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[5], Lilit Sahakyan[9], 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, F-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, 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, Spain

8 Organismal and Evolutionary Biology Research Program, Research Centre for Ecological Change, University of Helsinki, Finland

9 Institute of Geological Sciences, National Academy of Sciences of Republic of Armenia, Yerevan, Armenia

Correspondence to: Amy Cromartie (aec277@cornell.edu)

Abstract. Branched glycerol dialkyl glycerol tetraethers (brGDGTs) ~~serve as~~are critical molecular biomarkers for the quantitative reconstruction of past environments, ambient temperature and pH across various archives. ~~Despite their success,~~ N~~n~~umerous issues, however, 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 the MBT′~~’~~$_{5ME}$. This issue is especially significant in shallow lakes, wetlands, and peatlands~~ within semi-arid and arid regions~~, 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~~sources~~; however, these techniques are insufficient for detecting mixed ~~source~~ 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) was compiled to train, validate, test, and apply these algorithms to two sedimentary records. Our findings are corroborated by pollen, ~~and~~ non-pollen palynomorphs, and XRF obtained from the

same sedimentary core sequence~~from the identical records~~. These microfossil and geochemical proxies are utilized to discuss changes in provenance, hydrology, and ecology that influence ~~the distribution of~~ brGDGT~~s~~ provenance. Probability estimates derived from Random Forest with a sigmoid calibration are most effective in detecting changes in brGDGT ~~distribution~~provenance. Minor changes in the relative contributions of brGDGT~~s~~ provenance can significantly influence the distribution of brGDGT~~s~~, especially regarding the MBT'$_{5ME}$ index.~~ This study introduces a novel brGDGT wetland index aimed at monitoring potential biases arising from wetland development.~~

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 paleo-environmental reconstructions. The ubiquity of brGDGTs and their global correlations with temperature 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., 2012; Raberg et al., 2022). Researchers have identified brGDGTs across various depositional environments, such as peat, soils, loess, 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 continental paleoclimate.

~~Despite its potential, ~~A~~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., 2014 ; Naafs et al., 2017b; Dearing Crampton-Flood et al., 2020; Martínez-Sosa et al., 2020; Raberg et al., 2022b). In the context of lacustrine, wetland, and peat archives, the fractional abundance of brGDGTs produced in the 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, Loomis et al., 2012, Loomis et al., 2014a, Loomis et al., 2014b, 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 based on the widely recognized MBT′~~’~~$_{5ME}$ index. 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, concerning 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 (~~-~~Pablo Martínez-Sosa et al., 2021).

Furthermore, brGDGTs distributions within 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). ~~The potential differential distributions resulting from depositional environments underscore the influence of changes in provenance or hydrological conditions on brGDGT-based environmental reconstructions.~~

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, 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 ~~source~~ changes on the 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 the 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 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 their ability to handle data with multiple variables for a few subjects while examining non-linear relationships within the datasets (Bzdok et al., 2018). Martínez-Sosa et al. (2023) models proved effective at identifying shifts in provenance; 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 lacustrine, peat, and soil depositional environments, including mixed contexts, utilizing a new global brGDGT database, machine learning techniques, as well as environmental reconstructions based on pollen, non-pollen palynomorphs, and XRF datasets.

Two approaches are employed to achieve this objective. First, we use probability estimates derived from machine learning to identify changes in ~~sourcing~~ provenance over time extending on the~~. This study extends the~~ work of Martínez-Sosa et al., (2023). ~~), who employed supervised machine learning to identify changes in brGDGT sources using classification~~

~~models based on modern samples~~. Rather than employing discrete classification, as they did, we utilize the probability estimates from these classification algorithms to analyze the contributions from differential ~~sources~~ 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 machine learning 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 (CART), ensemble trees (RF), linear models (LR), margin-based classifiers (SVM), and instance-based lazy learners (K-NN) to evaluate performance. The best-performing model was then chosen to apply to two down-core sedimentary ~~We implement five algorithms: K-Nearest Neighbor, Support Vector Machines (SVM), Logistic Regression (LR), Classification and Regression Trees (CART), and Random Forest (RF). These~~sequences. These are employed using Python and scikit-learn to identify intervals where downcore records are predominantly influences by *in-situ* lake brGDGTs, mineral soils, and peatlands, as well as combinations of these elements.

Secondly, to ensure accurate identification of provenance changes, comparisons are conducted with published pollen, ~~and~~ non-pollen palynomorphs (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 has 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 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 machine learning results through the integration of brGDGT-based reconstructions with pollen, ~~and~~ NPPs, and XRF; ii.) demonstrate how these complementary proxies can aid in identifying potential hydrological, ecological, and depositional ~~shifts~~changes that may cause provenance shifts, thus introducing~~e~~ bias in brGDGT reconstructions. This study demonstrates that alterations in provenance and hydrology can significantly influence the distribution of brGDGTs and, consequently, established 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 lake (n=591), soil (n = 1197), and peat (n=532) depositional categories (Baxter et al. 2019; Cao et al. 2020; Chen et al., 2021; Dearing Crampton-Flood et al., 2020; Dang et al., 2018; De Jonge et al., 2014b; Ding et al., 2015 ; Dugerdil et al., 2021 ; Guo et al. 2020 ; Halffman et al. 2022 ; Jaeschke et al. 2018 ; Kirkels et al. 2020 ; Kou et al., 2022 ; Li et al., 2017 ; Li et al., 2018 ; Martin et al., 2019; 2020 ; Martínez-Sosa et al., 2021 ; Naafs et al., 2017 ; Ning et al., 2019 ; Pérez-Angel et al., 2020 ; Qian et al., 2019 ; Raberg et al., 2021 ; Robles et al., 2021 ; Rao et al., 2022 ; Russell et al., 2018 ; Stefanescu et al., 2021 ; Véquaud et al., 2021a ; 2021b ; Wang et al., 2016 ; 2018 ; 2020a ; 2020b ; 2021 ; Weber et al. 2018 ; Wu et al. 2020 ; Xiao et al., 2015 ; Yang et al., 2015 ; Yao et al., 2020 ; Fig. 1, full data on https://github.com/amycromartie/ProbbrGDGT). 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 fifteen distinct brGDGT structural isomers were sourced from the original authors or recalculated from the initial datasets (https://github.com/amycromartie/ProbbrGDGT). 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 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 ⅚ isomer, also referred to as IIIa'' (Weber et al., 2015) were excluded due to limited -data~~publication~~. The original author's description was utilized to categorize the samples into a classification index (i.e., soil, lake, peat). Suspended particulate matter (SPM), moss polsters, 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.

