Utilizing probability estimates from machine learning and pollen to understand the depositional influences on branched GDGT in wetlands, peatlands, and lakes

Amy Cromartie[1], Cindy De Jonge[2], Guillemette Ménot[3], Mary Robles[4,5], Lucas Dugerdil[3,5], Odile Peyron[5], Marta Rodrigo-Gámiz[6], Jon Camuera[7], Maria Jose Ramos-Roman[8], Gonzalo Jiménez-Moreno[6], Claude Colombié[9], Lilit Sahakyan[10], and Sébastien Joannin[5]

[1]Université Côte d'Azur, CNRS, CEPAM, UMR 7264, 06300 Nice, France
[2]Geological Institute, ETH Zürich, 8092 Zurich, Switzerland
[3]ENS de Lyon, Université Lyon 1, CNRS, UMR 5276 LGL-TPE, 69364 Lyon, France
[4]Aix-Marseille Univ., CNRS, IRD, INRAE, Coll France, UMR 34 CEREGE, 13545 Aix-en-Provence, France
[5]ISEM, Univ. Montpellier, CNRS, IRD, 34090 Montpellier, France
[6]Department of Stratigraphy and Paleontology, University of Granada, 18071 Granada, Spain
[7]Unit of Botany, Faculty of Pharmacy, Complutense University of Madrid, 28040 Madrid, Spain
[8]Instituto de Investigación en Cambio Global Universidad Rey Juan Carlos 28933, Madrid, Spain
[9]Univ Lyon, UCBL, ENSL, UJM, CNRS, LGL-TPE, Villeurbanne, F-69622 France
[10]Institute of Geological Sciences, National Academy of Sciences of Republic of Armenia, Yerevan 0019, Armenia

Correspondence: Amy Cromartie (aec277@cornell.edu)

Received: 11 February 2025 – Discussion started: 28 February 2025
Revised: 19 June 2025 – Accepted: 23 June 2025 – Published:

Abstract. Branched glycerol dialkyl glycerol tetraethers (brGDGTs) are critical molecular biomarkers for the quantitative reconstruction of past environments, ambient temperature, and pH across various archives. However, numerous issues persist that limit their application. The distribution of brGDGTs varies significantly based on provenance, resulting in biases in environmental reconstructions that rely on fractional abundances and derived indices, such as MBT'_{5ME}. This issue is especially significant in shallow lakes, wetlands, and peatlands, where ecosystems are sensitive to diverse environmental and climatic factors. Recent advancements, such as machine learning techniques, have been developed to identify changes in provenance; however, these techniques are insufficient for detecting mixed environments. The probability estimates derived from five machine learning algorithms are employed here to detect provenance changes in brGDGT downcore records and to identify periods of mixed provenance. A new global modern database ($n = 2031$ TS1) was compiled to train, validate, test, and apply these algorithms to two sedimentary records. Our findings are corroborated by pollen, non-pollen palynomorphs, and X-ray fluorescence (XRF) obtained from the same sedimentary core sequence. These microfossil and geochemical proxies are utilized to discuss changes in provenance, hydrology, and ecology that influence brGDGT provenance. Probability estimates derived from random forest with a sigmoid calibration are most effective in detecting changes in brGDGT provenance. Minor changes in the relative contributions of brGDGT provenance can significantly influence the distribution of brGDGT, especially regarding the MBT'_{5ME} index.

1 Introduction

Branched glycerol dialkyl glycerol tetraethers (brGDGTs), first identified in peat sequences (Weijers et al., 2006), have demonstrated significant potential as a quantitative proxy for paleoenvironmental reconstructions. The ubiquity of brGDGTs and their global correlations with tempera-

ture and pH, notably across different archive types, positions them as a valuable tool for paleoclimate reconstructions (among others, Weijers et al., 2007; Peterse et al., 2012; Loomis et al., 2011; Raberg et al., 2022a, b). Researchers have identified brGDGTs across various depositional environments, such as peat, soils, loess, and fossilized bones and lacustrine, marine, and river sediments (e.g., Weijers et al., 2006, 2007; De Jonge et al., 2014a; Warden et al., 2016; Naafs et al., 2017a, b; Dillon et al., 2018; Baker et al., 2019) at differing geological timescales, indicating their widespread potential as a proxy for reconstructing the continental paleoclimate.

A key challenge in utilizing brGDGT-based reconstructions in continental settings is the temperature-independent variability in fractional abundance (FA) distribution across these environments (De Jonge et al., 2014b; Naafs et al., 2017b; Dearing Crampton-Flood et al., 2020; Martínez-Sosa et al., 2021; Raberg et al., 2022b). In the context of lacustrine, wetland, and peat archives, the fractional abundance of brGDGTs produced in aquatic environments and surrounding soils varies (Tierney and Russell, 2009; Tierney et al., 2010; Zink et al., 2010; Buckles et al., 2014; Loomis et al., 2011, 2012, 2014a, b; Li et al., 2016; Russell et al., 2018; Dang et al., 2018). Potential changes in provenance thus result in distribution differences that may lead to inaccuracies in paleoenvironmental reconstructions. This includes paleotemperature reconstructions derived from the widely recognized index based on the methylation of branched tetraether (MBT) of 5-methyl (MBT'_{5ME}). This index measures the degree of methylation of the 5-methyl brGDGTs, distinguishing it from the 6-methyl brGDGTs to establish calibrations that exhibit a stronger correlation with mean annual air temperature (MAAT) (De Jonge et al., 2014a). The MBT'_{5ME} index has been successfully utilized as grounds for various global temperature calibrations because of its strong correlation to temperature in modern samples that include lakes, peats, and soils (e.g., De Jonge et al., 2014a; Hopmans et al., 2016; Naafs et al., 2017a; Dearing Crampton-Flood et al., 2020; Martínez-Sosa et al., 2021; Véquaud et al., 2022). Provenance changes may introduce bias to temperature reconstructions based on the MBT'_{5ME} index due to its value generally being higher in soils than in lakes (Martínez-Sosa et al., 2021).

Furthermore, brGDGT distributions in these depositional environments may be influenced by distinct environmental characteristics. Soil chemistry, particularly pH, can influence the 5-methyl brGDGTs (De Jonge et al., 2021, 2024). In certain lakes, the 6-methyl brGDGTs exhibit a stronger correlation with mean annual air temperature compared to the 5-methyl brGDGTs, which contrasts with the catchment soils (Dang et al., 2018). In peatlands, the MBT' and MBT'_{5ME} values are higher in dry sites compared to those that are waterlogged (Rao et al., 2022). Factors influencing the distribution of brGDGTs in lakes include lake stratification and redox conditions (Weber et al., 2018), salinity (Wang et al., 2021), conductivity (Tierney et al., 2010; Raberg et al., 2022b), dissolved oxygen (Wu et al., 2021), and water depth (Stefanescu et al., 2021), amongst others. In soils, vegetation and vegetation-mediated factors such as soil temperature (Liang et al., 2019, 2023), soil moisture (Menges et al., 2014; Dang et al., 2016), precipitation (Dugerdil et al., 2021a), and soil chemistry (Dang et al., 2016; De Jonge et al., 2021) all influence distributional changes. BrGDGT distributions in peat may vary in response to flooding, the drying of peatlands, and alterations in the water table (Rao et al., 2022; Ofiti et al., 2024). The potential differential distributions resulting from depositional environments underscore the influence of changes in provenance or hydrological conditions on brGDGT-based environmental reconstructions.

BrGDGT reconstruction in Quaternary downcore lacustrine records indicates that changes in depositional and mixed provenance significantly affect environmental reconstructions (i.e., Martin et al., 2019; Robles et al., 2022; Ramos-Román et al., 2022; d'Oliveira et al., 2023; Acharya et al., 2023). As climatic or successional changes occur concurrently with temperature variations, isolating the effects of provenance changes on MBT'_{5ME} is challenging. Several indexes and ratios have been developed to detect brGDGT provenance change. The BIT index (Hopmans et al., 2004), and later the IIIa / IIa ratio (Xiao et al., 2016), for example, were designed to identify terrestrial organic input in marine sediments. Although useful in marine contexts, these indexes have had limited success in lacustrine terrestrial environments (e.g., Martin et al., 2020). Ternary diagrams are commonly used to visualize brGDGT (e.g., Russell et al., 2018), enabling a comparison between fossil and modern datasets. These diagrams, however, reduce the data size to three variables, limiting their usefulness in isolating the influence of provenance change on the individual brGDGT isomers. Recently, Martínez-Sosa et al. (2023) employed supervised machine learning (ML) to identify changes in provenance using classification models based on modern samples. Their success highlights the power that ML applications can have in solving difficult issues. ML applications differ from traditional statistics applications by focusing on prediction rather than inference (Bzdok et al., 2018). ML's power over these conventional methods lies in its ability to handle data with multiple variables for a few subjects while examining nonlinear relationships within the datasets (Bzdok et al., 2018). The models of Martínez-Sosa et al. (2023) proved effective at identifying shifts in provenance, but a limitation of their study, however, is the inability to detect periods of mixed provenance.

This paper aims to correct that by introducing a strategy for identifying provenance changes across depositional lacustrine, peat, and soil environments, including mixed contexts, utilizing a new global brGDGT database and ML techniques, as well as environmental reconstructions based on pollen, non-pollen palynomorphs (NPPs), and X-ray fluo-

rescence (XRF) datasets. Two approaches are employed to achieve this objective.

First, we use probability estimates derived from ML to identify changes in provenance over time, extending the work of Martínez-Sosa et al. (2023). Rather than employing discrete classification, as they did, we utilize the probability estimates from these classification algorithms to analyze the contributions from differential provenance at any specific time. This method enhances prior approaches by recognizing environments that integrate brGDGTs from multiple inputs and depositional settings – and thus multiple provenances – that have not fully transitioned to a new depositional state. The probability estimates are derived from the classification of modern samples ($n = 2301$), categorized into three groups: soil, peat, and lake, utilizing both previously published and new datasets. We test five popular parametric and non-parametric ML models based on their ability to handle small tabular datasets and produce reliable probability estimates when calibrated (Malley et al., 2012; Wang et al., 2019). Models utilizing different structures were chosen, including simple tree-based algorithms (classification and regression trees), ensemble trees (random forests), linear models (logistic regression), margin-based classifiers (support vector machines), and instance-based lazy learners (K-nearest neighbors) to evaluate performance. The best-performing model was then chosen to apply to two downcore sedimentary sequences. These are employed using Python and scikit-learn to identify intervals where downcore records are predominantly influenced by in situ lake brGDGTs, mineral soils, and peatlands, as well as combinations of these elements.

Second, to ensure the accurate identification of provenance changes, comparisons are conducted with published pollen, NPPs, and XRF from extensive sediment records and variations in brGDGT distribution (i.e., Robles et al., 2022; Camuera et al., 2018, 2019; Ramos-Román et al., 2018; Rodrigo-Gámiz et al., 2022). The records are situated in the semi-arid mid-latitude zones, where water bodies are subject to temporal variations. Aquatic pollen and NPPs have previously been used to verify changes in provenance in brGDGT communities from fossil records (i.e., Robles et al., 2022; d'Oliveira et al., 2023; Ramos-Román et al., 2022; Barhoumi et al., 2023). In addition, we also compare our results with XRF core scanning data from the same sedimentary sequence. Utilizing these proxies allows for an independent comparison of outputs to (i) confirm ML results through the integration of brGDGT-based reconstructions with pollen, NPPs, and XRF and (ii) demonstrate how these complementary proxies can aid in identifying potential hydrological, ecological, and depositional changes that may cause provenance shifts, thus introducing bias in brGDGT reconstructions. This study demonstrates that alterations in provenance and hydrology can significantly influence the distribution of brGDGTs and, consequently, establish indices like MBT'_{5ME}, while also offering novel methodologies for identifying changes in global paleorecords.

2 Materials and methods

2.1 GDGT databases

2.1.1 Building a new modern sample database

This study compiles published brGDGT databases for depositional lake ($n = 591$), soil ($n = 1197$), and peat ($n = 532$) categories (Baxter et al., 2019; Cao et al., 2020; Chen et al., 2021; Dang et al., 2018; De Jonge et al., 2014b; Dearing Crampton-Flood, 2020; Dearing Crampton-Flood et al., 2020; Ding et al., 2015; Dugerdil et al., 2021a, b; Guo et al., 2020a, b; Halffman et al., 2022; Jaeschke, 2018; Jaeschke et al., 2018; Kirkels et al., 2020; Kou et al., 2022; Li et al., 2017; Liu et al., 2020; Martin et al., 2019, 2020; Martínez-Sosa et al., 2021; Naafs, 2017; Naafs et al., 2017a, b; Ning et al., 2019; Pérez-Angel et al., 2020; Qian et al., 2019; Raberg et al., 2021a, b; Rao et al., 2020, 2022; Robles et al., 2022; Russell et al., 2018; Sinninghe Damste et al., 2020; Stefanescu et al., 2020, 2021; Véquaud et al., 2021b, a; Wang et al., 2016, 2020a, 2018, 2020b; Wang and Liu, 2021; Weber et al., 2018; Wu et al., 2021; Xiao et al., 2015; Yang, 2020; Yang et al., 2015; Yao et al., 2020; Fig. 1, full data at https://doi.org/10.17632/tr8tppy9fz.1, Cromartie, 2025a). Round robin test results show that results from multiple laboratories can be integrated into a single database (De Jonge et al., 2024). Results were included only when chromatography enabled the separate quantification of 5- and 6-methyl brGDGTs (i.e., De Jonge et al., 2014b). The fractional abundances of 15 distinct brGDGT structural isomers were sourced from the original authors or recalculated from the initial datasets. We enhanced the training dataset for certain published datasets by obtaining data with greater precision from the original authors, where fractional abundances had been rounded to two decimal places. We incorporated the fractional abundances of individual downcore samples from Naafs et al. (2017a) to enhance the sample size of our peat analysis. This facilitated the development of a more robust model for assessing brGDGT distribution across various types. All samples originate from terrestrial environments (Fig. 1). The 7-methyl (Ding et al., 2016) or the 5/6 isomer, also referred to as IIIa″ (Weber et al., 2015), were excluded due to limited data. The original authors' description was utilized to categorize the samples into a classification index (i.e., soil, lake, peat). Suspended particulate matter (SPM), moss pollsters, marine, and river samples were excluded. Latitude and longitude data were converted to decimal degrees as required. The Köppen–Geiger classification of each modern sample was done with the kgcpy library (Yu et al., 2024) in Python to assess the climate distribution.

