP – 40.68 g m^{-2} yr^{-1}) are the two warm periods with the highest OCAR (see Figure 4.7), which also have the steepest gradients of temperature change (Kaufman, McKay, Routson, Erb, Dätwyler, et al. 2020; C. E. Rasmussen and Williams 2006; Sune O. Rasmussen et al. 2014). This may indicate that a rapid temperature change initiates a high accumulation of carbon in the lakes at the

beginning of these warm phases and then decreases over time as the biological activity in the lakes increases.

However, our entire observation period was not covered by more than half of our collected sediment cores (n=17), which partially limits the interpretation of individual sediment cores. Still we can identify numerous sediment cores (Co1309, Co1412, PG1111, PG1228, PG1238, and PG2208) that show a strong positive correlation (Pearson r-value above 0.5) between OCAR and temperature (see Table D2). In particular, OCAR values for Co1412 (Lake Emanda) follow temperature variations throughout the observation period with the highest r-value of 0.8503.

We also observe synchrony with high r-values from the pair PG2133 and PG2208 (0.4599 and 0.5027, respectively) originating from the same lake (Lake Bolshoe Toko) but different positions within the lake. Similarly, the pair PG1755 and PG1756 (Lake Billyakh) show positive correlation with close individual r-values (0.3478 and 0.2341, respectively), but are not statistically significant with p-values greater than 0.05. In contrast, sediment cores PG1111 and PG1341 both come from Lake Lama but do not show a similar correlation with r-values of 0.8065 (PG1111) and 0.1813 (PG1341). The main difference from metadata perspective is that the first two pairs were part of the same expedition (Lake Billyakh – Yakutia 2005; Lake Bolshoe Toko – Yakutia 2013), whereas PG1111 (Norilsk/Taymyr 1993) and PG1341 (Norilsk 1997) are from two different expeditions of two different years. Therefore, even though they are from the same lake, comparing them may not be fair as the collection method of the sediment cores may have affected the results (Pfalz et al. 2021). The accuracy of laboratory analysis further improved over time between the retrievals of the two cores, which may have contributed to the observed differences in results.

Despite the complexity of individual limnological studies of lake systems, our collected dataset showed a positive correlation between OCAR and temperature with statistical significance for 11 out of 28 sediment cores. Even if there is a given heterogeneity amongst lake systems, we can conclude that temperature alone can explain OCAR variability within a lake. The 11 sediment cores are highly diverse and vary significantly in several aspects, as they share no common feature. They differ in location, vegetation surrounding the lakes, permafrost influence, catchment size, drilling distance from the shore, water depth at drilling site, lake area, lake volume, drilling device used, climate zone, and lake type. This would confirm our general understanding of the independence of the relationship between temperature as the sole driver and OCAR from other factors. However, the strength of the correlation depends both directly and indirectly on each contributing factor. Other environmental factors may weakened or amplified the strength of the temperature signal during our observation period. It is possible that temperature affected the remaining 17 cores, but local factors may have obscured the signal in the sediment.

Many of the processes known to influence the production and accumulation of organic carbon are subject to change due to modern climate warming (Larsen, Andersen, and Hessen 2011; Biskaborn, Smith, et al. 2019). While understanding the long-term effects of temperature over thousands of years

on OCAR is the main focus of our research, the short-term effects over couple hundred years can produce drastic results (Kastowski, Hinderer, and Vecsei 2011; Heathcote et al. 2015; Q. Li et al. 2021). Many sediment cores in our collection (n=21) have at least one surface sample pointing to the industrial era of the last 250 years (Toynbee 1884). However, only six of those cores have more than two measurements, while only three (PG1111, PG2208, Tel2006) have more than ten measurements. Because of this poor resolution, we do not have enough evidence to explain recent changes in OCARs. However, we have included these recent measurements in our Late Holocene samples to allow a comparison over a longer period.

Nevertheless, we must also consider the potential contribution of the modeled paleo-temperature and age determination data as an influencing factor in our interpretation. While the TraCE dataset provides an excellent tool for reconstruction, its development relied on global climate models that may not reflect the spatial variability required for our analysis. One solution would be the refinement of the reconstructed temperature through more local proxies and input parameters or downscaling of the TraCE dataset (S. C. Brown et al. 2020; Karger et al. 2023). Although LANDO is a more advanced age-depth modeling technique that combines multiple age-depth modeling software, it faces the same challenge as any other age-depth modeling software: Modeling software relies on age controls to establish an age-depth relationship. However, due to an insufficient number of age controls (cf. Blaauw et al. 2018) or greater uncertainty in the age determination data, a resulting age-depth model may not represent the exact absolute age. To circumvent this issue, we included the 2σ confidence intervals of these age-depth models and their resulting sedimentation rate in our OCAR calculations. We applied these intervals to the weighted mean age derived from four or five modeling software for every OCAR measurement. The consequence was an increase in the 2σ uncertainty intervals for each OCAR measurement, but overall a more accurate representation given the inherent uncertainty.

4.4.2 Spatial heterogeneity of lake carbon accumulation

For 11 sediment cores examined in our study we found statistical significance between OCAR and temperate, while the remaining sediment cores showed no such relationship, likely due to the complex nature of limnological studies and heterogeneity between lake systems. Given the spatial extent of our research, we must also consider unique local factors that may have affected the results. While we have sourced metadata and data sediment cores used in this study from published research articles (Table 4.1) that further provide in-depth analyses and interpretations, we will focus on three important unifying aspects: vegetation, permafrost, and geomorphology.

The vegetation reconstruction for around 21 000 years BP suggest that the oldest cores (n = 9) were mostly surrounded by tundra (Table D3). The tundra biome is diverse and can present itself as an expansive landscape with mostly herbaceous plants, or as a mix of small trees and shrubs, such as Betula, Alnus, and Salix (Dallmeyer et al. 2022). The lack of significant abundance of evergreen trees – compared to boreal forests present in later reconstructions (11 700, 8200 and 0 years BP – Figure 4.2) –

may have contributed to the overall low carbon accumulation of lakes in the Last Deglaciation and Last Glacial Maximum.

To test this notion, we looked for sediment cores that remained in the same tundra biome throughout the entire vegetation reconstruction. We found that tundra vegetation surrounded one sediment cores (PG1351– Lake El'gygytgyn) for the longest time, with carbon accumulation rates averaging below 2.85 g m^{-2} yr^{-1} (mean value: 0.44 g m^{-2} yr^{-1}). Melles et al. (2007) attribute the relatively low carbon content to the decomposition of organic matter due to the high oxygen content of the bottom water, but also a limited supply of terrestrial organic matter. The authors continue to determine a limited vegetation cover in the tundra-dominated catchment as main reason for the low carbon accumulation in the sediment (Melles et al. 2007).

In contrast, during the Bølling Allerød and Younger Dryas, most of the lake catchment areas in our data collection shift from tundra vegetation to boreal forests (Table D3). Some even remain in boreal forest into the Late Holocene (n=9), presumably fueled by the Holocene Thermal Maximum around 8200 years BP (Kaufman et al. 2004; Wanner et al. 2015). Other catchment areas (EN18208, EN18218) transition back to tundra vegetation immediately at the end of the Younger Dryas, or at the end of the Early Holocene (LS-9, PG1228). OCAR values for sediment cores surrounded by boreal forest show strong variability in magnitude and incline. However, in most cases we observe an increase in OCAR values at the onset of higher vegetation cover, especially during warmer periods (Bølling Allerød, Early Holocene).

As an example of vegetation transition after the Younger Dryas, we looked at the lake catchment for sediment core PG1437 (Lake Lyadhej-To). There, the vegetation reconstruction indicates a transition from tundra to boreal forest at the beginning of the Early Holocene, which then lasted until the Late Holocene. Despite the Early Holocene being recognized as a warm period that allowed the boreal forest to expand northward (Tarasov et al. 2000; P. M. Anderson et al. 2010), our current data suggest that the carbon input from vegetation only affected the lake at the onset of the Early Holocene. Between 11 700 and 11 000 years BP mean OCAR increased to about 240 g m^{-2} yr^{-1}, but then decreased to 39 g m^{-2} yr^{-1} around 8200 years BP, with values even dropping to 12 g m^{-2} yr^{-1} at 7700 years BP. The gradual northward expansion of boreal forests may have resulted in not fully established forests to provide an increased amount of organic carbon, which could explain the observed phenomenon for PG1437. However, this further supports our theory that a steep temperature gradient leads to more OCAR rather than sustained higher temperature.

While temperature is a major factor in vegetation change, we tested whether there is a direct relationship between OCAR and vegetation. The mean OCAR for lakes surrounded by boreal forest and tundra was 31.17 g m^{-2} yr^{-1} and 20.8 g m^{-2} yr^{-1}, respectively. We saw the lowest mean OCAR in the "ice and polar dessert" biome with 5.3 g m^{-2} yr^{-1} and the highest mean OCAR in the "grassland and dry shrubland" biome with 50.17 g m^{-2} yr^{-1}. However, we found no general correlation between mean OCAR and catchment vegetation in our data collection, with an r-value of only 0.047 (Figure D3). A possible

explanation for this might be a delayed vegetation response to a warming climate in the establishing phase of the boreal forest (Chapin and Starfield 1997; Ernakovich et al. 2014; Zona et al. 2014). Following the reasoning of the dynamic vegetation module of JSBACH3 used in the vegetation reconstruction, we assume that trees live longer (up to 50 years) than grass (up to 1 year) (Dallmeyer et al. 2022). This would mean that as the trees grow, there would be less organic material available for transport to the lakes as they use their resources to grow. However, following the harmonization process by (Dallmeyer, Claussen, and Brovkin 2019), the grassland and dry shrubland biome has a similar minimum total vegetation coverage as the tundra biome, but differs by having more growing degree days, even as in boreal forest. This may indicate that vegetation with more growing days but short lifespans generally has a higher OCAR. We therefore have to assume that temperature drives both catchment vegetation and carbon accumulation in lakes, but at different times.

Despite the spatial heterogeneity of permafrost in the northern hemisphere (Mishra et al. 2021), our dataset contains a majority of cores (n = 26) located in areas with some degree of permafrost presence. Permafrost is perennially frozen ground that stays at or below 0 °C for at least two consecutive years (French 2007). Estimates of stored carbon within permafrost in the northern hemisphere range from around 1460 to 1600 PgC (petagrams of carbon, i.e., one billion metric tons of carbon) (Hugelius et al. 2014; Meredith et al. 2019; Schuur et al. 2022). Estimates by Lindgren et al. (2016) on the extent of permafrost during the Last Glacial Maximum suggest that permafrost had previously affected these areas as well. The two remaining sediment cores currently and previously unaffected by permafrost are Co1309 (Lake Ladoga) and PER3 (Lake Pernatoye). In addition, five sediment cores from our data collection originate from thermokarst lakes (Table D1) that form as a direct result of permafrost thawing (Olefeldt et al. 2016).

Based on the mean OCAR, three sediment cores out of these five thermokarst lake cores (LS-9 – Lake Dolgoe Ozero, PG1984 – Lake Sysy-Kyuele, and PG2023 – Lake Kyuntyunda) accumulate on average less than 60 g m^{-2} yr^{-1} (19.85 g m^{-2} yr^{-1}, 9.92 g m^{-2} yr^{-1}, and 24.78 g m^{-2} yr^{-1}, respectively). The other two cores (PG1746 – Lake Temje and PG2201 – Lake Malaya Chabyda) show significantly higher values (62.31 g m^{-2} yr^{-1} and 129.58 g m^{-2} yr^{-1}, respectively). Except for PG1984 (min/max values: 2.98 to 18.75 g m^{-2} yr^{-1}), the remaining cores are prone to strong fluctuations in OCAR in the minimum to maximum range:

PG2201: 20.64 to 275.04 g m^{-2} yr^{-1}

PG1746: 13.26 to 89.3 g m^{-2} yr^{-1}

LS-9: 8.3 to 52.74 g m^{-2} yr^{-1}

PG2023: 0.43 to 57.35 g m^{-2} yr^{-1}

Compared to a global collection of OCAR values for lakes by Mendonça et al. (2017) (Figure 4.8), these fluctuation are still within the range of previously observed values. However, they do not indicate that the permafrost degradation process would directly contribute to a high OCAR in these lake types. The only exception being PG2201 (Lake Malaya Chabyda), where Hughes-Allen et al. (2021)

associate the higher burial rates with increased bioproductivity in the lake. However, they also acknowledge that nutrient availability from the catchment, compact lake morphology, higher rates of sedimentation, and less exposure to warmer and oxygen-rich shallow waters further contributed to the higher OCAR values (Hughes-Allen et al. 2021). In the remaining cases, increased microbial activity may contribute to greater emission of greenhouse gases, resulting in less accumulation of carbon in the sediment (Serikova et al. 2019; in ’t Zandt, Liebner, and Welte 2020).

Lakes indirectly affected by permafrost, i.e. non-thermokarst lakes with permafrost in the catchment area, as well as lakes outside of permafrost zones show a similarly diverse picture. The uniqueness of a lake, given by its catchment area, lake volume and shape, its origin, and inflow parameters, can influence the carbon accumulation within the lake. Table D1 summarizes the standard parameter of the lakes in our collection we were able to collect. However, when we looked at the correlation between OCAR and these lake-specific attributes, we found no correlation between them (Figure D3). However, this shows the importance of limnological studies, as examining a wide variety of lakes would give us a better understanding of the accumulation process in Arctic lakes, since we cannot derive holistic statements from a limited number of lakes.

In this study, we primarily focused on the climatic impacts on lake sediment, which resulted in our assumptions partially overlooking the direct influence of microbial activity and oxygen levels in the water column. The limited availability of both current and historical data contributed to this situation. But given previous experiment (e.g., Z. Li et al. (2017), Velthuis et al. (2018)), we still assume that a temperature change has a direct influence on the microbial community. The effects of occurrence and interaction between different primary producers can contribute to a lead-lag relationship and needs further investigation when considering longer time scales. In meromictic lakes, the presence of low oxygen levels can create anaerobic/anoxic conditions affecting the in-lake carbon cycle, which in turn can skew the amount of deposited carbon. Obtaining additional sedimentological data on redox conditions is crucial for this analysis, but as many of the original studies did not include such data, we have to assume that oxygen levels in our lakes have changed on both short- and long-term basis. This means that future studies on individual lakes have the opportunity to link OCAR to redox conditions, providing further insight into the relationship between OCAR and microbial activity.

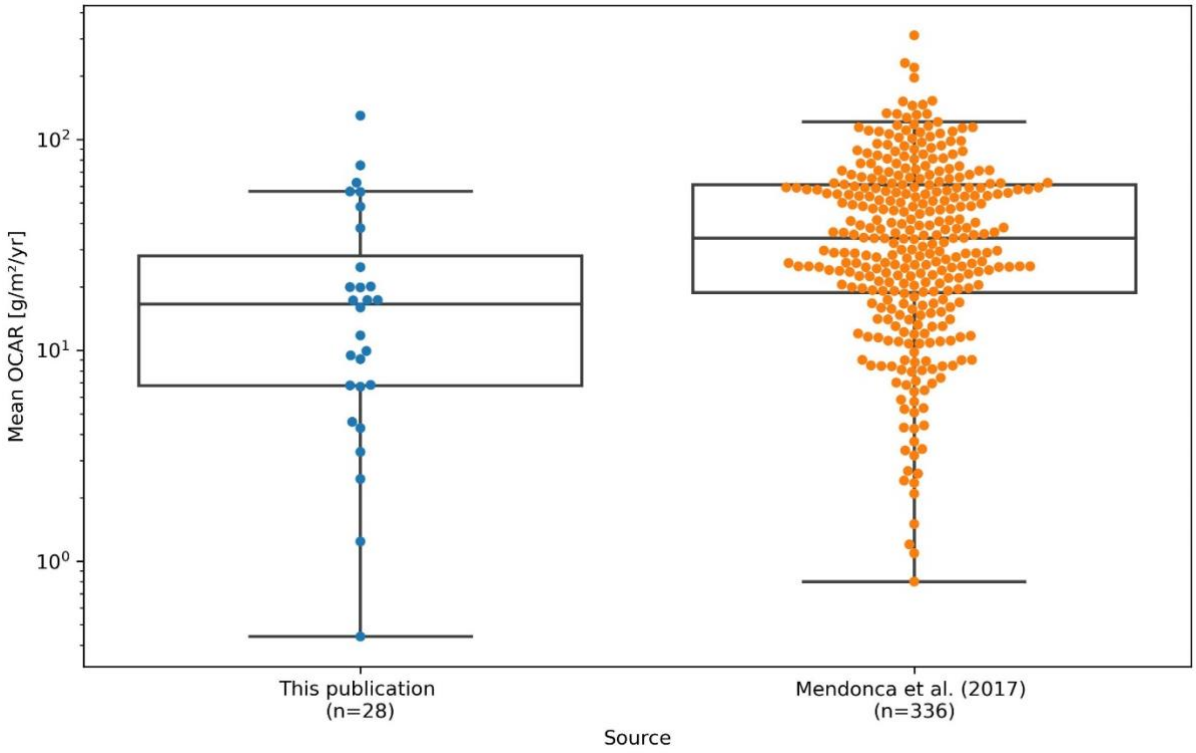

Figure 4.8: Comparison of mean OCAR values between lakes from our study (blue dots, left) and global lake compilation by Mendonça et al. (2017) (orange dots, right) with associated boxplots

4.4.3 Method selection for predicting dry bulk density

Understanding the relationship between TOC and DBD has been essential in predicting DBD values. While we observed a logarithmic trend between values in our data collection (Figure D4), the existing literature agreed that a log-linear model would best describe their relationship (Menounos 1997; Dean and Gorham 1998; Campbell et al. 2000; Avnimelech et al. 2001; Lan et al. 2015). However, we faced the challenge of high bulk density values occurring at low organic carbon values.

While low TOC values are common in northern lakes (Sobek et al. 2014), our log-linear model produced unrealistic DBD results, which exceeded physically possible values (Figure D4). We found the highest DBD values in the deeper part of sediment core PG1341 at a depth of 1461 cm to 1883 cm, where compaction also most likely had a major impact on the material. We found that extrapolating given empirical equations from the literature to lower organic carbon values would result in a similar outcome (Figure D4). To enable a more realistic representation, we assume that future models will have to take these special cases with different degrees of sediment compaction into account. We expect that a better understanding of the occurring compaction in sediment cores and its influence on the core composition will improve DBD predictions.

Due to the overestimation of dry bulk density for low TOC values of samples from deeper parts of sediment cores, and the apparent clustering of predicted values between 0 and 100 mg/g TOC, we considered several alternative prediction method to our log-linear model. As we evaluated these methods based on best-fit prediction metrics such as the R^2 score, we found that the supposedly best performing

methods also showed signs of overfitting. Overfitting means that a model shows a strong bias towards seen data allowing only little room for variability for interpreting unseen data (Bilbao and Bilbao 2017; Ying 2019). Despite reducing the potential bias through cross validation, we found that the small size of available data mainly contributed to the overfitting in three cases (Random Forest Regression, Gradient Boosting, and Extreme Gradient Boosting). An increase in sample size could alleviate overfitting, as hyperparameter tuning produces more reliable results (Ying 2019). However, in our case, having only 211 data points (80% of the total amount of training data points) available for hyperparameter tuning resulted in no visible improvements. We still assume that Random Forest Regression and Gradient Boosting methods have the potential to outperform log-linear models if more data is available. Other methods we tested may not be suitable for the prediction of water content and dry bulk density.

4.5 Conclusion

The purpose of the study was to determine whether there is connection between carbon accumulation in northern lakes and temperature changes that have occurred over the past 21 000 years. We found a slightly positive relationship between OCAR and temperature among our data collection, for which we generated more data using our log-linear model that superseded other data science techniques. While our dataset was diverse in terms of location, age of the sediment, permafrost areas, lake parameters, and catchment vegetation, we generally saw the highest OCAR values occurring during Bølling Allerød (14 300 to 12 700 years BP – 38.07 g m^{-2} yr^{-1}) and the Early Holocene (11 700 to 8200 years BP – 40.68 g m^{-2} yr^{-1}). This could indicate that rapid warming events lead to high levels of carbon accumulation in lakes. As warming progresses, this effect appears to change with lower accumulation rates, presumably due to increased microbial activity triggering carbon dioxide and methane outgassing. While we achieved promising results with data from 28 sediment cores, more data from northern lakes would help us to build a greater level of confidence and accuracy on the matter.

5

Synthesis

5.1 Navigating the interdisciplinary landscape: Geoscience and data science in dialogue

In recent years, the success of interdisciplinary research has sparked the shift towards collaboration across multiple disciplines, highlighting the importance of a collaborative environment over individual domain expertise. Computer science has played a pivotal role in driving advancements across various fields. In biology and medicine, it has enabled breakthroughs in areas such as drug discovery and protein folding (Vamathevan et al. 2019; Jumper et al. 2021). In astronomy and astrophysics, machine learning has contributed to the detection of transient objects and the modeling of thermal evolution on Mars (Miranda et al. 2022; Agarwal et al. 2021). Likewise, in geoscience and climate science, computer science has supported initiatives such as early earthquake warning and the study of trough development in permafrost landscapes (Münchmeyer et al. 2021; Rettelbach et al. 2021).

Within the main research chapters (Chapter 2 to 4) of this dissertation, data science played a central role in developing tools and deriving insights. Data science, as a sub field of computer science, applies computational methods and techniques to explore and analyze complex datasets to extract valuable information, uncover patterns, and derive meaningful insights (Kempler and Mathews 2017; Han, Kamber, and Pei 2012). Specifically, Chapter 2 demonstrated the successful extraction and integration of metadata and measurement data from multiple lakes across northern Eurasia into a newly designed database, which served as the foundation for the subsequent investigations in Chapters 3 and 4.

Balancing the technical requirements of database design with the need for an easily understandable cross-domain vocabulary was crucial. This ensured that geoscientists could effectively contribute their expectations and insights in a clear and accessible manner. By bringing together knowledge from data science and limnology[10], our research has significantly enhanced our ability to revive lake-specific datasets from 1994 to 2018 and give them new meaning in our broader regional analysis. While considering a broad array of proxies in the database with four biotic and three abiotic proxy groups (Figure 2.3 and Figure 2.5), we also emphasized simplicity in the database design. This involved deconstructing datasets to identify the most common denominator or intersection between them and finding the simplest unit for each proxy. This approach also ensured the flexibility to

[10] This includes paleo-limnology

accommodate future additions of proxies. By providing standardized templates for data acquisition, we created a data product consisting of a data collection of unique standardized datasets (n=55), which ultimately accelerated their integration into the database.

Integrating datasets into a database further offers an advantage over a loose collection of data, as it enables accessible data transformation into different standards, such as the Linked Paleo Data (LiPD) format (McKay and Emile-Geay 2016; Emile-Geay et al. 2018), through transformation pipelines. This facilitates the conversion of datasets and improves their interoperability and usability. As research projects interconnect, researchers can convert datasets to align with individual database designs or use existing databases, enabling the exchange of data among different research projects. By using databases, researchers gain the advantage of querying large volumes of data, conducting quantitative analyses, and deriving comprehensive interpretations on a global scale.

The data product itself represents a significant contribution to the scientific community, addressing a critical gap as many databases and data collections tend to focus primarily on the western Arctic (Canada, Greenland, and Alaska), often due to limited accessibility to datasets from Russia (PAGES 2k Consortium 2017; H. S. Sundqvist et al. 2014; McKay and Kaufman 2014). To address this, we have made our standardized datasets available on the PANGAEA online repository, ensuring that other scientists can access this valuable information. The research presented in this dissertation adds to the existing body of knowledge about the Arctic, complementing insights gained from other regions and contributing to a deeper understanding of past environmental changes in the Arctic as a whole. For example, Chapter 3 revealed that significant shifts in sedimentation rates at the Pleistocene–Holocene boundary are not consistent across all the lakes in our dataset. This suggests that Arctic lakes may not have uniform responses to major climatic events, indicating that spatial heterogeneity has already played a larger role historically in shaping past lake dynamics. Chapter 4 further supports this assumption and shows that although the relationship between carbon accumulation and temperature varies across lakes over the past 21,000 years – highlighting existing heterogeneity – a predominant trend of a positive association between carbon in lake sediments and paleotemperature can be detected. However, by obtaining more data from a larger number of lake systems, we can categorize lakes more effectively and then create tailored trajectories for each individual lake.

The availability of many proxies generally enhances environmental reconstruction by offering a range of interpretations and ultimately reducing the risk of reinforcing a specific favored paradigm through reliance on a single proxy (H. H. Birks and Birks 2006). As we have been working with heterogeneous multi-proxy data from multiple sediment cores in the Arctic, harmonizing raw data was the most effective approach to ensure data across multiple proxies remain unaltered to alleviate bias even further. Our conceptual framework, accommodating the diverse origin of data from various lake systems, facilitated both numerical comparisons between individual proxies and individual lakes, enabling us to observe and analyze a broader range of variability in our comparative analysis (Figure 2.8).

Including data from multiple lake systems additionally provided the advantage of establishing a robust testing ground for our age-depth modeling approach (LANDO) in Chapter 3. LANDO itself is a blueprint for connecting research software from different programming languages in one convenient place. In the current version, LANDO offers the opportunity to access five different age-depth modeling software, but given the simple requirements, it modularity allows adding further modeling software from other authors in the future. However, these software authors must actively transition their code and move from proprietary to open-source in order for the science community to contribute (Kilamo et al. 2012; Heron, Hanson, and Ricketts 2013). Nevertheless, at the same time, authors can actively gain credit beyond their initial contribution by citing them as the original source. Working closely with the original contributors of age-depth modeling software can accelerate its integration into LANDO, benefiting both sides through mutual debugging opportunity. This was the case during this dissertation with all age-depth modeling software integrated into LANDO, which provided opportunity to collaborate with Andrew Parnell, Andrew Dolman, Maarten Blaauw, Bryan Lougheed, and Stephen Obrochta.

While the LANDO codebase requires some programming knowledge, there is an opportunity to convert LANDO into an online application, if deployed with sufficient computing and processing power on a web-based infrastructure to run successfully. This would make the application accessible for global usage, providing opportunities for geoscientists lacking sufficient computing resources or coding expertise. In its design, with its thin user interface and scalable computing system, LANDO is comparable to efforts of the PANGEO community (e.g., Odaka et al. 2020; Eynard-Bontemps et al. 2019) to develop global open-source tools. However, further optimization of LANDO's codebase would be needed to enable it to run efficiently on HPC (high-performance computing) platforms, ensuring faster and more scalable computational capabilities.

Relying on an array of age-depth modeling software has revealed a significant subjectivity to the incoming parameters, such as age determination and settings, which we highlighted with a case study in Chapter 3 (see Chapter 3.3.2 – "Inconsistent Sequence" – Case Study no. 2). To address this issue, we adopted the approach of using original data, allowing for direct comparisons with existing literature and models, and reducing potential biases introduced by different software settings. By comparing the results of individual software, we observed instances where adjustments were required (Chapter 3.2.5.2 – Detection and filtering of unreasonable models), but these modifications were justified based on geological evidence that were not explicitly part of the data but the author's interpretation.

It is crucial to acknowledge that while LANDO is an accessible and user-friendly data scientific tool, successful application still demands domain-specific geoscientific knowledge and expertise grounded in multi-proxy evidence (e.g., elemental data, such as XRF or organic data, and biological data, such as diatoms or pollen). Despite this requirement, leveraging the existing data through our approach allowed us to preserve valuable information while enhancing our understanding of sedimentation rates over the Holocene and Late Pleistocene. The ability to directly compare results and

make informed adjustments based on geological evidence has been a significant advantage of this project, contributing to a more robust and reliable age-depth modeling process.