2.1.2. Addition of new samples from Armenia

Thirty 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 from Lake Sevan at a depth of 2-3 cm, as previously discussed in Robles et al., (2022), along with 20 surface soil samples. For brGDGT extraction, each sample was first lyophilized, freeze-dried, and ground. Lipids were extracted from 0.5 to 1g 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 ions monitoring (SIM) of m/z of 1050, 1048, 1046, 1036, 1032, 1022, 1020, 1018, and 744 was used for the brGDGTs isomers and the internal C_{46} standard (Hopmans et al., 2016; Davtian et al., 2021; 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 under-representation of lake and peat samples (Fig. 1). Unbalanced datasets can result in considerable performance issues, such as 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 machine learning model using both the raw and resampled datasets, which ~~incorportaed~~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 dataset. 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 supplement Figure S3 & S4 and Table S1). 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 peat and lake samples, 219 and 159 synthetic samples were incorporated into the raw dataset, respectively.

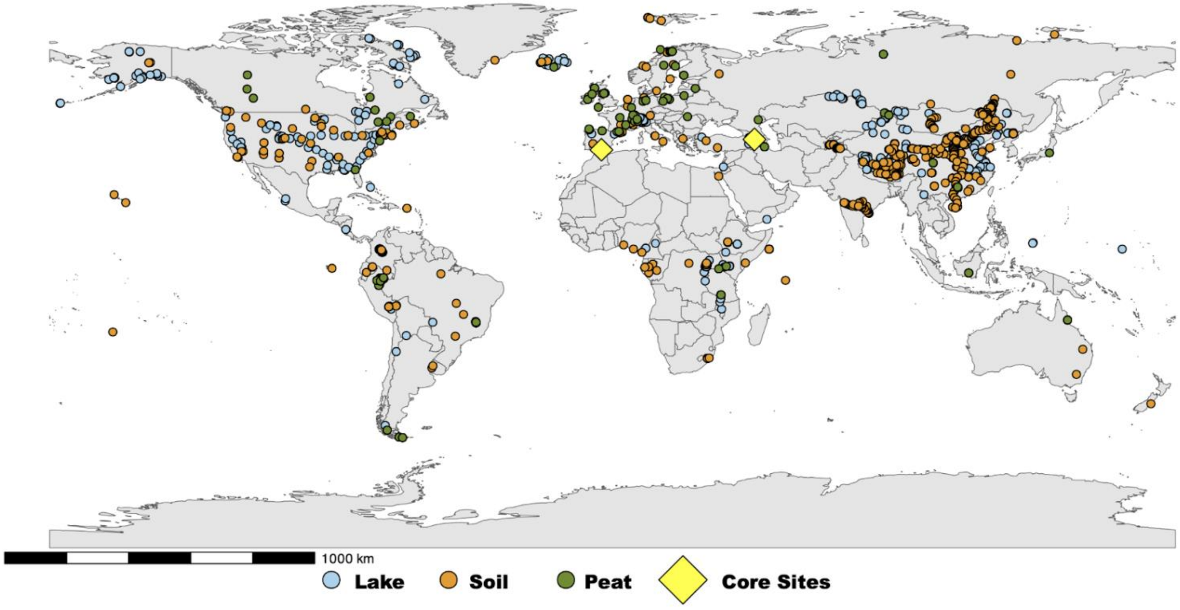


Figure 1: Map of modern sample locations used in the compiled database alongside the two sites designated for paleo-reconstructions. Map created with R package ggplot2 (Wickham, 2016)

2.2 Machine Learning models

2.2.1 Building probability and classification machines

In supervised classification problems, machine learning 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 (Geetha and Sendhilkumar, 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 machine learning 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 indicating the degree of likeness of an individual belongs to a category (Malley et al., 2012). Most machine learning 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~~ibid~~). 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. Firstly, 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. 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.~~We employed Python and scikit-learn (Pedregosa et al., 2011) for the machine learning analysis. Five commonly used supervised machine learning models were evaluated: k-nearest neighbors (KNN), support vector machine (SVM), logistic regression (LR), classification and regression trees (CART), and random forests (RF). The selection of these models is based on their capacity to produce calibrated probability estimates.~~