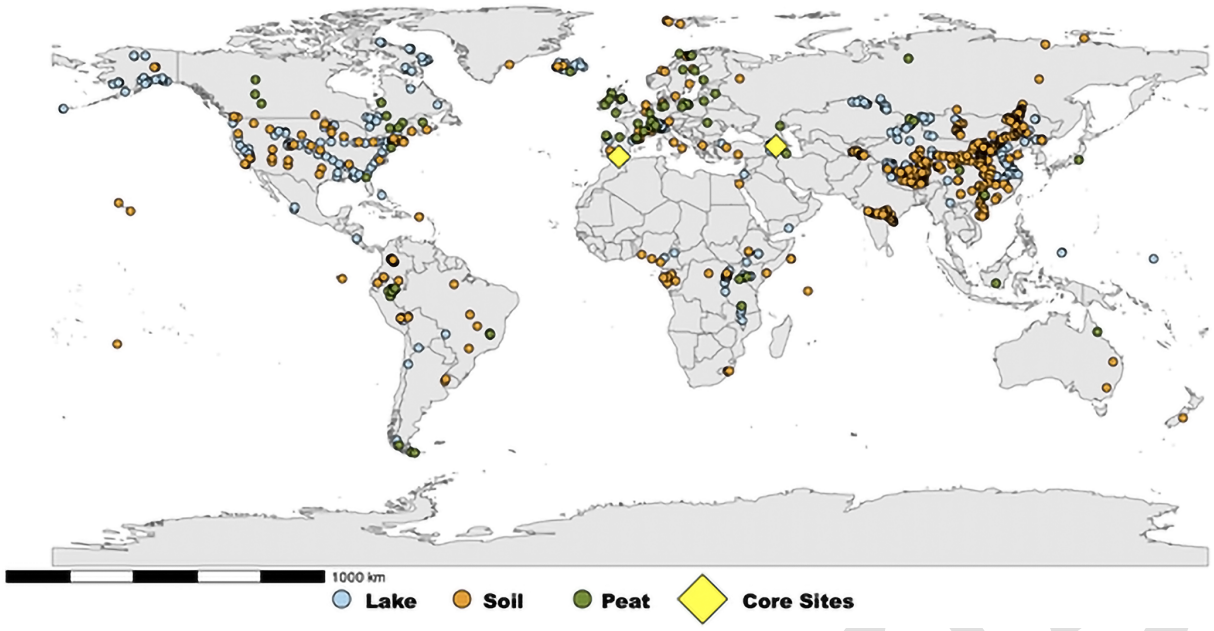

Figure 1. Map of modern sample locations used in the compiled database alongside the two sites designated for paleo-reconstructions. Map created with R package ggplot2 (Wickham, 2016).

2.1.2 Addition of new samples from Armenia

A total of 30 new surface samples from the country of Armenia were added to the global dataset to expand the database for semi-arid environments. Nine samples were collected from wetlands at a depth of 0–2 cm; one sample was collected from Lake Sevan at a depth of 2–3 cm, as previously discussed in Robles et al. (2022); and 20 surface soil samples were collected. For brGDGT extraction, each sample was first lyophilized, freeze-dried, and ground. Lipids were extracted from 0.5 to 1 g of sample in two rounds with a MARS 6 CEM microwave, using a 3 : 1 mixture of dichloromethane (DCM) and methanol (MeOH) (3 : 1). These samples were filtered with a silicon SPE cartridge with a mixture of hexane : DCM (1 : 1) and then DCM : MeOH (1 : 1) to separate the apolar and polar fraction, respectively. An internal standard of C_{46}, following Huguet et al. (2006), was added to the total lipid extract (TLE) prior to separating the fraction. The polar fraction was then analyzed on high-performance liquid chromatography with atmospheric pressure chemical ionization mass spectrometry (HPLC-APCI-MS, Agilent 1200) at LGL-TPE ENS, which allows separation of the 5- and 6-methyl GDGTs, following Hopmans et al. (2016). Selective ion monitoring (SIM) of m/z of 1050, 1048, 1046, 1036, 1032, 1022, 1020, 1018, and 744 was used for the brGDGT isomers and the internal C_{46} standard (Hopmans et al., 2016; Davtian et al., 2018; Huguet et al., 2006).

2.1.3 Resampling and balancing modern dataset

The distribution of modern brGDGT samples across classification datasets (i.e., soils, lakes, peats) was not uniform, with a predominance of soil samples and an underrepresentation of lake and peat samples (Fig. 1). Unbalanced datasets can result in considerable performance issues, such as the misclassification of data with limited sample sizes, which may prevent the learning algorithm from identifying general patterns within the datasets (He and Garcia, 2009). Consequently, a combination of downsampling and upsampling techniques was utilized for model comparisons (Fig. 2). This involved evaluating each ML model using both the raw and resampled datasets, which incorporated upsampled synthetic samples. In the resampled dataset, we initially preformed random downsampling of the soil samples in R to achieve a sample size of 750 from the original dataset. The synthetic minority oversampling technique (SMOTE) function from the R library smotefamily (Siriseriwan, 2019) was employed to upsample the peat and lake datasets. The SMOTE function is an oversampling technique that selects a sample from the minority dataset, identifies its nearest neighbor(s), and generates a new data point between the original pair (Siriseriwan, 2019). SMOTE was utilized to generate 1000 synthetic samples for the lake and peat datasets derived from the original datasets. The distribution of the SMOTE samples and original database were plotted, and a principal component analysis and Kolmogorov–Smirnov test were run to verify that no bias was introduced (results in Figs. S3 and S4 and Table S1 in the Supplement). Samples were randomly selected from

the synthetic dataset to adjust the raw datasets for lake and peat to a total of 750. In the case of the peat and lake samples, 219 and 159 synthetic samples were incorporated into the raw dataset, respectively.

2.2 Machine learning models

2.2.1 Building probability and classification machines

In supervised classification problems, ML algorithms utilize grouped attributes and features to identify patterns within human-curated datasets (Kalita, 2022). Samples in these datasets are typically assigned a label (class), target value, or dependent variable, which correspond to independent variables and features. The model utilizes this information to understand the relationships between the independent and dependent variables during the training process (Sendhilkumar and Geetha, 2023). The models are subsequently refined and evaluated for accuracy using a subset of the known classification dataset that has not been previously encountered by the model. A distinct validation set is employed to adjust the probability estimates. Numerous classification ML models employ probability estimates to determine the appropriate class (Murphy, 2012). When calibrated, these probability estimates can provide information that extends beyond merely identifying an individual's category but can also indicate the degree of likeness of an individual belonging to a category (Malley et al., 2012). Most ML algorithms, when initially deployed, lack calibration for precise probability predictions. Calibration is essential to ensure that the empirical probability is both valid and accurate (Dawid, 1985). In the absence of calibration, certain model outputs may push probability estimates toward 0 or 1, necessitating correction through calibration (Niculescu-Mizil and Caruana, 2005). Typically, either sigmoid ("Platt scaling") or isotonic regression is employed for calibration on a validation dataset that the model has not previously encountered (Niculescu-Mizil and Caruana, 2005). Subsequent to these steps, the model may be utilized to predict a class within a dataset where the classification remains unknown.

Five diverse algorithms were tested based on various methodological and practical reasons. First, we choose algorithms that could produce reliable probability estimates and have been widely utilized and validated (Malley et al., 2012; Wang et al., 2019). Algorithms were also chosen by performance on smaller tabular datasets, low computing resource requirements, and their availability in the scikit-learn Python library, which is available publicly for download (Pedregosa et al., 2011). These methods were chosen over more complex deep-learning methods, which often underperform on small tabular datasets (Grinsztajn et al., 2022) and require significant time and expertise for hyper-tuning (Mohammed and Kora, 2023); and other complex ensemble methods, which can require more computing resources without increased accuracy. Logistic regression (LR) is a parametric model akin to linear regression in its functionality, yet it is more appropriate for classification tasks (i.e., binary outcomes) (Hilbe, 2016). The analysis relies on the likelihood of an event occurring and the alignment of predictor response variables within the probability distribution (Hilbe, 2016). The algorithm is inherently calibrated for precise probability outputs due its foundational reliance on probability.

K-nearest neighbor (KNN), support vector machine (SVM), classification and regression tree (CART), and random forest (RF) are non-parametric models demonstrated to be effective in probability estimation following calibration (i.e., Niculescu-Mizil and Caruana, 2005; Kruppa et al., 2014; Dankowski and Ziegler, 2016; Cearns et al., 2020). KNN functions as a "lazy learner" by determining the distance between data points according to the characteristics of the training dataset (Geetha and Sendhilkumar, 2023). The "K" in KNN refers to a small positive integer that determines the number of neighbors taken into account when predicting the category of a data point (Geetha and Sendhilkumar, 2023). KNN is extensively employed in palaeosciences for paleoclimate regression issues, particularly through the modern analog technique (Simpson, 2007). The support vector machine (SVM) is a model that positions data items in n-dimensional space based on n features (Geetha and Sendhilkumar, 2023). Classification is achieved by identifying a hyperplane in the dimensional space that distinguishes between the classes (Geetha and Sendhilkumar, 2023).

CART and RF are tree-based learning algorithms. Trees are formed through three fundamental steps: (1) binary splits are selected, (2) a determination is made about whether the node is terminal or requires further splitting, and (3) a class is assigned to the terminal leaf node (Bell, 1999). RF is founded on the principles of natural variability and randomness inherent in trees, where both the variables and the individual elements exhibit a degree of randomness (Genuer and Poggi, 2020). RF classification problems utilize a committee of decision trees that collectively vote to determine the predicted class (Hastie et al., 2009). In the classification problems, each vote corresponds to a classification in the terminal node of the tree (Malley et al., 2012), with the majority vote determining the final classification outcome. The probability estimates are derived by calculating the fraction of votes from each tree to determine the predicted class probability.

2.2.2 Verification, tuning, and calibration of models

The data were split into a $60:20:20$ training, testing, and validation set. This provided enough data to train the model with high accuracy and ensure that testing and calibration could occur on datasets that were previously unseen during training. The models underwent testing and hyperparameter tuning using a k-fold cross-validation approach, incorporating 10 data splits and a parameter grid with the test dataset. K-fold cross-validation involves partitioning the data into

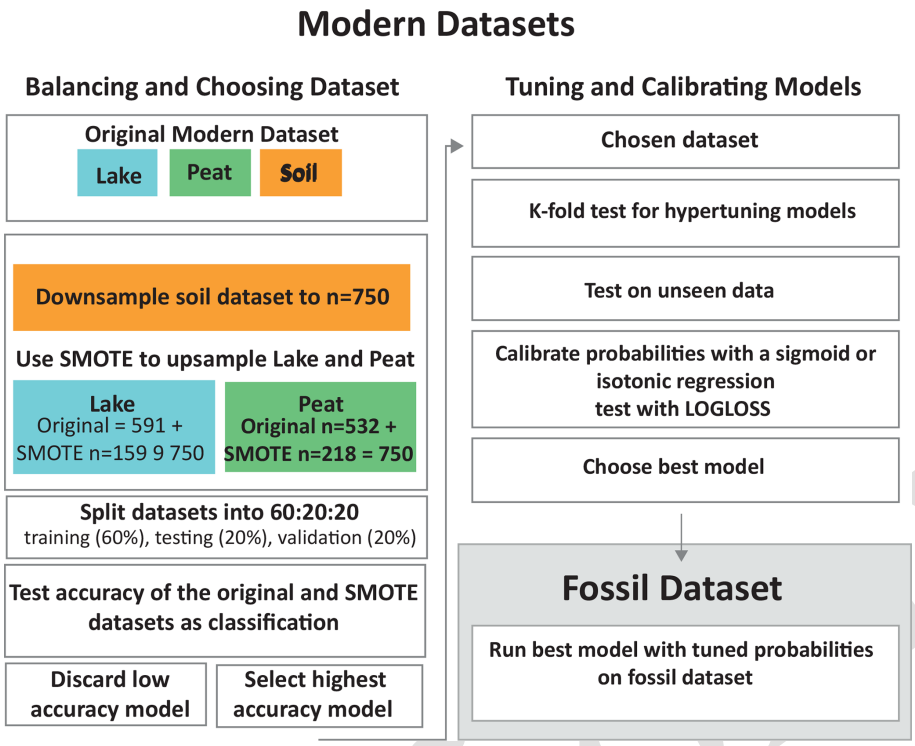

Figure 2. Illustration of the methods employed in this study for testing, tuning, and validating the datasets and models.

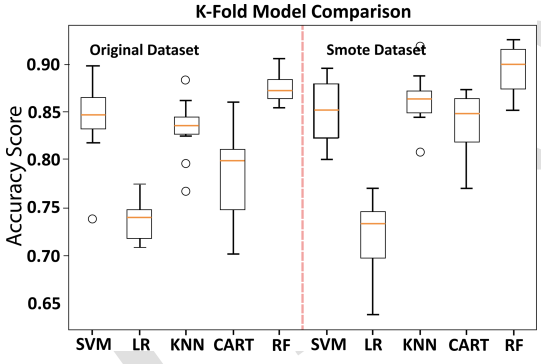

Figure 3. Comparison of k-fold testing models between datasets, utilizing k-fold cross-validation for classification on both SMOTE and original datasets. The k-fold comparison utilized 10 splits across supported vector machine (SVM), logistic regression (LR), K-nearest neighbor (KNN), classification and regression tree (CART), and random forest (RF).

equal-sized subsets, which are then utilized k times, with $k-1$ subsets used for training and one subset reserved for validation (Jung, 2018). The performance is evaluated by averaging each k iteration. The parameter grid facilitates the iteration over a finite set of values to identify optimal variables for tuning. After tuning, all models and datasets were retested for accuracy. The distribution was subsequently plotted, and the mean F_1 accuracy results were computed (Fig. 3).

2.2.3 Probability estimate calibration and application of classification machines

Instead of solely predicting the class (e.g., soil, peat, lake), we are employing the probability estimate output generated by the classification algorithms as a proxy for provenance change. The probability output enables the estimation of the likelihood that a specific sample belongs to a particular class, thus facilitating the identification of periods of mixed provenance. Due to the lack of calibration in the default probability estimates of the algorithms employed, we applied sigmoid and isotonic regression on the validation dataset to rectify any distortion and then assessed the effectiveness using the log-loss function in scikit-learn. Log loss is employed in probability scenarios where the likelihood of an event being true is represented as 1, equally true as 0.5, and false as 0 (Manzali et al., 2017). In log loss, a greater divergence between the predicted value and the actual value results in a higher log-loss score (Dembla, 2020). A lower score indicates greater accuracy in predictions. Log-loss scores were subsequently compared across models to evaluate performance. To estimate the 95 % confidence intervals for each downcore record, we performed 500 bootstrap resampling on the probability predictions. These were computed separately for each record to reflect their individual variance.

2.3 Application of models, downcore pollen, non-pollen palynomorph, XRF, and brGDGT analysis

To assess the accuracy of the probability estimates on the downcore record, five machine-learning models were applied to two published brGDGT records that included datasets of pollen and non-pollen palynomorphs (NPPs). Aquatic pollen and NPPs provide critical insights into alterations in lake or wetland ecology (e.g., Cromartie et al., 2020; Robles et al., 2022). We selected two records: one from Armenia in the southern Caucasus (Vanevan peat: 40°12′8.83″ N, 45°40′24.03″ E; Robles et al., 2022) and another from southern Spain (Padul paleolake: 37°00′39″ N, 3°36′14″ W; Camuera et al., 2018, 2019; Ramos-Román et al., 2018; Rodrigo-Gámiz et al., 2022), both situated at comparable latitudes in Eurasia.