Chapter 2 and Chapter 3 established the groundwork for Chapter 4, incorporating both the data from the standardized data collection from the database and the results of the age-depth modeling in the form of the corresponding sedimentation rates. Chapter 4 primarily centers on geoscientific aspects, while also exploring the potential of machine learning techniques to outperform traditional methods. While we have generally found that certain machine learning regression techniques performed well when pretested, the small number of data points used for training often limited their overall performance. A larger sample size often correlates with better model performance (e.g., Fassnacht et al. 2014; Wisz et al. 2008). However, newer techniques emphasize using less data but incorporating deep learning techniques (Karpatne et al. 2019; Zhu et al. 2020; Ng et al. 2020). Given that limnology frequently faces constraints leading to a limited number of samples for interpretation, it might be worthwhile to consider focusing on these innovative approaches in the future to improve resilience and futureproofing.

5.2 Delving into the depths: Uncovering past lake dynamics

To improve reliability and applicability of our findings on past lake dynamics, leading to robust and more widely relevant results, this dissertation focused on a multi-proxy and multi-site approach. Throughout all chapters, we have learned that dealing with the heterogeneity and uniqueness of lakes is the most difficult challenge to overcome in a multi-site study. While individual studies demonstrate the complexity of uncovering the history of a single lake, integrating multiple datasets with unique conditions and circumstances within the collection poses even greater challenges.

We have addressed this challenge with a number of different workarounds. Firstly, to address the diverse proxy resolutions resulting from various sampling techniques in individual studies, we commonly used bins or intervals to define the impacts of proxies. This has already proven to be an effective method in many studies, e.g. Erb et al. (2022), Mottl et al. (2021), Crema and Bevan (2020), and J. Müller and Joos (2020). However, the selection of bin sizes plays a crucial role in adjusting the time windows of observation and accurately defining specific moments in time for certain events to occur. Opting for a smaller bin size, such as 100 years instead of 1000 years, enables a more granular analysis, facilitating the identification of finer patterns and variations within the dataset. However, using very small bin sizes also introduces noise or random fluctuations that can obscure the underlying trends and make it challenging to distinguish meaningful patterns from random variations. On the other hand, larger bin sizes may oversimplify the data, averaging out important variations and potentially overlooking significant patterns or events.

Striking the right balance between bin size and data granularity is crucial to ensure accurate and meaningful interpretations in data analysis of past lake dynamics. For instance in Chapter 2 (“Harmonizing heterogeneous multi-proxy data from lake systems”), we continuously experimented with different bin sizes to align them with the available data per bin. However, this became more

challenging for data from the Late Pleistocene (11 700 to 21 00 years BP), as data points for proxies became scarcer, making it difficult to find suitable matches. Incorporating the Late Pleistocene in our analysis had the drawback of increasing the bin sizes compared to what they would have been if we had solely focused on the Holocene.

Chapter 4 ("Effect of temperature on carbon accumulation in northern lake systems over the past 21 000 years") followed a different approach by aggregating data points per geological periods. By adopting this method, we were able to take into account the larger spatial context of our research, which facilitated the identification of common factors affecting carbon accumulation in lakes. However, one drawback of this approach was that the extensive spatial coverage, spanning diverse climate and vegetation zones (referring to Chapter 1.2.1), sometimes led to correlations with these individual zones instead of the north-south gradient. The situation became even more ambiguous when accounting for different lake types and geomorphologies. A potential solution to address this issue would have been to expand the sample size by incorporating data from additional lakes of various types and locations, allowing for a more comprehensive understanding of the underlying mechanisms. Despite the challenges posed by the heterogeneous nature of the data collection, it is remarkable that we successfully identified temperature as one of the primary drivers for carbon accumulation in lakes.

Examining paleotemperature and carbon accumulation trends, we observed significant patterns correlating with climatic shifts. During the Last Glacial Maximum (22 000 to 18 000 years BP), we found mean OCAR values of 9.47 g m^{-2} yr^{-1} at temperatures averaging -12.75°C, ranging from -24.14°C to -0.51°C. This trend slightly increased during the Last Deglaciation (18 000 to 14 300 years BP) with an OCAR average of 10.53 g m^{-2} yr^{-1} and temperatures averaging -10.62°C, fluctuating between -20.58°C to 2.12°C. However, we observed higher OCAR values in the Bølling Allerød (14 300 to 12 700 years BP) and Early Holocene (11 700 to 8 200 years BP) periods, averaging 38.07 g m^{-2} yr^{-1} and 40.68 g m^{-2} yr^{-1} respectively. During these intervals, temperatures averaged -4.28°C (range: -14.51°C to 1.69°C) and -1.65°C (range: -12.02°C to 8.20°C). In contrast, the Late Holocene (4 200 years BP to present) presented the broadest temperature spectrum with an average of 3.37°C, spanning from -12.69°C to 14.76°C, yet its mean OCAR was 21.8 g m^{-2} yr^{-1}. We found that rapid warming events lead to high carbon accumulation in lakes. However, as warming progresses, this trend appears to change. One explanation, suggested by current observations, is that increased microbial activity might trigger greater outgassing as temperatures rise. While this interpretation is consistent with some current measurements and predictions (Walter et al. 2006; Jansen et al. 2022; Kosten et al. 2010; Kuhn et al. 2021), its clarity diminishes over longer timescales, which is the main focus of our study. Given this, we can cautiously hypothesize that northern lakes will increasingly contribute to atmospheric carbon dioxide and methane levels during ongoing climate change[11].

[11] Other studies found that carbon dioxide measurements seem to show more variability depending on the specific lake compared to methane measurements.

5.3 Conclusion

In this dissertation, we explored heterogeneous multi-proxy data from multiple sediment cores in the Arctic through a data-driven approach, aiming to enhance our knowledge of past environmental dynamics of lake systems in this critical region. Harnessing modern data-driven methodologies, our results have not only shed light on the intricacies of these lake systems but have also made significant advances in progressing our understanding of their historical and contemporary evolutions.

In Chapter 2, our development of a comprehensive conceptual model for heterogeneous multi-proxy data confirmed its value in homogenizing datasets and facilitating data transformation for comparative analyses. This model has not only enabled the seamless integration and standardization of diverse data sources but has also greatly enhanced our understanding of past environmental changes in the Arctic by allowing us to effectively use older datasets.

In Chapter 3, we successfully pioneered a tool within a multi-language Jupyter Notebook, bridging multiple modeling systems. This unified framework, combined with the introduction of an ensemble age–depth model, has significantly streamlined the age–depth determination process. As hypothesized, this approach boosted the accuracy and efficiency of sediment core dating, fostering greater comparability of results across various methods.

Lastly, in Chapter 4, our data-driven approach has unveiled a complex relationship between carbon accumulation rates and paleotemperature across a spectrum of lake types. As anticipated, our analysis revealed a pronounced correlation between higher temperatures and increased carbon accumulation rates in lakes, signifying a positive relationship between temperature and carbon balance. Importantly, the influence of factors like permafrost presence, vegetation composition, and specific lake attributes on this relationship was evident, underpinning the multifaceted dynamics at play.

In conclusion, the combined outcomes of this doctoral thesis underscore the invaluable role of modern tools and methodologies in deciphering the Arctic's intricate environmental narratives. Through our findings, we have not only expanded the boundaries of scientific knowledge but have also laid down robust foundations for future endeavors aiming to further dissect the relationships between environmental variables in polar regions.

References

Abadi, Martin, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, et al. 2016. "TensorFlow: A System for Large-Scale Machine Learning." In *12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16)*, 265–83. https://www.usenix.org/system/files/conference/osdi16/osdi16-abadi.pdf.

Abbott, Mark B., and Thomas W. Stafford. 1996. "Radiocarbon Geochemistry of Modern and Ancient Arctic Lake Systems, Baffin Island, Canada." *Quaternary Research* 45 (3): 300–311. https://doi.org/10.1006/qres.1996.0031.

Adams, Heather E., Byron C. Crump, and George W. Kling. 2010. "Temperature Controls on Aquatic Bacterial Production and Community Dynamics in Arctic Lakes and Streams." *Environmental Microbiology* 12 (5): 1319–33. https://doi.org/10.1111/j.1462-2920.2010.02176.x.

Agarwal, S., N. Tosi, P. Kessel, S. Padovan, D. Breuer, and G. Montavon. 2021. "Towards Constraining Mars' Thermal Evolution Using Machine Learning." *Earth and Space Science*, no. February. https://doi.org/10.1029/2020ea001484.

Ahmed, Moinuddin, Kevin J. Anchukaitis, Asfawossen Asrat, Hemant P. Borgaonkar, Martina Braida, Brendan M. Buckley, Ulf Büntgen, et al. 2013. "Continental-Scale Temperature Variability during the Past Two Millennia." *Nature Geoscience* 6 (5): 339–46. https://doi.org/10.1038/ngeo1797.

Alasadi, Suad A., and Wesam S. Bhaya. 2017. "Review of Data Preprocessing Techniques in Data Mining." *Journal of Engineering and Applied Sciences* 12 (16): 4102–7. https://doi.org/10.3923/jeasci.2017.4102.4107.

Ammann, C. M., M. G. Genton, and B. Li. 2010. "Technical Note: Correcting for Signal Attenuation from Noisy Proxy Data in Climate Reconstructions." *Climate of the Past* 6 (2): 273–79. https://doi.org/10.5194/cp-6-273-2010.

Amrhein, Daniel E. 2019. "How Large Are Temporal Representativeness Errors in Paleoclimatology?" *Climate of the Past Discussions*, no. February: 1–26. https://doi.org/10.5194/cp-2019-10.

Anderson, N. John, W. D'Andrea, and Sherilyn C. Fritz. 2009. "Holocene Carbon Burial by Lakes in SW Greenland." *Global Change Biology* 15 (11): 2590–98. https://doi.org/10.1111/j.1365-2486.2009.01942.x.

Anderson, N. John, A. J. Heathcote, Daniel R. Engstrom, D. B. Ryves, K. Mills, Y. T. Prairie, P. A. del Giorgio, et al. 2020. "Anthropogenic Alteration of Nutrient Supply Increases the Global Freshwater Carbon Sink." *Science Advances* 6 (16): 1–9. https://doi.org/10.1126/sciadv.aaw2145.

Anderson, Patricia M., and Anatoly V. Lozhkin. 2015. "Late Quaternary Vegetation of Chukotka (Northeast Russia), Implications for Glacial and Holocene Environments of Beringia." *Quaternary Science Reviews* 107: 112–28. https://doi.org/10.1016/j.quascirev.2014.10.016.

Anderson, Patricia M., Anatoly V. Lozhkin, Tatiana B. Solomatkina, and Thomas A. Brown. 2010. "Paleoclimatic Implications of Glacial and Postglacial Refugia for Pinus Pumila in Western Beringia." *Quaternary Research* 73 (2): 269–76. https://doi.org/10.1016/j.yqres.2009.09.008.

Anderson, Patricia M., Pavel Minyuk, Anatoly Lozhkin, Marina Cherepanova, Vladimir Borkhodoev, and Bruce Finney. 2015. "A Multiproxy Record of Holocene Environmental Changes from the Northern Kuril Islands (Russian Far East)." *Journal of Paleolimnology* 54 (4): 379–93. https://doi.org/10.1007/s10933-015-9858-y.

Andersson, Lennart, Anders Rantzer, and Carolyn Beck. 1999. "Model Comparison and Simplification." *International Journal of Robust and Nonlinear Control* 9 (3): 157–81. https://doi.org/10.1002/(SICI)1099-1239(199903)9:3<157::AID-RNC398>3.0.CO;2-8.

Andreev, Andrei A., Elena Raschke, Boris K. Biskaborn, Stuart A. Vyse, Jeremy Courtin, Thomas Böhmer, Kathleen Stoof-Leichsenring, Stefan Kruse, Lyudmila A. Pestryakova, and Ulrike Herzschuh. 2021. "Late Pleistocene to Holocene Vegetation and Climate Changes in Northwestern Chukotka (Far East Russia) Deduced from Lakes Ilirney and Rauchuagytgyn

Pollen Records." *Boreas* 50 (3): 652–70. https://doi.org/10.1111/bor.12521.
Andreev, Andrei A., Lyudmila S. Shumilovskikh, Larisa A. Savelieva, Raphael Gromig, Grigory B. Fedorov, Anna Ludikova, Bernd Wagner, Volker Wennrich, Dominik Brill, and Martin Melles. 2019. "Environmental Conditions in Northwestern Russia during MIS 5 Inferred from the Pollen Stratigraphy in a Sediment Core from Lake Ladoga." *Boreas* 48 (2): 377–86. https://doi.org/10.1111/bor.12382.
Andreev, Andrei A., Pavel E. Tarasov, Boris P. Ilyashuk, Elena A. Ilyashuk, Holger Cremer, Wolf-Dieter Hermichen, Frank Wischer, and Hans-Wolfgang Hubberten. 2005a. "Age Determinations on a Sediment Profile from Lake Lyadhej-To, PANGAEA [Data Set]." https://doi.org/10.1594/PANGAEA.728450.
———. 2005b. "Holocene Environmental History Recorded in Lake Lyadhej-To Sediments, Polar Urals, Russia." *Palaeogeography, Palaeoclimatology, Palaeoecology* 223 (3): 181–203. https://doi.org/10.1016/j.palaeo.2005.04.004.
Andreev, Andrei A., Pavel E. Tarasov, Vladimir A. Klimanov, Martin Melles, O. M. Lisitsyna, and Hans-Wolfgang Hubberten. 2004. "Vegetation and Climate Changes around the Lama Lake, Taymyr Peninsula, Russia during the Late Pleistocene and Holocene." *Quaternary International* 122 (1 SPEC. ISS.): 69–84. https://doi.org/10.1016/j.quaint.2004.01.032.
Andreev, Andrei A., Pavel E. Tarasov, Christine Siegert, Tobias Ebel, Vladimir A. Klimanov, Martin Melles, Anatoly A. Bobrov, Alexandr Y. Dereviagin, David J. Lubinski, and Hans-Wolfgang Hubberten. 2003a. "Table 1. Radiocarbon Dating on Profile PG1228, PANGAEA [Data Set]." https://doi.org/10.1594/PANGAEA.726591.
———. 2003b. "Late Pleistocene and Holocene Vegetation and Climate on the Northern Taymyr Peninsula, Arctic Russia." *Boreas* 32 (3): 484–505. https://doi.org/10.1080/03009480310003388.
Appleby, Peter G. 2008. "Three Decades of Dating Recent Sediments by Fallout Radionuclides: A Review." *Holocene* 18 (1): 83–93. https://doi.org/10.1177/0959683607085598.
Ascough, Philippa, Gordon Cook, and Andrew Dugmore. 2005. "Methodological Approaches to Determining the Marine Radiocarbon Reservoir Effect." *Progress in Physical Geography* 29 (4): 532–47. https://doi.org/10.1191/0309133305pp461ra.
Asikainen, Celeste A., Pierre Francus, and Julie Brigham-Grette. 2007. "Sedimentology, Clay Mineralogy and Grain-Size as Indicators of 65 Ka of Climate Change from El'gygytgyn Crater Lake, Northeastern Siberia." *Journal of Paleolimnology* 37 (1): 105–22. https://doi.org/10.1007/s10933-006-9026-5.
Austin, William E. N., Edouard Bard, J. B. Hunt, Dick Kroon, and J. D. Peacock. 1995. "The 14C Age of the Icelandic Vedde Ash: Implications for Younger Dryas Marine Reservoir Age Corrections." *Radiocarbon* 37 (1): 53–62. https://doi.org/10.1017/S0033822200014788.
Avnimelech, Yoram, Gad Ritvo, Leon E. Meijer, and Malka Kochba. 2001. "Water Content, Organic Carbon and Dry Bulk Density in Flooded Sediments." *Aquacultural Engineering* 25 (1): 25–33. https://doi.org/10.1016/S0144-8609(01)00068-1.
Ballinger, Thomas J., James E. Overland, Muyin Wang, Uma S. Bhatt, Edward Hanna, S. Kim, R. L. Thoman, and John E. Walsh. 2020. "Arctic Report Card 2020: Surface Air Temperature." In *The NOAA Arctic Report Card*. https://doi.org/10.25923/gcw8-2z06.
Baltzer, Jennifer L., Tyler Veness, Laura E. Chasmer, Anastasia E. Sniderhan, and William L. Quinton. 2014. "Forests on Thawing Permafrost: Fragmentation, Edge Effects, and Net Forest Loss." *Global Change Biology* 20 (3): 824–34. https://doi.org/10.1111/gcb.12349.
Bao, Rui, Ann P. McNichol, Jordon D. Hemingway, Mary C. Lardie Gaylord, and Timothy I. Eglinton. 2019. "Influence of Different Acid Treatments on the Radiocarbon Content Spectrum of Sedimentary Organic Matter Determined by RPO/Accelerator Mass Spectrometry." *Radiocarbon* 61 (2): 395–413. https://doi.org/10.1017/RDC.2018.125.
Bartosiewicz, Maciej, Anna Przytulska, Jean François Lapierre, Isabelle Laurion, Moritz F. Lehmann, and Roxane Maranger. 2019. "Hot Tops, Cold Bottoms: Synergistic Climate Warming and Shielding Effects Increase Carbon Burial in Lakes." *Limnology And Oceanography Letters* 4 (5): 132–44. https://doi.org/10.1002/lol2.10117.
Bartsch, Annett, Angelika Hofler, Christine Kroisleitner, and Anna Maria Trofaier. 2016. "Land Cover Mapping in Northern High Latitude Permafrost Regions with Satellite Data: Achievements and Remaining Challenges." *Remote Sensing* 8 (12). https://doi.org/10.3390/rs8120979.
Batini, Carlo, Cinzia Cappiello, Chiara Francalanci, and Andrea Maurino. 2009. "Methodologies for

Data Quality Assessment and Improvement." *ACM Computing Surveys* 41 (3). https://doi.org/10.1145/1541880.1541883.
Batini, Carlo, and Monica Scannapieca. 2006. *Data Quality*. Data-Centric Systems and Applications. Springer Berlin Heidelberg. https://doi.org/10.1007/3-540-33173-5.
Baud, Alexandre, Jean Philippe Jenny, Pierre Francus, and Irene Gregory-Eaves. 2021. "Global Acceleration of Lake Sediment Accumulation Rates Associated with Recent Human Population Growth and Land-Use Changes." *Journal of Paleolimnology* 66 (4): 453–67. https://doi.org/10.1007/s10933-021-00217-6.
Baumer, Marlene M., Bernd Wagner, Hanno Meyer, Niklas Leicher, Matthias Lenz, Grigory Fedorov, Lyudmila A. Pestryakova, and Martin Melles. 2021. "Climatic and Environmental Changes in the Yana Highlands of North-Eastern Siberia over the Last c. 57 000 Years, Derived from a Sediment Core from Lake Emanda." *Boreas* 50 (1): 114–33. https://doi.org/10.1111/bor.12476.
Bayer, Michael. 2012. "SQLAlchemy." In *The Architecture of Open Source Applications Volume II: Structure, Scale, and a Few More Fearless Hacks*, edited by Amy Brown and Greg Wilson. aosabook.org. http://aosabook.org/en/sqlalchemy.html.
Beck, Hylke E., Niklaus E. Zimmermann, Tim R. McVicar, Noemi Vergopolan, Alexis Berg, and Eric F. Wood. 2018. "Present and Future Köppen-Geiger Climate Classification Maps at 1-Km Resolution." *Scientific Data* 5: 1–12. https://doi.org/10.1038/sdata.2018.214.
Bergen, Karianne J., Paul A. Johnson, Maarten V. De Hoop, and Gregory C. Beroza. 2019. "Machine Learning for Data-Driven Discovery in Solid Earth Geoscience." *Science* 363 (6433). https://doi.org/10.1126/science.aau0323.
Bertino, Elisa, and Hyo-Sang Lim. 2010. "Assuring Data Trustworthiness - Concepts and Research Challenges." In , 1–12. https://doi.org/10.1007/978-3-642-15546-8_1.
Bilbao, Imanol, and Javier Bilbao. 2017. "Overfitting Problem and the Over-Training in the Era of Data: Particularly for Artificial Neural Networks." In *2017 Eighth International Conference on Intelligent Computing and Information Systems (ICICIS)*, 2018-Janua:173–77. IEEE. https://doi.org/10.1109/INTELCIS.2017.8260032.
Birks, Harry John Betteley. 2012. "Overview of Numerical Methods in Palaeolimnology." In *Tracking Environmental Change Using Lake Sediments*, 19–92. Springer.
Birks, Hilary Helen, and Harry John Betteley Birks. 2006. "Multi-Proxy Studies in Palaeolimnology." *Vegetation History and Archaeobotany* 15 (4): 235–51. https://doi.org/10.1007/s00334-006-0066-6.
Biskaborn, Boris K., Amy Forster, Gregor Pfalz, Lyudmila A. Pestryakova, Kathleen Stoof-Leichsenring, Jens Strauss, Tim Kröger, and Ulrike Herzschuh. 2023. "Diatom Responses and Geochemical Feedbacks to Environmental Changes at Lake Rauchuagytgyn (Far East Russian Arctic)." *Biogeosciences* 20 (9): 1691–1712. https://doi.org/10.5194/bg-20-1691-2023.
Biskaborn, Boris K., Ulrike Herzschuh, D. Bolshiyanov, L. Savelieva, and Bernhard Diekmann. 2012a. "Environmental Variability in Northeastern Siberia during the Last ~13,300yr Inferred from Lake Diatoms and Sediment-Geochemical Parameters." *Palaeogeography, Palaeoclimatology, Palaeoecology* 329–330: 22–36. https://doi.org/10.1016/j.palaeo.2012.02.003.
Biskaborn, Boris K., Ulrike Herzschuh, Dmitry Y. Bolshiyanov, Larisa A. Savelieva, and Bernhard Diekmann. 2012b. "(Table 1) Age Determination of Sediment Core PG1984, PANGAEA [Data Set]." https://doi.org/10.1594/PANGAEA.776407.
Biskaborn, Boris K., Ulrike Herzschuh, Dmitry Y. Bolshiyanov, Larisa A. Savelieva, R. Zibulski, and Bernhard Diekmann. 2013a. "Age Determination of Sediment Core PG1972-1 (09-Tik-03), PANGAEA [Data Set]." https://doi.org/10.1594/PANGAEA.780526.
———. 2013b. "Late Holocene Thermokarst Variability Inferred from Diatoms in a Lake Sediment Record from the Lena Delta, Siberian Arctic." *Journal of Paleolimnology* 49 (2): 155–70. https://doi.org/10.1007/s10933-012-9650-1.
Biskaborn, Boris K., Ulrike Herzschuh, Dmitry Y. Bolshiyanov, Georg Schwamborn, and Bernhard Diekmann. 2013a. "Age Determination of Sediment Core PG1975-1 (09-Tik-05), PANGAEA [Data Set]." https://doi.org/10.1594/PANGAEA.780385.
———. 2013b. "Thermokarst Processes and Depositional Events in a Tundra Lake, Northeastern Siberia." *Permafrost and Periglacial Processes* 24 (3): 160–74. https://doi.org/10.1002/ppp.1769.

Biskaborn, Boris K., J. P. Lanckman, H. Lantuit, Kirsten Elger, D. A. Streletskiy, W. L. Cable, and V. E. Romanovsky. 2015. “The New Database of the Global Terrestrial Network for Permafrost (GTN-P).” *Earth System Science Data* 7 (2): 745–59. https://doi.org/10.5194/essd-7-245-2015.

Biskaborn, Boris K., Larisa Nazarova, Tim Kröger, Liudmila A. Pestryakova, Liudmila S. Syrykh, Gregor Pfalz, Ulrike Herzschuh, and Bernhard Diekmann. 2021. “Late Quaternary Climate Reconstruction and Lead-Lag Relationships of Biotic and Sediment-Geochemical Indicators at Lake Bolshoe Toko, Siberia.” *Frontiers in Earth Science* 9 (August): 1–22. https://doi.org/10.3389/feart.2021.737353.

Biskaborn, Boris K., Larisa Nazarova, Lyudmila A. Pestryakova, Liudmila S. Syrykh, Kim Funck, Hanno Meyer, Bernhard Chapligin, et al. 2019. “Spatial Distribution of Environmental Indicators in Surface Sediments of Lake Bolshoe Toko, Yakutia, Russia.” *Biogeosciences* 16 (20): 4023–49. https://doi.org/10.5194/bg-16-4023-2019.

Biskaborn, Boris K., Sharon L. Smith, Jeannette Noetzli, Heidrun Matthes, Gonçalo Vieira, Dmitry A. Streletskiy, Philippe Schoeneich, et al. 2019. “Permafrost Is Warming at a Global Scale.” *Nature Communications* 10 (1): 264. https://doi.org/10.1038/s41467-018-08240-4.

Biskaborn, Boris K., Dmitry A. Subetto, L. A. Savelieva, P. S. Vakhrameeva, A. Hansche, Ulrike Herzschuh, Juliane Klemm, et al. 2016a. “Late Quaternary Vegetation and Lake System Dynamics in North-Eastern Siberia: Implications for Seasonal Climate Variability.” *Quaternary Science Reviews* 147: 406–21. https://doi.org/10.1016/j.quascirev.2015.08.014.

Biskaborn, Boris K., Dmitry A. Subetto, Larisa A. Savelieva, P. S. Vakhrameeva, A. Hansche, Ulrike Herzschuh, Juliane Klemm, et al. 2016b. “Radiocarbon Age Determination on Composite Core PG2023, PANGAEA [Data Set].” https://doi.org/10.1594/PANGAEA.848897.

Bjune, Anne E., Inger Greve Alsos, Jo Brendryen, Mary E. Edwards, Haflidi Haflidason, Maren S. Johansen, Jan Mangerud, et al. 2021. “Rapid Climate Changes during the Lateglacial and the Early Holocene as Seen from Plant Community Dynamics in the Polar Urals, Russia.” *Journal of Quaternary Science* 00 (July): jqs.3352. https://doi.org/10.1002/jqs.3352.

Blaauw, Maarten. 2010. “Methods and Code for ‘classical’ Age-Modelling of Radiocarbon Sequences.” *Quaternary Geochronology* 5 (5): 512–18. https://doi.org/10.1016/j.quageo.2010.01.002.

———. 2012. “Out of Tune: The Dangers of Aligning Proxy Archives.” *Quaternary Science Reviews* 36: 38–49. https://doi.org/10.1016/j.quascirev.2010.11.012.

———. 2021. “Clam: Classical Age-Depth Modelling of Cores from Deposits.” https://cran.r-project.org/package=clam.

Blaauw, Maarten, and J. Andrés Christen. 2011. “Flexible Paleoclimate Age-Depth Models Using an Autoregressive Gamma Process.” *Bayesian Analysis* 6 (3): 457–74. https://doi.org/10.1214/11-BA618.

Blaauw, Maarten, J. Andres Christen, and Marco Antonio Aquino-López. 2021. “Rbacon: Age-Depth Modelling Using Bayesian Statistics.” https://cran.r-project.org/package=rbacon.

Blaauw, Maarten, J. Andrés Christen, K. D. Bennett, and Paula J. Reimer. 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, Maarten, and Einar Heegaard. 2012. “Estimation of Age-Depth Relationships.” In , 379–413. https://doi.org/10.1007/978-94-007-2745-8_12.

Bouchard, Frédéric, Lauren A. MacDonald, Kevin W. Turner, Joshua R. Thienpont, Andrew S. Medeiros, Boris K. Biskaborn, Jennifer Korosi, Roland I. Hall, Reinhard Pienitz, and Brent B. Wolfe. 2017. “Paleolimnology of Thermokarst Lakes: A Window into Permafrost Landscape Evolution.” *Arctic Science* 3 (2): 91–117. https://doi.org/10.1139/as-2016-0022.

Box, Jason E., William T. Colgan, Torben Røjle Christensen, Niels Martin Schmidt, Magnus Lund, Frans Jan W. Parmentier, Ross Brown, et al. 2019. “Key Indicators of Arctic Climate Change: 1971-2017.” *Environmental Research Letters* 14 (4). https://doi.org/10.1088/1748-9326/aafc1b.

Bradley, Raymond S. 2015. *Paleoclimatology: Reconstructing Climates of the Quaternary Second Edition. Paleoclimatology : Reconstructing Climates of the Quaternary.* 3rd ed. Oxford: Elsevier. https://doi.org/10.1029/eo081i050p00613-01.