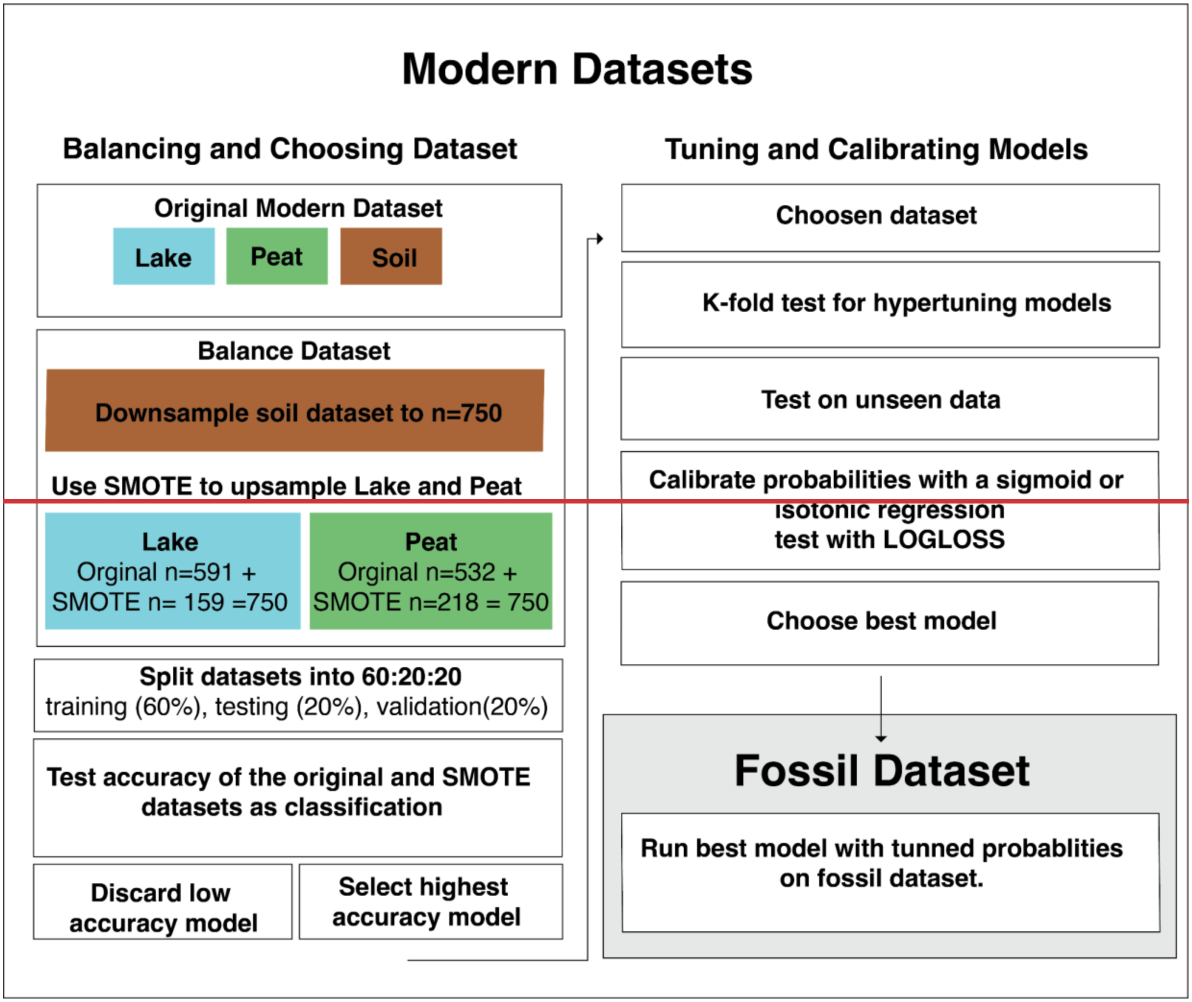
Logistic regression (LR) is a parametric model akin to linear regression in its functionality, yet it is more appropriate for classification tasks (i.e., binary outcomes) (Hilbe, 2016). The analysis relies on the likelihood of an event occurring and the alignment of predictor response variables within the probability distribution (Hilbe, 2016). The algorithm is inherently calibrated for precise probability outputs due its foundational reliance on probability.

K-nearest neighbors (KNN), support vector machine (SVM), classification and regression trees (CART), and random forests (RF) are non-parametric models demonstrated to be effective in probability estimation following calibration (i.e., Niculescu-Mizil and Caruana, 2005; Kruppa et al., 2013; Dankowski and Ziegler, 2016, Cearns et al., 2020). K-nearest neighbor (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~~ibid~~). KNN is extensively employed in palaeosciences for paleoclimate regression issues, particularly through the modern analog technique (Simpson, 2007). Supported Vector Machines (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~~ibid~~).

Classification and Regression Trees (CART) and Random Forests (RF) are tree-based learning algorithms. Trees are formed through three fundamental steps: (1) binary splits are selected, (2) a determination is made regarding the node is terminal or requires further splitting; and (3) a class is assigned to the terminal leaf node (Bell, 1999). Random Forest (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, 2009). In 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 was 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 ~~The raw and SMOTE datasets were divided into training, testing and validation sets in a 60:20:20 ratio (Fig. 2). The~~training. The models underwent testing and hyperparameter tuning using a k-fold cross-validation approach, incorporating ten data splits and a parameter grid with the test dataset. K-fold cross-validation involves partitioning the data into equal-sized subsets, which are then utilized k times, with k - 1 subsets used for training and one subset reserved for validation (Jung, 2017). 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 F1 accuracy results were computed (Fig. 3).