The extraction methods for brGDGTs are detailed in the original articles by Robles et al. (2022) and Rodrigo-Gámiz et al. (2022). We revisited the original chromatograms of Robles et al. (2022) to investigate the presence of the IIIa″ isomer, which had not been published previously, to verify the brGDGT-based ML lake probability output. The IIIa″ isomer was reported in Rodrigo-Gámiz et al. (2022). Robles et al. (2022) provide identification and counting methods for aquatic pollen and NPPs in the Vanevan peat, while Ramos-Román et al. (2018) and Camuera et al. (2019) address similar methods for the Padul paleolake. Additionally, we recalculated the reconstructed water depth based on aquatic pollen and NPPs. The analysis relies on the raw datasets employing the original equations established by Robles et al. (2022) and Camuera et al. (2019). Instead of applying a smoothing technique to the water-depth reconstruction, as done by Camuera et al. (2019) on the original 200 000-year-old sequence, we retained the original sample-to-sample curve for clarity to compare to the shorter brGDGT sequence. These fossils of semi-aquatic plants and fungal and fern spores are generally representative of local change, particularly in wetlands (Gill et al., 2009; Tunno and Mensing, 2017), which strengthens their usage as local indicators. The percentages of the aquatic and NPP taxa were calculated by summing up all relevant pollen types for each record and dividing each taxon by the total sum. We calculated and re-calculated key brGDGT-based indices (Table 1) to compare our ML results with the brGDGT record as well as the aquatic pollen and NPPs. In addition, we also compared our results with the principal component analysis on the XRF datasets, also taken from the same cores, published in Robles et al. (2022) and Camuera et al. (2018). The descriptions of this analysis can be found in the original publications.

2.4 Descriptive statistics

The programming languages R (R core team, 2020) and Python (Python Software Foundation, Python Language Reference, version 3.7.3, available at http://www.python.org, last access: 24 April 2024) were utilized alongside ggplot2 (Wickham, 2016) and matplotlib (Hunter, 2007) to visualize the results. Redundancy analysis (RDA) was performed using the vegan package (Oksanen et al., 2019). RDA was employed in two capacities: (i) to compare the fractional abundances of the brGDGTs in the global modern dataset with the authors' descriptive categories (i.e., soil, peat, lake) and (ii) to compare the probability estimate results from the Vanevan and Padul records with the pollen and NPPs. In this analysis, we downsampled the pollen record to align with the brGDGT resolution, selecting samples that were no more than 100 years apart. Bayesian change-point analyses were conducted on the brGDGT-based ML lake probability results using the bcp package (Erdman and Emerson, 2007) in R to identify significant shifts in depositional environments.

3 Results

3.1 New modern brGDGT dataset

The raw database compiled for this study comprised a total of 2282 TS2 samples (591 from lakes, 532 from peats, and 1177 TS3 from soils). This addition includes 319 lake samples to the database of Martínez-Sosa et al. (2021), 62 peat samples to Naafs et al. (2017b), and 450 soil samples to the Dearing Crampton-Flood et al. (2020) datasets. Subsequent to the compilation of this dataset, additional datasets have been published (e.g., Raberg et al., 2022b; Martínez-Sosa et al., 2023) that are not incorporated into our dataset. Figures 1 and S1 in the Supplement illustrate the distribution of the brGDGT datasets.

3.2 Model accuracy and log loss

In classification mode, all models demonstrated mean accuracy F_1 scores, which measures the predictive accuracy of the models between 0.72 and 0.90 (Fig. 3). The study compared the performance of various classification models, with the SMOTE dataset showing superior results over the raw unbalanced dataset for SVM, KNN, CART, and RF (Table 2). The raw dataset showed better performance with LR, while the SMOTE dataset improved probability estimates for SVM and RF but decreased for LR, KNN, and CART. The sigmoid calibration improved probabilities for RF, CART, and KNN but decreased for SVM and LR. The isotonic calibration improved probabilities for KNN and CART but decreased for SVM, LR, and RF over uncalibrated probabilities. The sigmoid function outperformed the isotonic function on both datasets for SVM, LR, and RF. The RF model with the SMOTE dataset had the highest accuracy and the lowest log-loss score for sigmoid and uncalibrated probabilities and was therefore chosen for our downcore analysis as the best-performing model.

Table 1. The brGDGT indices employed in this study.

Index	Formulae	Citation
MBT'_{5ME}	$MBT'_{5ME} = ([Ia] + [Ib] + [Ic]) / ([Ia] + [Ib] + [Ic] + [IIa] + [IIb] + [IIc] + [IIIa])$	De Jonge et al. (2014b)
CBT'	$CBT = {}^{10}\log([Ic] + [IIa'] + [IIb'] + [IIc'] + [IIIa'] + [IIIb'] + [IIIc']) / ([Ia] + [IIa] + [IIIa])$	De Jonge et al. (2014b)

Table 2. Evaluation of accuracy across various models to determine optimal performance. The mean accuracy results were derived from a k-fold evaluation of the models, focusing on the classification of data categories (i.e., lake, peat, soil) using 10 splits. Log loss was computed for the probability estimates following their calibration, using either a sigmoid or an isotonic function. For these functions, values approaching 0 signify improved performance. Bold values represent the best-performing models once calibrated.

Model and database	F_1 mean accuracy	Standard deviation	Log loss uncalibrated	Log loss sigmoid	Log loss isotonic
SVM	0.84	0.04	0.46	0.51	0.93
SVM_SMOTE	0.85	0.03	0.40	0.50	0.64
LR	0.74	0.02	0.66	0.68	0.75
LR_SMOTE	0.72	0.04	0.69	0.70	0.73
KNN	0.83	0.03	1.17	0.46	0.80
KNN_SMOTE	0.86	0.03	2.08	0.41	0.41
CART	0.79	0.05	0.91	0.59	0.64
CART_SMOTE	0.84	0.03	3.95	0.55	0.55
RF	0.88	0.01	0.35	**0.34**	0.48
RF_SMOTE	**0.89**	0.03	**0.31**	**0.3**	0.56

3.3 Downcore analysis

3.3.1 Downcore probability estimates and change-point analysis

The Padul record showed mean probabilities for lake, peat, and soil, with lake having the highest probability in 68 out of 93 samples and peat in 25 out of 93 (Fig. 4, column 1). Vanevan's mean probability was 0.87 for lake, 0.05 for soil, and 0.08 for peat, with lake samples having the highest probability in 44 out of 46 samples, peat in 2 out of 46, and no samples having the highest probability in soil (Fig. 4, column 2).

3.3.2 Probability analysis by change-point phases

For the Padul record, change-points were detected at 12 837, 20 627, and 29617 cal BP for lake probabilities, dividing it into phases 1–4 (Fig. 4, column 1). Change-points in the Vanevan record were identified at 2043, 3577, 4628, 5061, and 8592 cal BP, based on lake probabilities (Fig. 4, column 2). Change-point analysis of the Padul record indicates that brGDGT-based ML lake probabilities peak in phase 3, while in phases 1, 2, and 4, these probabilities vary between soil and peat. BrGDGT-based ML soil probabilities are the highest in phase 1, while peat probabilities are consistent, primarily in phases 4 and 2 (Fig. 4, column 2). The probabilities of Vanevan brGDGT-based ML lake probabilities are elevated across the entire record, peaking during phases 5, 6, and 2, while peat probabilities are elevated in phases 4 and 1, and soil probabilities exhibit fluctuations with peat and lake, predominantly in phases 4 and 3 (Fig. 4, column 2).

3.4 RDA analysis of modern and downcore pollen, NPP, and brGDGTs

3.4.1 Modern samples

The RDA analysis reveals the association of brGDGTs with each depositional unit (i.e., soil, lake, peat) in the global modern database and the downcore probability predictions. Most variance can be explained across RDA-1 (30.3), where peat and soil sources sit in contrast to lakes in the modern database (Fig. 5a). BrGDGTs Ia and IIa are more clearly associated with peat and soil depositional environments, while the rest have a stronger association with lake and soil environments (Fig. 5a). Comparisons between depositional environment probability estimates, aquatic pollen, and NPPs reveal relationships between these variables in the downcore record. Pollen and NPP associations between peat and lake probabilities include Cyperaceae pollen, while algae such as *Pediastrum*, *Botryococcus*, and *Myriophyllum* have associations with lake probabilities. For soil, spores and algae are associated with these depositional environments. Hdv-200 and Cyperaceae have the highest explanatory power for peat; monolete spores and *Polypodium* for soil; and *Botryococcus*, *Myriophyllum*, and *Pediastrum* for lakes (Fig. 5b and c).

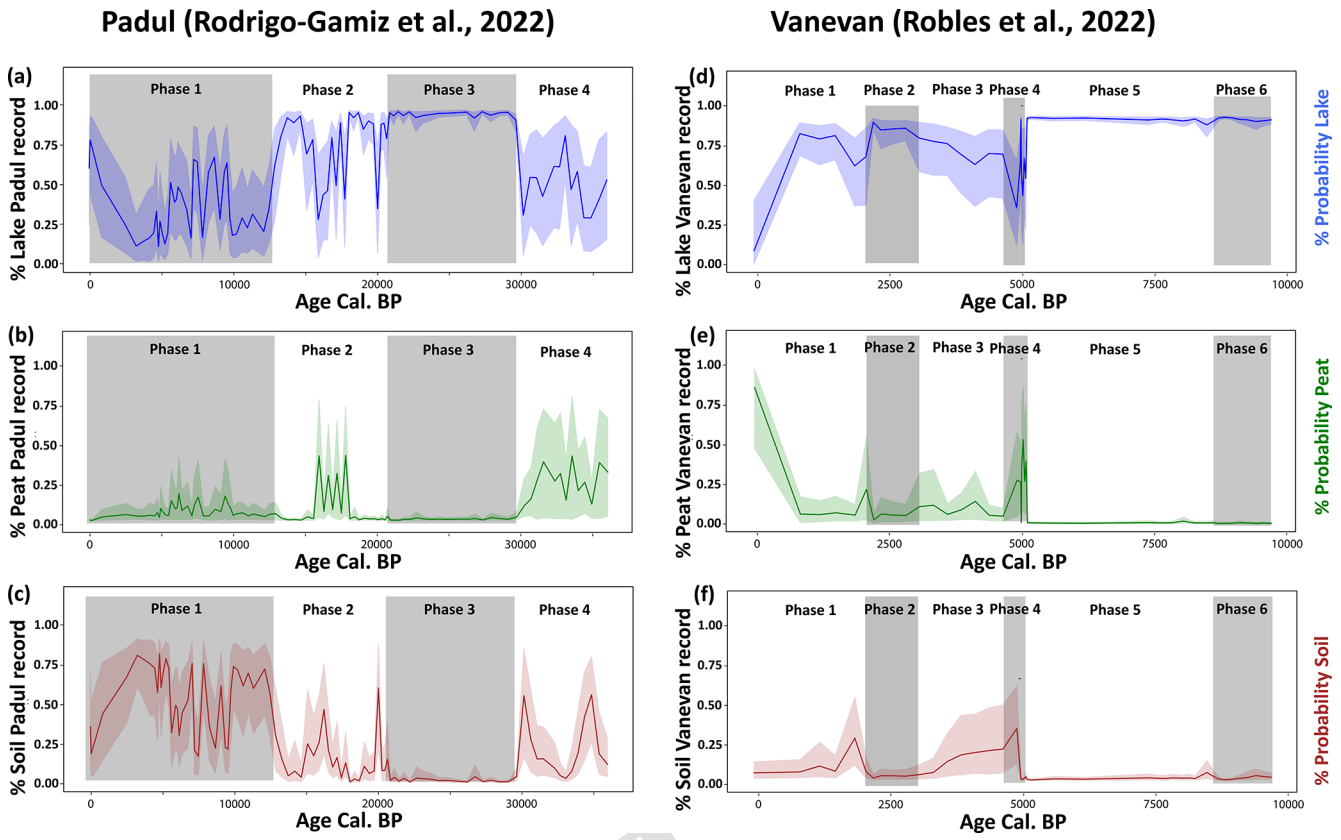

Figure 4. Downcore probability estimates with 95 % confidence intervals and change-point breaks from random forest (RF) on the SMOTE dataset with a sigmoid calibration. Results from the Padul (1) and Vanevan (2) records are broken down by lake probabilities (blue curves – **a**), peat probabilities (green curves – **b**), and soil probabilities (brown curves – **c**). Highlighted gray and white boxes indicate change-point mean breaks, identifying phases. Probability estimates from other models can be found in the Supplement (Figs. S9–S11).

4 Discussion

4.1 Probability estimates for chosen models and application to downcore records

4.1.1 Model accuracy

The F_1 score evaluates the accuracy of a model's predictions of both precision (how many predicted positives were positive) and recall (from all the positives, how many positives the model predicted) and can balance between understanding false positives and false negatives (Boozary et al., 2025). This score allows for a more robust accuracy when measuring each model. Many things may explain differences in F_1 scores across our models. For example, KNN, SVM, and CART models are prone to overfitting (Huang et al., 2005; Berk, 2008; Jadhav and Channe, 2016), which may have accounted for their lower F_1 scores (Table 1). RF generally does not overfit due to its ability to handle noise in the datasets (Parmar et al., 2019), which may result in a higher F_1 score. While LR does not typically overfit, the lower F_1 score may be due to its assumptions of linearity (Nick and Campbell 2007), which may be problematic if there is no clear division in the dataset. The balanced versus unbalanced datasets may have also impacted performance. RF generally handles unbalanced datasets well (Anaissi et al., 2013), and the SMOTE dataset only offered marginal improvements to the F_1 score, while CART's F_1 score was significantly improved with the balanced SMOTE datasets. For the log-loss scores, logistic regression is already calibrated (Kull et al., 2017) so calibrating may result in a lower log-loss score, with both sigmoid and isotonic calibration. SVM does not produce true probabilities by default and needs to be calibrated for these results (Kull et al., 2017). By calibrating them with a sigmoid or isotonic regression, the output turns to true probabilities which may result in a lower log-loss score. For KNN, CART, and RF, calibration improved the models on both datasets.