Brauer, Achim. 2004. “Annually Laminated Lake Sediments and Their Palaeoclimatic Relevance.” In , 109–27. https://doi.org/10.1007/978-3-662-10313-5_7.

Brock, Fiona, Thomas Higham, Peter Ditchfield, and Christopher Bronk Ramsey. 2010. “Current

Pretreatment Methods for AMS Radiocarbon Dating at the Oxford Radiocarbon Accelerator Unit (Orau).” *Radiocarbon* 52 (1): 103–12. https://doi.org/10.1017/S0033822200045069.
Bronk Ramsey, Christopher. 1995. “Radiocarbon Calibration and Analysis of Stratigraphy: The OxCal Program.” *Radiocarbon* 37 (2): 425–30. https://doi.org/10.1017/s0033822200030903.
———. 2008. “Deposition Models for Chronological Records.” *Quaternary Science Reviews* 27 (1–2): 42–60. https://doi.org/10.1016/j.quascirev.2007.01.019.
———. 2009. “Dealing with Outliers and Offsets in Radiocarbon Dating.” *Radiocarbon* 51 (3): 1023–45. https://doi.org/10.1017/s0033822200034093.
Bronk Ramsey, Christopher, and Sharen Lee. 2013. “Recent and Planned Developments of the Program OxCal.” *Radiocarbon* 55 (2): 720–30. https://doi.org/10.1017/s0033822200057878.
Brown, Stuart C., Tom M.L. Wigley, Bette L. Otto-Bliesner, and Damien A. Fordham. 2020. “StableClim, Continuous Projections of Climate Stability from 21000 BP to 2100 CE at Multiple Spatial Scales.” *Scientific Data* 7 (1): 1–13. https://doi.org/10.1038/s41597-020-00663-3.
Cadena-Vela, Susana, Jose-Norberto Mazón, and Andrés Fuster-Guilló. 2020. “Defining a Master Data Management Approach for Increasing Open Data Understandability.” In *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 11878 LNCS:169–78. https://doi.org/10.1007/978-3-030-40907-4_17.
Cai, Li, and Yangyong Zhu. 2015. “The Challenges of Data Quality and Data Quality Assessment in the Big Data Era.” *Data Science Journal* 14: 1–10. https://doi.org/10.5334/dsj-2015-002.
Campbell, I. D., C. Campbell, D. H. Vitt, D. Kelker, L. D. Laird, D. Trew, B. Kotak, D. LeClair, and S. Bayley. 2000. “A First Estimate of Organic Carbon Storage in Holocene Lake Sediments in Alberta, Canada.” *Journal of Paleolimnology* 24 (4): 395–400. https://doi.org/10.1023/A:1008103605817.
Chapin, F. Stuart, and Anthony M. Starfield. 1997. “Time Lags and Novel Ecosystems in Response to Transient Climatic Change in Arctic Alaska.” *Climatic Change* 35 (4): 449–61. https://doi.org/10.1023/A:1005337705025.
Chatterjee, Sourav. 2021. “A New Coefficient of Correlation.” *Journal of the American Statistical Association* 116 (536): 2009–22. https://doi.org/10.1080/01621459.2020.1758115.
Chen, Peter Pin Shan. 1975. “The Entity-Relationship Model: Toward a Unified View of Data.” *ACM SIGIR Forum* 10 (3): 9. https://doi.org/10.1145/1095277.1095279.
Chen, Tianqi, and Carlos Guestrin. 2016. “XGBoost.” In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 13-17-Augu:785–94. New York, NY, USA: ACM. https://doi.org/10.1145/2939672.2939785.
Chen, Wei-Chia, Ammar Tareen, and Justin B. Kinney. 2018. “Density Estimation on Small Datasets.” *Physical Review Letters* 121 (16): 160605. https://doi.org/10.1103/PhysRevLett.121.160605.
Chernokulsky, A. V., F. A. Kozlov, O. G. Zolina, O. N. Bulygina, and V. A. Semenov. 2018. “Climatology of Precipitation of Different Genesis in Northern Eurasia.” *Russian Meteorology and Hydrology* 43 (7): 425–35. https://doi.org/10.3103/S1068373918070014.
Chollet, François, and others. 2015. “Keras: The Python Deep Learning Library.” *Https://Keras.Io*. https://keras.io.
Choudhary, Sonal, Aimeric Blaud, A Mark Osborn, Malcolm C Press, and Gareth K. Phoenix. 2016. “Nitrogen Accumulation and Partitioning in a High Arctic Tundra Ecosystem from Extreme Atmospheric N Deposition Events.” *Science of The Total Environment* 554–555 (0): 303–10. https://doi.org/10.1016/j.scitotenv.2016.02.155.
Ciarletta, Daniel J., Justin L. Shawler, Christopher Tenebruso, Christopher J. Hein, and Jorge Lorenzo-Trueba. 2019. “Reconstructing Coastal Sediment Budgets From Beach- and Foredune-Ridge Morphology: A Coupled Field and Modeling Approach.” *Journal of Geophysical Research: Earth Surface* 124 (6): 1398–1416. https://doi.org/10.1029/2018JF004908.
Clark, Peter U., Arthur S. Dyke, Jeremy D. Shakun, Anders E Carlson, Jorie Clark, Barbara Wohlfarth, Jerry X Mitrovica, Steven W Hostetler, and A Marshall McCabe. 2009. “The Last Glacial Maximum.” *Science* 325 (5941): 710–14. https://doi.org/10.1126/science.1172873.
Clow, David W., Sarah M. Stackpoole, Kristine L. Verdin, David E. Butman, Zhiliang Zhu, David P. Krabbenhoft, and Robert G. Striegl. 2015. “Organic Carbon Burial in Lakes and Reservoirs of the Conterminous United States.” *Environmental Science and Technology* 49 (13): 7614–22. https://doi.org/10.1021/acs.est.5b00373.

Codd, Edgar F. 1970. “A Relational Model of Data for Large Shared Data Banks.” *Communications of the ACM* 13 (6): 377–87. https://doi.org/10.1145/357980.358007.
Cohen, Andrew S. 2003. *Paleolimnology: The History and Evolution of Lake Systems*. New York: Oxford University Press.
Colman, S. M., G. A. Jones, Meyer Rubin, J. W. King, J. A. Peck, and W. H. Orem. 1996. “AMS Radiocarbon Analyses from Lake Baikal, Siberia: Challenges of Dating Sediments from a Large, Oligotrophic Lake.” *Quaternary Science Reviews* 15 (7): 669–84. https://doi.org/10.1016/0277-3791(96)00027-3.
Conze, Ronald, Henning Lorenz, Damian Ulbricht, Kirsten Elger, and Thomas Gorgas. 2017. “Utilizing the International Geo Sample Number Concept in Continental Scientific Drilling during ICDP Expedition COSC-1.” *Data Science Journal* 16: 1–8. https://doi.org/10.5334/dsj-2017-002.
Corner, Geoffrey D., Vasili V. Kolka, Vladimir Y. Yevzerov, and Jakob J. Møller. 2001. “Postglacial Relative Sea-Level Change and Stratigraphy of Raised Coastal Basins on Kola Peninsula, Northwest Russia.” *Global and Planetary Change* 31 (1–4): 155–77. https://doi.org/10.1016/S0921-8181(01)00118-7.
Courtin, Jérémy, Andrei A. Andreev, Elena Raschke, Sarah Bala, Boris K. Biskaborn, Sisi Liu, Heike Zimmermann, et al. 2021. “Vegetation Changes in Southeastern Siberia During the Late Pleistocene and the Holocene.” *Frontiers in Ecology and Evolution* 9: 233. https://doi.org/10.3389/fevo.2021.625096.
Cramer, Wolfgang, Alberte Bondeau, F. Ian Woodward, I. Colin Prentice, Richard A. Betts, Victor Brovkin, Peter M. Cox, et al. 2001. “Global Response of Terrestrial Ecosystem Structure and Function to CO2 and Climate Change: Results from Six Dynamic Global Vegetation Models.” *Global Change Biology* 7 (4): 357–73. https://doi.org/10.1046/j.1365-2486.2001.00383.x.
Crema, Enrico R, and Andrew Bevan. 2020. “Inference From Large Sets of Radiocarbon Dates: Software and Methods.” *Radiocarbon* 00 (00): 1–17. https://doi.org/10.1017/rdc.2020.95.
Cremer, Holger, Bernd Wagner, Martin Melles, and Hans-Wolfgang Hubberten. 2001a. “Age Determination of Sediment Profile PG1214, PANGAEA [Data Set].” https://doi.org/10.1594/PANGAEA.734137.
———. 2001b. “The Postglacial Environmental Development of Raffles Sø, East Greenland: Inferences from a 10,000 Year Diatom Record.” *Journal of Paleolimnology* 26 (1): 67–87. https://doi.org/10.1023/A:1011179321529.
Curry, Patrick A., and Nils Moosdorf. 2019. “An Open Source Web Application for Distributed Geospatial Data Exploration.” *Scientific Data* 6 (October 2018): 1–7. https://doi.org/10.1038/sdata.2019.14.
Dakos, Vasilis, Blake Matthews, Andrew P. Hendry, Jonathan Levine, Nicolas Loeuille, Jon Norberg, Patrik Nosil, Marten Scheffer, and Luc De Meester. 2019. “Ecosystem Tipping Points in an Evolving World.” *Nature Ecology and Evolution* 3 (3): 355–62. https://doi.org/10.1038/s41559-019-0797-2.
Dallmeyer, Anne, Martin Claussen, and Victor Brovkin. 2019. “Harmonising Plant Functional Type Distributions for Evaluating Earth System Models.” *Climate of the Past* 15 (1): 335–66. https://doi.org/10.5194/cp-15-335-2019.
Dallmeyer, Anne, Thomas Kleinen, Martin Claussen, Nils Weitzel, Xianyong Cao, and Ulrike Herzschuh. 2022. “The Deglacial Forest Conundrum.” *Nature Communications* 13 (1): 1–10. https://doi.org/10.1038/s41467-022-33646-6.
Dask Development Team. 2016. “Dask: Library for Dynamic Task Scheduling.” https://dask.org.
DBeaver Community. 2020. “DBeaver.” 2020. https://dbeaver.io/.
Dean, Walter E., and Eville Gorham. 1998. “Magnitude and Significance of Carbon Burial in Lakes, Reservoirs, and Peatlands.” *Geology* 26 (6): 535–38. https://doi.org/10.1130/0091-7613(1998)026<0535:MASOCB>2.3.CO;2.
Dee, M. W., S. W.L. Palstra, A. Th Aerts-Bijma, M. O. Bleeker, S. De Bruijn, F. Ghebru, H. G. Jansen, et al. 2020. “Radiocarbon Dating at Groningen: New and Updated Chemical Pretreatment Procedures.” *Radiocarbon* 62 (1): 63–74. https://doi.org/10.1017/RDC.2019.101.
Dee, Sylvia G., Nathan J. Steiger, Julien Emile-Geay, and Gregory J. Hakim. 2016. “On the Utility of Proxy System Models for Estimating Climate States over the Common Era.” *Journal of Advances in Modeling Earth Systems* 8 (3): 1164–79. https://doi.org/10.1002/2016MS000677.

Denfeld, Blaize A., Helen M. Baulch, Paul A. del Giorgio, Stephanie E. Hampton, and Jan Karlsson. 2018. “A Synthesis of Carbon Dioxide and Methane Dynamics during the Ice-Covered Period of Northern Lakes.” *Limnology And Oceanography Letters* 3 (3): 117–31. https://doi.org/10.1002/lol2.10079.

Diekmann, Bernhard, Lyudmila A. Pestryakova, Larisa Nazarova, Dmitry A. Subetto, Pavel E. Tarasov, Georg Stauch, Arne Thiemann, et al. 2017. “Late Quaternary Lake Dynamics in the Verkhoyansk Mountains of Eastern Siberia: Implications for Climate and Glaciation History.” *Polarforschung* 86 (2): 97–110. https://doi.org/10.2312/polarforschung.86.2.97.

Diepenbroek, Michael, Hannes Grobe, Manfred Reinke, Uwe Schindler, Reiner Schlitzer, Rainer Sieger, and Gerold Wefer. 2002. “PANGAEA - an Information System for Environmental Sciences.” *Computers and Geosciences* 28 (10): 1201–10. https://doi.org/10.1016/S0098-3004(02)00039-0.

Dirksen, Veronika, Oleg Dirksen, Christel van den Bogaard, and Bernhard Diekmann. 2015. “Holocene Pollen Record from Lake Sokoch, Interior Kamchatka (Russia), and Its Paleobotanical and Paleoclimatic Interpretation.” *Global and Planetary Change* 134: 129–41. https://doi.org/10.1016/j.gloplacha.2015.07.010.

Dolman, Andrew M. 2022. “Hamstr: Hierarchical Accumulation Modelling with Stan and R.” https://github.com/EarthSystemDiagnostics/hamstr.

Dolman, Andrew M., and Thomas Laepple. 2018. “Sedproxy: A Forward Model for Sediment-Archived Climate Proxies.” *Climate of the Past* 14 (12): 1851–68. https://doi.org/10.5194/cp-14-1851-2018.

Ebel, Tobias, Martin Melles, and Frank Niessen. 1999. “Laminated Sediments from Levinson-Lessing Lake, Northern Central Siberia - A 30,000 Year Record of Environmental History?” In *Land-Ocean Systems in the Siberian Arctic: Dynamics and History*, edited by Heidemarie Kassens, H. A. Bauch, I. A. Dmitrenko, Hajo Eicken, Hans-Wolfgang Hubberten, Martin Melles, Jörn Thiede, and L. A. Timokhov, 425–36. Heidelberg: Springer Berlin.

Eggermont, Hilde, and Oliver Heiri. 2012. “The Chironomid-Temperature Relationship: Expression in Nature and Palaeoenvironmental Implications.” *Biological Reviews* 87 (2): 430–56. https://doi.org/10.1111/j.1469-185X.2011.00206.x.

Elger, Kirsten, Boris K. Biskaborn, Heinz Pampel, and Hugues Lantuit. 2016. “Open Research Data, Data Portals and Data Publication - An Introduction to the Data Curation Landscape.” *Polarforschung* 85 (2): 119–33. https://doi.org/10.2312/polfor.2016.009.

Elmasri, Ramez, and Shamkant B Navathe. 2009. *Grundlagen von Datenbanksystemen (Fundamentals of Database Systems)*. Pearson.

Emile-Geay, Julien, Deborah Khider, Daniel Garijo, Nicholas P. McKay, Yolanda Gil, Varun Ratnatkar, and Elizabeth Bradley. 2018. “The Linked Earth Ontology: A Modular, Extensible Representation of Open Paleoclimate Data.” *Earth Science Informatics*, 26. http://linked.earth.

Enders, Craig K. 2022. *Applied Missing Data Analysis, 2nd Ed. Applied Missing Data Analysis, 2nd Ed.* Methodology in the Social Sciences. New York, NY, US: The Guilford Press.

Erb, Michael P., Nicholas P. McKay, Nathan Steiger, Sylvia Dee, Chris Hancock, Ruza F. Ivanovic, Lauren J. Gregoire, and Paul Valdes. 2022. “Reconstructing Holocene Temperatures in Time and Space Using Paleoclimate Data Assimilation.” *Climate of the Past* 18 (12): 2599–2629. https://doi.org/10.5194/cp-18-2599-2022.

Ernakovich, Jessica G., Kelly A. Hopping, Aaron B. Berdanier, Rodney T. Simpson, Emily J. Kachergis, Heidi Steltzer, and Matthew D. Wallenstein. 2014. “Predicted Responses of Arctic and Alpine Ecosystems to Altered Seasonality under Climate Change.” *Global Change Biology* 20 (10): 3256–69. https://doi.org/10.1111/gcb.12568.

Eynard-Bontemps, Guillaume, Ryan Abernathey, Joseph Hamman, Aurelien Ponte, and Willi Rath. 2019. “The PANGEO Big Data Ecosystem and Its Use at CNES.” In *Proc. of the 2019 Conference on Big Data from Space (BiDS’2019), EUR 29660 EN*, 49–52. Luxembourg: ARRAY(0x558707436af0). https://doi.org/doi:10.2760/848593.

Fassnacht, F. E., F. Hartig, H. Latifi, C. Berger, J. Hernández, P. Corvalán, and B. Koch. 2014. “Importance of Sample Size, Data Type and Prediction Method for Remote Sensing-Based Estimations of Aboveground Forest Biomass.” *Remote Sensing of Environment* 154 (1): 102–14. https://doi.org/10.1016/j.rse.2014.07.028.

Felden, Janine, Lars Möller, Uwe Schindler, Robert Huber, Stefanie Schumacher, Roland Koppe,

Michael Diepenbroek, and Frank Oliver Glöckner. 2023. “PANGAEA - Data Publisher for Earth & Environmental Science.” *Scientific Data* 10 (1): 347. https://doi.org/10.1038/s41597-023-02269-x.

Ferguson, Adam R, Jessica L Nielson, Melissa H Cragin, Anita E Bandrowski, and Maryann E Martone. 2014. “Big Data from Small Data: Data-Sharing in the ‘long Tail’ of Neuroscience.” *Nature Neuroscience* 17 (11): 1442–47. https://doi.org/10.1038/nn.3838.

Ferland, Marie Eve, Yves T. Prairie, Cristian Teodoru, and Paul A. Del Giorgio. 2014. “Linking Organic Carbon Sedimentation, Burial Efficiency, and Long-Term Accumulation in Boreal Lakes.” *Journal of Geophysical Research: Biogeosciences* 119 (5): 836–47. https://doi.org/10.1002/2013JG002345.

Finkenbinder, M. S., Mark B. Abbott, B. P. Finney, J. S. Stoner, and J. M. Dorfman. 2015. “A Multi-Proxy Reconstruction of Environmental Change Spanning the Last 37,000 Years from Burial Lake, Arctic Alaska.” *Quaternary Science Reviews* 126: 227–41. https://doi.org/10.1016/j.quascirev.2015.08.031.

Food and Agriculture Organization [FAO]. 2012. “Global Ecological Zones for FAO Forest Reporting: 2010 Update.” *Forest Resources Assessment Working Paper 179*, 42. http://www.fao.org/docrep/017/ap861e/ap861e00.pdf.

Forster, Piers M., Christopher J. Smith, Tristram Walsh, William F. Lamb, Robin Lamboll, Mathias Hauser, Aurélien Ribes, et al. 2023. “Indicators of Global Climate Change 2022: Annual Update of Large-Scale Indicators of the State of the Climate System and Human Influence.” *Earth System Science Data* 15 (6): 2295–2327. https://doi.org/10.5194/essd-15-2295-2023.

French, Hugh M. 2007. *The Periglacial Environment*. Thrid. Vol. 9. West Sussex, England: John Wiley & Sons Ltd,. https://doi.org/10.1002/9781118684931.

Friedlingstein, Pierre, Matthew W. Jones, Michael O’Sullivan, Robbie M. Andrew, Dorothee C. E. Bakker, Judith Hauck, Corinne Le Quéré, et al. 2022. “Global Carbon Budget 2021.” *Earth System Science Data* 14 (4): 1917–2005. https://doi.org/10.5194/essd-14-1917-2022.

Friedlingstein, Pierre, Michael O’sullivan, Matthew W. Jones, Robbie M. Andrew, Luke Gregor, Judith Hauck, Corinne Le Quéré, et al. 2022. “Global Carbon Budget 2022.” *Earth System Science Data* 14 (11): 4811–4900. https://doi.org/10.5194/essd-14-4811-2022.

Fritz, Sherilyn C. 2008. “Deciphering Climatic History from Lake Sediments.” *Journal of Paleolimnology* 39 (1): 5–16. https://doi.org/10.1007/s10933-007-9134-x.

Fujiwara, Hayato, Jaroslav Hajek, and Olaf Till. 2021. “Octave Forge - The ‘Parallel’ Package.” https://octave.sourceforge.io/parallel/index.html.

Galloway, J N, F J Dentener, D G Capone, E W Boyer, R W Howarth, S P Seitzinger, G P Asner, et al. 2004. “Nitrogen Cycles: Past, Present, and Future.” *Biogeochemistry* 70 (2): 153–226. https://doi.org/10.1007/s10533-004-0370-0.

Garcia-Molina, Hector, Jeffrey D. Ullman, and Jennifer Widom. 2002. *Database Systems: The Complete Book*. An Alan R. Apt Book. Prentice Hall. https://books.google.de/books?id=96oZAQAAIAAJ.

Gaujoux, Renaud. 2020. “DoRNG: Generic Reproducible Parallel Backend for ‘foreach’ Loops.” https://cran.r-project.org/package=doRNG.

Glückler, Ramesh, Ulrike Herzschuh, Stefan Kruse, Andrei A. Andreev, Stuart Andrew Vyse, Bettina Winkler, Boris K. Biskaborn, Luidmila Pestryakova, and Elisabeth DIetze. 2021. “Wildfire History of the Boreal Forest of South-Western Yakutia (Siberia) over the Last Two Millennia Documented by a Lake-Sediment Charcoal Record.” *Biogeosciences* 18 (13): 4185–4209. https://doi.org/10.5194/bg-18-4185-2021.

Goring, S., J. W. Williams, J. L. Blois, S. T. Jackson, C. J. Paciorek, R. K. Booth, J. R. Marlon, M. Blaauw, and J. A. Christen. 2012. “Deposition Times in the Northeastern United States during the Holocene: Establishing Valid Priors for Bayesian Age Models.” *Quaternary Science Reviews* 48: 54–60. https://doi.org/10.1016/j.quascirev.2012.05.019.

Goswami, Bedartha. 2014. “Uncertainties in Climate Data Analysis.” University of Potsdam.

Grant, Luke, Inne Vanderkelen, Lukas Gudmundsson, Zeli Tan, Marjorie Perroud, Victor M. Stepanenko, Andrey V. Debolskiy, et al. 2021. “Attribution of Global Lake Systems Change to Anthropogenic Forcing.” *Nature Geoscience* 14 (11): 849–54. https://doi.org/10.1038/s41561-021-00833-x.

Greenacre, Michael. 2021. “Compositional Data Analysis.” *Annual Review of Statistics and Its*

Application 8: 271–99. https://doi.org/10.1146/annurev-statistics-042720-124436.
Gromig, Raphael, Bernd Wagner, Volker Wennrich, Grigory Fedorov, Larisa Savelieva, Elodie Lebas, Sebastian Krastel, et al. 2019. "Deglaciation History of Lake Ladoga (Northwestern Russia) Based on Varved Sediments." *Boreas* 48 (2): 330–48. https://doi.org/10.1111/bor.12379.
Grosse, Guido, Benjamin M. Jones, and Christopher D. Arp. 2013. *Thermokarst Lakes, Drainage, and Drained Basins*. *Treatise on Geomorphology*. Vol. 8. Elsevier Ltd. https://doi.org/10.1016/B978-0-12-374739-6.00216-5.
Gudasz, Cristian, David Bastviken, Kristin Steger, Katrin Premke, Sebastian Sobek, and Lars J. Tranvik. 2010. "Temperature-Controlled Organic Carbon Mineralization in Lake Sediments." *Nature* 466 (7305): 478–81. https://doi.org/10.1038/nature09186.
Gudasz, Cristian, Sebastian Sobek, David Bastviken, Birgit Koehler, and Lars J. Tranvik. 2015. "Temperature Sensitivity of Organic Carbon Mineralization in Contrasting Lake Sediments." *Journal of Geophysical Research: Biogeosciences* 120 (7): 1215–25. https://doi.org/10.1002/2015JG002928.
Guillemette, François, Eddie von Wachenfeldt, Dolly N. Kothawala, David Bastviken, and Lars J. Tranvik. 2017. "Preferential Sequestration of Terrestrial Organic Matter in Boreal Lake Sediments." *Journal of Geophysical Research: Biogeosciences* 122 (4): 863–74. https://doi.org/10.1002/2016JG003735.
Habibullah, Muzafar Shah, Badariah Haji Din, Siow Hooi Tan, and Hasan Zahid. 2022. "Impact of Climate Change on Biodiversity Loss: Global Evidence." *Environmental Science and Pollution Research* 29 (1): 1073–86. https://doi.org/10.1007/s11356-021-15702-8.
Hajdas, Irka, Philippa Ascough, Mark H. Garnett, Stewart J. Fallon, Charlotte L. Pearson, Gianluca Quarta, Kirsty L. Spalding, Haruka Yamaguchi, and Minoru Yoneda. 2021. "Radiocarbon Dating." *Nature Reviews Methods Primers* 1 (1): 62. https://doi.org/10.1038/s43586-021-00058-7.
Hammer, Kathrine Jul, Theis Kragh, and Kaj Sand-Jensen. 2019. "Inorganic Carbon Promotes Photosynthesis, Growth, and Maximum Biomass of Phytoplankton in Eutrophic Water Bodies." *Freshwater Biology* 64 (11): 1956–70. https://doi.org/10.1111/fwb.13385.
Han, Jiawei, Micheline Kamber, and Jian Pei. 2012. *Data Mining: Concepts and Techniques*. *Morgan Kaufmann Publishers, Elsevier*. 3rd ed. Vol. 3.
Harris, Charles R., K. Jarrod Millman, Stéfan J. van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, et al. 2020. "Array Programming with NumPy." *Nature* 585 (7825): 357–62. https://doi.org/10.1038/s41586-020-2649-2.
Harwart, Stefanie, Birgit Hagedorn, Martin Melles, and Ulrich Wand. 1999. "Lithological and Biochemical Properties in Sediments of Lama Lake as Indicators for the Late Pleistocene and Holocene Ecosystem Development of the Southern Taymyr Peninsula, Central Siberia." *Boreas* 28 (1): 167–80. https://doi.org/10.1111/j.1502-3885.1999.tb00212.x.
Haslett, John, and Andrew C. Parnell. 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. https://doi.org/10.1111/j.1467-9876.2008.00623.x.
He, Feng. 2011. "Simulating Transient Climate Evolution of the Last Deglaciation with CCSM3." University of Wisconsin Madison, Madison, WC, USA.
Head, Martin J., Brad Pillans, Jan A. Zalasiewicz, Brent Alloway, Alan G. Beu, Kim M. Cohen, Philip L. Gibbard, et al. 2021. "Formal Ratification of Subseries for the Pleistocene Series of the Quaternary System." *Episodes* 44 (3): 241–47. https://doi.org/10.18814/epiiugs/2020/020084.
Heathcote, Adam J., N. John Anderson, Yves T. Prairie, Daniel R. Engstrom, and Paul A. del Giorgio. 2015. "Large Increases in Carbon Burial in Northern Lakes during the Anthropocene." *Nature Communications* 6 (1): 10016. https://doi.org/10.1038/ncomms10016.
Heaton, Timothy J., Peter Köhler, Martin Butzin, Edouard Bard, Ron W. Reimer, William E. N. Austin, Christopher Bronk Ramsey, et al. 2020. "Marine20 - The Marine Radiocarbon Age Calibration Curve (0-55,000 Cal BP)." *Radiocarbon* 62 (4): 779–820. https://doi.org/10.1017/RDC.2020.68.
Heidorn, P. Bryan. 2008. "Shedding Light on the Dark Data in the Long Tail of Science." *Library Trends* 57 (2): 280–99. https://doi.org/10.1353/lib.0.0036.
Hein, Mette. 1997. "Inorganic Carbon Limitation of Photosynthesis in Lake Phytoplankton." *Freshwater Biology* 37 (3): 545–52. https://doi.org/10.1046/j.1365-2427.1997.00180.x.

Heron, Michael J., Vicki L Hanson, and Ian Ricketts. 2013. “Open Source and Accessibility: Advantages and Limitations.” *Journal of Interaction Science* 1 (1): 2. https://doi.org/10.1186/2194-0827-1-2.