Modern Datasets
Balancing and Choosing Dataset
Original Modern Dataset
Lake
Peat
Soil
Balance Dataset
Downsample soil dataset to n=750
Use SMOTE to upsample Lake and Peat
Lake
Orginal n=591 +
SMOTE n= 159 =750
Peat
Orginal n=532 +
SMOTE n=218 = 750
Split datasets into 60:20:20
training (60%), testing (20%), validation(20%)
Test accuracy of the original and SMOTE datasets as classification
Discard low accuracy model
Select highest accuracy model
Tuning and Calibrating Models
Choosen dataset
K-fold test for hypertuning models
Test on unseen data
Calibrate probabilities with a sigmoid or isotonic regression
test with LOGLOSS
Choose best model
Fossil Dataset
Run best model with tunned probablities on fossil dataset.

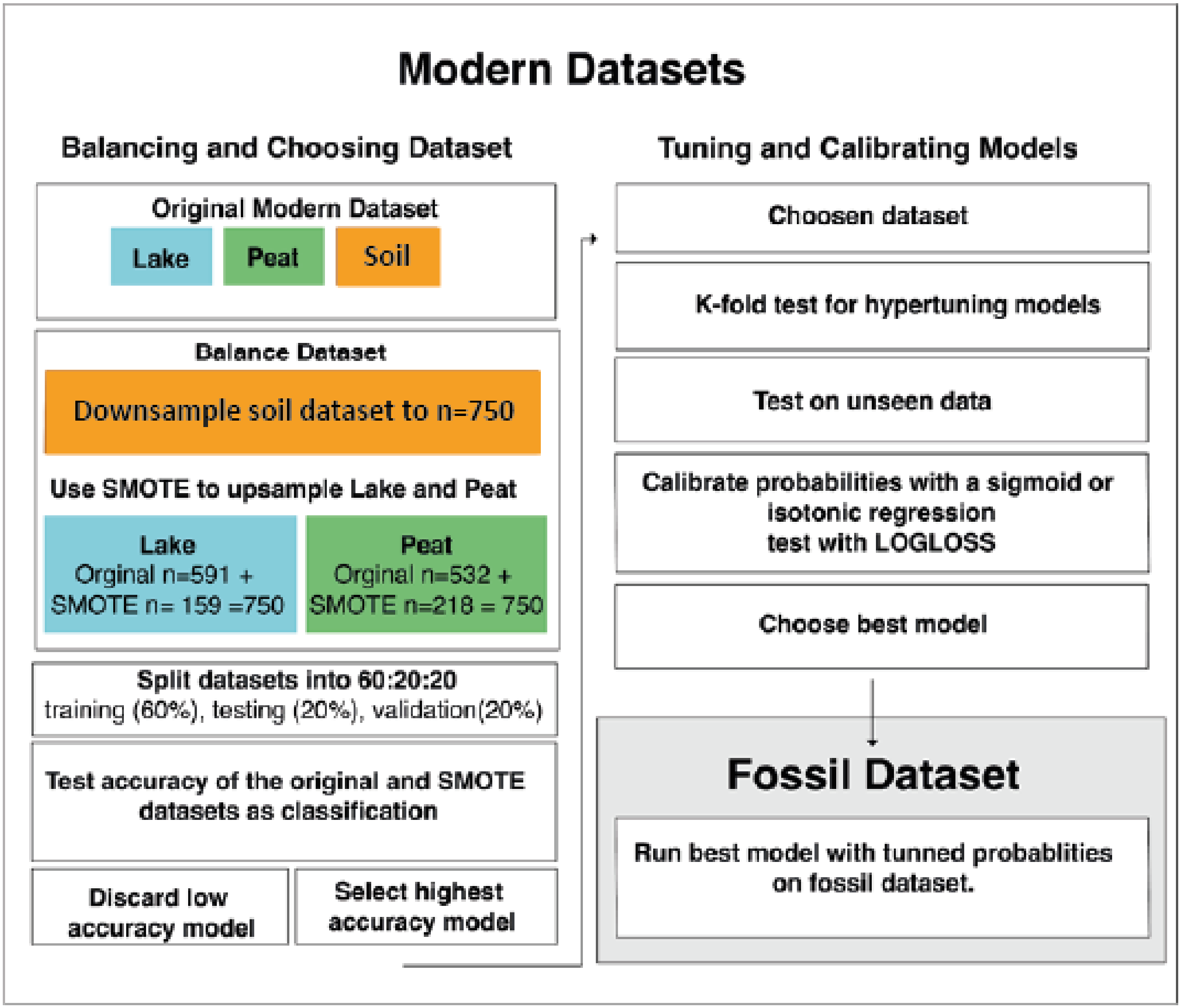

Modern Datasets
Balancing and Choosing Dataset
Tuning and Calibrating Models
Original Modern Dataset
Lake
Peat
Soil
Balance Dataset
Downsample soil dataset to n=750
Use SMOTE to upsample Lake and Peat
Lake
Orginal n=591 +
SMOTE n= 159 =750
Peat
Orginal n=532 +
SMOTE n=218 = 750
Split datasets into 60:20:20
training (60%), testing (20%), validation(20%)
Test accuracy of the original and SMOTE datasets as classification
Discard low accuracy model
Select highest accuracy model
Choosen dataset
K-fold test for hypertuning models
Test on unseen data
Calibrate probabilities with a sigmoid or isotonic regression
test with LOGLOSS
Choose best model
Fossil Dataset
Run best model with tunned probablities on fossil dataset.

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

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 ~~environmental~~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 ~~sourcing~~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 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, 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 brGDGTs 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 re-calculated 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, fungal and fern spores, generally are representative of local change particularly in wetlands (Gill et al., 2014; Tunno and Mensing 2017) which strengthens their usage as local indicators of change. Percentages of the aquatic and NPP taxa were calculated by summing 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 machine learning results with the brGDGT record as well as the aquatic pollen and NPPs. In addition, we also compared our results with the principal components 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.