The random forest model utilizing the SMOTE dataset achieves an F_1 score of 89 % (Table 2) in classification, which is lower than the 95 % F_1 score reported for the BigMac model by Martínez-Sosa et al. (2023). The difference in F_1 scores is anticipated as a result of differing training datasets and methodologies, including the incorporation of isoprenoid GDGTs by Martínez-Sosa et al. (2023), their es-

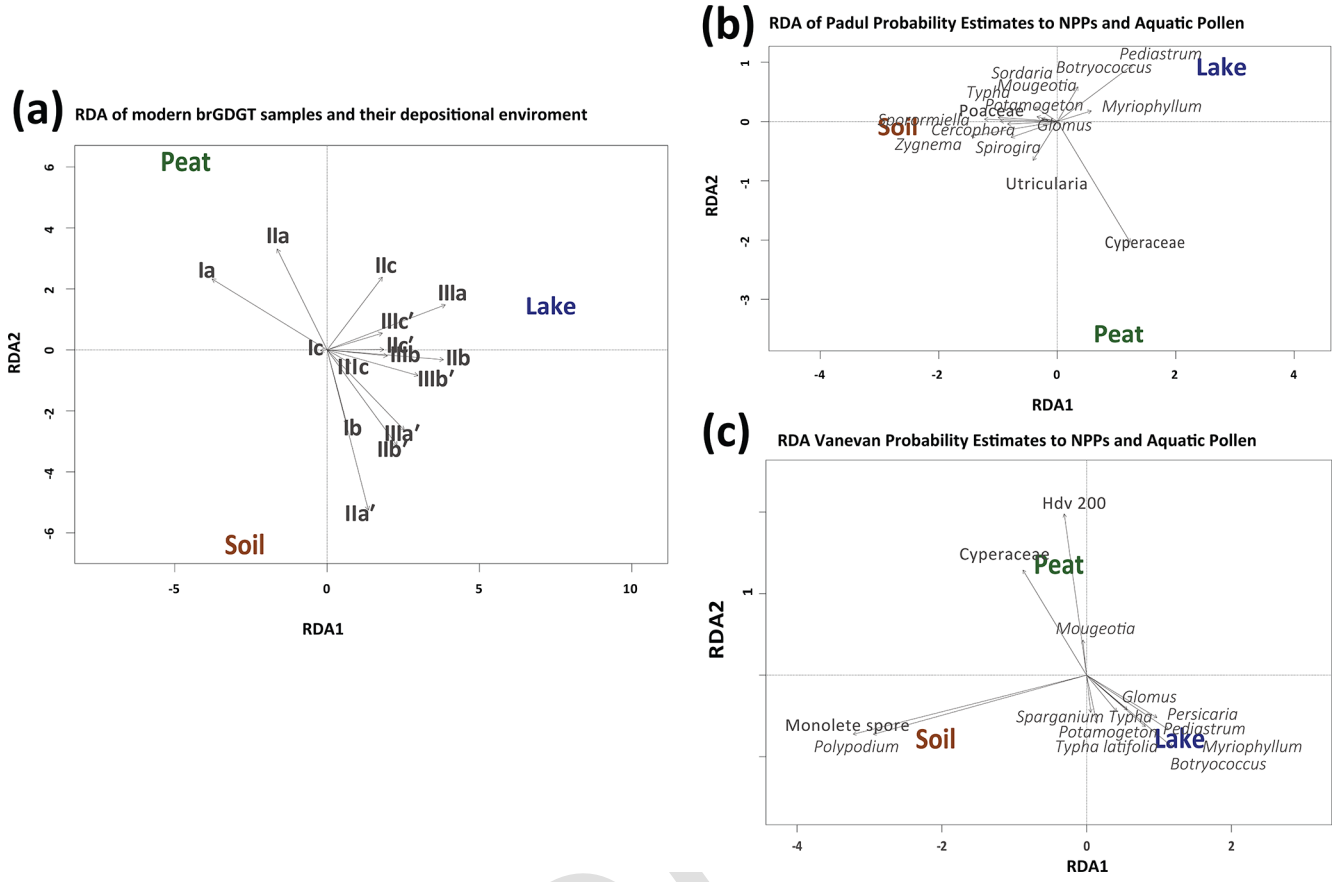

Figure 5. **(a)** RDA analysis of the modern fractional abundances of the global brGDGT database with their depositional environments. **(b)** Probability estimates results of the depositional environment compared with pollen and NPPs of the Padul record (i.e., Rodrigo-Gámiz et al., 2022; Camuera et al., 2019) and **(c)** the Vanevan record (i.e., Robles et al., 2022).

tablishment of curated clusters, and the application of classification models. Our models focus on probability estimate outputs instead of discrete classes, allowing for a nuanced understanding of shifts in provenance; thus, a lower F_1 score is permissible. A score of 89 % demonstrated a strong and precise model, despite being lower than expected. The difference between the new model and the BigMac model is also seen when applied to the downcore records: the BigMac model failed to predict soil classification for both cores, whereas our probabilities for soil were elevated in the Padul record (Figs. S6 and S7). The inclusion of additional soil samples in our database enhances soil identification during model training (BigMac $n = 192$, database in this paper $n = 750$).

4.1.2 Validating models with pollen, NPPs, XRF, and derived water-depth reconstructions

We evaluated the accuracy of our ML model for detecting provenance change by comparing the probability estimates it produced with published data on pollen, NPPs, XRF, and water-depth estimates derived from these proxies to check for quantitative similarities (Figs. 5 and 6). The comparison of GDGT-model variables with depositional environments indicates clear associations. Individual RDA analysis of both records associates Cyperaceae pollen, prevalent in wetland and peat contexts, with modern brGDGT samples obtained from depositional peat environments (Fig. 5). Downcore records indicate distinct associations between pollen and algae, specifically *Pediastrum* and *Botryococcus*, which are typically associated with open lakes, and the brGDGT-based ML probability estimates associated with the depositional lake environment (Fig. 5). Monolete spores in the Vanevan record, along with *Sordaria* and *Sporormiella* in the Padul record, are related to the brGDGT-trained ML probability estimates of depositional soil environments. In the Padul record, these spores were likely introduced through human activity (Ramos-Román et al., 2018) and may be associated with erosion into the lake. Pollen from semi-aquatic plants, including Cyperaceae and *Typha*, follow similar trends that align with brGDGT-trained ML lake probabilities, as well as increases in brGDGT-trained ML probabilities for peat and soil across both records (Figs. 6 and 7). The comparison of reconstructed water-depth results with the probabil-

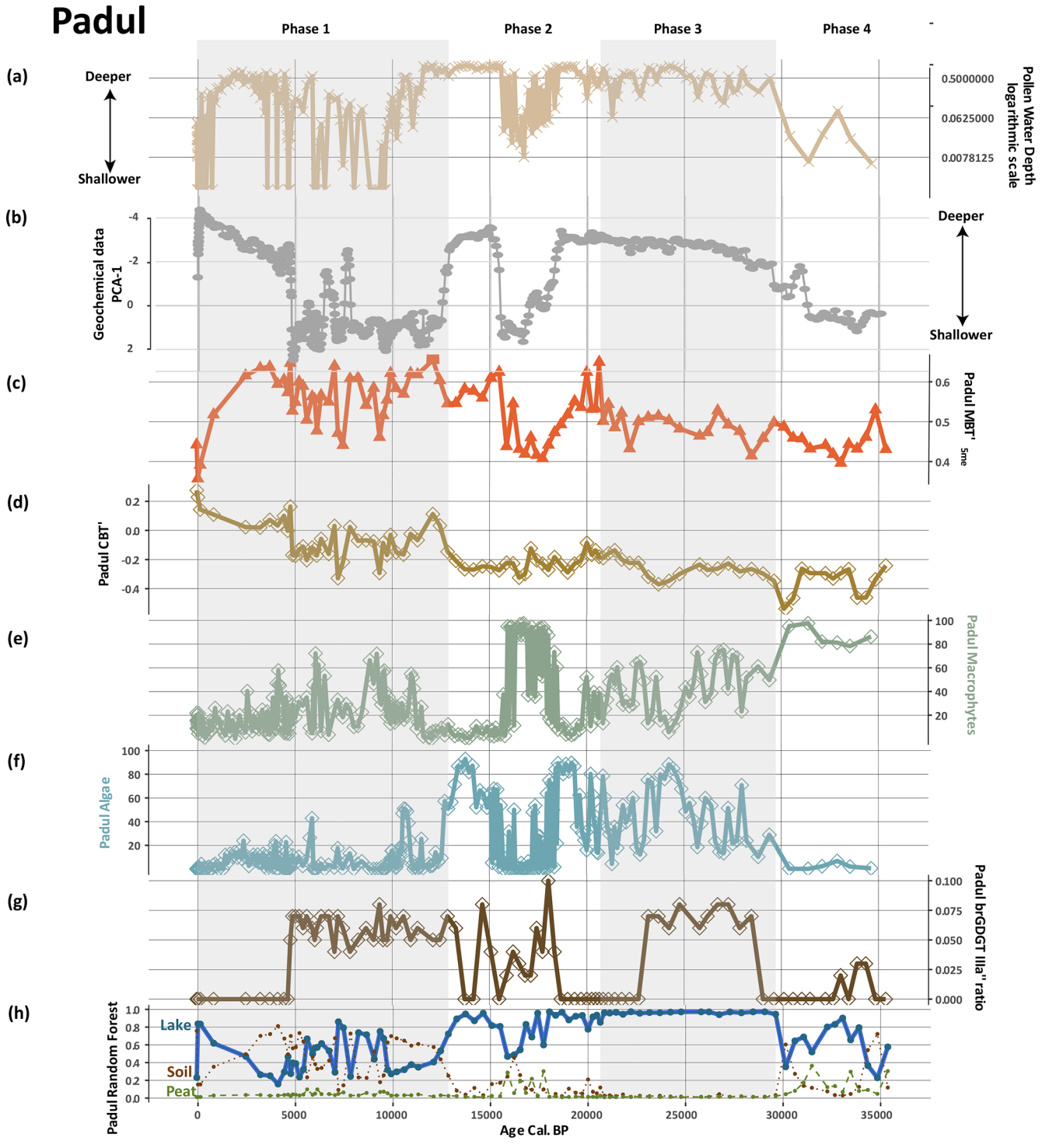

Figure 6. Comparison of probability estimates from the Padul record with aquatic pollen, NPPs, XRF, and brGDGT indexes (data from Ramos-Román et al., 2018; Camuera et al., 2018, 2019; Rodrigo-Gámiz et al., 2022). **(a)** Pollen- and NPP-based water-depth reconstructions. **(b)** Output from principal component analysis (PCA-1) from the XRF and geochemical data (Camuera et al., 2018). **(c)** The MBT'_{5ME} brGDGT index. **(d)** The CBT′ brGDGT index. **(e)** Selected aquatic plants including Cyperaceae and *Typha*. **(f)** Selected algae *Pediastrum*, *Botryococcus*, and *Mougeota* as well as aquatic plants Cyperaceae and *Typha*. **(g)** The IIIa″ brGDGT ratio. **(h)** Probability estimates for the depositional lake environments (this study).

ity estimates from the brGDGT-trained ML model for both cores indicates that the two proxies exhibit similar quantitative patterns of increase and decrease. Similarly, PCA data derived from XRF datasets, from the same core, follow these trends. Robles et al. (2022) associate higher lake levels with the PCA-1 associated with positive loadings (P, K, Al, Mg, Si, Ti, Fe) and negative loadings (S) with lower lake levels. This suggests that our models accurately identify changes in sourcing and hydrology across both records (Fig. 7a). Robles et al. (2022) interpreted the water-depth changes for the Vanevan record based on aquatic pollen, NPPs, and XRF data identifying a shallow lake from 9700 to 9400 cal BP, a lake system from 9700 to 5100 cal BP, a transitional phase from 5100 to 4950 cal BP, and peatland development from 5100 cal BP to today. This aligns closely with our change-point phases, indicating elevated lake probabilities during phases 6 and 5, high peat probabilities during phase 4, and a rise in soil and peat probabilities from our brGDGT-trained ML model over the past 5000 years (Fig. 7e).

Principal component analysis (PCA) was also conducted on the XRF dataset on the Padul core and the authors used this as a proxy for lake-level change. Camuera et al. (2018) attributed negative loadings of PCA-1 with higher lake levels (Ca, Sr, Si, A, MS) and positive loadings with lower lake levels (Fe, S, Br, TOC, C/N). The ML-brGDGT-based probability estimates, the PCA output from geochemical data including XRF, and the pollen-based water-depth reconstructions are in alignment with Padul (Fig. 6), indicating the model's accuracy. The probability estimates in the Padul record exhibit trends analogous to the Vanevan results, with an alignment with water depth, as indicated by pollen and NPPs (Fig. 7b). The estimates derive from the pollen data and XRF data (Camuera et al., 2018, 2019), indicating a low water stand in phase 4, a high water stand in phase 3, a fluctuating high to low to high stand in phase 2, and a high fluctuating to low to high stand in phase 1. The observed trends are reflected in our brGDGT-based ML lake probability estimates. Similar to the Vanevan record, the Padul record predominantly features samples with brGDGT-based ML lake probabilities assigned to lakes. However, there is greater variation among categorical types. This is evident in phases 4 and 3 (where peat and soil probabilities are combined with lake probabilities) and in phase 1 (where notable fluctuations occur in soil and lake probabilities).

4.2 Environmental controls and depositional shifts in downcore brGDGT records

4.2.1 Identifying provenance changes in downcore records (and their impact on MBT'_{5ME})

The ML-probability estimates may be interpreted as originating from either a dominant or mixed-sourced sedimentary environment. During periods of high brGDGT-based ML lake probabilities, the findings indicate that brGDGTs are produced in situ in shallow lakes and wetlands, alongside contributions from other sources. This result is unexpected for the Vanevan record, as the brGDGTs from the last 5000 years exhibit a closer alignment with soil samples when plotted on a ternary diagram (Robles et al., 2022). We compared our results with the classification provided by the BigMac ML model from Martínez-Sosa et al. (2023), which classified most samples as a depositional lake environment, similar to our results, while the samples at 5000 cal BP for Vanevan were categorized as soil and peat rather than soil and lake. In the Padul record between 35 000 and 30 000 cal BP and from 10 000 cal BP to the present, differences between models include samples classified as soil (our model) rather than peat or lake (Fig. S3). A potential bias in both models may arise from samples in the global database categorized as lakes but with substantial contributions of brGDGTs from depositional soil or peat environments.

Second, our findings indicate that the identification of lacustrine brGDGTs produced in situ from depositional lake environments using a model provides more nuance than the quantification of the IIIa″ brGDGT isomers. Rodrigo-Gámiz et al. (2022) identified IIIa″ in the Padul record, which is attributed to in situ brGDGT lake production. Here, the ratio of brGDGTs IIIa″ aligns with the brGDGT-based ML lake-probability estimates for this record (Fig. 6). The IIIa″ isomer is completely absent from the Vanevan record; however, the brGDGT-based ML lake probability estimates approach 100 %. This supports earlier studies indicating that the IIIa″ isomer is not universally found in all lake systems (i.e., Weber et al., 2015; Dang et al., 2018). However, the lack of discussion regarding the absence of the IIIa″ isomer in both modern and downcore records is notable and warrants attention in future investigations.

The probability and RDA results underscore the need for multivariate methods in the analyses of depositional environments. RDA analysis of the global brGDGT database reveals a distinct separation along RDA-2 between 5- and 6- methyl pentamethylated brGDGTs in various modern depositional environments (Fig. 5a). This is observed between IIa and IIa′ associated with depositional peat and soil environments, respectively (Fig. 6). Martínez-Sosa et al. (2023) identified IIa′ as the most significant isomer for provenance classification using their random forest model, a finding that aligns with our models. Furthermore, brGDGT Ia exhibits a stronger association with depositional peat environments compared to other tetramethylated brGDGTs linked to lake environments, highlighting the need to advance beyond ternary diagrams for provenance identification.