Hersbach, Hans, Bill Bell, Paul Berrisford, Shoji Hirahara, András Horányi, Joaquín Muñoz-Sabater, Julien Nicolas, et al. 2020. “The ERA5 Global Reanalysis.” *Quarterly Journal of the Royal Meteorological Society* 146 (730): 1999–2049. https://doi.org/10.1002/qj.3803.

Hippel, Barbara von, Kathleen R. Stoof-Leichsenring, Luise Schulte, Peter Seeber, Boris K. Biskaborn, Bernhard Diekmann, Martin Melles, Lyudmila A. Pestryakova, and Ulrike Herzschuh. 2021. “Long-Term Fungus–Plant Co-Variation from Multi-Site Sedimentary Ancient DNA Metabarcoding in Siberia.” https://doi.org/10.1101/2021.11.05.465756.

Hoff, Ulrike, Boris K. Biskaborn, Veronika G. Dirksen, Oleg Dirksen, Gerhard Kuhn, Hanno Meyer, Larisa Nazarova, Alexandra Roth, and Bernhard Diekmann. 2015. “Holocene Environment of Central Kamchatka, Russia: Implications from a Multi-Proxy Record of Two-Yurts Lake.” *Global and Planetary Change* 134 (November): 101–17. https://doi.org/10.1016/j.gloplacha.2015.07.011.

Hoff, Ulrike, Oleg Dirksen, Veronika Dirksen, Ulrike Herzschuh, Hans-Wolfgang Hubberten, Hanno Meyer, Christel van den Bogaard, and Bernhard Diekmann. 2012. “Late Holocene Diatom Assemblages in a Lake-Sediment Core from Central Kamchatka, Russia.” *Journal of Paleolimnology* 47 (4): 549–60. https://doi.org/10.1007/s10933-012-9580-y.

Hoff, Ulrike, Oleg Dirksen, Veronika Dirksen, Gerhard Kuhn, Hanno Meyer, and Bernhard Diekmann. 2014. “Holocene Freshwater Diatoms: Palaeoenvironmental Implications from South Kamchatka, Russia.” *Boreas* 43 (1): 22–41. https://doi.org/10.1111/bor.12019.

Hogg, Alan G., Timothy J. Heaton, Quan Hua, Jonathan G. Palmer, Chris S.M. Turney, John Southon, Alex Bayliss, et al. 2020. “SHCal20 Southern Hemisphere Calibration, 0-55,000 Years Cal BP.” *Radiocarbon* 62 (4): 759–78. https://doi.org/10.1017/RDC.2020.59.

Hollaway, M. J., P. A. Henrys, R. Killick, A. Leeson, and J. Watkins. 2021. “Evaluating the Ability of Numerical Models to Capture Important Shifts in Environmental Time Series: A Fuzzy Change Point Approach.” *Environmental Modelling and Software* 139 (February): 104993. https://doi.org/10.1016/j.envsoft.2021.104993.

Hollis, J. M., J. Hannam, and P. H. Bellamy. 2012. “Empirically-Derived Pedotransfer Functions for Predicting Bulk Density in European Soils.” *European Journal of Soil Science* 63 (1): 96–109. https://doi.org/10.1111/j.1365-2389.2011.01412.x.

Huang, Zhi, and Shawn W. Laffan. 2009. “Sensitivity Analysis of a Decision Tree Classification to Input Data Errors Using a General Monte Carlo Error Sensitivity Model.” *International Journal of Geographical Information Science* 23 (11): 1433–52. https://doi.org/10.1080/13658810802634949.

Hugelius, G., J. Strauss, S. Zubrzycki, J. W. Harden, E. A.G. Schuur, C. L. Ping, L. Schirrmeister, et al. 2014. “Estimated Stocks of Circumpolar Permafrost Carbon with Quantified Uncertainty Ranges and Identified Data Gaps.” *Biogeosciences* 11 (23): 6573–93. https://doi.org/10.5194/bg-11-6573-2014.

Hughes-Allen, Lara, Frédéric Bouchard, Christine Hatté, Hanno Meyer, Lyudmila A. Pestryakova, Bernhard Diekmann, Dmitry A. Subetto, and Boris K. Biskaborn. 2021. “14,000-Year Carbon Accumulation Dynamics in a Siberian Lake Reveal Catchment and Lake Productivity Changes.” *Frontiers in Earth Science* 9 (November): 1–19. https://doi.org/10.3389/feart.2021.710257.

Huntley, Brian. 2012. “Reconstructing Palaeoclimates from Biological Proxies: Some Often Overlooked Sources of Uncertainty.” *Quaternary Science Reviews* 31: 1–16. https://doi.org/10.1016/j.quascirev.2011.11.006.

IPCC. 2014. *Climate Change 2014: Synthesis Report. Contribution of Working Groups I, II and III to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*. Edited by Core Writing Team, L.A. Meyer, and R.K. Pachauri. Geneva, Switzerland.

———. 2023. “Climate Change 2023: Synthesis Report. A Report of the Intergovernmental Panel on Climate Change. Contribution of Working Groups I, II and III to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change [Core Writing Team, H. Lee and J.” Geneva, Switzerland. https://www.ipcc.ch/report/ar6/syr/downloads/report/IPCC_AR6_SYR_LongerReport.pdf.

Iturrate-Garcia, Maitane, Monique M.P.D. Heijmans, L. Hans C. Cornelissen, Fritz H. Schweingruber,

Pascal A. Niklaus, and Gabriela Schaepman-Strub. 2020. “Plant Trait Response of Tundra Shrubs to Permafrost Thaw and Nutrient Addition.” *Biogeosciences* 17 (20): 4981–98. https://doi.org/10.5194/bg-17-4981-2020.

Jansen, Joachim, Richard Iestyn Woolway, Benjamin M. Kraemer, Clément Albergel, David Bastviken, Gesa A. Weyhenmeyer, Rafael Marcé, et al. 2022. “Global Increase in Methane Production under Future Warming of Lake Bottom Waters.” *Global Change Biology* 28 (18): 5427–40. https://doi.org/10.1111/gcb.16298.

Joblib Development Team. 2020. “Joblib: Running Python Functions as Pipeline Jobs.” https://joblib.readthedocs.io/.

Jumper, John, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, et al. 2021. “Highly Accurate Protein Structure Prediction with AlphaFold.” *Nature* 596 (7873): 583–89. https://doi.org/10.1038/s41586-021-03819-2.

Kageyama, Masa, J Annan, P Braconnot, C Brierley, J Fidel, J Hargreaves, S P P Harrison, et al. 2021. “The Contributions of PMIP to the IPCC Assessment Reports.” *Past Global Changes Magazine* 29 (2): 68–69. https://doi.org/10.22498/pages.29.2.68.

Kaliappan, Jayakumar, Kathiravan Srinivasan, Saeed Mian Qaisar, Karpagam Sundararajan, Chuan Yu Chang, and C. Suganthan. 2021. “Performance Evaluation of Regression Models for the Prediction of the COVID-19 Reproduction Rate.” *Frontiers in Public Health* 9 (September): 1–12. https://doi.org/10.3389/fpubh.2021.729795.

Kalugin, Ivan, Andrei Daryin, Lyubov Smolyaninova, Andrei A. Andreev, Bernhard Diekmann, and Oleg Khlystov. 2007. “800-Yr-Long Records of Annual Air Temperature and Precipitation over Southern Siberia Inferred from Teletskoye Lake Sediments.” *Quaternary Research* 67 (3): 400–410. https://doi.org/10.1016/j.yqres.2007.01.007.

Karger, Dirk Nikolaus, Michael P. Nobis, Signe Normand, Catherine H. Graham, and Niklaus E. Zimmermann. 2023. “CHELSA-TraCE21k - High-Resolution (1 Km) Downscaled Transient Temperature and Precipitation Data since the Last Glacial Maximum.” *Climate of the Past* 19 (2): 439–56. https://doi.org/10.5194/cp-19-439-2023.

Karpatne, Anuj, Imme Ebert-Uphoff, Sai Ravela, Hassan Ali Babaie, and Vipin Kumar. 2019. “Machine Learning for the Geosciences: Challenges and Opportunities.” *IEEE Transactions on Knowledge and Data Engineering* 31 (8): 1544–54. https://doi.org/10.1109/TKDE.2018.2861006.

Kastowski, Martin, Matthias Hinderer, and Adam Vecsei. 2011. “Long-Term Carbon Burial in European Lakes: Analysis and Estimate.” *Global Biogeochemical Cycles* 25 (3): 1–12. https://doi.org/10.1029/2010GB003874.

Kaufman, Darrell S., T. A. Ager, N. John Anderson, Patricia M. Anderson, J. T. Andrews, P. J. Bartlein, Linda B. Brubaker, et al. 2004. “Holocene Thermal Maximum in the Western Arctic (0 - 180° W).” *Quaternary Science Reviews* 23 (5–6): 529–60. https://doi.org/10.1016/j.quascirev.2003.09.007.

Kaufman, Darrell S., and Ellie Broadman. 2023. “Revisiting the Holocene Global Temperature Conundrum.” *Nature* 614 (7948): 425–35. https://doi.org/10.1038/s41586-022-05536-w.

Kaufman, Darrell S., Nicholas P. McKay, Cody C. Routson, Michael P. Erb, Basil A.S. Davis, Oliver Heiri, Samuel Jaccard, et al. 2020. “A Global Database of Holocene Paleotemperature Records.” *Scientific Data* 7 (115): 1–34. https://doi.org/10.1038/s41597-020-0445-3.

Kaufman, Darrell S., Nicholas P. McKay, Cody Routson, Michael P. Erb, Christoph Dätwyler, Philipp S. Sommer, Oliver Heiri, and Basil Davis. 2020. “Holocene Global Mean Surface Temperature, a Multi-Method Reconstruction Approach.” *Scientific Data* 7 (1): 1–13. https://doi.org/10.1038/s41597-020-0530-7.

Kempler, Steve, and Tiffany Mathews. 2017. “Earth Science Data Analytics: Definitions, Techniques and Skills.” *Data Science Journal* 16 (147): 1–8. https://doi.org/10.5334/dsj-2017-006.

Khazin, L. B., I. V. Khazina, S. K. Krivonogov, Ya V. Kuzmin, A. A. Prokopenko, S. Yi, and G. S. Burr. 2016. “Holocene Climate Changes in Southern West Siberia Based on Ostracod Analysis.” *Russian Geology and Geophysics* 57 (4): 574–85. https://doi.org/10.1016/j.rgg.2015.05.012.

Khider, Deborah, Julien Emile-Geay, Nicholas P. McKay, Y. Gil, D. Garijo, V. Ratnakar, M. Alonso-Garcia, et al. 2019. “PaCTS 1.0: A Crowdsourced Reporting Standard for Paleoclimate Data.” *Paleoceanography and Paleoclimatology* 34 (10): 1570–96. https://doi.org/10.1029/2019PA003632.

Kilamo, Terhi, Imed Hammouda, Tommi Mikkonen, and Timo Aaltonen. 2012. “From Proprietary to Open Source—Growing an Open Source Ecosystem.” *Journal of Systems and Software* 85 (7): 1467–78. https://doi.org/https://doi.org/10.1016/j.jss.2011.06.071.

Killick, Rebecca, and Idris A Eckley. 2014. “Changepoint: An R Package for Changepoint Analysis.” *Journal of Statistical Software* 58 (3): 1–19. http://www.jstatsoft.org/v58/i03/.

Killick, Rebecca, Kaylea Haynes, and Idris A Eckley. 2016. “Changepoint: An R Package for Changepoint Analysis.” https://cran.r-project.org/package=changepoint.

Klump, J. Val, David N. Edgington, Liba Granina, and Charles C. Remsen. 2020. “Estimates of the Remineralization and Burial of Organic Carbon in Lake Baikal Sediments.” *Journal of Great Lakes Research* 46 (1): 102–14. https://doi.org/10.1016/j.jglr.2019.10.019.

Kluyver, Thomas, Benjamin Ragan-Kelley, Fernando Pérez, Brian Granger, Matthias Bussonnier, Jonathan Frederic, Kyle Kelley, et al. 2016. “Jupyter Notebooks -- a Publishing Format for Reproducible Computational Workflows.” In *Positioning and Power in Academic Publishing: Players, Agents and Agendas*, edited by F Loizides and B Schmidt, 87–90. https://doi.org/10.3233/978-1-61499-649-1-87.

Kokorowski, H. D., Patricia M. Anderson, C. J. Mock, and A. V. Lozhkin. 2008. “A Re-Evaluation and Spatial Analysis of Evidence for a Younger Dryas Climatic Reversal in Beringia.” *Quaternary Science Reviews* 27 (17–18): 1710–22. https://doi.org/10.1016/j.quascirev.2008.06.010.

Kolka, V. V., O. P. Korsakova, N. B. Lavrova, Tatyana S. Shelekhova, A. N. Tolstobrova, D. S. Tolstobrov, and E. N. Zaretskaya. 2018. “Small Lakes Bottom Sediments Stratigraphy and Paleogeography of the Onega Bay West Coast of the White Sea in the Late Glacial and Holocene.” *Geomorphology RAS*, no. 2: 48–59. https://doi.org/10.7868/S0435428118020049.

Kolka, V V, O P Korsakova, Tatyana S. Shelekhova, and A N Tolstobrova. 2015. “Reconstruction of the Relative Level of the White Sea during the Lateglacial – Holocene According to Lithological, Diatom Analyses and Radiocarbon Dating of Small Lakes Bottom Sediments in the Area of the Chupa Settlement (North Karelia, Russia),” 255–68.

Konapala, Goutam, Ashok K. Mishra, Yoshihide Wada, and Michael E. Mann. 2020. “Climate Change Will Affect Global Water Availability through Compounding Changes in Seasonal Precipitation and Evaporation.” *Nature Communications* 11 (1): 1–10. https://doi.org/10.1038/s41467-020-16757-w.

Kosten, Sarian, Fábio Roland, David M.L. Da Motta Marques, Egbert H. Van Nes, Néstor Mazzeo, Leonel Da S.L. Sternberg, Marten Scheffer, and Jon J. Cole. 2010. “Climate-Dependent CO2 Emissions from Lakes.” *Global Biogeochemical Cycles* 24 (2): 1–7. https://doi.org/10.1029/2009GB003618.

Kottek, Markus, Jürgen Grieser, Christoph Beck, Bruno Rudolf, and Franz Rubel. 2006. “World Map of the Köppen-Geiger Climate Classification Updated.” *Meteorologische Zeitschrift* 15 (3): 259–63. https://doi.org/10.1127/0941-2948/2006/0130.

Kuang, Xueyuan, Frederik Schenk, Rienk Smittenberg, Petter Hällberg, and Qiong Zhang. 2021. “Seasonal Evolution Differences of East Asian Summer Monsoon Precipitation between Bølling-Allerød and Younger Dryas Periods.” *Climatic Change* 165 (1–2). https://doi.org/10.1007/s10584-021-03025-z.

Kublitskiy, Yuriy, Marianna Kulkova, Olga Druzhinina, Dmitry A. Subetto, Miglė Stančikaitė, Laura Gedminienė, and Khikmatulla Arslanov. 2020. “Geochemical Approach to the Reconstruction of Sedimentation Processes in Kamyshovoye Lake (SE Baltic, Russia) during the Late Glacial and Holocene.” *Minerals* 10 (9): 764. https://doi.org/10.3390/min10090764.

Kuhn, McKenzie A., Lauren M. Thompson, Johanna C. Winder, Lucas P. P. Braga, Andrew J. Tanentzap, David Bastviken, and David Olefeldt. 2021. “Opposing Effects of Climate and Permafrost Thaw on CH4 and CO2 Emissions From Northern Lakes.” *AGU Advances* 2 (4): 1–16. https://doi.org/10.1029/2021av000515.

Lacourse, Terri, and Konrad Gajewski. 2020. “Current Practices in Building and Reporting Age-Depth Models.” *Quaternary Research* 96 (1): 28–38. https://doi.org/10.1017/qua.2020.47.

Lan, Jianghu, Hai Xu, Bin Liu, Enguo Sheng, Jiangtao Zhao, and Keke Yu. 2015. “A Large Carbon Pool in Lake Sediments over the Arid/Semiarid Region, NW China.” *Chinese Journal of Geochemistry* 34 (3): 289–98. https://doi.org/10.1007/s11631-015-0047-5.

Larsen, Søren, Tom Andersen, and Dag O. Hessen. 2011. “Climate Change Predicted to Cause Severe

Increase of Organic Carbon in Lakes." *Global Change Biology* 17 (2): 1186–92. https://doi.org/10.1111/j.1365-2486.2010.02257.x.
Latif, Atif, Fidan Limani, and Klaus Tochtermann. 2019. "A Generic Research Data Infrastructure for Long Tail Research Data Management." *Data Science Journal* 18 (1): 1–11. https://doi.org/10.5334/dsj-2019-017.
Lehnherr, Igor, Vincent L. St Louis, Martin Sharp, Alex S. Gardner, John P. Smol, Sherry L. Schiff, Derek C.G. Muir, et al. 2018. "The World's Largest High Arctic Lake Responds Rapidly to Climate Warming." *Nature Communications* 9 (1): 1–9. https://doi.org/10.1038/s41467-018-03685-z.
Li, Qian, Sébastien Gogo, Fabien Leroy, Christophe Guimbaud, and Fatima Laggoun-Défarge. 2021. "Response of Peatland CO2 and CH4 Fluxes to Experimental Warming and the Carbon Balance." *Frontiers in Earth Science* 9 (June): 1–13. https://doi.org/10.3389/feart.2021.631368.
Li, Zhongqiang, Liang He, Huan Zhang, Pablo Urrutia-Cordero, Mattias K. Ekvall, Johan Hollander, and Lars Anders Hansson. 2017. "Climate Warming and Heat Waves Affect Reproductive Strategies and Interactions between Submerged Macrophytes." *Global Change Biology* 23 (1): 108–16. https://doi.org/10.1111/gcb.13405.
Lindgren, Amelie, Gustaf Hugelius, Peter Kuhry, Torben R. Christensen, and Jef Vandenberghe. 2016. "GIS-Based Maps and Area Estimates of Northern Hemisphere Permafrost Extent during the Last Glacial Maximum." *Permafrost and Periglacial Processes* 27 (1): 6–16. https://doi.org/10.1002/ppp.1851.
Lougheed, Bryan C., H. J.L. Van Der Lubbe, and Gareth R. Davies. 2016. "87Sr/86Sr as a Quantitative Geochemical Proxy for 14C Reservoir Age in Dynamic, Brackish Waters: Assessing Applicability and Quantifying Uncertainties." *Geophysical Research Letters* 43 (2): 735–42. https://doi.org/10.1002/2015GL066983.
Lougheed, Bryan C., and Stephen P. Obrochta. 2019. "A Rapid, Deterministic Age-Depth Modeling Routine for Geological Sequences With Inherent Depth Uncertainty." *Paleoceanography and Paleoclimatology* 34 (1): 122–33. https://doi.org/10.1029/2018PA003457.
Lowe, J. John, and Michael Walker. 2014. *Reconstructing Quaternary Environments*. Routledge. https://doi.org/10.4324/9781315797496.
Lozhkin, A. V., P. M. Anderson, T. V. Matrosova, and P. S. Minyuk. 2007. "The Pollen Record from El'gygytgyn Lake: Implications for Vegetation and Climate Histories of Northern Chukotka since the Late Middle Pleistocene." *Journal of Paleolimnology* 37 (1): 135–53. https://doi.org/10.1007/s10933-006-9018-5.
Lozhkin, Anatoly, Patricia M. Anderson, Pavel Minyuk, Julya Korzun, Thomas Brown, Alexander Pakhomov, Valeriya Tsygankova, Sergei Burnatny, and Alexey Naumov. 2018. "Implications for Conifer Glacial Refugia and Postglacial Climatic Variation in Western Beringia from Lake Sediments of the Upper Indigirka Basin." *Boreas* 47 (3): 938–53. https://doi.org/10.1111/bor.12316.
Lozhkin, Anatoly, Marina Cherepanova, Patricia M. Anderson, Pavel Minyuk, Bruce Finney, Alexander Pakhomov, Thomas Brown, Julya Korzun, and Valeriya Tsigankova. 2020. "Late Holocene History of Tokotan Lake (Kuril Archipelago, Russian Far East): The Use of Lacustrine Records for Paleoclimatic Reconstructions from Geologically Dynamic Settings." *Quaternary International* 553: 104–17. https://doi.org/https://doi.org/10.1016/j.quaint.2020.05.023.
Lozhkin, Anatoly, Pavel Minyuk, Marina Cherepanova, Patricia M. Anderson, and Bruce Finney. 2017. "Holocene Environments of Central Iturup Island, Southern Kuril Archipelago, Russian Far East." *Quaternary Research (United States)* 88 (1): 23–38. https://doi.org/10.1017/qua.2017.21.
Lu, Yanwei, Asim Biswas, Minyi Wen, and Bingcheng Si. 2021. "Predicting Bulk Density in Deep Unsaturated Soils Based on Multiple Scale Decomposition." *Geoderma* 385 (July 2020): 114859. https://doi.org/10.1016/j.geoderma.2020.114859.
Mackay, Anson W., Elena V. Bezrukova, Melanie J. Leng, Miriam Meaney, Ana Nunes, Natalia Piotrowska, Angela Self, et al. 2012. "Aquatic Ecosystem Responses to Holocene Climate Change and Biome Development in Boreal, Central Asia." *Quaternary Science Reviews* 41: 119–31. https://doi.org/10.1016/j.quascirev.2012.03.004.
Maher, Louis J, Oliver Heiri, and André F Lotter. 2012. "Assessment of Uncertainties Associated with Palaeolimnological Laboratory Methods and Microfossil Analysis." In *Tracking Environmental*

Change Using Lake Sediments: Data Handling and Numerical Techniques, edited by H John B Birks, André F Lotter, Steve Juggins, and John P Smol, 143–66. Dordrecht: Springer Netherlands. https://doi.org/10.1007/978-94-007-2745-8_6.

Manley, Gordon. 1953. "The Mean Temperature of Central England, 1698–1952." *Quarterly Journal of the Royal Meteorological Society* 79 (340): 242–61. https://doi.org/10.1002/qj.49707934006.

Martin, Hinz, Clemens Schmid, Daniel Knitter, and Carolin Tietze. 2021. "OxcAAR: Interface to 'OxCal' Radiocarbon Calibration." https://cran.r-project.org/package=oxcAAR.

Martín, Miguel Ángel, Miguel Reyes, and F. Javier Taguas. 2017. "Estimating Soil Bulk Density with Information Metrics of Soil Texture." *Geoderma* 287: 66–70. https://doi.org/10.1016/j.geoderma.2016.09.008.

Mathews, George M., and John Vial. 2017. "Overcoming Model Simplifications When Quantifying Predictive Uncertainty," March, 1–37. http://arxiv.org/abs/1703.07198.

Mauro, Andrea De, Marco Greco, and Michele Grimaldi. 2016. "A Formal Definition of Big Data Based on Its Essential Features." *Library Review* 65 (3): 122–35. https://doi.org/10.1108/LR-06-2015-0061.

McKay, Nicholas P., and Julien Emile-Geay. 2016. "Technical Note: The Linked Paleo Data Framework – a Common Tongue for Paleoclimatology." *Climate of the Past* 12 (4): 1093–1100. https://doi.org/10.5194/cp-12-1093-2016.

McKay, Nicholas P., Julien Emile-Geay, and Deborah Khider. 2021. "GeoChronR – an R Package to Model, Analyze, and Visualize Age-Uncertain Data." *Geochronology* 3 (1): 149–69. https://doi.org/10.5194/gchron-3-149-2021.

McKay, Nicholas P., and Darrell S. Kaufman. 2014. "An Extended Arctic Proxy Temperature Database for the Past 2,000 Years." *Scientific Data* 1: 1–10. https://doi.org/10.1038/sdata.2014.26.

Melaku Melese, Solomon. 2016. "Effect of Climate Change on Water Resources." *Journal of Water Resources and Ocean Science* 5 (1): 14. https://doi.org/10.11648/j.wros.20160501.12.

Melles, Martin, Julie Brigham-Grette, Olga Yu Glushkova, Pavel S. Minyuk, Norbert R. Nowaczyk, and Hans-Wolfgang Hubberten. 2007. "Sedimentary Geochemistry of Core PG1351 from Lake El'gygytgyn-a Sensitive Record of Climate Variability in the East Siberian Arctic during the Past Three Glacial-Interglacial Cycles." *Journal of Paleolimnology* 37 (1): 89–104. https://doi.org/10.1007/s10933-006-9025-6.

Melles, Martin, Julie Brigham-Grette, Pavel S. Minyuk, Norbert R. Nowaczyk, Volker Wennrich, Robert M. DeConto, Patricia M. Anderson, et al. 2012. "2.8 Million Years of Arctic Climate Change from Lake El'gygytgyn, NE Russia." *Science* 337 (6092): 315–20. https://doi.org/10.1126/science.1222135.

Mendonça, Raquel, Roger A. Müller, David Clow, Charles Verpoorter, Peter Raymond, Lars J. Tranvik, and Sebastian Sobek. 2017. "Organic Carbon Burial in Global Lakes and Reservoirs." *Nature Communications* 8 (1): 1–6. https://doi.org/10.1038/s41467-017-01789-6.

Menounos, Brian. 1997. "The Water Content of Lake Sediments and Its Relationship to Other Physical Parameters: An Alpine Case Study." *Holocene* 7 (2): 207–12. https://doi.org/10.1177/095968369700700208.

Meredith, Michael, Martin Sommerkorn, Sandra Cassotta, Chris Derksen, Alexey Ekaykin, Anne Hollowed, Gary Kofinas, et al. 2019. "Polar Regions." In *IPCC Special Report on the Ocean and Cryosphere in a Changing Climate*, edited by H.-O. Pörtner, D. C. Roberts, V. Masson-Delmotte, P. Zhai, M. Tignor, E. Poloczanska, K. Mintenbeck, et al., 203–320. Cambridge, UK: Cambridge Univ. Press.

Messager, Mathis Loïc, Bernhard Lehner, Günther Grill, Irena Nedeva, and Oliver Schmitt. 2016. "Estimating the Volume and Age of Water Stored in Global Lakes Using a Geo-Statistical Approach." *Nature Communications* 7: 1–11. https://doi.org/10.1038/ncomms13603.

Meyer, Michael F., Stephanie G. Labou, Alli N. Cramer, Matthew R. Brousil, and Bradley T. Luff. 2020. "The Global Lake Area, Climate, and Population Dataset." *Scientific Data* 7 (1): 1–12. https://doi.org/10.1038/s41597-020-0517-4.

Meyers, Philip A., and R Ishiwatari. 1993. "Lacustrine Organic Geochemistry - an Overview of Indicators of Organic-Matter Sources and Diagenesis in Lake-Sediments." *Organic Geochemistry* 20 (7): 867–900. https://doi.org/10.1016/0146-6380(93)90100-p.

Michalak, Anna M. 2016. "Study Role of Climate Change in Extreme Threats to Water Quality."