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 1: Table of brGDGT indices employed in this study.

2.4 Descriptive statistics

The programming languages R (R core team) and Python (Python Software Foundation. Python Language Reference, version 3.7.3. available at http://www.python.org) 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 author's descriptive categories (i.e., soil, peat, lake), and ii.) To compare the probability estimates 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 brGDGTs based ML lake probability results using the bcp package (Erdman and Emerson, 2007;2008; Wang and Emerson, 2015) 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 samples (591 from lakes, 532 from peats, and 1177 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. (2019) 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 in our dataset. Fig. 1 and Fig. S1 illustrate the distribution of the brGDGT datasets.

3.2. Model accuracy and Log loss

In classification mode, all models demonstrated mean accuracy F1 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, KNN, but decreased for SVM and LR. The isotonic calibration improved probabilities for KNN, 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~~. The RF model with the SMOTE dataset and sigmoid calibration~~ and therefore was chosen for our downcore analysis as the best performing model.

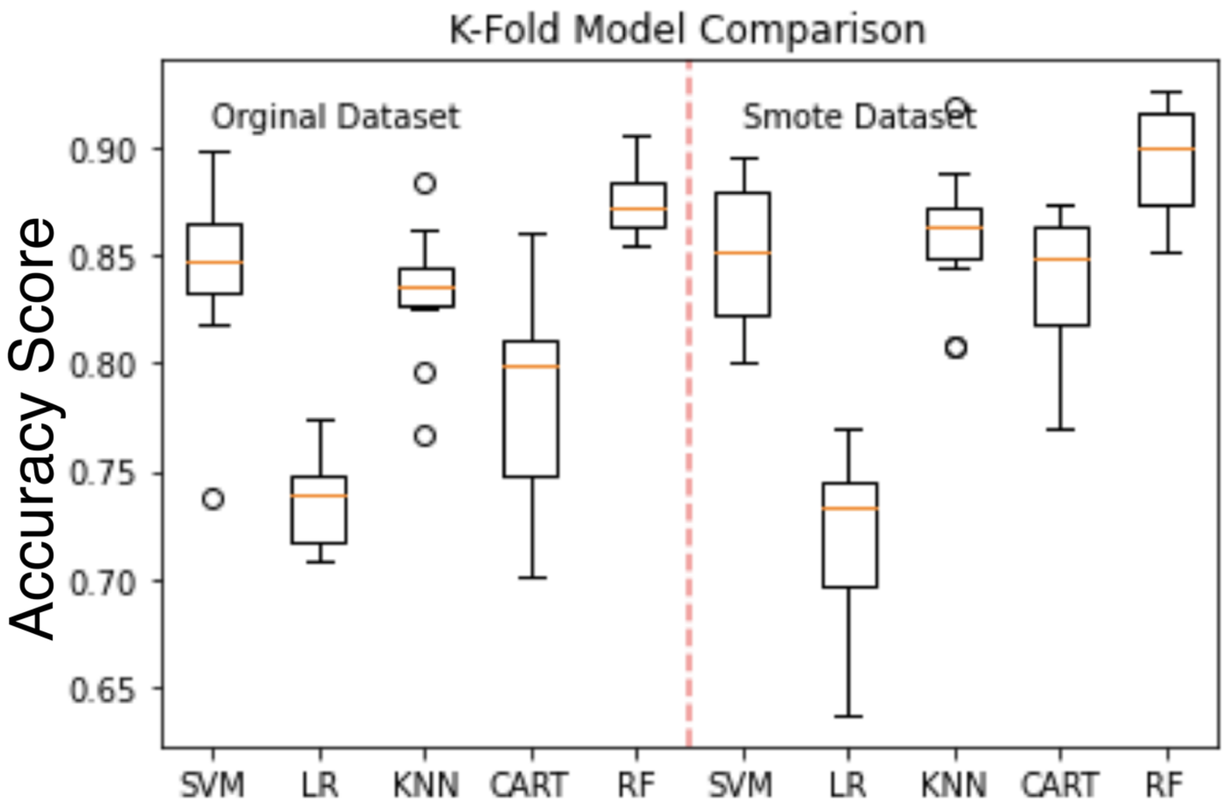


Figure 3: Comparison of k-fold testing models between datasets, utilizing k-fold cross-validation for classification on both SMOTE and original datasets. The k-fold comparison utilized 10 splits across Supported Vector Machines (SVM), Logistic Regression (LR), K-Nearest Neighbor (KNN), Classification and Regression Tree (CART), and Random Forest (RF).