The results of our models indicate that variations in brGDGT provenance, even in mixed sedimentary environments, significantly influence widely used indexes like MBT'_{5ME} (Figs. 6 and 7). Pollen and NPPs provide independent confirmation of the impact that these changes have on MBT'_{5ME}, particularly when analyzed alongside data from our new brGDGTs global database. In this database,

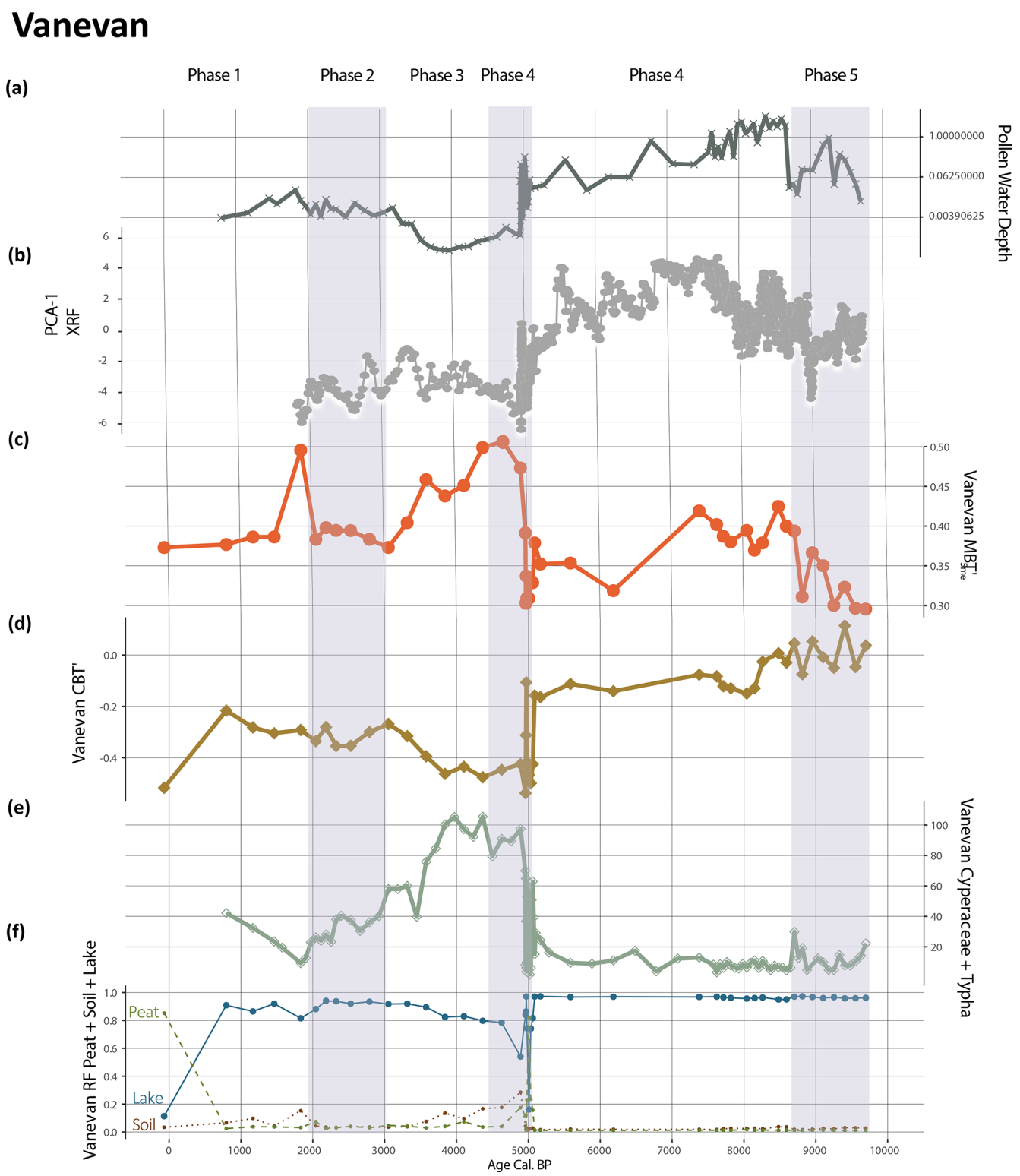

Figure 7. Comparison between the probability estimates on the Vanevan record and the aquatic pollen, NPPs, XRF, and brGDGT indexes (data from Robles et al., 2022). **(a)** Pollen- and NPP-based water-depth reconstructions. **(b)** PCA output on XRF datasets. **(c)** The MBT'_{5ME} brGDGT index. **(d)** The CBT' brGDGT index. **(e)** The selected algae and aquatic plants for the Vanevan record are *Pediastrum*, *Botryococcus*, *Mougeota*, Cyperaceae, and *Typha*. **(f)** Probability estimates for the depositional lake environments on both records.

soil (0.56) and peat (0.58) exhibit higher mean MBT'_{5ME} values compared to the lower values observed in lakes (0.39) (Fig. S5). Both the Vanevan and Padul records indicate that MBT'_{5ME} values are elevated during periods characterized by high brGDGT-based ML soil and peat probabilities, while they are reduced during periods of high lake probabilities. The pollen-based water-depth reconstructions, serving as an independent proxy, exhibit trends analogous to both the MBT'_{5ME} and brGDGT-based ML probability estimates. These changes are documented in additional proxies from these records, including XRF and sediment analysis (e.g., Robles et al., 2022; Camuera et al., 2018), highlighting the necessity of identifying the appropriate depositional contexts.

Our findings indicate that even minor changes in provenance can affect the MBT'_{5ME} and cyclization of branched tetraether (CBT') indexes. Where increased brGDGT-based ML soil probabilities occur in the Vanevan and Padul records, they do not reach the threshold, indicative of a complete depositional environment shift, instead indicating mixed provenance (Figs. 6 and 7). The Vanevan record indicates that the large shifts in MBT'_{5ME} and CBT' occur with increased inputs of soil and peat brGDGTs during phases 3 and 4. In the Padul record, variations in CBT' correlate with increases in brGDGT-based soil ML probabilities, particularly during phase 2.

The co-occurrence of aquatic pollen, NPPs, and MBT'_{5ME} variations indicates that provenance, rather than temperature, drives these changes. Increases in MBT'_{5ME} during phases 3 and 4 of the Vanevan record correspond to shifts in aquatic pollen, indicating a transition from lake to peatland, driven by a local catchment fire event (Leroyer et al., 2016; Robles et al., 2022). The observed changes are inconsistent with regional climate reconstructions (i.e., Joannin et al., 2014; Cromartie et al., 2020), confirming that provenance change is the primary driver of this alteration.

4.2.2 Environmental drivers of provenance changes

The impact of provenance changes on MBT'_{5ME} highlights various factors that can alter the distribution of brGDGTs over time, making it a crucial aspect for environmental reconstructions. The environmental changes that cause a change in GDGT provenance will also affect the environmental chemistry. While large pH changes have the potential to impact MBT'_{5ME} values in soils, muted pH changes in soils and the impact on GDGTs produced in lakes are less well constrained. The introduction of soil brGDGTs into a lake, even in small amounts, can alter the MBT'_{5ME} distribution and also potentially introduce pH-related changes.

Rodrigo-Gámiz et al. (2022) identified a relationship in the Padul record between increases in reconstructed pH and MAAT variability within the upper 116 cm, approximately correlating to the last 5000 years. They associated this with a dried ephemeral lake and suggested potential bias in the MBT'_{5ME} reconstruction. In this section of the Padul record, high brGDGT-based ML soil probabilities align with increased CBT' and MBT'_{5ME} values, indicating the contribution of soil-derived brGDGTs and potentially pH to this variation (Fig. 7). Soil provenance changes appear to exert a greater influence on MBT'_{5ME} compared to peat and are more prevalent in the record; however, brGDGT-based ML probability estimates for these depositional environments overlap in certain sections of both records.

Our results also highlight the impact of both sudden and gradual depositional changes on the distribution of brGDGTs, driven by hydrological and ecological shifts. Hydrological changes, including variations in water depth in lakes (Stefanescu et al., 2021) and alterations in water-table levels in peat (Ofiti et al., 2024), have been demonstrated to affect brGDGT distribution. The water-depth equations for the Padul and Vanevan records incorporate Cyperaceae pollen at one end of the equation. Cyperaceae is typically associated with the development of wetland ecosystems and the process of lake shallowing. A correlation is observed between MBT'_{5ME} and Cyperaceae for the Padul record (-0.52, p-value: < 0.001). There is a correlation between MBT'_{5ME} and the water-depth reconstruction for the Vanevan record (0.40, p-value: 0.006). Alterations in hydrology influence shifts in ecological communities, which may or may not be driven by climate. Our results indicate that wetland development, resulting from ecological shifts such as the introduction of aquatic plants and/or lake shallowing, can influence MBT'_{5ME} and the distribution of brGDGTs (Figs. 6 and 7).

4.2.3 Considerations for application to sedimentary sequences

The samples from the modern training dataset come from diverse modern environmental contexts, making our script applicable to most paleoenvironmental reconstructions. The probability outputs on the Padul core, which is 36 000 years old, align with the pollen and XRF water-depth reconstructions during the last glacial period (Fig. 6), confirming the model's usefulness for records extending beyond the Holocene. Distributions of samples across the Köppen–Geiger climate gradient are not balanced (Fig. S2), with temperate environments well represented, but tropical, arid, and arctic conditions are underrepresented. Considering this, caution is urged when applying this model in these environments. In addition, caution must be taken for records in deep time, where no current modern analogs exist.

The log-loss score of 0.31 for RF with a sigmoid calibration suggests that the downcore predicted probabilities accurately detect sediment change, even with mixed provenance. The 95 % confidence intervals on the downcore predictions, however, can vary throughout the sediment sequence, and care must be taken when applying the models to ensure the records' accuracy. Due to the nature of the modern database,

which relies on the correct sedimentary context identified by the original authors, some accuracy uncertainties are possible, even on a well-trained model.

4.3 Limitations and future directions

Our findings indicate that multivariate methods, such as machine learning, are needed for analyzing brGDGT distributions. To effectively utilize these tools, a standardized collection of datasets across research groups is essential, along with increased datasets from a variety of environments. A notable limitation of our study was our reliance on the published sample name. The identification of depositional environments was successful with these names; however, variations within a depositional environment, for example, shallow versus deep lakes, remain inadequately represented. To effectively utilize these tools, it is essential to collect additional information, including water depth, salinity, pH, and redox conditions, in a standardized manner across research teams.

This study demonstrates the necessity of multi-proxy approaches to comprehend the influence of ecological, hydrological, and depositional changes on brGDGT-based reconstructions. Numerous studies presently employ pollen-based climate reconstructions in conjunction with brGDGT reconstructions (e.g., Watson et al., 2018; Martin et al., 2019; Dugerdil et al., 2021c; Robles et al., 2022, 2023; Stefanescu et al., 2021). The findings indicate that aquatic pollen, NPPs, and XRF provide valuable insights for understanding biases introduced by alterations in depositional environments and provenance.

Many downcore studies are derived from smaller lakes, wetlands, and peatlands (e.g., Martin et al., 2019; Dugerdil et al., 2021c; Robles et al., 2022, 2023; Ramos-Román et al., 2022; Acharya et al., 2023; Barhoumi et al., 2024). This study emphasizes the importance of comprehending changes in depositional environments over geological time and advocates for studies in smaller lakes, wetlands, and peatlands. There is a particular necessity for the improved classification of wetland environments in the brGDGT literature.

5 Conclusion

This study demonstrated that multivariate methods enhance the understanding of how provenance changes affect brGDGT distributions and MBT'_{5ME}. A new database of modern samples ($n = 2301$ samples) has been utilized to apply probability estimates from five machine learning algorithms to downcore sediments, facilitating the identification of changes in brGDGT provenance across depositional lake, soil, and peat environments. Utilizing calibrated probability estimates enhances the identification of the provenance of brGDGTs, including those originating from mixed sources.

The results indicate that alterations in provenance, depositional environments, and hydrology, particularly the transition from open lakes to wetlands and variations in water depth, can substantially influence the brGDGT signal. The introduction of soil-derived brGDGTs, even in minimal quantities, significantly influences the brGDGT distribution and MBT'_{5ME}.

This study confirms that independent proxies, including aquatic pollen, non-pollen palynomorphs, and XRF, can effectively quantify hydrological and ecological changes, thereby influencing the gradual depositional alterations that may affect brGDGT distribution. Our models can accurately and independently identify changes in provenance and are applicable to existing global paleoenvironmental datasets. We suggest that complementary environmental proxies, including fossil pollen, non-pollen palynomorphs, XRF, diatoms, and testate amoebae, among others, are essential for confirming changes in provenance in brGDGT environmental reconstructions.

Code availability. The code and data for this project are publicly available at https://doi.org/10.5281/zenodo.17459703 (Cromartie, 2025b).

Data availability. Data for this paper are available at https://doi.org/10.17632/tr8tppy9fz.1 (Cromartie, 2025a).

Supplement. The supplement related to this article is available online at [the link will be implemented upon publication].

Author contributions. AC conceptualized the project. AC and CDJ performed the data curation. AC created the methodology. AC wrote the software and did the formal analysis. AC, CDJ, GM, LD, MR, MJRR, and SJ provided the validation. AC, SJ, GM, and CDJ provided the funding acquisition. AC, SJ, and GM did the project administration. MR, MRG, JC, MJRR, and GJM provided the resources. SJ, GM, LS, and OP provided the supervision. AC prepared the original draft. AC, CDJ, and GM wrote the subsequent drafts. AC, CDJ, GM, MR, LD, OP, MRG, JC, MJRR, GJM, CC, LS, and SJ did the reviewing and editing.

Competing interests. The contact author has declared that none of the authors has any competing interests.

Disclaimer. Publisher's note: Copernicus Publications remains neutral with regard to jurisdictional claims made in the text, published maps, institutional affiliations, or any other geographical representation in this paper. While Copernicus Publications makes every effort to include appropriate place names, the final responsibility lies with the authors.

Acknowledgements. We would like to thank the reviewers for their thoughtful comments that improved this paper. We would like to thank members of the GDGT community who provided additional information in order to complete the GDGT database. ChatGPT was utilized to help streamline and improve performance for both the R and Python code utilized in this project. This is an ISEM contribution ISEM 2025-145.

Financial support. This research was funded by a National Science Foundation Graduate Research Fellowship (DGE-1650441) for A. Cromartie, a Chateaubriand "Make Our Planet Great Again" fellowship from the Embassy of France in the United States for A. Cromartie, and an iSite Muse mobility grant from the University of Montpellier for A. Cromartie. This research was funded, in whole or in part, by ANR, grant ANR-22-CE27-0018-02. A CC-BY public copyright license has been applied by the authors to the present document and will be applied to all subsequent versions up to the Author Accepted Manuscript arising from this submission, in accordance with the grant's open access conditions. CDJ received an SNSF PRIMA (grant no. 179783).

Review statement. This paper was edited by Petr Kuneš and reviewed by Joseph B. Novak and one anonymous referee.

References

Acharya, S., Zech, R., Strobel, P., Bliedtner, M., Prochnow, M., and De Jonge, C.: Environmental controls on the distribution of GDGT molecules in Lake Höglwörth, Southern Germany, Org. Geochem., 186, https://doi.org/10.1016/j.orggeochem.2023.104689, 2023.

Anaissi, A., Kennedy, P. J., Goyal, M., and Catchpoole, D. R.: A balanced iterative random forest for gene selection from microarray data, BMC Bioinform., 14, 104689, https://doi.org/10.1186/1471-2105-14-261, 2013.

Baker, A., Blyth, A. J., Jex, C. N., Mcdonald, J. A., Woltering, M., and Khan, S. J.: Glycerol dialkyl glycerol tetraethers (GDGT) distributions from soil to cave: Refining the speleothem paleothermometer, Org. Geochem., 136, 103890, https://doi.org/10.1016/j.orggeochem.2019.06.011, 2019.

Barhoumi, C., Ménot, G., Joannin, S., Ali, A. A., Ansanay-Alex, S., Golubeva, Y., Subetto, D., Kryshen, A., Drobyshev, I., and Peyron, O.: Temperature and fire controls on vegetation dynamics in Northern Ural (Russia) boreal forests during the Holocene based on brGDGT and pollen data, Quaternary Sci. Rev., 305, 108014, https://doi.org/10.1016/j.quascirev.2023.108, 2023.

Baxter, A. J., Hopmans, E. C., Russell, J. M., and Sinninghe Damsté, J. S.: Bacterial GMGTs in East African lake sediments: Their potential as palaeotemperature indicators, Geochim. Cosmochim. Acta, 259, 155–169, https://doi.org/10.1016/j.gca.2019.05.039, 2019.