Nature 535 (7612): 349–50. https://doi.org/10.1038/535349a.
Microsoft Corporation, and Stephen Weston. 2020a. "DoSNOW: Foreach Parallel Adaptor for the 'snow' Package." https://cran.r-project.org/package=doSNOW.
Microsoft Corporation, and Steve Weston. 2020b. "DoParallel: Foreach Parallel Adaptor for the 'parallel' Package." https://cran.r-project.org/package=doParallel.
———. 2020c. "Foreach: Provides Foreach Looping Construct." https://cran.r-project.org/package=foreach.
Middelboe, Mathias, and Claus Lundsgaard. 2003. "Microbial Activity in the Greenland Sea: Role of DOC Lability, Mineral Nutrients and Temperature." *Aquatic Microbial Ecology* 32 (2): 151–63. https://doi.org/10.3354/ame032151.
Miller, Gifford H., Richard B. Alley, Julie Brigham-Grette, Joan J. Fitzpatrick, Leonid Polyak, Mark C. Serreze, and James W.C. White. 2010. "Arctic Amplification: Can the Past Constrain the Future?" *Quaternary Science Reviews* 29 (15–16): 1779–90. https://doi.org/10.1016/j.quascirev.2010.02.008.
Miner, Kimberley R., Merritt R. Turetsky, Edward Malina, Annett Bartsch, Johanna Tamminen, A. David McGuire, Andreas Fix, Colm Sweeney, Clayton D. Elder, and Charles E. Miller. 2022. "Permafrost Carbon Emissions in a Changing Arctic." *Nature Reviews Earth and Environment* 3 (1): 55–67. https://doi.org/10.1038/s43017-021-00230-3.
Miranda, Nicolas, J. C. Freytag, J. Nordin, R. Biswas, V. Brinnel, C. Fremling, M. Kowalski, A. Mahabal, S. Reusch, and J. van Santen. 2022. "SNGuess: A Method for the Selection of Young Extragalactic Transients." *Astronomy & Astrophysics* 665 (September): A99. https://doi.org/10.1051/0004-6361/202243668.
Mishra, Umakant, Gustaf Hugelius, Eitan Shelef, Yuanhe Yang, Jens Strauss, Alexey Lupachev, Jennifer W. Harden, et al. 2021. "Spatial Heterogeneity and Environmental Predictors of Permafrost Region Soil Organic Carbon Stocks." *Science Advances* 7 (9): 1–13. https://doi.org/10.1126/sciadv.aaz5236.
Mottl, Ondřej, John-Arvid Grytnes, Alistair W.R. Seddon, Manuel J. Steinbauer, Kuber P. Bhatta, Vivian A. Felde, Suzette G.A. Flantua, and H. John B. Birks. 2021. "Rate-of-Change Analysis in Paleoecology Revisited: A New Approach." *Review of Palaeobotany and Palynology* 293 (October): 104483. https://doi.org/10.1016/j.revpalbo.2021.104483.
Müller, Jurek, and Fortunat Joos. 2020. "Global Peatland Area and Carbon Dynamics from the Last Glacial Maximum to the Present – A Process-Based Model Investigation." *Biogeosciences* 17 (21): 5285–5308. https://doi.org/10.5194/bg-17-5285-2020.
Müller, Stefanie, Pavel E. Tarasov, Andrei A. Andreev, and Bernhard Diekmann. 2008. "(Table 2) Radiocarbon Dates from Lake Billyakh, PANGAEA [Data Set]." https://doi.org/10.1594/PANGAEA.708169.
———. 2009. "Late Glacial to Holocene Environments in the Present-Day Coldest Region of the Northern Hemisphere Inferred from a Pollen Record of Lake Billyakh, Verkhoyansk Mts, NE Siberia." *Climate of the Past* 5 (1): 73–84. https://doi.org/10.5194/cp-5-73-2009.
Müller, Stefanie, Pavel E. Tarasov, Andrei A. Andreev, Thomas Tütken, Steffi Gartz, and Bernhard Diekmann. 2010. "Late Quaternary Vegetation and Environments in the Verkhoyansk Mountains Region (NE Asia) Reconstructed from a 50-Kyr Fossil Pollen Record from Lake Billyakh." *Quaternary Science Reviews* 29 (17–18): 2071–86. https://doi.org/10.1016/j.quascirev.2010.04.024.
Münchmeyer, Jannes, Dino Bindi, Ulf Leser, and Frederik Tilmann. 2021. "The Transformer Earthquake Alerting Model: A New Versatile Approach to Earthquake Early Warning." *Geophysical Journal International* 225 (1): 646–56. https://doi.org/10.1093/gji/ggaa609.
Muster, Sina. 2018. "Arctic Freshwater – A Commons Requires Open Science." In *Arctic Summer College Yearbook*, 107–20. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-66459-0_9.
Nazarova, Larisa, Herman Lüpfert, Dmitry A. Subetto, Lyudmila A. Pestryakova, and Bernhard Diekmann. 2013a. "(Table 1) Radiocarbon Dates from the Lake Temje Sediment Core, PANGAEA [Data Set]." https://doi.org/10.1594/PANGAEA.802677.
———. 2013b. "Holocene Climate Conditions in Central Yakutia (Eastern Siberia) Inferred from Sediment Composition and Fossil Chironomids of Lake Temje." *Quaternary International* 290–291: 264–74. https://doi.org/10.1016/j.quaint.2012.11.006.

Ng, Wartini, Budiman Minasny, Wanderson de Sousa Mendes, and José Alexandre Melo Demattê. 2020. “The Influence of Training Sample Size on the Accuracy of Deep Learning Models for the Prediction of Soil Properties with Near-Infrared Spectroscopy Data.” *SOIL* 6 (2): 565–78. https://doi.org/10.5194/soil-6-565-2020.

Nicolle, Marie, Maxime Debret, Nicolas Massei, Christophe Colin, Anne Devernal, Dmitry Divine, Johannes P. Werner, Anne Hormes, Atte Korhola, and Hans W. Linderholm. 2018. “Climate Variability in the Subarctic Area for the Last 2 Millennia.” *Climate of the Past* 14 (1): 101–16. https://doi.org/10.5194/cp-14-101-2018.

Niephaus, Fabio, Tim Felgentreff, and Robert Hirschfeld. 2019. “Towards Polyglot Adapters for the GraalVM.” *ACM International Conference Proceeding Series*. https://doi.org/10.1145/3328433.3328458.

Nowaczyk, N. R., P. Minyuk, M. Melles, J. Brigham-Grette, O. Glushkova, M. Nolan, A. V. Lozhkin, T. V. Stetsenko, P. M. Andersen, and S. L. Forman. 2002. “Magnetostratigraphic Results from Impact Crater Lake El’gygytgyn, Northeastern Siberia: A 300 Kyr Long High-Resolution Terrestrial Palaeoclimatic Record from the Arctic.” *Geophysical Journal International* 150 (1): 109–26. https://doi.org/10.1046/j.1365-246X.2002.01625.x.

Obu, Jaroslav, Sebastian Westermann, Annett Bartsch, Nikolai Berdnikov, Hanne H. Christiansen, Avirmed Dashtseren, Reynald Delaloye, et al. 2019. “Northern Hemisphere Permafrost Map Based on TTOP Modelling for 2000–2016 at 1 km2 Scale.” *Earth-Science Reviews* 193 (April): 299–316. https://doi.org/10.1016/j.earscirev.2019.04.023.

Odaka, Tina Erica, Anderson Banihirwe, Guillaume Eynard-Bontemps, Aurelien Ponte, Guillaume Maze, Kevin Paul, Jared Baker, and Ryan Abernathey. 2020. “The Pangeo Ecosystem: Interactive Computing Tools for the Geosciences: Benchmarking on HPC.” *Communications in Computer and Information Science* 1190 CCIS: 190–204. https://doi.org/10.1007/978-3-030-44728-1_12.

Olefeldt, David, S. Goswami, Guido Grosse, D. Hayes, Gustaf Hugelius, P. Kuhry, A. David McGuire, et al. 2016. “Circumpolar Distribution and Carbon Storage of Thermokarst Landscapes.” *Nature Communications* 7: 1–11. https://doi.org/10.1038/ncomms13043.

Olsen, Jesper, Philippa Ascough, Bryan C. Lougheed, and Peter Rasmussen. 2017. “Radiocarbon Dating in Estuarine Environments.” In , 141–70. https://doi.org/10.1007/978-94-024-0990-1_7.

PAGES 2k Consortium. 2017. “A Global Multiproxy Database for Temperature Reconstructions of the Common Era.” *Scientific Data* 4 (1): 170088. https://doi.org/10.1038/sdata.2017.88.

Palagushkina, Olga, Sebastian Wetterich, Boris K. Biskaborn, Larisa Nazarova, Lutz Schirrmeister, Josefine Lenz, Georg Schwamborn, and Guido Grosse. 2017. “Diatom Records and Tephra Mineralogy in Pingo Deposits of Seward Peninsula, Alaska.” *Palaeogeography, Palaeoclimatology, Palaeoecology* 479: 1–15. https://doi.org/10.1016/j.palaeo.2017.04.006.

Palladino, Mario, Nunzio Romano, Edoardo Pasolli, and Paolo Nasta. 2022. “Developing Pedotransfer Functions for Predicting Soil Bulk Density in Campania.” *Geoderma* 412 (May 2021). https://doi.org/10.1016/j.geoderma.2022.115726.

Pannekoek, Jeroen, Sander Scholtus, and Mark Van der Loo. 2013. “Automated and Manual Data Editing: A View on Process Design and Methodology.” *Journal of Official Statistics* 29 (4): 511–37. https://doi.org/10.2478/jos-2013-0038.

Parker, D. E., T. P. Legg, and C. K. Folland. 1992. “A New Daily Central England Temperature Series, 1772–1991.” *International Journal of Climatology* 12 (4): 317–42. https://doi.org/10.1002/joc.3370120402.

Parnell, Andrew C., Caitlin E. Buck, and Thinh K. Doan. 2011. “A Review of Statistical Chronology Models for High-Resolution, Proxy-Based Holocene Palaeoenvironmental Reconstruction.” *Quaternary Science Reviews* 30 (21–22): 2948–60. https://doi.org/10.1016/j.quascirev.2011.07.024.

Parnell, Andrew C., John Haslett, Judy R.M. Allen, Caitlin E. Buck, and Brian Huntley. 2008. “A Flexible Approach to Assessing Synchroneity of Past Events Using Bayesian Reconstructions of Sedimentation History.” *Quaternary Science Reviews* 27 (19–20): 1872–85. https://doi.org/10.1016/j.quascirev.2008.07.009.

Pearson, Richard G., Steven J. Phillips, Michael M. Loranty, Pieter S.A. Beck, Theodoros Damoulas, Sarah J. Knight, and Scott J. Goetz. 2013. “Shifts in Arctic Vegetation and Associated Feedbacks under Climate Change.” *Nature Climate Change* 3 (7): 673–77.

https://doi.org/10.1038/nclimate1858.
Pedregosa, Fabian, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, et al. 2012. “Scikit-Learn: Machine Learning in Python,” no. May 2014. https://doi.org/10.1007/s13398-014-0173-7.2.
Peel, M. C., B. L. Finlayson, and T. A. McMahon. 2007. “Updated World Map of the Köppen-Geiger Climate Classification.” *Hydrology and Earth System Sciences* 11 (5): 1633–44. https://doi.org/10.5194/hess-11-1633-2007.
Peng, Bo, Gao Wang, Jun Ma, Man Chong Leong, Chris Wakefield, James Melott, Yulun Chiu, Di Du, and John N. Weinstein. 2018. “SoS Notebook: An Interactive Multi-Language Data Analysis Environment.” *Bioinformatics* 34 (21): 3768–70. https://doi.org/10.1093/bioinformatics/bty405.
Pennington, Deana, Imme Ebert-Uphoff, Natalie Freed, Jo Martin, and Suzanne A. Pierce. 2020. “Bridging Sustainability Science, Earth Science, and Data Science through Interdisciplinary Education.” *Sustainability Science* 15 (2): 647–61. https://doi.org/10.1007/s11625-019-00735-3.
Pfalz, Gregor, Bernhard Diekmann, Johann-Christoph Freytag, and Boris K. Biskaborn. 2021. “Harmonizing Heterogeneous Multi-Proxy Data from Lake Systems.” *Computers and Geosciences* 153: 11. https://doi.org/10.1016/j.cageo.2021.104791.
Pfalz, Gregor, Bernhard Diekmann, Johann-Christoph Freytag, Liudmila S. Syrykh, Dmitry A. Subetto, and Boris K. Biskaborn. 2022. “Improving Age–Depth Relationships by Using the LANDO (‘Linked Age and Depth Modeling’) Model Ensemble.” *Geochronology* 4 (1): 269–95. https://doi.org/10.5194/gchron-4-269-2022.
Piotrowska, Natalia, Andrzej Bluszcz, Dieter Demske, Wojciech Granoszewski, and Georg Heumann. 2004. “Extraction and AMS Radiocarbon Dating of Pollen from Lake Baikal Sediments.” *Radiocarbon* 46 (1): 181–87. https://doi.org/10.1017/S0033822200039503.
———. 2005. “Age Determination of Lake Baikal Sediment Cores, PANGAEA [Data Set].” https://doi.org/10.1594/PANGAEA.856103.
Pisaric, M.F.J., G.M. MacDonald, A.A. Velichko, and L.C. Cwynar. 2001. “The Lateglacial and Postglacial Vegetation History of the Northwestern Limits of Beringia, Based on Pollen, Stomate and Tree Stump Evidence.” *Quaternary Science Reviews* 20 (1–3): 235–45. https://doi.org/10.1016/S0277-3791(00)00120-7.
PostgreSQL Global Development Group. 2018. “PostgreSQL 11.” 2018. https://www.postgresql.org/.
Qin, Lu, Lirong Lin, Shuwen Ding, Ceng Yi, Jiazhou Chen, and Zhengchao Tian. 2022. “Evaluation of Pedotransfer Functions for Predicting Particle Density of Soils with Low Organic Matter Contents.” *Geoderma* 416 (March): 115812. https://doi.org/10.1016/j.geoderma.2022.115812.
R Core Team. 2019. “Package ‘ Parallel ,’” 1–14.
Raab, Alexandra, Martin Melles, Glenn W. Berger, Birgit Hagedorn, and Hans-Wolfgang Hubberten. 2003. “Non-Glacial Paleoenvironments and the Extent of Weichselian Ice Sheets on Severnaya Zemlya, Russian High Arctic.” *Quaternary Science Reviews* 22 (21–22): 2267–83. https://doi.org/10.1016/S0277-3791(03)00139-2.
Raissi, Maziar, and George Em Karniadakis. 2018. “Hidden Physics Models: Machine Learning of Nonlinear Partial Differential Equations.” *Journal of Computational Physics* 357 (March): 125–41. https://doi.org/10.1016/j.jcp.2017.11.039.
Rasmussen, C. E., and C. K. I. Williams. 2006. *Gaussian Processes for Machine Learning*. *The Journal of Machine Learning Research*. Vol. 11. https://doi.org/10.1142/S0129065704001899.
Rasmussen, S. Olander, K. K. Andersen, A. M. Svensson, J. P. Steffensen, B. M. Vinther, H. B. Clausen, M. L. Siggaard-Andersen, et al. 2006. “A New Greenland Ice Core Chronology for the Last Glacial Termination.” *Journal of Geophysical Research Atmospheres* 111 (6): 1–16. https://doi.org/10.1029/2005JD006079.
Rasmussen, Sune O., Matthias Bigler, Simon P. Blockley, Thomas Blunier, Susanne L. Buchardt, Henrik B. Clausen, Ivana Cvijanovic, et al. 2014. “A Stratigraphic Framework for Abrupt Climatic Changes during the Last Glacial Period Based on Three Synchronized Greenland Ice-Core Records: Refining and Extending the INTIMATE Event Stratigraphy.” *Quaternary Science Reviews* 106: 14–28. https://doi.org/10.1016/j.quascirev.2014.09.007.
Raynolds, Martha K., Donald A. Walker, Andrew Balser, Christian Bay, Mitch Campbell, Mikhail M. Cherosov, Fred J.A. Daniëls, et al. 2019. “A Raster Version of the Circumpolar Arctic Vegetation Map (CAVM).” *Remote Sensing of Environment* 232 (January). https://doi.org/10.1016/j.rse.2019.111297.

Reback, Jeff, Wes McKinney, jbrockmendel, Joris Van den Bossche, Tom Augspurger, Phillip Cloud, gfyoung, et al. 2020. “Pandas-Dev/Pandas: Pandas 1.1.5,” December. https://doi.org/10.5281/ZENODO.4309786.

Rebba, Ramesh, Sankaran Mahadevan, and Shuping Huang. 2006. “Validation and Error Estimation of Computational Models.” *Reliability Engineering and System Safety* 91 (10–11): 1390–97. https://doi.org/10.1016/j.ress.2005.11.035.

Reimer, Paula J., William E. N. Austin, Edouard Bard, Alex Bayliss, Paul G. Blackwell, Christopher Bronk Ramsey, Martin Butzin, et al. 2020. “The IntCal20 Northern Hemisphere Radiocarbon Age Calibration Curve (0-55 Cal KBP).” *Radiocarbon*, August, 1–33. https://doi.org/10.1017/RDC.2020.41.

Reschke, Maria, Torben Kunz, and Thomas Laepple. 2019. “Comparing Methods for Analysing Time Scale Dependent Correlations in Irregularly Sampled Time Series Data.” *Computers and Geosciences* 123: 65–72. https://doi.org/10.1016/j.cageo.2018.11.009.

Reschke, Maria, Kira Rehfeld, and Thomas Laepple. 2019. “Empirical Estimate of the Signal Content of Holocene Temperature Proxy Records.” *Climate of the Past* 15 (2): 521–37. https://doi.org/10.5194/cp-15-521-2019.

Rethemeyer, J, M Gierga, S Heinze, A Stolz, A Wotte, P Wischhöfer, S Berg, JO Melchert, and A Dewald. 2019. “Current Sample Preparation and Analytical Capabilities of the Radiocarbon Laboratory at CologneAMS.” *Radiocarbon* 61 (5): 1449–60. https://doi.org/10.1017/rdc.2019.16.

Rettelbach, Tabea, Moritz Langer, Ingmar Nitze, Benjamin Jones, Veit Helm, Johann-Christoph Freytag, and Guido Grosse. 2021. “A Quantitative Graph-Based Approach to Monitoring Ice-Wedge Trough Dynamics in Polygonal Permafrost Landscapes.” *Remote Sensing* 13 (16): 3098. https://doi.org/10.3390/rs13163098.

Roiha, Toni, Sari Peura, Mathieu Cusson, and Milla Rautio. 2016. “Allochthonous Carbon Is a Major Regulator to Bacterial Growth and Community Composition in Subarctic Freshwaters.” *Scientific Reports* 6 (September): 1–12. https://doi.org/10.1038/srep34456.

Rothwell, R. Guy, and Ian W. Croudace. 2015. “Micro-XRF Studies of Sediment Cores.” *Micro-XRF Studies of Sediment Cores: Applications of a Non-Destructive Tool for the Environmental Sciences* 17: 25–35. https://doi.org/10.1007/978-94-017-9849-5.

Rubel, Franz, Katharina Brugger, Melanie Walter, Janna R. Vogelgesang, Yuliya M. Didyk, Su Fu, and Olaf Kahl. 2018. “Geographical Distribution, Climate Adaptation and Vector Competence of the Eurasian Hard Tick Haemaphysalis Concinna.” *Ticks and Tick-Borne Diseases* 9 (5): 1080–89. https://doi.org/10.1016/j.ttbdis.2018.04.002.

Rudaya, Natalia A. 2020. “Radiocarbon Dates of Sediment Core Tel2006 Lake Teletskoye, Altai Mountains, Southeastern West Siberia, Russia, PANGAEA [Data Set].” https://doi.org/10.1594/PANGAEA.914417.

Rudaya, Natalia A., Larisa Nazarova, Larisa Frolova, Olga Palagushkina, Vasiliy Soenov, Xianyong Cao, Liudmila S. Syrykh, Ivan Grekov, Demberel Otgonbayar, and Batbayar Bayarkhuu. 2021. “The Link between Climate Change and Biodiversity of Lacustrine Inhabitants and Terrestrial Plant Communities of the Uvs Nuur Basin (Mongolia) during the Last Three Millennia.” *Holocene* 31 (9): 1443–58. https://doi.org/10.1177/09596836211019093.

Rudaya, Natalia A., Larisa Nazarova, Danis Nourgaliev, Olga Palagushkina, Dmitry Papin, and Larisa Frolova. 2012. “Mid-Late Holocene Environmental History of Kulunda, Southern West Siberia: Vegetation, Climate and Humans.” *Quaternary Science Reviews* 48: 32–42. https://doi.org/10.1016/j.quascirev.2012.06.002.

Rudaya, Natalia A., Larisa Nazarova, Elena Novenko, Andrei A. Andreev, Ivan Kalugin, Andrei Daryin, Valery Babich, Hong Chun Li, and Pavel Shilov. 2016. “Quantitative Reconstructions of Mid- to Late Holocene Climate and Vegetation in the North-Eastern Altai Mountains Recorded in Lake Teletskoye.” *Global and Planetary Change* 141: 12–24. https://doi.org/10.1016/j.gloplacha.2016.04.002.

Sadeghi, Behnam. 2022. “Chatterjee Correlation Coefficient: A Robust Alternative for Classic Correlation Methods in Geochemical Studies- (Including ‘TripleCpy’ Python Package).” *Ore Geology Reviews* 146 (May): 104954. https://doi.org/10.1016/j.oregeorev.2022.104954.

Sadeghi Eshkevari, Soheil, Martin Takáč, Shamim N. Pakzad, and Majid Jahani. 2021. “DynNet: Physics-Based Neural Architecture Design for Nonlinear Structural Response Modeling and Prediction.” *Engineering Structures* 229 (April 2020).

https://doi.org/10.1016/j.engstruct.2020.111582.
Savelieva, Larisa A., Andrei A. Andreev, Raphael Gromig, Dmitry A. Subetto, Grigory B. Fedorov, Volker Wennrich, Bernd Wagner, and Martin Melles. 2019. "Vegetation and Climate Changes in Northwestern Russia during the Lateglacial and Holocene Inferred from the Lake Ladoga Pollen Record." *Boreas* 48 (2): 349–60. https://doi.org/10.1111/bor.12376.
Schleusner, Philipp, Boris K. Biskaborn, Frank Kienast, Juliane Wolter, Dmitry A. Subetto, and Bernhard Diekmann. 2015. "Basin Evolution and Palaeoenvironmental Variability of the Thermokarst Lake El'gene-Kyuele, Arctic Siberia." *Boreas* 44 (1): 216–29. https://doi.org/10.1111/bor.12084.
Schuur, Edward A. G., A. David McGuire, Guido Grosse, J W Harden, D J Hayes, Gustaf Hugelius, C D Koven, and P. Kuhry. 2015. "Climate Change and the Permafrost Carbon Feedback." *Nature* 520 (January 2016): 171–79. https://doi.org/10.1038/nature14338.
Schuur, Edward A G, Benjamin Walter Abbott, Roisin Commane, Jessica G. Ernakovich, Eugénie S. Euskirchen, Gustaf Hugelius, Guido Grosse, et al. 2022. "Permafrost and Climate Change: Carbon Cycle Feedbacks From the Warming Arctic." *Annual Review of Environment and Resources* 16 (1): 1–29.
Seabold, Skipper, and Josef Perktold. 2010. "Statsmodels: Econometric and Statistical Modeling with Python." In *9th Python in Science Conference*.
Sebastian-Coleman, Laura. 2013. *Measuring Data Quality for Ongoing Improvement*. *Measuring Data Quality for Ongoing Improvement*. https://doi.org/10.1016/C2011-0-07321-0.
Serikova, S., O. S. Pokrovsky, H. Laudon, I. V. Krickov, A. G. Lim, R. M. Manasypov, and J. Karlsson. 2019. "High Carbon Emissions from Thermokarst Lakes of Western Siberia." *Nature Communications* 10 (1): 1–7. https://doi.org/10.1038/s41467-019-09592-1.
Sharma, Sapna, Kevin Blagrave, John J. Magnuson, Catherine M. O'Reilly, Samantha Oliver, Ryan D. Batt, Madeline R. Magee, et al. 2019. "Widespread Loss of Lake Ice around the Northern Hemisphere in a Warming World." *Nature Climate Change* 9 (3): 227–31. https://doi.org/10.1038/s41558-018-0393-5.
Sharma, Sapna, Michael F. Meyer, Joshua Culpepper, Xiao Yang, Stephanie Hampton, Stella A. Berger, Matthew R. Brousil, et al. 2020. "Integrating Perspectives to Understand Lake Ice Dynamics in a Changing World." *Journal of Geophysical Research: Biogeosciences* 125 (8): 1–18. https://doi.org/10.1029/2020JG005799.
Shelekhova, Tatyana S., N. B. Lavrova, O. V. Lazareva, and Yu. S. Tikhonova. 2021. "Paleogeographic Conditions Of Sedimentation In The Small Lakes Of Western Karelia In The Holocene." In *Routes Of Evolutionary Geography - Issue 2*, 449–54. http://eg.igras.ru/wp-content/uploads/2021/12/Velichko2021.pdf.
Shelekhova, Tatyana S., N. B. Lavrova, and Dmitry A. Subetto. 2021. "Reconstruction of Paleogeographic Conditions in the Late Glacial-Holocene in Central Karelia Based on Comprehensive Analysis of Sediments from the Lake Yuzhnoe Haugilampi." *Известия Русского Географического Общества* 153 (6): 73–89. https://doi.org/10.31857/S0869607121060070.
Shelekhova, Tatyana S., and Nadezhda Lavrova. 2020. "Paleogeographic Reconstructions of the Northwest Karelia Region Evolution in the Holocene Based on the Study of Small Lake Sediments." *Proceedings of the Karelian Research Centre of the Russian Academy of Sciences*, no. 9: 101. https://doi.org/10.17076/lim1268.
Shelekhova, Tatyana S., Yuliya Tikhonova, and Oksana Lazareva. 2021. "Late Glacial and Holocene Natural Environment Dynamics and Evolution of Lake Okunozero, South Karelia: Micropalaeontological Data." *Proceedings of the Karelian Research Centre of the Russian Academy of Sciences* 55 (4): 134. https://doi.org/10.17076/lim1319.
Smol, John P. 2016. "Arctic and Sub-Arctic Shallow Lakes in a Multiple-Stressor World: A Paleoecological Perspective." *Hydrobiologia* 778 (1): 253–72. https://doi.org/10.1007/s10750-015-2543-3.
Sobek, S., N. John Anderson, S. M. Bernasconi, and T. Del Sontro. 2014. "Low Organic Carbon Burial Efficiency in Arctic Lake Sediments." *Journal of Geophysical Research: Biogeosciences* 119 (6): 1231–43. https://doi.org/10.1002/2014JG002612.
Stall, Shelley, Lynn Yarmey, Reid Boehm, Helena Cousijn, Patricia Cruse, Joel Cutcher-Gershenfeld, Robin Dasler, et al. 2018. "Advancing FAIR Data in Earth, Space, and Environmental Science."