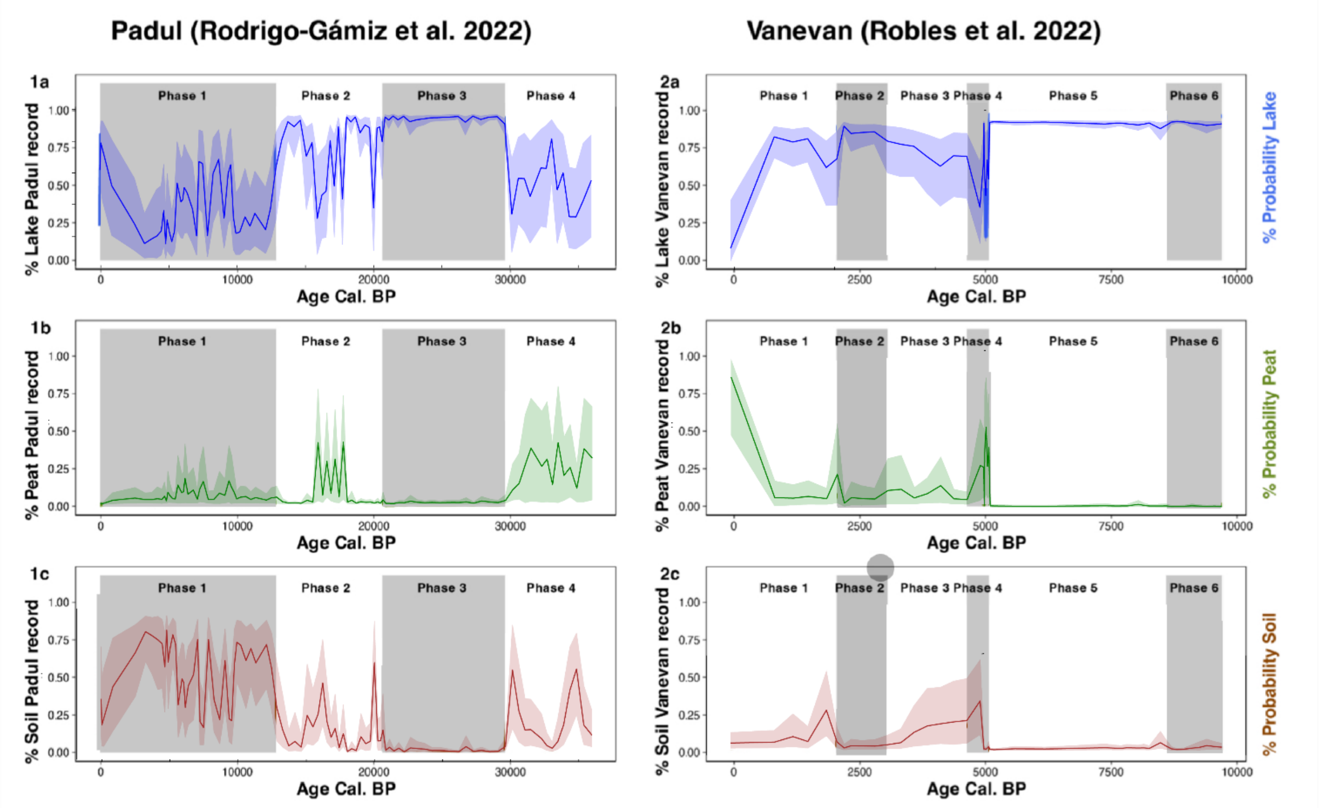
Model and Database	F1 Mean Accuracy	Standard Deviation	LogLoss uncalibrated	LogLoss Sigmoid	LogLoss Isotonic
SVM	0.84	0.04	0.46	0.51	0.93
SVM_SMOTE	0.85	0.03	0.40	0.50	0.64
LR	0.74	0.02	0.66	0.68	0.75
LR_SMOTE	0.72	0.04	0.69	0.70	0.73

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

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 isotonic function. For these functions, values approaching 0 signify improved performance.

3.3 Downcore analysis

3.3.1 Downcore probability estimates and changepoint analysis


Padul (Rodrigo-Gámiz et al. 2022)
Vanevan (Robles et al. 2022)
1a
Phase 1
Phase 2
Phase 3
Phase 4
% Lake Padul record
Age Cal. BP
2a
Phase 1
Phase 2
Phase 3
Phase 4
Phase 5
Phase 6
% Lake Vanevan record
% Probability Lake
1b
% Peat Padul record
2b
% Peat Vanevan record
% Probability Peat
1c
% Soil Padul record
2c
% Soil Vanevan record
% Probability Soil
0.00
0.25
0.50
0.75
1.00
0
10000
20000
30000
2500
5000
7500
10000