Bell, J. F.: Tree-based methods, in: Machine Learning Methods for Ecological Applications, edited by: Fielding, A. H., Springer US, Boston, MA, 89–105, https://doi.org/10.1007/978-1-4615-5289-5_3, 1999.

Berk, R. A.: Support vector machines, Statistical Learning from a Regression Perspective, 1–28, ISBN 978-3-030-40188-7, https://doi.org/10.1007/978-3-030-40189-4, 2008.

Boozary, P., Sheykhan, S., GhorbanTanhaei, H., and Magazzino, C.: Enhancing customer retention with machine learning: A comparative analysis of ensemble models for accurate churn prediction, Int. J. Inf. Manage. Data Insights, 5, 100331, https://doi.org/10.1016/j.jjimei.2025.100331, 2025.

Buckles, L. K., Weijers, J. W. H., Tran, X.-M., Waldron, S., and Sinninghe Damsté, J. S.: Provenance of tetraether membrane lipids in a large temperate lake (Loch Lomond, UK): implications for glycerol dialkyl glycerol tetraether (GDGT)-based palaeothermometry, Biogeosciences, 11, 5539–5563, https://doi.org/10.5194/bg-11-5539-2014, 2014.

Bzdok, D., Altman, N., and Krzywinski, M.: Statistics versus machine learning, Nature Methods, 15, 233–234, https://doi.org/10.1038/nmeth.4642, 2018.

Camuera, J., Jiménez-Moreno, G., Ramos-Román, M. J., García-Alix, A., Toney, J. L., Anderson, R. S., Jiménez-Espejo, F., Kaufman, D., Bright, J., Webster, C., Yanes, Y., Carrión, J. S., Ohkouchi, N., Suga, H., Yamame, M., Yokoyama, Y., and Martínez-Ruiz, F.: Orbital-scale environmental and climatic changes recorded in a new $\sim 200,000$-year-long multiproxy sedimentary record from Padul, southern Iberian Peninsula, Quaternary Sci. Rev., 198, 91–114, https://doi.org/10.1016/j.quascirev.2018.08.014, 2018.

Camuera, J., Jiménez-Moreno, G., Ramos-Román, M. J., García-Alix, A., Toney, J. L., Anderson, R. S., Jiménez-Espejo, F., Bright, J., Webster, C., Yanes, Y., and Carrión, J. S.: Vegetation and climate changes during the last two glacial-interglacial cycles in the western Mediterranean: A new long pollen record from Padul (southern Iberian Peninsula), Quaternary Sci. Rev., 205, 86–105, https://doi.org/10.1016/j.quascirev.2018.12.013, 2019.

Cao, J., Rao, Z., Shi, F., and Jia, G.: Ice formation on lake surfaces in winter causes warm-season bias of lacustrine brGDGT temperature estimates, Biogeosciences, 17, 2521–2536, https://doi.org/10.5194/bg-17-2521-2020, 2020.

Cearns, M., Hahn, T., Clark, S., and Baune, B. T.: Machine learning probability calibration for high-risk clinical decision-making, Aust. New Zeal. J. Psychiat , 54, 123–126, 2020.

Chen, C., Bai, Y., Fang, X., Zhuang, G., Khodzhiev, A., Bai, X., and Murodov, A.: Evaluating the potential of soil bacterial tetraether proxies in westerlies dominating western Pamirs, Tajikistan and implications for paleoenvironmental reconstructions, Chem. Geol., 559, 119908, https://doi.org/10.1016/j.chemgeo.2020.119908, 2021.

Cromartie, A.: Data for: Utilizing Probability Estimates from Machine Learning and Pollen to Understand the Depositional Influences on Branched GDGT in Wetlands, Peatlands, and Lakes, Mendeley Data [data set], 1, https://doi.org/10.17632/tr8tppy9fz.1, 2025a.

Cromartie, A.: amycromartie/ProbbrGDGT: ProbbrGDGT (1.1.2), Zenodo [code], https://doi.org/10.5281/zenodo.17459703, 2025b.

Cromartie, A., Blanchet, C., Barhoumi, C., Messager, E., Peyron, O., Ollivier, V., Sabatier, P., Etienne, D., Karakhanyan, A., Khatchadourian, L., Smith, A. T., Badalyan, R., Perello, B., Lindsay, I., and Joannin, S.: The vegetation, climate, and fire his-

tory of a mountain steppe: A Holocene reconstruction from the South Caucasus, Shenkani, Armenia, Quaternary Sci. Rev., 246, 106485, https://doi.org/10.1016/j.quascirev.2020.106485, 2020.

Dang, X., Yang, H., Naafs, B. D. A., Pancost, R. D., and Xie, S.: Evidence of moisture control on the methylation of branched glycerol dialkyl glycerol tetraethers in semi-arid and arid soils, Geochim. Cosmochim. Ac., 189, 24–36, https://doi.org/10.1016/j.gca.2016.06.004, 2016.

Dang, X., Ding, W., Yang, H., Pancost, R. D., Naafs, B. D. A., Xue, J., Lin, X., Lu, J., and Xie, S.: Different temperature dependence of the bacterial brGDGT isomers in 35 Chinese lake sediments compared to that in soils, Org. Geochem., 119, 72–79, https://doi.org/10.1016/j.orggeochem.2018.02.008, 2018.

Dankowski, T. and Ziegler, A.: Calibrating random forests for probability estimation, Stat. Med., 35, 3949–3960, 2016.

Davtian, N., Bard, E., Ménot, G., and Fagault, Y.: The importance of mass accuracy in selected ion monitoring analysis of branched and isoprenoid tetraethers, Org. Geochem., 118, 58–62, https://doi.org/10.1016/j.orggeochem.2018.01.007, 2018.

Dawid, A. P.: Calibration-Based Empirical Probability, Ann. Stat., 13, 1251–1274, https://doi.org/10.1214/aos/1176349736, 1985.

Dearing Crampton-Flood, E.: Data for: BayMBT: A Bayesian calibration model for branched glycerol dialkyl glycerol tetraethers in soils and peats (V1), Mendeley Data [data set], https://doi.org/10.17632/tyv9pbv3jd.1, 2020.

Dearing Crampton-Flood, E., Tierney, J. E., Peterse, F., Kirkels, F. M. S. A., and Damsté, J. S. S.: BayMBT: A Bayesian calibration model for branched glycerol dialkyl glycerol tetraethers in soils and peats, Geochimica et Cosmochimica Acta, 268, 142–159, https://doi.org/10.1016/j.gca.2019.09.043, 2020.

De Jonge, C., Stadnitskaia, A., Hopmans, E. C., Cherkashov, G., Fedotov, A., and Sinninghe Damsté, J. S.: In situ produced branched glycerol dialkyl glycerol tetraethers in suspended particulate matter from the Yenisei River, Eastern Siberia, Geochim. Cosmochim. Ac., 125, 104706, https://doi.org/10.1016/j.gca.2013.10.031, 2014a.

De Jonge, C., Hopmans, E. C., Zell, C. I., Kim, J.-H., Schouten, S., and Damsté, J. S. S.: Occurrence and abundance of 6-methyl branched glycerol dialkyl glycerol tetraethers in soils: Implications for palaeoclimate reconstruction, Geochim. Cosmochim. Ac., 141, 97–112, 2014b.

De Jonge, C., Kuramae, E. E., Radujković, D., Weedon, J. T., Janssens, I. A., and Peterse, F.: The influence of soil chemistry on branched tetraether lipids in mid- and high latitude soils: Implications for brGDGT-based paleothermometry, Geochim. Cosmochim. Ac., 310, 95–112, https://doi.org/10.1016/j.gca.2021.06.037, 2021.

De Jonge, C., Guo, J., Hällberg, P., Griepentrog, M., Rifai, H., Richter, A., Ramirez, E., Zhang, X., Smittenberg, R. H., Peterse, F., Boeckx, P., and Dercon, G.: The impact of soil chemistry, moisture and temperature on branched and isoprenoid GDGTs in soils: A study using six globally distributed elevation transects, Org. Geochem., 187, 104706, https://doi.org/10.1016/j.orggeochem.2023.104706, 2024.

Dembla, G.: Intuition behind Log-loss score, Towards Data Science, https://towardsdatascience.com/intuition-behind-log-loss-score-4e0c9979680a/ (last access: October 2024), 2020.

Dillon, J. T., Lash, S., Zhao, J., Smith, K. P., van Dommelen, P., Scherer, A. K., and Huang, Y.: Bacterial tetraether lipids in ancient bones record past climate conditions at the time of disposal, J. Archaeolog. Sci., 96, 45–56, https://doi.org/10.1016/j.jas.2018.05.009, 2018.

Ding, S., Schwab, V. F., Ueberschaar, N., Roth, V.-N., Lange, M., Xu, Y., Gleixner, G., and Pohnert, G.: Identification of novel 7-methyl and cyclopentanyl branched glycerol dialkyl glycerol tetraethers in lake sediments, Org. Geochem., 102, 52–58, 2016.

Ding, S., Xu, Y., Wang, Y., He, Y., Hou, J., Chen, L., and He, J.-S.: Distribution of branched glycerol dialkyl glycerol tetraethers in surface soils of the Qinghai–Tibetan Plateau: implications of brGDGTs-based proxies in cold and dry regions, Biogeosciences, 12, 3141–3151, https://doi.org/10.5194/bg-12-3141-2015, 2015.

d'Oliveira, L., Dugerdil, L., Ménot, G., Evin, A., Muller, S. D., Ansanay-Alex, S., Azuara, J., Bonnet, C., Bremond, L., Shah, M., and Peyron, O.: Reconstructing 15 000 years of southern France temperatures from coupled pollen and molecular (branched glycerol dialkyl glycerol tetraether) markers (Canroute, Massif Central), Clim. Past, 19, 2127–2156, https://doi.org/10.5194/cp-19-2127-2023, 2023.

Dugerdil, L., Joannin, S., Peyron, O., Jouffroy-Bapicot, I., Vannière, B., Boldgiv, B., Unkelbach, J., Behling, H., and Ménot, G.: Climate reconstructions based on GDGT and pollen surface datasets from Mongolia and Baikal area: calibrations and applicability to extremely cold–dry environments over the Late Holocene, Clim. Past, 17, 1199–1226, https://doi.org/10.5194/cp-17-1199-2021, 2021a.

Dugerdil, L., Joannin, S., Peyron, O., Jouffroy-Bapicot, I., Vannière, B., Boldgiv, B., Behling, H., and Ménot, G.: New Mongolian-Siberian pollen and brGDGT surface dataset: local calibration for paleoclimate reconstructions, PANGEA [data set], https://doi.org/10.1594/PANGAEA.933664, 2021b.

Dugerdil, L., Ménot, G., Peyron, O., Jouffroy-Bapicot, I., Ansanay-Alex, S., Antheaume, I., Behling, H., Boldgiv, B., Develle, A. L., Grossi, V., Magail, J., Makou, M., Robles, M., Unkelbach, J., Vannière, B., and Joannin, S.: Late Holocene Mongolian climate and environment reconstructions from brGDGTs, NPPs and pollen transfer functions for Lake Ayrag: Paleoclimate implications for Arid Central Asia, Quaternary Sci. Rev., 273, 107235, https://doi.org/10.1016/j.quascirev.2021.107235, 2021c.

Erdman, C. and Emerson, J. W.: bcp: an R package for performing a Bayesian analysis of change point problems, Journal of statistical Software, 23, 1–13, https://doi.org/10.18637/jss.v023.i03, 2007.

Genuer, R., Poggi, J.-M.: Random forests with R, 1st ed., Springer Cham, X–98, https://doi.org/10.1007/978-3-030-56485-8_3, 2020.

Gill, J. L., Williams, J. W., Jackson, S. T., Lininger, K. B., and Robinson, G. S.: Pleistocene Megafaunal Collapse, Novel Plant Communities, and Enhanced Fire Regimes in North America, Science, 326, 1100–1103, https://doi.org/10.1126/science.1179504, 2009.

Grinsztajn, L., Oyallon, E., and Varoquaux, G.: Why do tree-based models still outperform deep learning on typical tabular data?, in: Advances in Neural Information Processing Systems 35 (NeurIPS 2022), vol. 35, edited by: Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., and Oh, A., 507–520, ISBN 978-1-7138-7108-8,

https://proceedings.neurips.cc/paper_files/paper/2022/file/0378c7692da36807bdec87ab043cdadc-Paper-Datasets_and_Benchmarks.pdf (last access: 14 March 2025), 2022.

Guo, J., Glendell, M., Meersmans, J., Kirkels, F., Middelburg, J. J., and Peterse, F.: Assessing branched tetraether lipids as tracers of soil organic carbon transport through the Carminowe Creek catchment (southwest England), Biogeosciences, 17, 3183–3201, https://doi.org/10.5194/bg-17-3183-2020, 2020a.

Guo, J., Glendell, M., Meersmans, J., Kirkels, F. M. S. A., Middelburg, J. J., and Peterse, F.: Branched tetraether lipids in Carminowe Creek catchment (southwest England), PANGEA [data set], https://doi.org/10.1594/PANGAEA.918523, 2020b.

Halffman, R., Lembrechts, J., Radujković, D., De Gruyter, J., Nijs, I., and De Jonge, C.: Soil chemistry, temperature and bacterial community composition drive brGDGT distributions along a subarctic elevation gradient, Org. Geochem., 163, 104346, https://doi.org/10.1016/j.orggeochem.2021.104346, 2022.

Hastie, T., Tibshirani, R., Friedman, J.: The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second., Springer, 744 pp., ISBN 978-0-387-84857-0, 2009.

He, H. and Garcia, E. A.: Learning from Imbalanced Data, IEEE Transactions on Knowledge and Data Engineering, 21, 1263–1284, https://doi.org/10.1109/TKDE.2008.239, 2009.

Hilbe, J. M.: Practical guide to logistic regression, CRC Press, ISBN 978-1-4987-0958-3, 2016.

Hopmans, E. C., Weijers, J. W. H., Schefuß, E., Herfort, L., Sinninghe Damsté, J. S., and Schouten, S.: A novel proxy for terrestrial organic matter in sediments based on branched and isoprenoid tetraether lipids, Earth Planet. Sc. Lett., 224, 107–116, https://doi.org/10.1016/j.epsl.2004.05.012, 2004.

Hopmans, E. C., Schouten, S., and Damsté, J. S. S.: The effect of improved chromatography on GDGT-based palaeoproxies, Org. Geochem., 93, 1–6, 2016.

Huang, K., Yang, H., King, I., and Lyu, M. R.: Local Learning vs. Global Learning: An Introduction to Maxi-Min Margin Machine, Springer Nature, https://doi.org/10.1007/10984697_5, 2005.

Huguet, C., Hopmans, E. C., Febo-Ayala, W., Thompson, D. H., Sinninghe Damsté, J. S., and Schouten, S.: An improved method to determine the absolute abundance of glycerol dibiphytanyl glycerol tetraether lipids, Org. Geochem., 37, 1036–1041, https://doi.org/10.1016/j.orggeochem.2006.05.008, 2006.

Hunter, J. D.: Matplotlib: A 2D graphics environment, Comput. Sci. Eng., 9, 90–95, https://doi.org/10.1109/MCSE.2007.55, 2007.

Jadhav, S. D. and Channe, H.: Comparative study of K-NN, naive Bayes and decision tree classification techniques, Int. J. Sci. Res., 5, 1842–1845, 2016.