Eos 99 (November): 1–9. https://doi.org/10.1029/2018EO109301.
Strunk, Astrid, Jesper Olsen, Hamed Sanei, Arka Rudra, and Nicolaj K. Larsen. 2020. "Improving the Reliability of Bulk Sediment Radiocarbon Dating." *Quaternary Science Reviews* 242 (August): 106442. https://doi.org/10.1016/j.quascirev.2020.106442.
Su, Z., W. Timmermans, Y. Zeng, J. Schulz, V. O. John, R. A. Roebeling, P. Poli, et al. 2018. "An Overview of European Efforts in Generating Climate Data Records." *Bulletin of the American Meteorological Society* 99 (2): 349–59. https://doi.org/10.1175/BAMS-D-16-0074.1.
Subetto, Dmitry A., Larisa Nazarova, Lyudmila A. Pestryakova, Liudmila S. Syrykh, A. V. Andronikov, Boris K. Biskaborn, Bernhard Diekmann, D. D. Kuznetsov, T. V. Sapelko, and I. M. Grekov. 2017. "Paleolimnological Studies in Russian Northern Eurasia: A Review." *Contemporary Problems of Ecology* 10 (4): 327–35. https://doi.org/10.1134/S1995425517040102.
Sun, Siao, Jean-Luc Bertrand-Krajwewski, Anders Lynggaard-Jensen, Joep van den Broeke, Florian Edthofer, Mria do Ceu Almeida, Alvaro Silva Ribeiro, and Jose Menaia. 2011. "Literature Review of Data Validation Methods." *Science and Technology* 47 (2): 95–102.
Sundqvist, H. S., Darrell S. Kaufman, Nicholas P. McKay, N. L. Balascio, J. P. Briner, L. C. Cwynar, H. P. Sejrup, et al. 2014. "Arctic Holocene Proxy Climate Database - New Approaches to Assessing Geochronological Accuracy and Encoding Climate Variables." *Climate of the Past* 10 (4): 1605–31. https://doi.org/10.5194/cp-10-1605-2014.
Sundqvist, Maja K., Zhanfeng Liu, Reiner Giesler, and David A. Wardle. 2014. "Plant and Microbial Responses to Nitrogen and Phosphorus Addition across an Elevational Gradient in Subarctic Tundra." *Ecology* 95 (7): 1819–35. https://doi.org/10.1890/13-0869.1.
Syrykh, Liudmila S., Dmitry A. Subetto, and Larisa Nazarova. 2021. "Paleolimnological Studies on the East European Plain and Nearby Regions: The PaleoLake Database." *Journal of Paleolimnology* 65 (3): 369–75. https://doi.org/10.1007/s10933-020-00172-8.
Tarasov, Pavel E., V. S. Volkova, T. Webb, J. Guiot, Andrei A. Andreev, L. G. Bezusko, T. V. Bezusko, et al. 2000. "Last Glacial Maximum Biomes Reconstructed from Pollen and Plant Macrofossil Data from Northern Eurasia." *Journal of Biogeography* 27 (3): 609–20. https://doi.org/10.1046/j.1365-2699.2000.00429.x.
Tchebakova, N. M., E. Parfenova, and A. J. Soja. 2009. "The Effects of Climate, Permafrost and Fire on Vegetation Change in Siberia in a Changing Climate." *Environmental Research Letters* 4 (4). https://doi.org/10.1088/1748-9326/4/4/045013.
Telford, Richard J. 2019. "Review and Test of Reproducibility of Subdecadal Resolution Palaeoenvironmental Reconstructions from Microfossil Assemblages." *Quaternary Science Reviews* 222 (October). https://doi.org/10.1016/j.quascirev.2019.105893.
Telford, Richard J., Einar Heegaard, and Harry John Betteley Birks. 2004. "The Intercept Is a Poor Estimate of a Calibrated Radiocarbon Age." *Holocene* 14 (2): 296–98. https://doi.org/10.1191/0959683604hl707fa.
Teorey, Toby J., Stephen Buxton, Lowell Fryman, Ralf Hartmut Güting, Terry Halpin, Jan L. Harrington, William H. Inmon, Sam S. Lightstone, Jim Melton, and Tony Morgan. 2008. *Database Design: Know It All*. Morgan Kaufmann.
Thanos, Costantino. 2017. "Research Data Reusability: Conceptual Foundations, Barriers and Enabling Technologies." *Publications* 5 (1): 2. https://doi.org/10.3390/publications5010002.
Tolle, Kristin M., D. Stewart W. Tansley, and Anthony J. G. Hey. 2009. *The Fourth Paradigm: Data-Intensive Scientific Discovery*. Microsoft Research. https://www.microsoft.com/en-us/research/wp-content/uploads/2009/10/Fourth_Paradigm.pdf.
Tolstobrov, Dmitry, Alyona Tolstobrova, Vasily Kolka, Olga Korsakova, and Dmitry A. Subetto. 2018. "Putative Records Of The Holocene Tsunami In Lacustrine Bottom Sediments Near The Teriberka Settlement (Kola Peninsula, Russia)." *Proceedings of the Karelian Research Centre of the Russian Academy of Sciences* 06 (9): 92. https://doi.org/10.17076/lim865.
Tolstobrova, Alena, Dmitry Tolstobrov, Vasily Kolka, and Olga Korsakova. 2016. "Late Glacial And Postglacial History Of Lake Osinovoye (Kola Region) Inferred From Ssdimentary Diatom Assemblages." *Proceedings of the Karelian Research Centre of the Russian Academy of Sciences* 89 (5): 106. https://doi.org/10.17076/lim305.
Touzé-Peiffer, Ludovic, Anouk Barberousse, and Hervé Le Treut. 2020. "The Coupled Model Intercomparison Project: History, Uses, and Structural Effects on Climate Research." *WIREs*

Climate Change 11 (4). https://doi.org/10.1002/wcc.648.
Toynbee, Arnold. 1884. *The Industrial Revolution*. Beacon Paperback, No. 32. Beacon Press. https://books.google.co.uk/books?id=NU0VeWQSrKYC.
Trachsel, Mathias, and Richard J. Telford. 2017. "All Age–Depth Models Are Wrong, but Are Getting Better." *Holocene* 27 (6): 860–69. https://doi.org/10.1177/0959683616675939.
Vamathevan, Jessica, Dominic Clark, Paul Czodrowski, Ian Dunham, Edgardo Ferran, George Lee, Bin Li, et al. 2019. "Applications of Machine Learning in Drug Discovery and Development." *Nature Reviews Drug Discovery* 18 (6): 463–77. https://doi.org/10.1038/s41573-019-0024-5.
Velthuis, Mandy, Sarian Kosten, Ralf Aben, Garabet Kazanjian, Sabine Hilt, Edwin T.H.M. Peeters, Ellen van Donk, and Elisabeth S. Bakker. 2018. "Warming Enhances Sedimentation and Decomposition of Organic Carbon in Shallow Macrophyte-Dominated Systems with Zero Net Effect on Carbon Burial." *Global Change Biology* 24 (11): 5231–42. https://doi.org/10.1111/gcb.14387.
Virtanen, Pauli, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, et al. 2020. "SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python." *Nature Methods*. https://doi.org/10.1038/s41592-019-0686-2.
Vonk, J. E., S. E. Tank, W. B. Bowden, I. Laurion, W. F. Vincent, P. Alekseychik, M. Amyot, et al. 2015. "Reviews and Syntheses: Effects of Permafrost Thaw on Arctic Aquatic Ecosystems." *Biogeosciences* 12 (23): 7129–67. https://doi.org/10.5194/bg-12-7129-2015.
Vuglinsky, Valery, and Dmitry Valatin. 2018. "Changes in Ice Cover Duration and Maximum Ice Thickness for Rivers and Lakes in the Asian Part of Russia." *Natural Resources* 09 (03): 73–87. https://doi.org/10.4236/nr.2018.93006.
Vyse, Stuart A., Ulrike Herzschuh, Andrei A. Andreev, Lyudmila A. Pestryakova, Bernhard Diekmann, Simon J. Armitage, and Boris K. Biskaborn. 2020a. "Age Determination of Sediment Core EN18208 from Lake Ilirney, Chukotka, Far East Russia, PANGAEA [Data Set]." https://doi.org/10.1594/PANGAEA.921228.
———. 2020b. "Geochemical and Sedimentological Responses of Arctic Glacial Lake Ilirney, Chukotka (Far East Russia) to Palaeoenvironmental Change since ~51.8 Ka BP." *Quaternary Science Reviews* 247: 106607. https://doi.org/10.1016/j.quascirev.2020.106607.
Vyse, Stuart A., Ulrike Herzschuh, Gregor Pfalz, Lyudmila A. Pestryakova, Bernhard Diekmann, Norbert Nowaczyk, and Boris K. Biskaborn. 2021. "Sediment and Carbon Accumulation in a Glacial Lake in Chukotka (Arctic Siberia) during the Late Pleistocene and Holocene: Combining Hydroacoustic Profiling and down-Core Analyses." *Biogeosciences* 18 (16): 4791–4816. https://doi.org/10.5194/bg-18-4791-2021.
Wagner, Bernd, and Martin Melles. 2008. "A Holocene Seabird Record from Raffles Sø Sediments, East Greenland, in Response to Climatic and Oceanic Changes." *Boreas* 30 (3): 228–39. https://doi.org/10.1111/j.1502-3885.2001.tb01224.x.
Wagner, Bernd, Martin Melles, Jürgen Hahne, Frank Niessen, and Hans-Wolfgang Hubberten. 2000a. "Age Determination of Sediment Core PG1205, PANGAEA [Data Set]." https://doi.org/10.1594/PANGAEA.734962.
———. 2000b. "Holocene Climate History of Geographical Society Ø, East Greenland — Evidence from Lake Sediments." *Palaeogeography, Palaeoclimatology, Palaeoecology* 160 (1): 45–68. https://doi.org/https://doi.org/10.1016/S0031-0182(00)00046-8.
Walker, Mike, Phil Gibbard, Martin J. Head, Max Berkelhammer, Svante Björck, Hai Cheng, Les C. Cwynar, et al. 2019. "Formal Subdivision of the Holocene Series/Epoch: A Summary." *Journal of the Geological Society of India* 93 (2): 135–41. https://doi.org/10.1007/s12594-019-1141-9.
Walker, Mike, Sigfus Johnsen, Sune Olander Rasmussen, Jørgen-Peder Steffensen, Trevor Popp, Philip Gibbard, Wim Hoek, et al. 2008. "The Global Stratotype Section and Point (GSSP) for the Base of the Holocene Series/Epoch (Quaternary System/Period) in the NGRIP Ice Core." *Episodes* 31 (2): 264–67. https://doi.org/10.18814/epiiugs/2008/v31i2/016.
Walter, K. M., S. A. Zimov, J. P. Chanton, D. Verbyla, and F. S. Chapin. 2006. "Methane Bubbling from Siberian Thaw Lakes as a Positive Feedback to Climate Warming." *Nature* 443 (7107): 71–75. https://doi.org/10.1038/nature05040.
Walther, Gian-reto, Eric Post, Peter Convey, Annette Menzel, Camille Parmesank, Trevor J C Beebee, Jean-marc Fromentin, Ove Hoegh-guldberg I, and Franz Bairlein. 2002. "Ecological Response to Recent Climate Cnahge." *Nature* 416: 389–95.

Wang, Richard Y, Mostapha Ziad, and Yang W Lee. 2001. *Data Quality*. Vol. 23. Advances in Database Systems. Boston: Kluwer. https://doi.org/10.1007/b116303.
Wanner, H., H. Wanner, L. Mercolli, L. Mercolli, M. Grosjean, M. Grosjean, and S. P. Ritz. 2015. "Holocene Climate Variability and Change; a Data-Based Review." *Journal of the Geological Society* 172 (2): 254–63. https://doi.org/10.1144/jgs2013-101.
Weiskopf, Sarah R., Madeleine A. Rubenstein, Lisa G. Crozier, Sarah Gaichas, Roger Griffis, Jessica E. Halofsky, Kimberly J.W. Hyde, et al. 2020. "Climate Change Effects on Biodiversity, Ecosystems, Ecosystem Services, and Natural Resource Management in the United States." *Science of the Total Environment* 733. https://doi.org/10.1016/j.scitotenv.2020.137782.
Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the Tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.
Wilke, Thomas, Bernd Wagner, Bert Van Bocxlaer, Christian Albrecht, Daniel Ariztegui, Diana Delicado, Alexander Francke, et al. 2016. "Scientific Drilling Projects in Ancient Lakes: Integrating Geological and Biological Histories." *Global and Planetary Change* 143: 118–51. https://doi.org/10.1016/j.gloplacha.2016.05.005.
Wilkinson, Mark D., Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, et al. 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, John W., Eric C. Grimm, Jessica L. Blois, Donald F. Charles, Edward B. Davis, Simon J. Goring, Russell W. Graham, et al. 2018. "The Neotoma Paleoecology Database, a Multiproxy, International, Community-Curated Data Resource." *Quaternary Research (United States)* 89 (1): 156–77. https://doi.org/10.1017/qua.2017.105.
Windirsch, Torben, Guido Grosse, Mathias Ulrich, Bruce C. Forbes, Mathias Göckede, Juliane Wolter, Marc Macias-Fauria, Johan Olofsson, Nikita Zimov, and Jens Strauss. 2022. "Large Herbivores on Permafrost— a Pilot Study of Grazing Impacts on Permafrost Soil Carbon Storage in Northeastern Siberia." *Frontiers in Environmental Science* 10 (August): 1–17. https://doi.org/10.3389/fenvs.2022.893478.
Wisz, M. S., R. J. Hijmans, J. Li, A. T. Peterson, C. H. Graham, and A. Guisan. 2008. "Effects of Sample Size on the Performance of Species Distribution Models." *Diversity and Distributions* 14 (5): 763–73. https://doi.org/10.1111/j.1472-4642.2008.00482.x.
Wolfe, Alexander P. 1996. "A High-Resolution Late-Glacial and Early Holocene Diatom Record from Baffin Island, Eastern Canadian Arctic." *Canadian Journal of Earth Sciences* 33 (6): 928–37. https://doi.org/10.1139/e96-070.
Wolfe, Brent B., Thomas W.D. Edwards, and Ramon Aravena. 1999. "Changes in Carbon and Nitrogen Cycling during Tree-Line Retreat Recorded in the Isotopic Content of Lacustrine Organic Matter, Western Taimyr Peninsula, Russia." *Holocene* 9 (2): 215–22. https://doi.org/10.1191/095968399669823431.
Wolfe, Brent B., Thomas W.D. Edwards, Ramon Aravena, Steven L. Forman, Barry G. Warner, Andrei A. Velichko, and Glen M. MacDonald. 2000. "Holocene Paleohydrology and Paleoclimate at Treeline, North-Central Russia, Inferred from Oxygen Isotope Records in Lake Sediment Cellulose." *Quaternary Research* 53 (3): 319–29. https://doi.org/10.1006/qres.2000.2124.
Wright, Alexander J., Robin J. Edwards, Orson van de Plassche, Maarten Blaauw, Andrew C. Parnell, Klaas van der Borg, Arie F.M. de Jong, Helen M. Roe, Katherine Selby, and Stuart Black. 2017. "Reconstructing the Accumulation History of a Saltmarsh Sediment Core: Which Age-Depth Model Is Best?" *Quaternary Geochronology* 39: 35–67. https://doi.org/10.1016/j.quageo.2017.02.004.
Ying, Xue. 2019. "An Overview of Overfitting and Its Solutions." *Journal of Physics: Conference Series* 1168 (2). https://doi.org/10.1088/1742-6596/1168/2/022022.
Zaharia, Matei, Reynold S. Xin, Patrick Wendell, Tathagata Das, Michael Armbrust, Ankur Dave, Xiangrui Meng, et al. 2016. "Apache Spark: A Unified Engine for Big Data Processing." *Communications of the ACM* 59 (11): 56–65. https://doi.org/10.1145/2934664.
Zander, Paul D., Sönke Szidat, Darrell S. Kaufman, Maurycy Żarczyński, Anna I. Poraj-Górska, Petra Boltshauser-Kaltenrieder, and Martin Grosjean. 2020. "Miniature Radiocarbon Measurements (<

150 Mg C) from Sediments of Lake Żabińskie, Poland: Effect of Precision and Dating Density on Age–Depth Models." *Geochronology* 2 (1): 63–79. https://doi.org/10.5194/gchron-2-63-2020.
Zandt, Michiel H. in 't, Susanne Liebner, and Cornelia U. Welte. 2020. "Roles of Thermokarst Lakes in a Warming World." *Trends in Microbiology* 28 (9): 769–79. https://doi.org/10.1016/j.tim.2020.04.002.
Zeng, Yijian, Zhongbo Su, Iakovos Barmpadimos, Adriaan Perrels, Paul Poli, K. Folkert Boersma, Anna Frey, et al. 2019. "Towards a Traceable Climate Service: Assessment of Quality and Usability of Essential Climate Variables." *Remote Sensing* 11 (10). https://doi.org/10.3390/rs11101186.
Zhang, Yunlin, Erik Jeppesen, Xiaohan Liu, Boqiang Qin, Kun Shi, Yongqiang Zhou, Sidinei Magela Thomaz, and Jianmin Deng. 2017. "Global Loss of Aquatic Vegetation in Lakes." *Earth-Science Reviews* 173 (May): 259–65. https://doi.org/10.1016/j.earscirev.2017.08.013.
Zhdanova, A. N., E. P. Solotchina, P. A. Solotchin, S. K. Krivonogov, and I. V. Danilenko. 2017. "Reflection of Holocene Climatic Changes in Mineralogy of Bottom Sediments from Yarkovsky Pool of Lake Chany (Southern West Siberia)." *Russian Geology and Geophysics* 58 (6): 692–701. https://doi.org/10.1016/j.rgg.2016.07.005.
Zhu, Yuehua, Muli Yang, Cheng Deng, and Wei Liu. 2020. "Fewer Is More: A Deep Graph Metric Learning Perspective Using Fewer Proxies." *Advances in Neural Information Processing Systems* 2020-Decem (NeurIPS): 1–12.
Zolitschka, Bernd, Pierre Francus, Antti E.K. Ojala, and Arndt Schimmelmann. 2015. "Varves in Lake Sediments - a Review." *Quaternary Science Reviews* 117: 1–41. https://doi.org/10.1016/j.quascirev.2015.03.019.
Zona, D., D. A. Lipson, J. H. Richards, G. K. Phoenix, A. K. Liljedahl, M. Ueyama, C. S. Sturtevant, and W. C. Oechel. 2014. "Delayed Responses of an Arctic Ecosystem to an Extreme Summer: Impacts on Net Ecosystem Exchange and Vegetation Functioning." *Biogeosciences* 11 (20): 5877–88. https://doi.org/10.5194/bg-11-5877-2014.
Zwart, J. A., Z. J. Hanson, J. S. Read, M. N. Fienen, A. F. Hamlet, D. Bolster, and S. E. Jones. 2019. "Cross-Scale Interactions Dictate Regional Lake Carbon Flux and Productivity Response to Future Climate." *Geophysical Research Letters* 46 (15): 8840–51. https://doi.org/10.1029/2019GL083478.

Appendix A

Code and data availability for each chapter

Chapter 2:

This chapter used multiple Python scripts for visualization, database connection, calculation, and interpolation. The codes are available at GitHub (https://github.com/GPawi/MAYHEM). We provide a SQL script to create a blank database following the introduced conceptual data model in the same repository. Likewise, readers can find further files containing accessible links to the used datasets and contact details for unpublished data in there. Contact details comprise name of research group and personal communication address of working group leader.

Chapter 3:

The LANDO code is accessible on GitHub/Zenodo (https://github.com/GPawi/LANDO; https://doi.org/10.5281/zenodo.5734333). We provide five example spreadsheets in the repository for users to test the application. The dataset with all dating points used in this study, including their references, is accessible via PANGAEA (https://doi.org/10.1594/PANGAEA.945777).

Chapter 4:

The age determination data and its metadata used for this study are available on Pangaea: https://doi.org/10.1594/PANGAEA.945777**.** Readers can find proxy data published on Pangaea (https://pangaea.de/) by searching for each unique CoreID. Codes are available on Zenodo (https://doi.org/10.5281/zenodo.7997700).

Appendix B

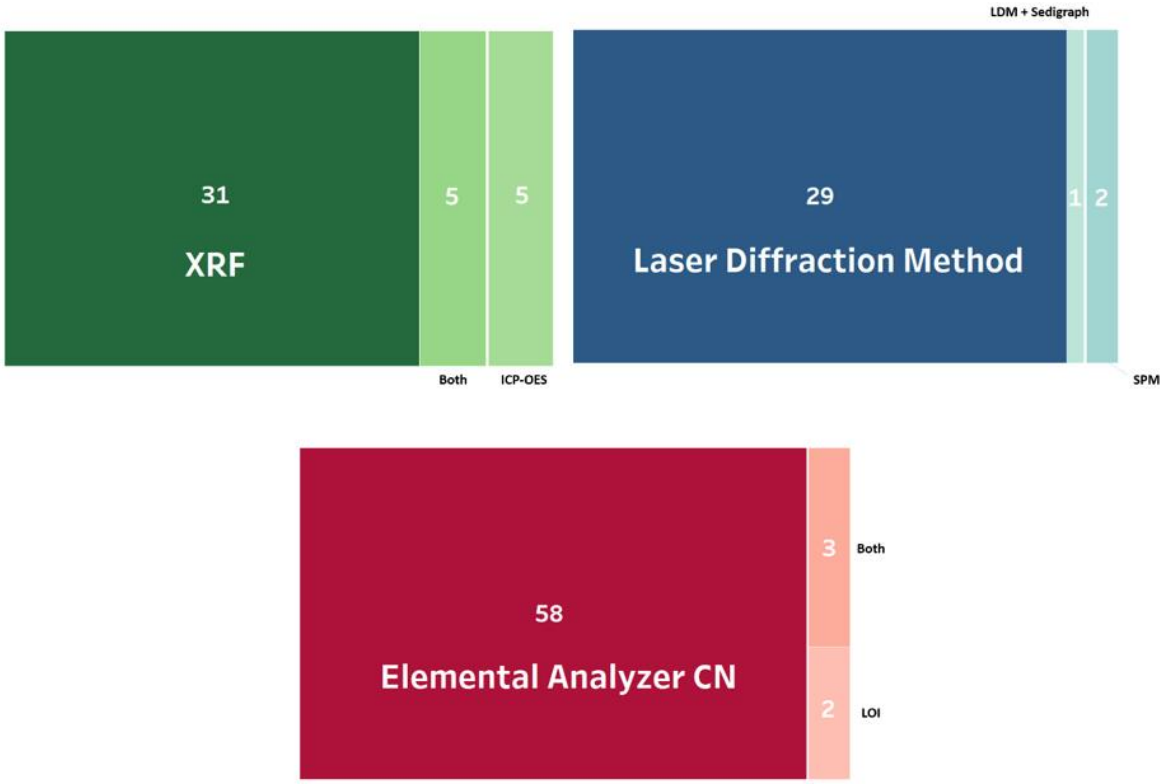

Figure B1: Treemaps showing the number of methods used for the proxy analysis of elements (upper left corner, green treemap), grain size (upper right corner, blue treemap) and organic carbon (lower red treemap) in the data collection of 70 unique sediment cores. For elemental data (green treemap), we found that 31 studies used primarily X-ray fluorescence (XRF). Five other studies worked solely with inductively coupled plasma optical emission spectrometry (ICP-OES) to determine elements. Five datasets provided data both from XRF and ICP-OES analysis. Grain size determination methods (blue treemap) included laser diffraction method (LDM), standard sieve-pipette method (SPM), or a combination of LDM and Sedigraph. LDM was the most used technique for grain size within the data collection (29 datasets). Data showed that scientists decided to use SPM for two sediment cores and both LDM and Sedigraph for one core. 63 studies conducted the analysis of the organic content within the sediment (red treemap) with either an elemental analyzer, or through Loss-On-Ignition (LOI). 58 datasets used the elemental analyzer for organic carbon determination. A small proportion focused solely on LOI (two datasets). Three datasets used a combination of LOI and elemental analyzer.

List B1: The following listing contains the core-specific and measurement-specific entities with their attributes for the reference implementation. This listing includes a short explanation, the PostgreSQL data type, and - in some cases - an example. Underlined attributes are the primary keys of the corresponding entity. Entities '*participant'*, '*citation'* and '*storage'* are bridge tables.

1. scientist:

- scientistid — sequential ID numbering; *int4*
- firstname — fully written-out first name of scientist plus the initials of his/her middle names; *varchar* - example: "*George E.A."*
- lastname — fully written-out last name of scientist; *varchar*
- email — email address for contacting scientist; *varchar*
- orcid — if provided by the scientist, the fully written-out https URI of the "Open Researcher and Contributor IDentifier", also known as ORCID; *varchar(19)* - example: 0000-0003-1218-177X

2. expedition:

- expeditionname — name of the field campaign / expedition (mostly the region name) or the corresponding project name in English with Latin letters; *varchar* - example: "*Chukotka"*
- expeditionyear — four digits resembling the year of the expedition; *int4*

❖ *Foreign Key: scientistid*

3. climateclassification:

- climatezone — abbreviation of climate zonation according to Köppen-Geiger, based on Beck et al. (2018); *varchar(3)* - example: "*Dfc"*
- czdefinition — short definition of individual climate zone; *varchar* - example: "*Cold - Without Dry Season - Cold Summer"*

4. vegetationclassification:

- vegetationzone — description of the vegetation in the surrounding area, based on classification of vegetation types according to Biskaborn et al. (2015), which is based on the Global Ecological Zones (GEZ) spatial dataset (Food and Agriculture Organization [FAO] 2012) and the Bioclimatic subzones in the Circumpolar Arctic (Raynold et al., 2019); *varchar* - example*: "Shrub Tundra"*
- vzdefinition — short definition of individual vegetation zone; *varchar* - example: "*Less than 5% of ground covered by lichens, mosses and vascular plants – Other location"*

5. lakeclassification

- laketype — name of lake type based on Grosse et al. (2013), Lowe and Walker (2014), and Olefeldt et al. (2016); *varchar*
- lakedefinition — short definition of lake origin based on their formation; *varchar* – example: *"Closed depression formed by settlement of the ground following thawing of ice-rich permafrost or the melting of massive ice"*

6. lake:

- lakeid — unique lake identifier derived from the HydroLAKES database (Messager et al., 2016; Meyer et al., 2020). In case of absence, unique identifier will be assigned; *varchar*
- sitename — name of the lake in English with Latin letters; *varchar*
- country — country in which the lake is located in English with Latin letters; *varchar*
- lakedepth — deepest point of the lake in unit [m]; *numeric (6,2)*
- lakeextent — approximate area of the lake in unit [km²]; *float8*
- catchmentarea — approximate area of the catchment supplying the lake with water in unit [km²]; *float8*

❖ *Foreign Key: climatezone, laketypeID, vegetationzone*

7. drilling:

- coreid — unique identifier of cumulative core segments from one drilling location; *varchar*
- latitude — geographical coordinates in unit [decimal degrees, °] from 0° to 90°; *numeric (9,6)*
- longitude — geographical coordinates in unit [decimal degrees, °] from -180° to 180°; *numeric (9,6)*
- waterdepth — specific water depth at drilling location in unit [m]; *numeric (5,2)*
- corelength — cumulative length of core in unit [m]; *numeric (5,2)*
- drillingdevice — name of drilling device with focus on drilling mechanism, such as push-tube corer, piston corer, hammer action corer, gravity corer, Russian peat corer, etc.; *varchar* - example: "*UWITEC hammer action gravity corer"*

❖ *Foreign Keys: sitename, expeditionname, expeditionyear*

8. publication:

- pubid — sequential ID numbering; *int4*
- pubshort — short text citation similar to the in-text APA citation style; *varchar* - example: "*Raab, 2003"* or "*Hahne & Melles, 1999"* or "*Diekmann et al., 2016"*
- citation — citation of publication in reference list APA citation style; *varchar*
- type — type definition - corresponding publication has to be either of type "*journal*", "*book chapter*", "*report*" or "*thesis*"; *varchar (12)*
- doi — digital object identifier of publication; *varchar*

9. source

- fileid — sequential ID numbering; *int4*
- repository — name of the source - either the name of the repository from which the file originates, 'internal', or 'unpublished'; *varchar*
- filename — fully written-out name of the file containing the information about the entity; *varchar*
- accessible — DOI or link to an accessible file containing information about the entity, or name of person, who could provide this file; *varchar*

10. measurement:

- measurementid — composite key, which combines the unique CoreID with the composite depth of an analytical measurement, plus in some cases indicates, whether measurement was a repeated determination; *varchar* - example: "*PG1755 25.5*" *or "PG2133 18_duplicate1"*
- compositedepth — midpoint centimeter with maximum two decimal digits of corresponding analytical measurement; *float8*

❖ *Foreign Key: coreid*

11. chironomid:

- chironomid_taxa — species name of chironomid found in one sample at a specific composite depth within one core; *varchar*
- chironomid_count — raw counts of chironomid head capsules, no percentages; *numeric*

❖ *Foreign Key: measurementid*

12. diatom:

- diatom_taxa — species name of diatom found in one sample at a specific composite depth within one core; *varchar*
- diatom_count — raw counts of diatom valves, no percentages; *numeric*

❖ *Foreign Key: measurementid*

13. pollen:

- pollen_taxa — pollen taxa of grain found in one sample at a specific composite depth within one core; *varchar*
- pollen_count — raw counts of pollen grains, no percentages; *numeric*

❖ *Foreign Key: measurementid*

14. organic:

- tn — total nitrogen content of sediment sample at a specific composite depth in unit [wt%]; *numrange*
- tc — total carbon content of sediment sample at a specific composite depth in unit [wt%]; *numrange*
- toc — total organic carbon content of sediment sample at a specific composite depth in unit [wt%]; *numrange*
- d13c — ratio of stable isotopes ^{13}C to ^{12}C within one sediment sample at a specific composite depth against the Vienna PDB reference in unit [per mil]; *numrange*
- water_content — water content of sediment sample at a specific composite depth, determined by drying the sample in unit [wt%]; *numrange*

❖ *Foreign Key: measurementid*

15. grainsize:

- total_clay — total percentage of clay within one sediment sample a specific composite depth in unit [Vol-%]; *numeric (4,2)*
- total_silt — total percentage of silt within one sediment sample a specific composite depth in unit [Vol-%]; *numeric (4,2)*
- fine_silt — percentage of the fine silt fraction within one sediment sample a specific composite depth in unit [Vol-%]; *numeric (4,2)*
- medium_silt — percentage of the medium silt fraction within one sediment sample a specific composite depth in unit [Vol-%]; *numeric (4,2)*
- coarse_silt — percentage of the coarse silt fraction within one sediment sample a specific composite depth in unit [Vol-%]; *numeric (4,2)*
- total_sand — total percentage of sand within one sediment sample a specific composite depth in unit [Vol-%]; *numeric (4,2)*
- fine_sand — percentage of the fine sand fraction within one sediment sample a specific composite depth in unit [Vol-%]; *numeric (4,2)*
- medium_sand — percentage of the medium sand fraction within one sediment sample a specific composite depth in unit [Vol-%]; *numeric (4,2)*
- coarse_sand — percentage of the coarse sand fraction within one sediment sample a specific composite depth in unit [Vol-%]; *numeric (4,2)*
- total_gravel — total percentage of gravel within one sediment sample a specific composite depth in unit [Vol-%]; *numeric (4,2)*

❖ *Foreign Key: measurementid*

16. element:

- element_name — XRF data will be displayed as "_Area", "_DArea", and "_ChiSqr", while ICP-OES data is indicated through "_percent" and "_ppm"; *varchar* - example: *"Al_Area", "Al_DArea", "Al_ChiSqr", "Al2O3_percent", or "Al_ppm"*
- element_value — XRF data is given in unit [counts per second, cps], ICP-OES data will either be in unit [%] or [ppm]; *numrange*

❖ *Foreign Key: measurementid*

17. mineral:

- mineral_name — fully written-out name of mineral from XRD analysis, no abbreviations; *varchar*
- mineral_wavelength — corresponding wavelength from the selected XRD peak for the mineral in unit [Ångström, Å]; *numeric*
- mineral_intensity — intensity of the selected XRD peak in unit [counts per second, cps]; *int4*

❖ *Foreign Key: measurementid*

18. agedetermination:

- thickness — thickness of the layer from which the sample for age determination was taken from; *numeric*
- labid — unique sample identifier that was provided by the laboratory for age determination; *varchar*
- lab_location — name of city, where laboratory that conducted the analysis resides; *varchar*
- material_category — one of the eight categories that describes the material best, based on the categories from age depth model "Undatable" (Lougheed and Obrochta 2019); *varchar (22)*:
  14C marine fossil - 14C terrestrial fossil - 14C sediment – tephra - tie point – paleomag - U/Th - other
- material_description — short description of the used material; *varchar*
- material_weight — weight of the material that was used for radiocarbon dating in unit [µg]; *int4*
- age — uncalibrated radiocarbon age in unit [years Before Present (before 1950 AD), uncal yr BP], or Lead-210/Cesium-137 ages as negative values in unit [years Before Present (before 1950 AD), yr BP]; *numrange*
- age_error — error of the uncalibrated radiocarbon age in unit [years, uncal yr], or Lead-210/Cesium-137 age in unit [years, yr]; *numeric (7,2)*
- pretreatment_dating — some determinations have a special pretreatment of specific fractions, for instance, "SOL" stands for humic acid fraction, which are alkali-soluble; *varchar* - example: *"SOL", "RES", "TOC", "AAA", "A", "ABA", or "OSL"*
- reservoir_age — additional reservoir effect of dates, also known as hard-water effect or age offset, that was added by the author for correction in unit [yr], if not stated, it is assumed to be 0 yr; *numeric (7,2)*
- reservoir_error — error of reservoir age stated by the author in unit [yr], if not stated, it is assumed to be 0 yr; *numeric (7,2)*

❖ *Foreign Key: measurementid*

19. modeloutput

- modeloutput_median — median calibrated age from one age depth model for a specific composite depth in unit [cal yr BP]; *numeric (8,2)*
- modeloutput_mean — mean calibrated age from one age depth model for a specific composite depth in unit [cal yr BP]; *numeric (8,2)*
- lower_2_sigma — lower 2σ boundary as calibrated age from one age depthmodel for a specific composite depth in unit [cal yr BP]; *numeric (8,2)*
- lower_1_sigma — lower 1σ boundary as calibrated age from one age depthmodel for a specific composite depth in unit [cal yr BP]; *numeric (8,2)*
- upper_1_sigma — upper 1σ boundary as calibrated age from one age depth model for a specific composite depth in unit [cal yr BP]; *numeric (8,2)*
- upper_2_sigma — upper 2σ boundary as calibrated age from one age depth model for a specific composite depth in unit [cal yr BP]; *numeric (8,2)*
- model_name — name of the applied age depth modeling software / package; *varchar* – example: *"Undateable", "Bacon", "Bchron", "OxCal", or "Clam"*

❖ *Foreign Key: measurementid*

20. storage

- entity — name of measurement-specific entity; *varchar*

❖ *Foreign Key: coreid, fileID*

21. participant

❖ *Foreign Key: coreid, scientistid*

22. citation

❖ *Foreign Key: coreid, pubid*

Appendix C

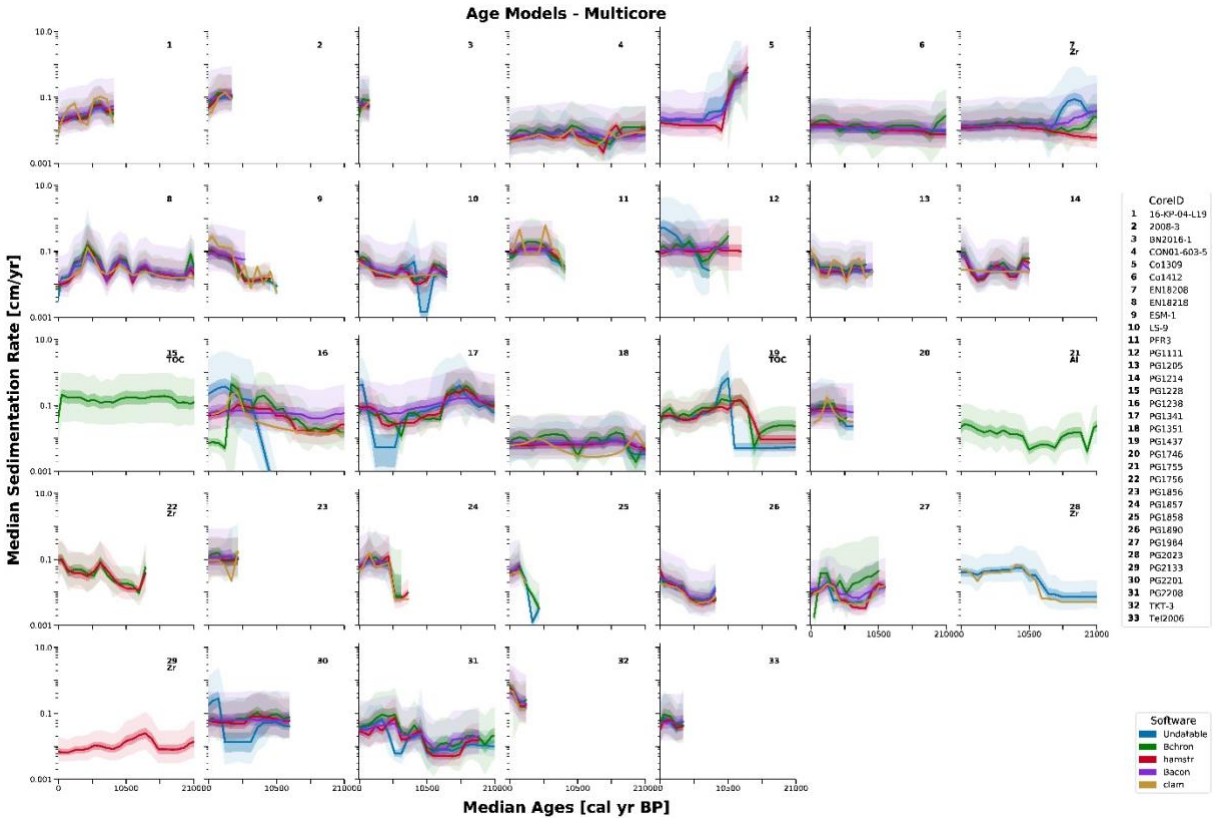

Figure C1: Optimized models for 33 published sediment cores displayed for each modeling system as weighted average median sedimentation rate (in centimeter per year, cm yr^{-1} – y axis) binned into 1000-year bins (in calibrated years Before Present, cal. yr BP, i.e. before 1950 CE – x axis) for the last 21 000 years. Bold lines indicate the weighted average median sedimentation rate for all models, while shaded areas are their respective 1σ and 2σ ranges in the same colors with different opacities. Each grid cell contains the unique core identifier of each involved sediment core. In seven cases, the letters below each number give the name of the independent proxy used for optimization process.

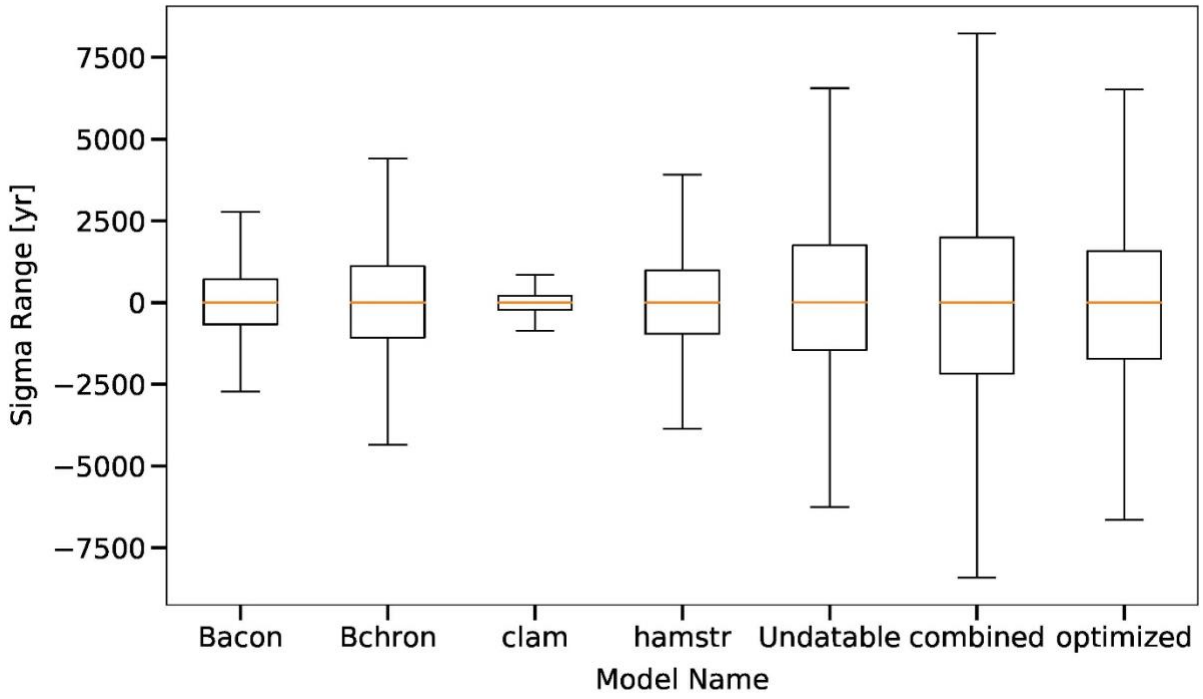

Figure C2: Boxplot representing the overall two-sigma ranges (in years) for each model within our data collection of 55 sediment cores. We examined the modeling results over the entire length of each individual sediment core.

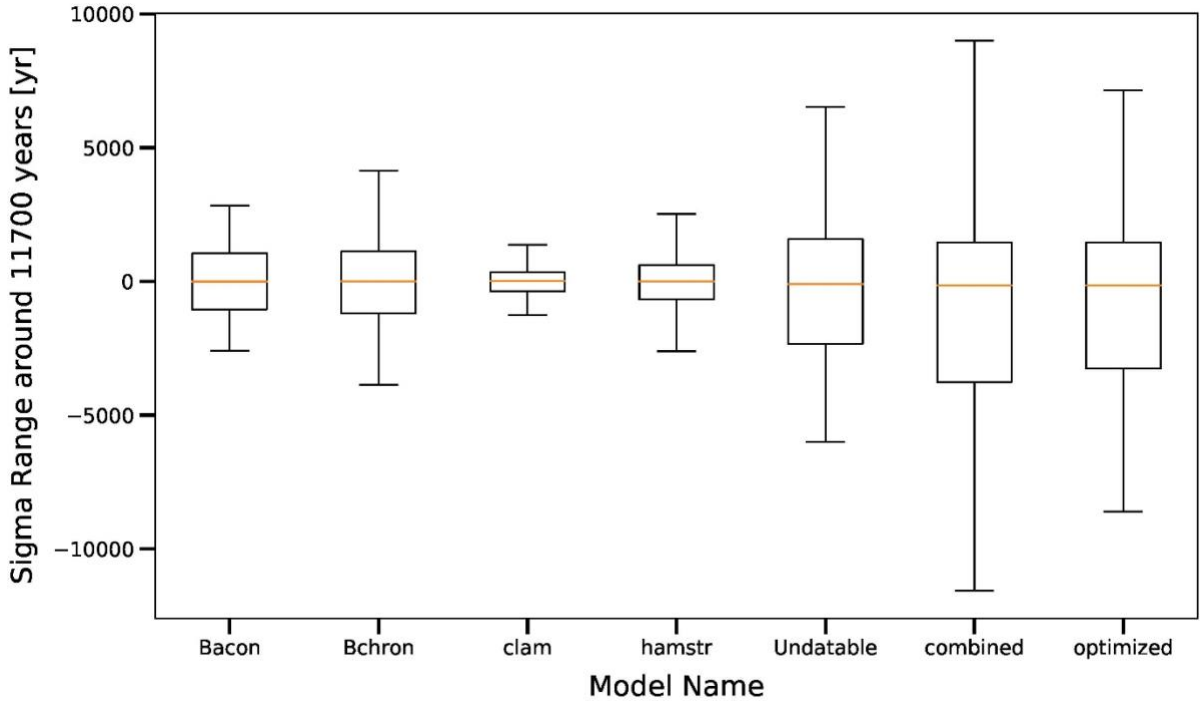

Figure C3: Boxplot representing the 2σ ranges (in years) of each model around 11 700 yr BP for our data collection of 55 sediment cores. We examined the period from 11 600 to 11 800 yr BP to enable a comparison with the 100-year-binned model results.

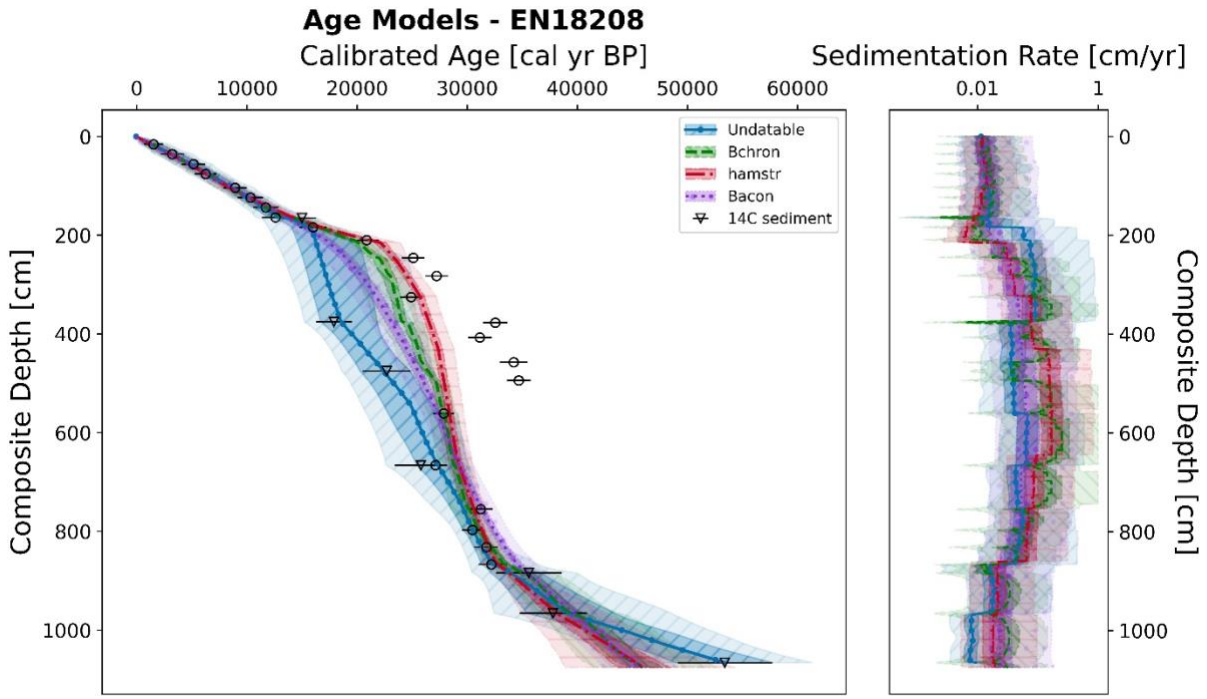

Figure C4 – Color vision deficiency plot for sediment core EN18208 (OSL and ^{14}C data from Vyse et al. (2020b)) generated by LANDO. Equivalent to panel (a) of Figure 3.3 in the main publication, the left plot shows the age-depth models for EN18208, whereas the right plot displays the results from the sedimentation rate calculation for each modeling system. The difference to Figure 3 is that each modeling system has received a different line style and shading to help differentiate between the models. Instead of representing the median age and median sedimentation rate of all models by solid lines, the various line styles shall support the interpretation of age-depth models for people with color vision deficiency. Furthermore, each shading characterizes both 1σ and 2σ ranges for the individual models with decreasing opacities, respectively.

Table C1: Overview of the screening results for whether modeling systems were able to use age determination data to create an age-depth model. Label "CHECK" refers to a successful modeling process, while label "FAIL" indicates an unsuccessful process.

CoreID	Undatable	Bchron	hamstr	Bacon	clam
16-KP-04-L19	CHECK	CHECK	CHECK	CHECK	CHECK
2008-3	CHECK	CHECK	CHECK	CHECK	CHECK
BC2008	CHECK	CHECK	CHECK	CHECK	FAIL
BL02-2007	CHECK	CHECK	CHECK	CHECK	CHECK
BN2016-1	CHECK	CHECK	CHECK	CHECK	CHECK
Chupa-8	CHECK	CHECK	CHECK	CHECK	CHECK
Co1309	CHECK	CHECK	CHECK	CHECK	FAIL
Co1412	CHECK	CHECK	CHECK	CHECK	FAIL
CON01-603-5	CHECK	CHECK	CHECK	CHECK	CHECK
Dolgoe2012	CHECK	CHECK	CHECK	CHECK	CHECK
EN18208	CHECK	CHECK	CHECK	CHECK	FAIL
EN18218	CHECK	CHECK	CHECK	CHECK	CHECK
ESM-1	CHECK	CHECK	CHECK	CHECK	CHECK
KAS-1	CHECK	CHECK	CHECK	CHECK	CHECK
Korzhino2010	FAIL	CHECK	CHECK	CHECK	CHECK
LENDERY180-4	CHECK	CHECK	CHECK	CHECK	CHECK
LENDERY192	CHECK	CHECK	CHECK	CHECK	CHECK
LENDERY200-1	CHECK	CHECK	CHECK	CHECK	CHECK
LENDERY203-3	CHECK	CHECK	CHECK	CHECK	CHECK
LOT83-7	CHECK	CHECK	CHECK	CHECK	CHECK
LS-9	CHECK	CHECK	CHECK	CHECK	CHECK
Maloye-1	CHECK	CHECK	CHECK	CHECK	CHECK
MC2006	CHECK	CHECK	CHECK	CHECK	CHECK
Muan2018	CHECK	CHECK	CHECK	CHECK	CHECK
Okun2018	FAIL	CHECK	CHECK	CHECK	CHECK
OSIN	CHECK	CHECK	CHECK	CHECK	CHECK
PER3	CHECK	CHECK	CHECK	CHECK	CHECK
PG1111	CHECK	CHECK	CHECK	CHECK	FAIL
PG1205	CHECK	CHECK	CHECK	CHECK	CHECK
PG1214	CHECK	CHECK	CHECK	CHECK	CHECK
PG1228	CHECK	CHECK	CHECK	CHECK	FAIL
PG1238	CHECK	CHECK	CHECK	CHECK	CHECK
PG1341	CHECK	CHECK	CHECK	CHECK	FAIL
PG1351	CHECK	CHECK	CHECK	CHECK	CHECK
PG1437	CHECK	CHECK	CHECK	CHECK	FAIL
PG1746	CHECK	CHECK	CHECK	CHECK	CHECK
PG1755	CHECK	CHECK	CHECK	CHECK	FAIL
PG1756	CHECK	CHECK	CHECK	CHECK	CHECK
PG1856	CHECK	CHECK	CHECK	CHECK	CHECK
PG1857	CHECK	CHECK	CHECK	CHECK	CHECK
PG1858	CHECK	CHECK	CHECK	CHECK	CHECK
PG1890	CHECK	CHECK	CHECK	CHECK	CHECK
PG1972	CHECK	CHECK	CHECK	CHECK	CHECK
PG1975	CHECK	CHECK	CHECK	CHECK	FAIL
PG1984	CHECK	CHECK	CHECK	CHECK	CHECK
PG2023	CHECK	CHECK	CHECK	CHECK	CHECK
PG2133	CHECK	CHECK	CHECK	CHECK	FAIL
PG2201	CHECK	CHECK	CHECK	CHECK	FAIL
PG2208	CHECK	CHECK	CHECK	CHECK	FAIL
Tel2006	CHECK	CHECK	CHECK	CHECK	FAIL
Teriberka17	CHECK	CHECK	CHECK	CHECK	CHECK
TKT-3	CHECK	CHECK	CHECK	CHECK	CHECK
TL-1-1	CHECK	CHECK	CHECK	CHECK	CHECK
TULOMA27	FAIL	CHECK	CHECK	CHECK	CHECK
UKhau2015	CHECK	CHECK	CHECK	CHECK	CHECK

Table C2: Runtime for each model for case study CS1 ("Continuously deposited sequence") and CS2 ("Inconsistent sequence") split into their individual steps. Both case studies ran ten times in our test setup. The presented values are the mean value and their standard deviation. Note: a) Within our test setup, we let *Bacon* adjust the default values automatically. For the "Continuously deposited sequence" case study (CS1), *Bacon* changed the accumulation rate prior mean (*"acc.mean"*) to 50 yr cm^{-1} and the thickness to 4. In the "Inconsistent sequence" case study (CS2), *Bacon* adjusted the *"acc.mean"* to the same value (50 yr cm^{-1}), but increased the thickness to 10 to account for the length of the sediment core, which resulted in a reduction of runtime. b) For both case studies, we used our "best fit" option within *clam*. For the "Inconsistent sequence" case study, LANDO could not find a best fit with the *clam* models, hence, our program skipped both "Aggregation" and "Sedimentation Rate Calculation" (SRC) step.

Case Study		**"Continuously deposited sequence" CS1**	**"Inconsistent sequence" CS2**
Length of selected sediment cores [m]		6.53	10.76
Execution time [s]			
Reservoir correction		39.87 ± 0.39	45.39 ± 0.55
Undatable			
	Preparation	0.40 ± 0.01	0.40 ± 0.02
	Execution	8.58 ± 0.16	10.39 ± 0.43
	Aggregation	0.36 ± 0.01	0.55 ± 0.01
	SRC	18.35 ± 0.18	34.62 ± 0.12
Bchron			
	Preparation	0.10 ± 0.00	0.10 ± 0.00
	Execution	166.89 ± 0.55	193.42 ± 1.94
	Aggregation	1.86 ± 0.02	3.07 ± 0.10
	SRC	18.82 ± 0.13	35.77 ± 0.95
hamstr			
	Preparation	0.10 ± 0.01	0.10 ± 0.01
	Execution	93.37 ± 0.58	118.90 ± 1.90
	Aggregation	2.93 ± 0.02	4.21 ± 0.01
	SRC	18.70 ± 0.08	35.48 ± 0.66
Bacon			
	Preparation	0.10 ± 0.01	0.11 ± 0.01
	Execution	[a)] 1220.62 ± 4.08	[a)] 657.46 ± 3.48
	Aggregation	2.99 ± 0.01	4.25 ± 0.03
	SRC	18.71 ± 0.14	36.07 ± 0.91
clam			
	Preparation	0.05 ± 0.01	0.05 ± 0.01
	Execution	193.72 ± 2.88	217.64 ± 4.43
	Aggregation	1.96 ± 0.06	[b)] 0.04 ± 0.00
	SRC	19.39 ± 0.40	[b)] 0.04 ± 0.00
Overall execution time [min]		30.46 ± 0.16	23.30 ± 0.26

Appendix D

Supplementary Equations

Equations following the notation of Kaliappan et al. (2021), where "P" is the predicted value, "A" stands for the actual value, "N" represents the total number of data points.

a) Mean Absolute Error: $MAE = \frac{1}{N}\sum_{i=1}^{N}|P_i - A_i|$

b) Relative Absolute Error: $RAE = \frac{\left[\sum_{i=1}^{N}(P_i - A_i)^2\right]^{1/2}}{\left[\sum_{i=1}^{N}{A_i}^2\right]^{1/2}}$

c) Mean Square Error: $MSE = \sum_{i=1}^{N}\frac{(P_i - A_i)^2}{N}$

d) Root Mean Squared Error: $RMSE = \sqrt{\sum_{i=1}^{N}\frac{(P_i - A_i)^2}{N}}$

e) Root Relative Squared Error: $RRSE = \sqrt{\frac{\sum_{i=1}^{N}(P_i - A_i)^2}{\sum_{i=1}^{N}(A_i - \bar{A})^2}}$ with $\bar{A} = \frac{1}{N}\sum_{i=1}^{N}A_i$

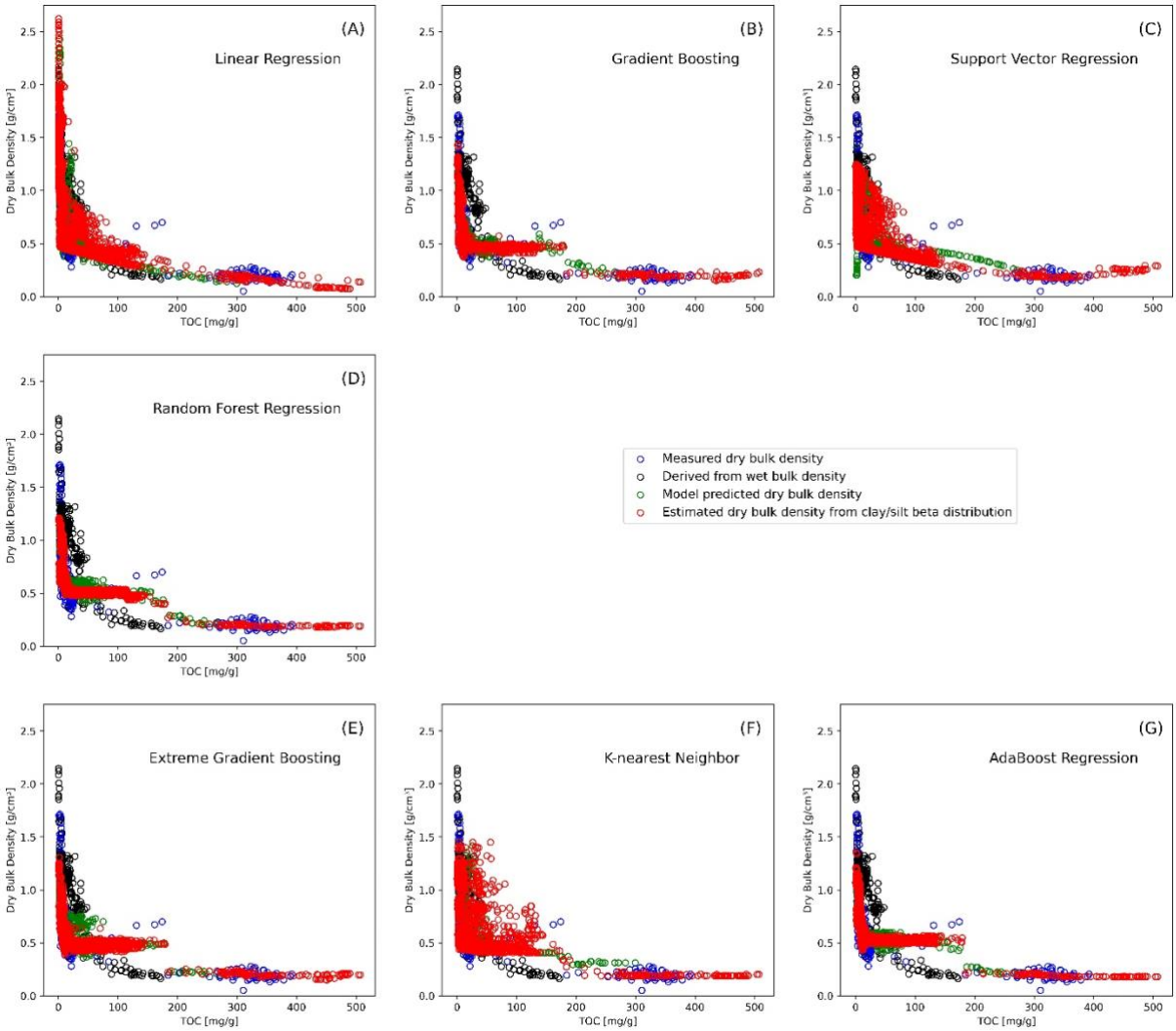

Figure D1: Model prediction results for dry bulk density against total organic carbon (TOC) for all methods used in this study. Directly measured dry bulk density (blue circles) and dry bulk density derived from measured wet bulk density (black circles) data are the same across all subplots. Green circles represents predicted dry bulk density values for each prediction method, where all input values were available. Red circles are estimated mean dry bulk density values for each prediction method with grain size data based on beta distribution for clay and silt.