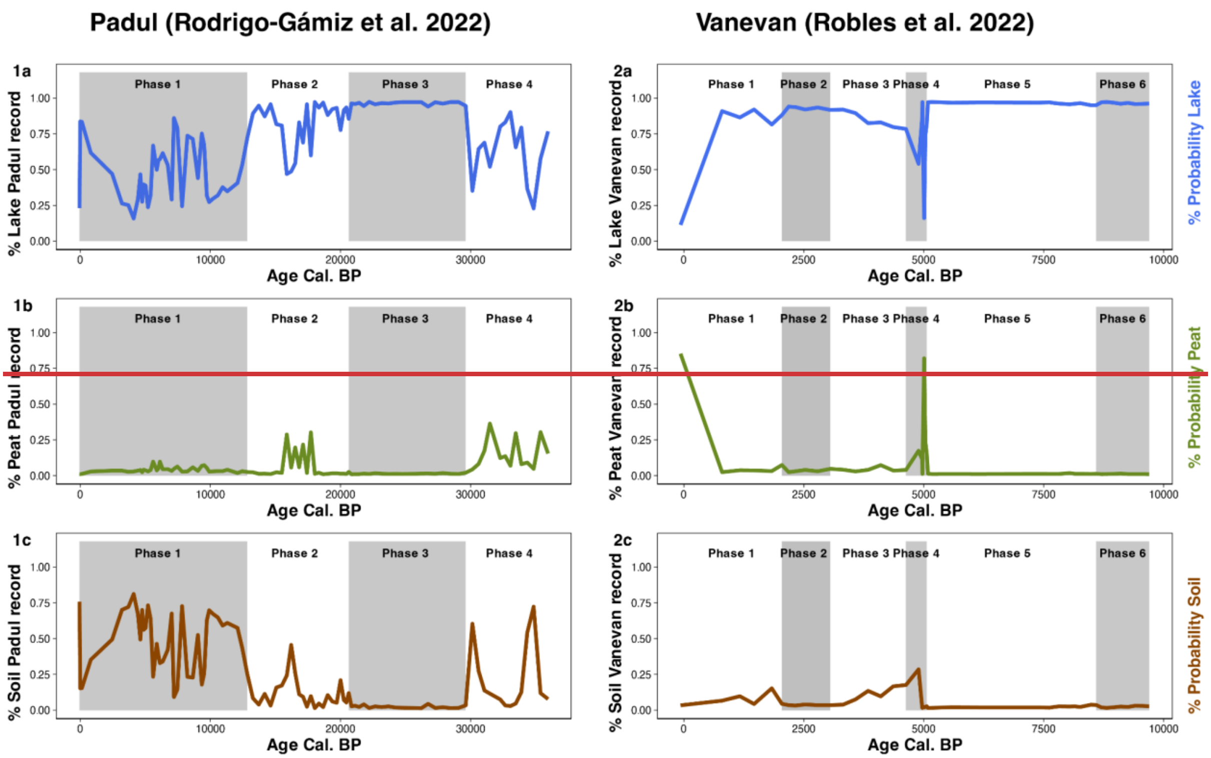


Figure 4: Downcore probability estimates with 95% confidence intervals ~~with~~ and changepoint breaks from Random Forests (RF) on the SMOTE dataset with a sigmoid calibration. Results from the Padul (1) and Vanevan (2) records are broken down by lake probabilities (blue curves - a), peat probabilities (green curves -b), and soil probabilities (brown curves - c). Highlighted grey and white boxes indicate changepoint mean breaks identifying phases. Probability estimates from other models can be found in Supplement 1. (Fig. S9~~5~~ – S11~~8~~)

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 .87 for lake, .05 for soil, and .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 changepoint phases

For the Padul record, changepoints were detected at 12837 cal. BP, 20627 cal. BP, and 29617 cal. BP for lake probabilities dividing it into phases 1 - 4 (Fig. 4, column 1) and changepoints in the Vanevan record were identified at 2043 cal. BP, 3577 cal. BP, 4628 cal. BP, 5061 cal. BP, and 8592 cal. BP based lake probabilities (Fig. 4, column 2). Changepoint analysis of the Padul record indicates that brGDGTs based ML lake probabilities peak in phase 3, while in phases 1, 2, and 4, these probabilities vary between soil and peat. BrGDGTs-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 on modern and downcore pollen, NPP, and brGDGTs

3.4.1 Modern samples

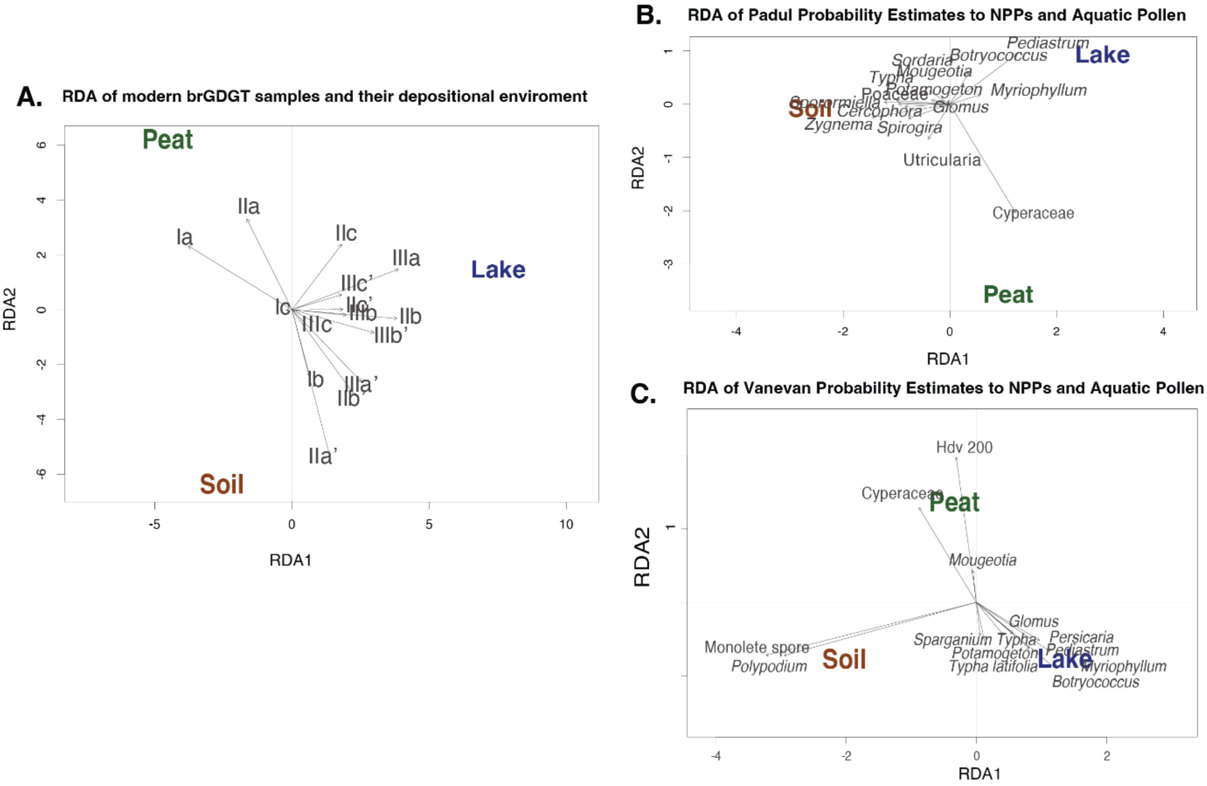


Figure 5: (A) RDA analysis of the modern fractional abundances of the global brGDGT database with their depositional environments. (B) Probability estimates results of the depositional environment compared with pollen and NPPs of Padul record (i.e., Rodrigo-Gámiz et al., 2022; Camuera et al., 2019) and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

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 like *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 & 5c).