Jaeschke, A., Rethemeyer, J., Lappé, M., Schouten, S., Boeckx, P., and Schefuß, E.: Influence of land use on distribution of soil n-alkane δD and brGDGTs along an altitudinal transect in Ethiopia: Implications for (paleo)environmental studies, Org. Geochem., 124, 77–87, https://doi.org/10.1016/j.orggeochem.2018.06.006, 2018.

Jaeschke, A.: Distribution of soil n-alkane δD and branched glycerol dialkyl glycerol tetraether lipids (brGDGT) along an altitudinal transect in Ethiopia, PANGAEA [data set], https://doi.org/10.1594/PANGAEA.895602, 2018.

Joannin, S., Ali, A. A., Ollivier, V., Roiron, P., Peyron, O., Chevaux, S., Nahapetyan, S., Tozalakyan, P., Karakhanyan, A., and Chataigner, C.: Vegetation, fire and climate history of the Lesser Caucasus: a new Holocene record from Zarishat fen (Armenia), J. Quaternary Sci., 29, 70–82, https://doi.org/10.1002/jqs.2679, 2014.

Jung, Y.: Multiple predicting K-fold cross-validation for model selection, J. Nonparam. Stat., 30, 197–215, https://doi.org/10.1080/10485252.2017.1404598, 2018.

Kalita, J.: Machine learning: Theory and practice, Chapman and Hall/CRC, ISBN 9781003002611, https://doi.org/10.1201/9781003002611, 2022.

Kirkels, F. M. S. A., Panton, C., Galy, V., West, A. J., Feakins, S. J., and Peterse, F.: From Andes to Amazon: Assessing Branched Tetraether Lipids as Tracers for Soil Organic Carbon in the Madre de Dios River System, JGR Biogeosciences, 125, e2019JG005270, https://doi.org/10.1029/2019JG005270, 2020.

Kou, Q., Zhu, L., Ju, J., Wang, J., Xu, T., Li, C., and Ma, Q.: Influence of salinity on glycerol dialkyl glycerol tetraether-based indicators in Tibetan Plateau lakes: Implications for paleotemperature and paleosalinity reconstructions, Palaeogeogr. Palaeoclimatol. Palaeoecol., 601, 111127, https://doi.org/10.1016/j.palaeo.2022.111127, 2022.

Kruppa, J., Liu, Y., Biau, G., Kohler, M., König, I. R., Malley, J. D., and Ziegler, A.: Probability estimation with machine learning methods for dichotomous and multicategory outcome: theory, Biometric. J., 56, 534–563, 2014.

Kull, M., Silva Filho, T., and Flach, P.: Beta calibration: a well-founded and easily implemented improvement on logistic calibration for binary classifiers, in: Artificial intelligence and statistics, 623–631, https://proceedings.mlr.press/v54/kull17a.html (last access: April 2025), 2017.

Leroyer, C., Joannin, S., Aoustin, D., Ali, A. A., Peyron, O., Ollivier, V., Tozalakyan, P., Karakhanyan, A., and Jude, F.: Mid Holocene vegetation reconstruction from Vanevan peat (southeastern shore of Lake Sevan, Armenia), Quatern. Int., 395, 5–18, 2016.

Li, J., Naafs, B. D. A., Pancost, R. D., Yang, H., Liu, D., and Xie, S.: Distribution of branched tetraether lipids in ponds from Inner Mongolia, NE China: Insight into the source of brGDGTs, Org. Geochem., 112, 127–136, https://doi.org/10.1016/j.orggeochem.2017.07.005, 2017.

Li, J., Pancost, R. D., Naafs, B. D. A., Yang, H., Zhao, C., and Xie, S.: Distribution of glycerol dialkyl glycerol tetraether (GDGT) lipids in a hypersaline lake system, Org. Geochem., 99, 113–124, https://doi.org/10.1016/j.orggeochem.2016.06.007, 2016.

Liang, J., Russell, J. M., Xie, H., Lupien, R. L., Si, G., Wang, J., Hou, J., and Zhang, G.: Vegetation effects on temperature calibrations of branched glycerol dialkyl glycerol tetraether (brGDGTs) in soils, Org. Geochem., 127, 1–11, https://doi.org/10.1016/j.orggeochem.2018.10.010, 2019.

Liang, J., Richter, N., Xie, H., Zhao, B., Si, G., Wang, J., Hou, J., Zhang, G., and Russell, J. M.: Branched glycerol dialkyl glycerol tetraether (brGDGT) distributions influenced by bacterial community composition in various vegetation soils on the Tibetan Plateau, Palaeogeogr. Palaeocl. Palaeoecol., 611, 111358, https://doi.org/10.1016/j.palaeo.2022.111358, 2023.

Loomis, S. E., Russell, J. M., Ladd, B., Street-Perrott, F. A., and Sinninghe Damsté, J. S.: Calibration and application of the branched GDGT temperature proxy on East African lake sedi-

ments, Earth and Planetary Science Letters, 357–358, 277–288, https://doi.org/10.1016/j.epsl.2012.09.031, 2012.

Loomis, S. E., Russell, J. M., Eggermont, H., Verschuren, D., and Sinninghe Damsté, J. S.: Effects of temperature, pH and nutrient concentration on branched GDGT distributions in East African lakes: Implications for paleoenvironmental reconstruction, Organic Geochemistry, 66, 25–37, https://doi.org/10.1016/j.orggeochem.2013.10.012, 2014a.

Loomis, S. E., Russell, J. M., Heureux, A. M., D'Andrea, W. J., and Sinninghe Damsté, J. S.: Seasonal variability of branched glycerol dialkyl glycerol tetraethers (brGDGTs) in a temperate lake system, Geochimica et Cosmochimica Acta, 144, 173–187, https://doi.org/10.1016/j.gca.2014.08.027, 2014b.

Loomis, S. E., Russell, J. M., and Sinninghe Damsté, J. S.: Distributions of branched GDGTs in soils and lake sediments from western Uganda: Implications for a lacustrine paleothermometer, Org. Geochem., 42, 739–751, https://doi.org/10.1016/j.orggeochem.2011.06.004, 2011.

Malley, J. D., Kruppa, J., Dasgupta, A., Malley, K. G., and Ziegler, A.: Probability Machines: Consistent probability estimation using nonparametric learning machines, Meth. Inform. Med., 51, 74–81, https://doi.org/10.3414/ME00-01-0052, 2012.

Manzali, Y., Chahhou, M., and El Mohajir, M.: Impure decision trees for Auc and log loss optimization, in: Proceddings for: 2017 International Conference on Wireless Technologies, Embedded and Intelligent Systems (WITS), Fez, Morraco, 19–20 April 2017, 1–6, https://doi.org/10.1109/WITS.2017.7934675, 2017.

Martin, C., Ménot, G., Thouveny, N., Davtian, N., Andrieu-Ponel, V., Reille, M., and Bard, E.: Impact of human activities and vegetation changes on the tetraether sources in Lake St Front (Massif Central, France), Org. Geochem., 135, 38–52, https://doi.org/10.1016/j.orggeochem.2019.06.005, 2019.

Martin, C., Ménot, G., Thouveny, N., Peyron, O., Andrieu-Ponel, V., Montade, V., Davtian, N., Reille, M., and Bard, E.: Early Holocene Thermal Maximum recorded by branched tetraethers and pollen in Western Europe (Massif Central, France), Quaternary Sci. Rev., 228, 106109, https://doi.org/10.1016/j.quascirev.2019.106109, 2020.

Martínez-Sosa, P., Tierney, J. E., Stefanescu, I. C., Dearing Crampton-Flood, E., Shuman, B. N., and Routson, C.: A global Bayesian temperature calibration for lacustrine brGDGTs, Geochim. Cosmochim. Ac., 305, 87–105, https://doi.org/10.1016/j.gca.2021.04.038, 2021.

Martínez-Sosa, P., Tierney, J. E., Pérez-Angel, L. C., Stefanescu, I. C., Guo, J., Kirkels, F., Sepúlveda, J., Peterse, F., Shuman, B. N., and Reyes, A. V.: Development and Application of the Branched and Isoprenoid GDGT Machine Learning Classification Algorithm (BIGMaC) for Paleoenvironmental Reconstruction, Paleoceanogr. Paleoclimatol., 38, e2023PA004611, https://doi.org/10.1029/2023PA004611, 2023.

Menges, J., Huguet, C., Alcañiz, J. M., Fietz, S., Sachse, D., and Rosell-Melé, A.: Influence of water availability in the distributions of branched glycerol dialkyl glycerol tetraether in soils of the Iberian Peninsula, Biogeosciences, 11, 2571–2581, https://doi.org/10.5194/bg-11-2571-2014, 2014.

Mohammed, A. and Kora, R.: A comprehensive review on ensemble deep learning: Opportunities and challenges, J. King Saud Univers. – Comput. Inf. Sci., 35, 757–774, https://doi.org/10.1016/j.jksuci.2023.01.014, 2023.

Murphy, K. P.: Machine learning: a probabilistic perspective, MIT Press, ISBN 978-0-262-01802-9, 2012.

Naafs, B. D. A.: Global biomarker (GDGT) database for peatlands, PANGAEA [data set], https://doi.org/10.1594/PANGAEA.883765, 2017.

Naafs, B. D. A., Inglis, G. N., Zheng, Y., Amesbury, M. J., Biester, H., Bindler, R., Blewett, J., Burrows, M. A., del Castillo Torres, D., Chambers, F. M., Cohen, A. D., Evershed, R. P., Feakins, S. J., Gałka, M., Gallego-Sala, A., Gandois, L., Gray, D. M., Hatcher, P. G., Honorio Coronado, E. N., Hughes, P. D. M., Huguet, A., Könönen, M., Laggoun-Défarge, F., Lähteenoja, O., Lamentowicz, M., Marchant, R., McClymont, E., Pontevedra-Pombal, X., Ponton, C., Pourmand, A., Rizzuti, A. M., Rochefort, L., Schellekens, J., De Vleeschouwer, F., and Pancost, R. D.: Introducing global peat-specific temperature and pH calibrations based on brGDGT bacterial lipids, Geochim. Cosmochim. Ac., 208, 285–301, https://doi.org/10.1016/j.gca.2017.01.038, 2017a.

Naafs, B. D. A., Gallego-Sala, A. V., Inglis, G. N., and Pancost, R. D.: Refining the global branched glycerol dialkyl glycerol tetraether (brGDGT) soil temperature calibration, Org. Geochem., 106, 48–56, 2017b.

Nick, T. G. and Campbell, K. M.: Logistic regression, Topics in biostatistics, edited by: Ambrosius, W. T., Humana Press, Humana Totowa, NJ, XII-528, 273–301, ISBN 978-1-59745-530-5, 2007.

Niculescu-Mizil, A. and Caruana, R.: Predicting good probabilities with supervised learning, in: Proceedings of the 22nd international conference on Machine learning, Bonn, Germany, 7–11 August 2005, 625–632, https://doi.org/10.1145/1102351.1102430, 2005.

Ning, D., Zhang, E., Shulmeister, J., Chang, J., Sun, W., and Ni, Z.: Holocene mean annual air temperature (MAAT) reconstruction based on branched glycerol dialkyl glycerol tetraethers from Lake Ximenglongtan, southwestern China, Org. Geochem., 133, 65–76, https://doi.org/10.1016/j.orggeochem.2019.05.003, 2019.

Ofiti, N. O. E., Huguet, A., Hanson, P. J., and Wiesenberg, G. L. B.: Peatland warming influences the abundance and distribution of branched tetraether lipids: Implications for temperature reconstruction, Sci. Total Environ., 924, 171666, https://doi.org/10.1016/j.scitotenv.2024.171666, 2024.

Oksanen, J., Blanchet, F. G., Friendly, M., Kindt, R., Legendre, P., McGlinn, D., Minchin, P. R., O'Hara, R. B., Simpson, G. L., Solymos, P., Stevens, M. H. H., Szoecs, E., and Wagner, H.: vegan: Community Ecology Package, version 2.7-1, https://doi.org/10.32614/CRAN.package.vegan, 2019.

Parmar, A., Katariya, R., and Patel, V.: A Review on Random Forest: An Ensemble Classifier, in: Lecture Notes on Data Engineering and Communications Technologies, vol. 26, Springer Nature, https://doi.org/10.1007/978-3-030-03146-6_86, 2019.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V.,Vanderplas, J., Passos, A., and Cournapeau, D.: Scikit-learn: Machine learning in Python, J. Mach. Learn. Res, 12, 2825–2830, 2011.

Pérez-Angel, L. C., Sepúlveda, J., Molnar, P., Montes, C., Rajagopalan, B., Snell, K., Gonzalez-Arango, C., and Dildar, N.: Soil and air temperature calibrations using branched GDGTs for

the Tropical Andes of Colombia: Toward a pan-tropical calibration, Geochem. Geophys. Geosystems, 21, e2020GC008941, https://doi.org/10.1029/2020GC008941, 2020.

Peterse, F., van der Meer, J., Schouten, S., Weijers, J. W. H., Fierer, N., Jackson, R. B., Kim, J.-H., and Damsté, J. S. S.: Revised calibration of the MBT–CBT paleotemperature proxy based on branched tetraether membrane lipids in surface soils, Geochim. Cosmochim. Ac., 96, 215–229, 2012.

Qian, S., Yang, H., Dong, C., Wang, Y., Wu, J., Pei, H., Dang, X., Lu, J., Zhao, S., and Xie, S.: Rapid response of fossil tetraether lipids in lake sediments to seasonal environmental variables in a shallow lake in central China: Implications for the use of tetraether-based proxies, Org. Geochem., 128, 108–121, https://doi.org/10.1016/j.orggeochem.2018.12.007, 2019.

Raberg, J. H., Harning, D. J., Crump, S. E., de Wet, G., Blumm, A., Kopf, S., Geirsdóttir, Á., Miller, G. H., and Sepúlveda, J.: Revised fractional abundances and warm-season temperatures substantially improve brGDGT calibrations in lake sediments, Biogeosciences, 18, 3579–3603, https://doi.org/10.5194/bg-18-3579-2021, 2021a.

Raberg, J. H., Harning, D. J., Crump, S. E., de Wet, G. A., Blumm, A., Kopf, S., Geirsdóttir, Á., Miller, G. H., and Sepúlveda, J.: brGDGT distributions and environmental parameters of lake sediments from the Eastern Canadian Arctic and Iceland, 2003–2019, PANGEA [data set], https://doi.org/10.1594/PANGAEA.931003, 2021b.

Raberg, J. H., Flores, E., Crump, S. E., de Wet, G., Dildar, N., Miller, G. H., Geirsdóttir, Á., and Sepúlveda, J.: Intact Polar brGDGTs in Arctic Lake Catchments: Implications for Lipid Sources and Paleoclimate Applications, J. Geophys. Res.-Biogeo., 127, e2022JG006969, https://doi.org/10.1029/2022JG006969, 2022a.

Raberg, J. H., Miller, G. H., Geirsdóttir, Á., and Sepúlveda, J.: Near-universal trends in brGDGT lipid distributions in nature, Sci. Adv., 8, eabm7625, https://doi.org/10.1126/sciadv.abm7625, 2022b.