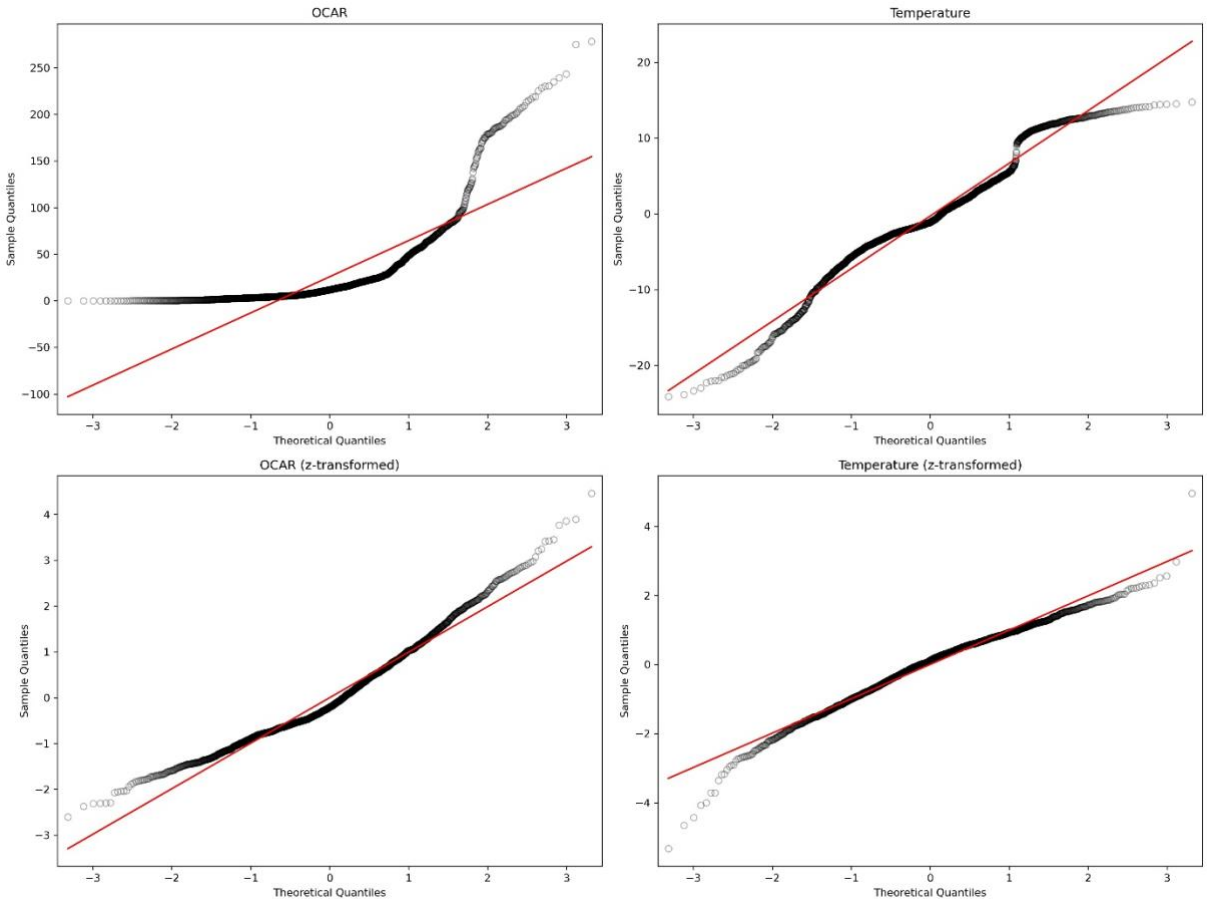

Figure D2: Q-Q plot for untransformed and z-transformed temperature and OCAR data to check for normality. More linearity means more likely to be normally distributed.

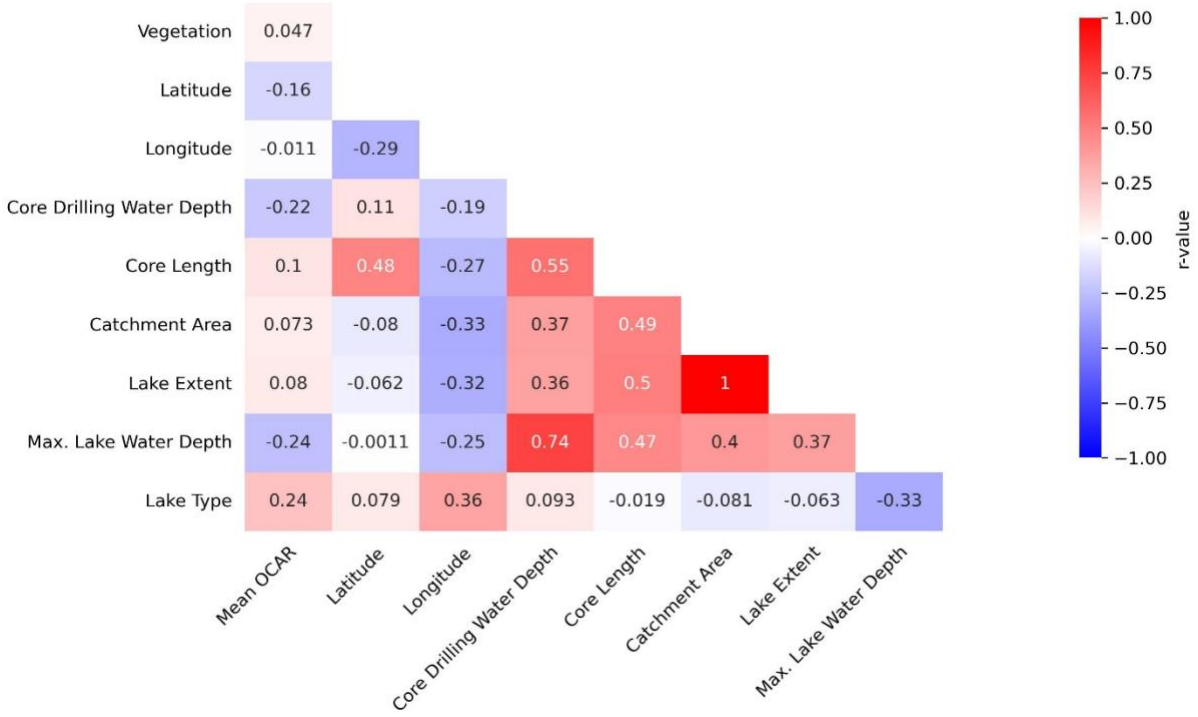

Figure D3: Correlation between OCAR and lake parameters plus the correlation between mean OCAR and vegetation

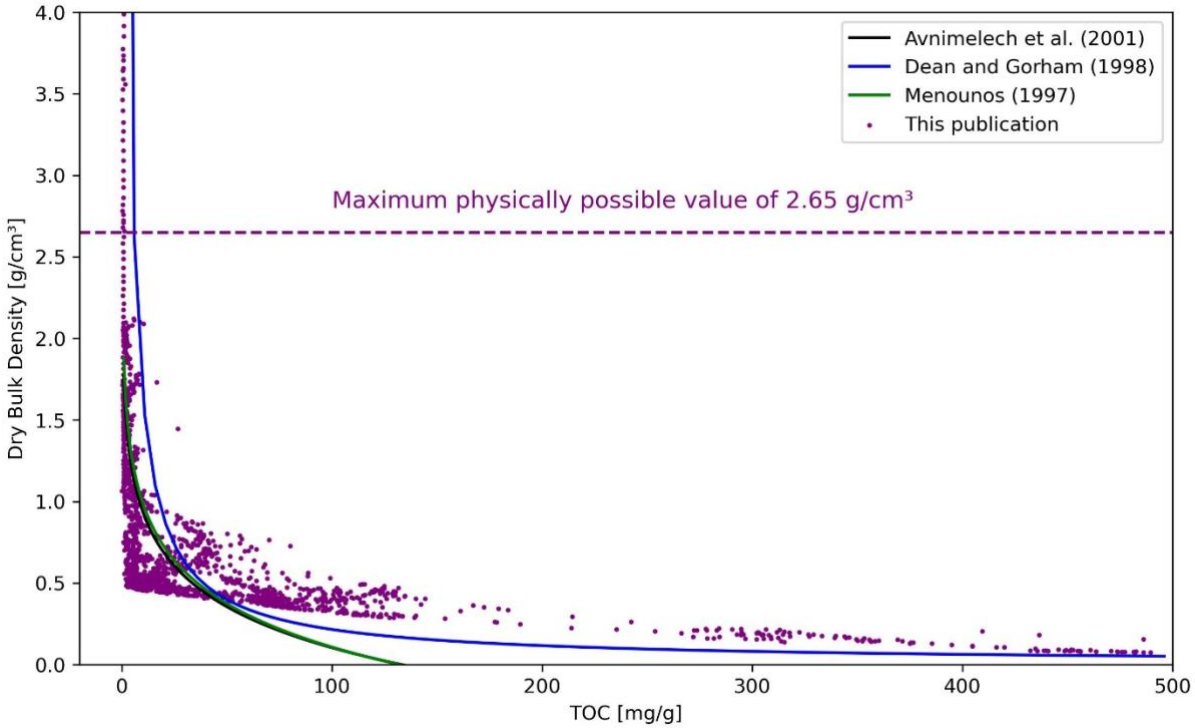

Figure D4: Comparison between the results of the log-linear model with existing formulas from the literature referenced in the publication

Table D1: Summary of attributes for sediment cores used in this study

CoreID	Lake Name	Lake Type	Latitude	Longitude	Lake Elevation [m. a. s. l.]	Lake Area [km^2]	Lake Volume [km^3]	Max. Lake Water Depth [m]	Core Drilling Water Depth [m]	Catchment Area [km^2]
BN2016-1	**Bayan Nuur**	Tectonic	50.01072	93.9745	929	31.83	0.6365	29	29	620.70
Co1309	**Ladoga**	Glacial	60.98894	30.69086	19	17515.89	798.4542	230	111	276335.45
Co1412	**Emanda**	Glacial	65.29415	135.759234	672	35.70	0.2474	15.33	14.7	171.52
EN18208	**Ilirney**	Glacial	67.3403	168.29567	428	29.59	0.4000	43.47	19	1144.34
EN18218	**Rauchuvagytgyn**	Glacial	67.78938	168.73352	620	6.26	0.0886	35.94	29.5	207.86
ESM-1	**East Sayan Mountains Lake**	Glacial	52.02695	101.05908	1985	0.13	0.0001	7	4.5	4.23
LS-9	**Dolgoe Ozero**	Thermokarst	71.8769	127.0722	19	0.76	0.0017	4	4	2.20
PER3	**Pernatoye**	Shoreline	50.0405	155.395167	2	0.83	0.0008	1.8	1.8	7.77
PG1111	**Lama**	Tectonic	69.548	90.211	28	319.38	17.0584	255	52.2	6517.80
PG1205	**Basalt**	Glacial	72.72333	-22.465	187	1.42	0.0107	21	21	9.42
PG1214	**Raffles**	Glacial	70.595	-21.535	48	0.25	0.0068	63	63	1.11
PG1228	**Levinson-Lessing**	Glacial	74.473	98.636	46	24.25	1.4624	120	108	498.09
PG1238	**Changeable**	Glacial	79.11833	95.125	7	11.61	0.0849	18	17.2	205.46
PG1341	**Lama**	Tectonic	69.548	90.211	28	319.38	17.0584	255	66	6517.80
PG1351	**El'gygytgyn**	Crater	67.50666	172.13833	496	118.77	13.8893	175	175	291.44
PG1437	**Lyadhej-To**	Glacial	68.250611	65.789333	145	1.87	0.0180	25	21	19.86
PG1746	**Temje**	Thermokarst	62.051902	129.480061	193	0.47	0.0001	0.7	0.7	33.55
PG1755	**Billyakh**	Glacial	65.272	126.74633	325	16.82	0.0755	24	7.8	137.77
PG1756	**Billyakh**	Glacial	65.2955	126.77516	325	16.82	0.0755	24	7.9	137.77
PG1857	**Two-Yurts**	Glacial	56.81675	160.06778	287	11.40	0.1989	28.21	25	211.15
PG1858	**Sigrid**	Glacial	56.826667	160.115	293	0.03	0.0001	5	2.5	0.09
PG1890	**Sokoch**	Glacial	53.253861	157.755382	497	0.46	0.0012	6.5	4.5	28.34
PG1984	**09-Tik-13 / Sysy-Kyuele**	Thermokarst	69.40485	123.82792	73	0.74	0.0007	2.4	2.4	10.75
PG2023	**Kyuntyunda**	Thermokarst	69.628367	123.649783	56	4.97	0.0101	4.2	2.9	47.34
PG2133	**Bolshoe Toko**	Glacial	56.05615	130.8678	914	83.75	2.5226	72.5	26	932.82
PG2201	**Malaya Chabyda**	Thermokarst	61.958483	129.408333	192	0.18	0.0001	3	2	16.11
PG2208	**Bolshoe Toko**	Glacial	56.05615	130.8678	914	83.75	2.5226	72.5	68.3	932.82
Tel2006	**Teletskoye**	Tectonic	51.716667	87.65	430	227.98	45.1185	332	90	20110.22

Table D2: Summary of different correlation metrics for temperature and OCAR for each sediment cores used in this study. We used an alpha value of 0.05. If the p-value is below the alpha value, then the correlation is statistically significant (green). If the p-value is above the alpha value, it is statistically insignificant (red).

CoreID	Pearson r-value	Pearson p-value	Spearman rho-value	Spearman p-value	Kendall tau-value	Kendall tau p-value
BN2016-1	0.07925	6.362000e-01	0.10780	5.195000e-01	0.075390	5.052000e-01
Co1309	0.50260	2.104000e-07	0.35910	3.516000e-04	0.253300	2.766000e-04
Co1412	0.85030	5.533000e-15	0.78160	2.098000e-11	0.583700	2.219000e-09
EN18208	0.22940	3.307000e-01	0.09323	6.958000e-01	0.063160	7.246000e-01
EN18218	0.29530	2.864000e-02	0.61650	5.398000e-07	0.411400	9.183000e-06
ESM-1	0.11000	5.292000e-01	0.03333	8.492000e-01	-0.001681	9.887000e-01
LS-9	-0.18530	1.885000e-01	-0.31830	2.146000e-02	-0.221600	2.075000e-02
PER3	0.10910	3.795000e-01	0.05507	6.580000e-01	0.027590	7.413000e-01
PG1111	0.80650	2.586000e-30	0.77550	9.981000e-27	0.573700	1.137000e-21
PG1205	-0.20840	1.165000e-01	-0.14230	2.865000e-01	-0.091350	3.111000e-01
PG1214	-0.13250	3.350000e-01	-0.07150	6.039000e-01	-0.045120	6.267000e-01
PG1228	0.53590	1.786000e-07	0.56370	2.874000e-08	0.400500	8.309000e-08
PG1238	0.67660	9.781000e-14	0.77190	1.351000e-19	0.556800	2.670000e-15
PG1341	0.18130	3.333000e-02	0.12900	1.315000e-01	0.084740	1.404000e-01
PG1351	-0.23440	8.805000e-02	-0.15750	2.555000e-01	-0.102700	2.728000e-01
PG1437	-0.04106	5.678000e-01	-0.01931	7.882000e-01	0.014340	7.654000e-01
PG1746	-0.04341	7.022000e-01	-0.01245	9.127000e-01	0.000633	9.934000e-01
PG1755	0.34780	2.945000e-01	0.34550	2.981000e-01	0.272700	2.830000e-01
PG1756	0.23050	2.379000e-01	0.20960	2.843000e-01	0.127000	3.564000e-01
PG1857	0.26560	1.290000e-01	0.28100	1.075000e-01	0.180000	1.343000e-01
PG1858	0.00384	9.825000e-01	-0.02857	8.706000e-01	-0.011760	9.208000e-01
PG1890	-0.01295	9.122000e-01	-0.02151	8.547000e-01	-0.017660	8.226000e-01
PG1984	0.18840	1.768000e-01	-0.03354	8.115000e-01	-0.026120	7.824000e-01
PG2023	0.23410	1.676000e-02	0.10490	2.891000e-01	0.069830	2.935000e-01
PG2133	0.45990	1.056000e-02	0.40160	2.784000e-02	0.301100	1.943000e-02
PG2201	0.46720	5.915000e-05	0.45100	1.133000e-04	0.318700	1.215000e-04
PG2208	0.50270	1.703000e-13	0.48530	1.470000e-12	0.344700	1.880000e-12
Tel2006	0.05455	3.783000e-01	0.02193	7.233000e-01	0.014080	7.337000e-01

Table D3: Summary of biomes for each sediment cores used in this study for a) snapshots in Figure 4.7 (at 21 000, 11 700, 8200, and 0 years BP) and b) mid-points of our seven periods in years BP according to the vegetation reconstruction by Dallmeyer et al. (2022). "n/a" means that there was no OCAR data available for this core at this specific time. "Not covered" means that for these sediment cores we had OCAR data, but the vegetation reconstruction did not explicitly cover the location with a vegetation grid cell.

CoreID	21 000	20000	16200	13500	12200	11 700	9900	8200	6200	2100	0
BN2016-1	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	Boreal
Co1309	n/a	n/a	n/a	Grass/Shrub	Not covered[1]	Not covered[1]	Boreal	Boreal	Boreal	Boreal	Boreal
Co1412	Tundra	Tundra	Tundra	Boreal	Boreal	Boreal	Boreal	Boreal	Boreal	Tundra	Tundra
EN18208	n/a	n/a	Tundra	Boreal	Boreal	Boreal	Tundra	Tundra	Tundra	Tundra	Tundra
EN18218	Tundra	Tundra	Tundra	Boreal	Boreal	Boreal	Tundra	Tundra	Tundra	Tundra	Tundra
ESM-1	n/a	n/a	n/a	n/a	n/a	n/a	n/a	Boreal	Boreal	Tundra	Tundra
LS-9	n/a	n/a	n/a	Boreal	Boreal	Boreal	Boreal	Tundra	Tundra	Tundra	Tundra
PER3	n/a	n/a	n/a	n/a	n/a	n/a	n/a	Not covered[2]	Not covered[2]	Not covered[2]	Not covered[2]
PG1111	n/a	n/a	n/a	n/a	Tundra	Boreal	Boreal	Boreal	Boreal	Tundra	Tundra
PG1205	n/a	n/a	n/a	n/a	n/a	n/a	Ice Desert	Ice Desert	Ice Desert	Ice Desert	Ice Desert
PG1214	n/a	n/a	n/a	n/a	n/a	n/a	Not covered[3]	Not covered[3]	Not covered[3]	Not covered[3]	Not covered[3]
PG1228	Tundra	Tundra	Tundra	Boreal	Boreal	Boreal	Boreal	Tundra	Tundra	Tundra	Tundra
PG1238	Ice Desert	Ice Desert	Tundra	Tundra	Not covered[4]	Not covered[4]	Not covered[4]	Not covered[4]	Not covered[4]	Not covered[4]	Not covered[4]
PG1341	Tundra	Tundra	Tundra	Boreal	Tundra	Boreal	Boreal	Boreal	Boreal	Tundra	Tundra
PG1351	Tundra	Tundra	Tundra	Boreal	Tundra	Tundra	Tundra	Tundra	Tundra	Tundra	Tundra
PG1437	Tundra	Tundra	Tundra	Tundra	Tundra	Tundra	Boreal	Boreal	Boreal	Tundra	Tundra
PG1746	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	Boreal	Tundra	Boreal
PG1755	Tundra	Tundra	Boreal	Boreal	Boreal	Boreal	Boreal	Boreal	Boreal	Boreal	Boreal
PG1756	n/a	n/a	n/a	Boreal	Boreal	Boreal	Boreal	Boreal	Boreal	Boreal	Boreal
PG1857	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	Tundra	Tundra
PG1858	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	Tundra	Tundra
PG1890	n/a	n/a	n/a	n/a	n/a	n/a	n/a	Not covered[2]	Not covered[2]	Not covered[2]	Not covered[2]
PG1984	n/a	n/a	n/a	n/a	n/a	Boreal	Boreal	Boreal	Boreal	Boreal	Boreal
PG2023	n/a	Tundra	Boreal	Boreal	Boreal	Boreal	Boreal	Boreal	Boreal	Boreal	Boreal
PG2133	Tundra	Tundra	Tundra	Boreal	Boreal	Boreal	Boreal	Boreal	Boreal	Tundra	Grass/ Shrub
PG2201	n/a	n/a	n/a	n/a	Boreal	Boreal	Boreal	Boreal	Boreal	Tundra	Boreal
PG2208	n/a	Tundra	Tundra	Boreal	Boreal	Boreal	Boreal	Boreal	Boreal	Tundra	Grass/ Shrub
Tel2006	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	Boreal	Boreal

[1] Most likely in the "Boreal Forest" biome
[2] Most likely in the "Grass & Dry Shrubland" biome
[3] Most likely in the "Ice & Polar Desert" biome
[4] Most likely in the "Tundra" biome

Acknowledgement

Creating knowledge is about sharing, challenging yourself and the status quo, and being bold and curious. By acknowledging the people who have shaped your journey along the way, you realize that your knowledge would be worthless without them. I am deeply grateful to the people in this acknowledgement and people I might have missed for their support, encouragement, challenges, and friendships.

First, I would like to thank my three HEIBRiDS supervisors Bernhard Diekmann, Johann-Christoph Freytag, and Boris Biskaborn. All three of them had faith in my potential, pushing me to become the best version of myself and complete this thesis. Bernhard Diekmann and Boris Biskaborn provided unwavering support from the Alfred Wegener Institute, while Johann-Christoph Freytag welcomed me into his group at HU Berlin, ensuring I could broaden my perspective from a unique angle. While working with me might not always have been straightforward, I cherish every moment and am deeply grateful for the life lessons they taught me, their advice and mentorship.

Being part of the HEIBRiDS graduate school has truly enriched my life. From meeting exceptionally talented individuals to having the honor of being the inaugural student representative, helping shape the graduate school from a student's perspective. I would like to thank Eirini Kouskoumvekaki and Johann-Christoph Freytag for guiding me through this task, and Elizabeth Robertson and Sergey Redyuk for being wonderful co-chairs. I am truly grateful to Sidd Agarwal, Nicolas Miranda, Tabea Rettelbach, and Jannes Münchmeyer for their friendship and support, and thankful for all the amazing people I met in the first three HEIBRiDS cohorts.

My academic journey began with the Alfred Wegener Institute during my master's thesis, and I believe there could not have been a better place to write my doctoral thesis. While the ARCLAKE group led by Boris was the smallest group in the Polar Terrestrial Environmental Systems section, all of the people who joined our group were exceptional. I could not have done it without Stuart Vyse, who is a true friend and a positive, joyful genius. In addition, Daria, Rebecca, Levent, Sebastian, Tim, Amy, Alex, and Maximilian made the group something extraordinary. Even though we worked in different sections, I am very grateful for the mentorship I received from Hugues Lantuit. His attitude and knowledge have always been a great inspiration and helped me to be a better communicator. However, there were also many more people working in the ENVI section whose support helped me to reach the finish line: Elizabeth Dietze, Thomas Opel, Andrew Dolman, Mareike Wieczorek, Hanno Meyer, Stefan Kruse, Andrej Andreev, Georg Schwamborn, Heike Zimmermann, Birgit Heim, Thomas Laepple, and Ulrike Herzschuh. In addition, I would like to thank Ramesh Glückler, Loeka Jongejans, Torben Windirsch, Raphaël Hébert, Jeremy Courtin, Michael Fritz, Lisa Grosfeld, and Barbara von Hippel for their moral support along the way. I also extend my gratitude to my collaborators Liudmila Syrykh and Dmitry A. Subetto and all the scientists who generously shared their data and promptly addressed my inquiries.

Lastly, this doctoral thesis would have not been possible without the support of my family. My parents Kerstin and Matthias raised me to walk, learn and explore, supported me in every corner of the world, and without their love and support, I would not be where I am today. From Christian and Susanne, my elder siblings, I have gained invaluable life lessons; their knowledge surpasses mine, ensuring their advice is something I can always depend on. But my biggest thanks are due to my dearest wife, Leonie, who has been my biggest supporter from day one, despite daily grudges, a pandemic, a move to another country, and the birth of our two beautiful daughters, Harriet and Georgia. Having her by my side made me strive for the best, go the extra mile, and make a dream come true. I am grateful to my extended family for believing in my work and me, encouraging me day after day and without the smiles of the children in our big family, some days would not have been so bright.

Finally, I would like to thank you for reading this.

Eidesstaatliche Erklärung

Hiermit versichere ich, dass ich die vorliegende Arbeit selbstständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel verwendet habe. Ich habe diese kumulative Dissertation am Alfred-Wegener-Institut Helmholtz Zentrum für Polar und Meeresforschung in Potsdam erarbeitet und in englischer Sprache angefertigt. Diese Dissertation wird erstmalig und ausschließlich an der Universität Potsdam eingereicht. Die dem Promotionsverfahren zugrundeliegende Promotionsordnung vom 18.09.2013 ist mir bekannt.

_____

Ort, Datum Gregor Pfalz