Ramos-Román, M. J., Jiménez-Moreno, G., Camuera, J., García-Alix, A., Anderson, R. S., Jiménez-Espejo, F. J., and Carrión, J. S.: Holocene climate aridification trend and human impact interrupted by millennial- and centennial-scale climate fluctuations from a new sedimentary record from Padul (Sierra Nevada, southern Iberian Peninsula), Clim. Past, 14, 117–137, https://doi.org/10.5194/cp-14-117-2018, 2018.

Ramos-Román, M. J., De Jonge, C., Magyari, E., Veres, D., Ilvonen, L., Develle, A. L., and Seppä, H.: Lipid biomarker (brGDGT)- and pollen-based reconstruction of temperature change during the Middle to Late Holocene transition in the Carpathians, Global Planet. Change, 215, 103859, https://doi.org/10.1016/j.gloplacha.2022.103859, 2022.

Rao, Z., Haichun, G., Cao, J., Shi, F., Jia, G., Li, Y., and Chen, F.: Consistent long-term Holocene warming trend at different elevations in the Altai Mountains in arid central Asia, Journal of Quaternary Science, 35, 1036–1045, https://doi.org/10.1002/jqs.3254, 2020.

Rao, Z., Guo, H., Wei, S., Cao, J., and Jia, G.: Influence of water conditions on peat brGDGTs: A modern investigation and its paleoclimatic implications, Chem. Geol., 606, 120993, https://doi.org/10.1016/j.chemgeo.2022.120993, 2022.

Robles, M., Peyron, O., Brugiapaglia, E., Ménot, G., Dugerdil, L., Ollivier, V., Ansanay-Alex, S., Develle, A. L., Tozalakyan, P., Meliksetian, K., Sahakyan, K., Sahakyan, L., Perello, B., Badalyan, R., Colombié, C., and Joannin, S.: Impact of climate changes on vegetation and human societies during the Holocene in the South Caucasus (Vanevan, Armenia): A multiproxy approach including pollen, NPPs and brGDGTs, Quaternary Sci. Rev., 277, 107297, https://doi.org/10.1016/j.quascirev.2021.107297, 2022.

Robles, M., Peyron, O., Ménot, G., Brugiapaglia, E., Wulf, S., Appelt, O., Blache, M., Vannière, B., Dugerdil, L., Paura, B., Ansanay-Alex, S., Cromartie, A., Charlet, L., Guédron, S., de Beaulieu, J.-L., and Joannin, S.: Climate changes during the Late Glacial in southern Europe: new insights based on pollen and brGDGTs of Lake Matese in Italy, Clim. Past, 19, 493–515, https://doi.org/10.5194/cp-19-493-2023, 2023.

Rodrigo-Gámiz, M., García-Alix, A., Jiménez-Moreno, G., Ramos-Román, M. J., Camuera, J., Toney, J. L., Sachse, D., Anderson, R. S., and Sinninghe Damsté, J. S.: Paleoclimate reconstruction of the last 36 kyr based on branched glycerol dialkyl glycerol tetraethers in the Padul palaeolake record (Sierra Nevada, southern Iberian Peninsula), Quaternary Sci. Rev., 281, 107434, https://doi.org/10.1016/j.quascirev.2022.107434, 2022.

Russell, J. M., Hopmans, E. C., Loomis, S. E., Liang, J., and Sinninghe Damsté, J. S.: Distributions of 5- and 6-methyl branched glycerol dialkyl glycerol tetraethers (brGDGTs) in East African lake sediment: Effects of temperature, pH, and new lacustrine paleotemperature calibrations, Org. Geochem., 117, 56–69, https://doi.org/10.1016/j.orggeochem.2017.12.003, 2018.

Sendhilkumar, S. and Geetha, T. V.: Machine Learning: Concepts, Techniques and Applications, 1st ed., CRC Press, Boca Raton, FL, ISBN 978-1-032-26829-3, 2023.

Simpson, G. L.: Analogue methods in palaeoecology: using the analogue package, J. Stat. Softw., 22, 1–29, 2007.

Sinninghe Damste, J., Baxter, A. J., Hopmans, E., and Russell, J.: Data for: Bacterial GMGTs in East African lake sediments: Their potential as palaeotemperature indicators (V1), Mendeley Data [data set], https://doi.org/10.17632/npsyv38f6t.1, 2020.

Siriseriwan, W.: A collection of oversampling techniques for class imbalance problem based on SMOTE, https://cran.r-project.org/web/packages/smotefamily/smotefamily.pdf (last access: 24 October 2024), 2019.

Stefanescu, I. C., Shuman, B. N., and Tierney, J. E.: Dataset from manuscript "Temperature and water depth effects on brGDT distributions in sub-alpine lakes of mid-latitude North America" (V1), Mendeley Data [data set], https://doi.org/10.15786/20.500.11919/7164, 2020.

Stefanescu, I. C., Shuman, B. N., and Tierney, J. E.: Temperature and water depth effects on brGDGT distributions in sub-alpine lakes of mid-latitude North America, Organic Geochemistry, 152, 104174, https://doi.org/10.1016/j.orggeochem.2020.104174, 2021.

R Core Team: R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, https://www.R-project.org (last access: 24 October 2024), 2020.

Tierney, J. E. and Russell, J. M.: Distributions of branched GDGTs in a tropical lake system: Implications for lacustrine application of the MBT/CBT paleoproxy, Org. Geochem., 40, 1032–1036, https://doi.org/10.1016/j.orggeochem.2009.04.014, 2009.

Tierney, J. E., Russell, J. M., Eggermont, H., Hopmans, E. C., Verschuren, D., and Sinninghe Damsté, J. S.: Environmental controls on branched tetraether lipid distributions in tropical East African lake sediments, Geochimica et Cosmochimica Acta, 74, 4902–4918, https://doi.org/10.1016/j.gca.2010.06.002, 2010.

Tunno, I. and Mensing, S. A.: The value of non-pollen palynomorphs in interpreting paleoecological change in the Great Basin (Nevada, USA), Quatern. Res., 87, 529–543, https://doi.org/10.1017/qua.2017.8, 2017.

Vèquaud, P., Derenne, S., Thibault, A., Anquetil, C., Bonanomi, G., Collin, S., Contreras, S., Nottingham, A. T., Sabatier, P., Salinas, N., Scott, W. P., Werne, J. P., and Huguet, A.: Development of global temperature and pH calibrations based on bacterial 3-hydroxy fatty acids in soils, Biogeosciences, 18, 3937–3959, https://doi.org/10.5194/bg-18-3937-2021, 2021a.

Véquaud, P., Derenne, S., Anquetil, C., Collin, S., Poulenard, J., Sabatier, P., and Huguet, A.: Influence of environmental parameters on the distribution of bacterial lipids in soils from the French Alps: Implications for paleo-reconstructions, Org. Geochem., 153, 104194, https://doi.org/10.1016/j.orggeochem.2021.104194, 2021b.

Véquaud, P., Thibault, A., Derenne, S., Anquetil, C., Collin, S., Contreras, S., Nottingham, A. T., Sabatier, P., Werne, J. P., and Huguet, A.: FROG: A global machine-learning temperature calibration for branched GDGTs in soils and peats, Geochim. Cosmochim. Ac., 318, 468–494, https://doi.org/10.1016/j.gca.2021.12.007, 2022.

Wang, H., An, Z., Lu, H., Zhao, Z., and Liu, W.: Calibrating bacterial tetraether distributions towards in situ soil temperature and application to a loess-paleosol sequence, Quat. Sci. Rev., 231, 106172, https://doi.org/10.1016/j.quascirev.2020.106172, 2020a.

Wang, H. and Liu, W.: Soil temperature and brGDGTs along an elevation gradient on the northeastern Tibetan Plateau: A test of soil brGDGTs as a proxy for paleoelevation, Chem. Geol., 566, 120079, https://doi.org/10.1016/j.chemgeo.2021.120079, 2021.

Wang, H., Liu, W., He, Y., Zhou, A., Zhao, H., Liu, H., Cao, Y., Hu, J., Meng, B., Jiang, J., Kolpakova, M., Krivonogov, S., and Liu, Z.: Salinity-controlled isomerization of lacustrine brGDGTs impacts the associated MBT5ME' terrestrial temperature index, Geochim. Cosmochim. Ac., 305, 33-48, https://doi.org/10.1016/j.gca.2021.05.004, 2021.

Wang, H., Liu, W., and Lu, H.: Appraisal of branched glycerol dialkyl glycerol tetraether-based indices for North China, Org. Geochem., 98, 118–130, https://doi.org/10.1016/j.orggeochem.2016.05.013, 2016.

Wang, M., Yang, H., Zheng, Z., and Tian, L.: Altitudinal climatic index changes in subtropical China indicated from branched glycerol dialkyl glycerol tetraethers proxies, Chem. Geol., 541, 119579, https://doi.org/10.1016/j.chemgeo.2020.119579, 2020b.

Wang, M., Zong, Y., Zheng, Z., Man, M., Hu, J., and Tian, L.: Utility of brGDGTs as temperature and precipitation proxies in subtropical China, Sci. Rep., 8, 194, https://doi.org/10.1038/s41598-017-17964-0, 2018.

Wang, X., Zhang, H. H., and Wu, Y.: Multiclass Probability Estimation With Support Vector Machines, J. Comput. Graph. Stat., 28, 117947, https://doi.org/10.1080/10618600.2019.1585260, 2019.

Warden, L., Kim, J.-H., Zell, C., Vis, G.-J., de Stigter, H., Bonnin, J., and Sinninghe Damsté, J. S.: Examining the provenance of branched GDGTs in the Tagus River drainage basin and its outflow into the Atlantic Ocean over the Holocene to determine their usefulness for paleoclimate applications, Biogeosciences, 13, 5719–5738, https://doi.org/10.5194/bg-13-5719-2016, 2016.

Watson, B. I., Williams, J. W., Russell, J. M., Jackson, S. T., Shane, L., and Lowell, T. V.: Temperature variations in the southern Great Lakes during the last deglaciation: Comparison between pollen and GDGT proxies, Quaternary Sci. Rev., 182, 586–595, https://doi.org/10.1016/j.quascirev.2017.12.011, 2018.

Weber, Y., De Jonge, C., Rijpstra, W. I. C., Hopmans, E. C., Stadnitskaia, A., Schubert, C. J., Lehmann, M. F., Sinninghe Damsté, J. S., and Niemann, H.: Identification and carbon isotope composition of a novel branched GDGT isomer in lake sediments: Evidence for lacustrine branched GDGT production, Geochim. Cosmochim. Ac., 154, 118–129, https://doi.org/10.1016/j.gca.2015.01.032, 2015.

Weber, Y., Damsté, J. S. S., Zopfi, J., De Jonge, C., Gilli, A., Schubert, C. J., Lepori, F., Lehmann, M. F., and Niemann, H.: Redox-dependent niche differentiation provides evidence for multiple bacterial sources of glycerol tetraether lipids in lakes, P. Natl. Acad. Sci. USA, 115, 10926–10931, https://doi.org/10.1073/pnas.1805186115, 2018.

Weijers, J. W. H., Schouten, S., Hopmans, E. C., Geenevasen, J. A. J., David, O. R. P., Coleman, J. M., Pancost, R. D., and Sinninghe Damsté, J. S.: Membrane lipids of mesophilic anaerobic bacteria thriving in peats have typical archaeal traits, Environ. Microbiol., 8, 648–657, 2006.

Weijers, J. W. H., Schouten, S., van den Donker, J. C., Hopmans, E. C., and Damsté, J. S. S.: Environmental controls on bacterial tetraether membrane lipid distribution in soils, Geochim. Cosmochim. Ac., 71, 703–713, 2007.

Wickham, H.: ggplot2: Elegant Graphics for Data Analysis, Springer-Verlag New York, https://doi.org/10.32614/RJ-2017-023, 2016.

Wu, J., Yang, H., Pancost, R. D., Naafs, B. D. A., Qian, S., Dang, X., Sun, H., Pei, H., Wang, R., Zhao, S., and Xie, S.: Variations in dissolved O_2 in a Chinese lake drive changes in microbial communities and impact sedimentary GDGT distributions, Chem. Geol., 579, 120348, https://doi.org/10.1016/j.chemgeo.2021.120348, 2021.

Xiao, W., Wang, Y., Zhou, S., Hu, L., Yang, H., and Xu, Y.: Ubiquitous production of branched glycerol dialkyl glycerol tetraethers (brGDGTs) in global marine environments: a new source indicator for brGDGTs, Biogeosciences, 13, 5883–5894, https://doi.org/10.5194/bg-13-5883-2016, 2016.

Xiao, W., Xu, Y., Ding, S., Wang, Y., Zhang, X., Yang, H., Wang, G., and Hou, J.: Global calibration of a novel, branched GDGT-based soil pH proxy, Org. Geochem., 89–90, 56–60, https://doi.org/10.1016/j.orggeochem.2015.10.005, 2015.

Yang, H., Lü, X., Ding, W., Lei, Y., Dang, X., and Xie, S.: The 6-methyl branched tetraethers significantly affect the performance of the methylation index (MBT') in soils from an altitudinal transect at Mount Shennongjia, Org. Geochem., 82, 42–53, https://doi.org/10.1016/j.orggeochem.2015.02.003, 2015.

Yang, H.: Data for: Rapid response of fossil tetraether lipids in lake sediments to seasonal environmental variables in a shallow lake in central China: Implications for the use of tetraether-based proxies (1), Mendeley data [data set], https://doi.org/10.17632/bv8wd6yh2f.1, 2020.

Yao, Y., Zhao, J., Vachula, R. S., Werne, J. P., Wu, J., Song, X., and Huang, Y.: Correlation between the ratio of 5-methyl hexamethylated to pentamethylated branched GDGTs (HP5) and water depth reflects redox variations in stratified lakes, Org. Geochem., 147, 104076, https://doi.org/10.1016/j.orggeochem.2020.104076, 2020.

Yu, X., Ascencio, J., and French, R.: Open-Source Climate Classification Package: kgcPy, in: 2024 IEEE 52nd Photovoltaic Specialist Conference (PVSC), 1085–1085, 9–14 June 2024, Settle, WA, https://doi.org/10.1109/PVSC57443.2024.10749138, 2024.

Zink, K.-G., Vandergoes, M. J., Mangelsdorf, K., Dieffenbacher-Krall, A. C., and Schwark, L.: Application of bacterial glycerol dialkyl glycerol tetraethers (GDGTs) to develop modern and past temperature estimates from New Zealand lakes, Org. Geochem., 41, 1060–1066, https://doi.org/10.1016/j.orggeochem.2010.03.004, 2010.

Remarks from the typesetter

TS1 This change is due to an accidental typo that reversed the numbers (2031 instead of 2301) when writing the abstract. This does not change any of the data that was used in the article and no data was modified in the database. This has been approved by all authors. The correct number (2301) is currently in the introduction and conclusion and has not been changed.

TS2 This is a mistake that happened prior to submission. We added 20 soil samples into the database to bring the total database number to: 2301 but this number was not updated and was not caught in peer-review. However, all the datasets, and data in the article is based the full dataset (2301). This update just brings the number consistent throughout the article. The final database number of 2301 is in the conclusion and introduction.

TS3 Prior to submitting the article we added an additional 20 samples to bring the soil sample count up to 1197 and the total db count up to 2301. Unfortunately, we did not catch that this section needed to be updated prior to submission and through peer-review. The correct number is on the methods section and the data in figure 2 is correct, and the total number of the db is correct in the conclusion and introduction. The data and information used in the article is based on the db number of 2301 and this change just reflects having an accurate number throughout the article.