4 Discussion:

4.1 Probability estimates for chosen models and application to downcore records

4.1.1 Model accuracy

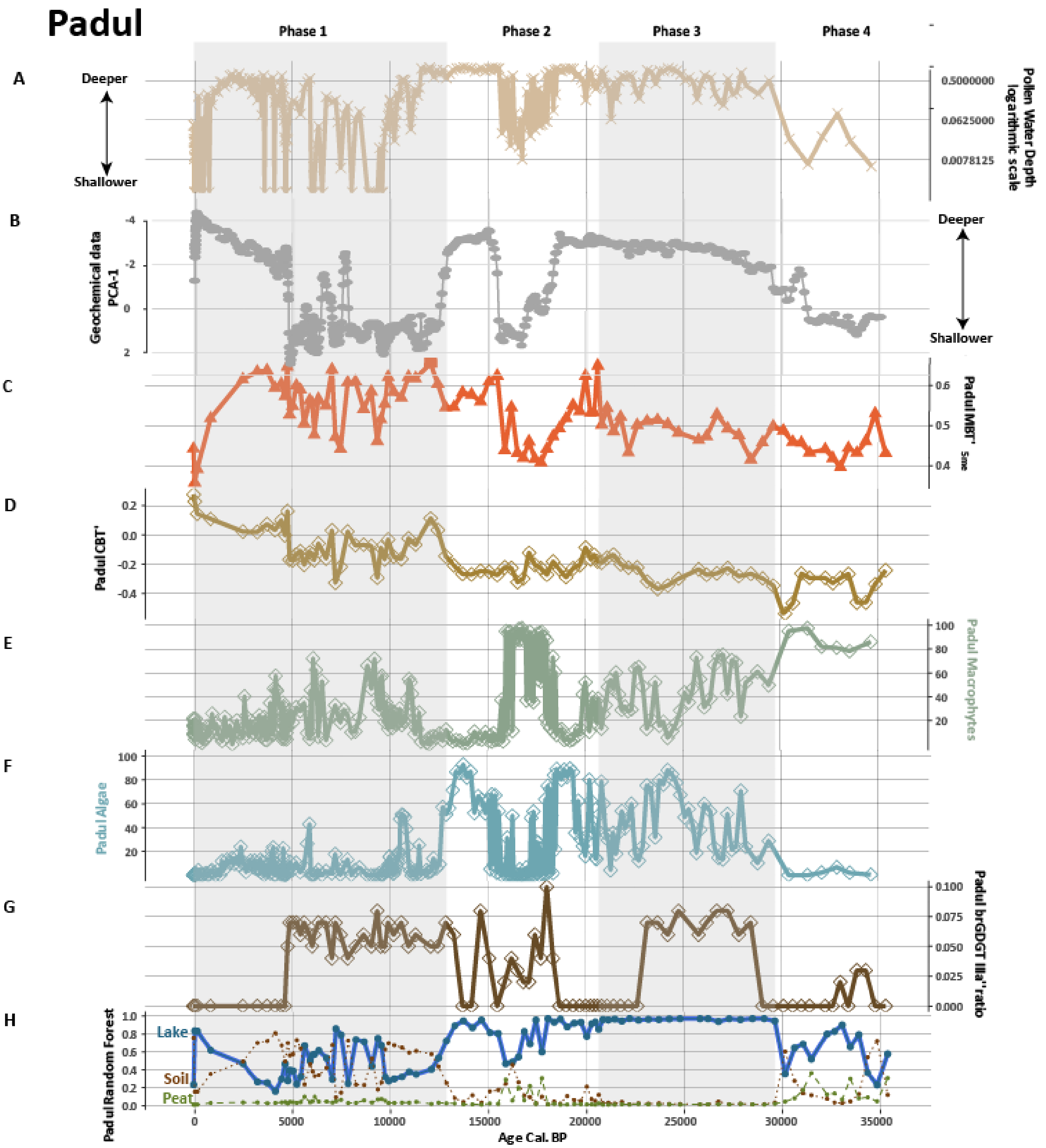
The F1 score evaluates the accuracy of a model's predictions of both precision (how many predicted positives were positive) and recall (from all the positives, how many positives did the model predict) and can balance between understanding false positives and false negatives (Boozary et al., 2025). This score allows for a more robust accuracy when measuring each model. Many things may explain differences in F1 scores across our models. For example, K-NN, SVM, and CART models are prone to overfitting (Huang et al., 2005; Berk, 2008 Jadjav and Channe, 2013), which may have accounted for their lower F1 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 F1 score. While LR does not typically overfit, the lower F1 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 F1 score, while CART's F1 score was significantly improved with the balanced SMOTE datasets. For the logloss scores, logistic regression is already calibrated (Kull et al., 2017a) so calibrating may result in a lower log loss score, with both a sigmoid and isotonic calibration. SVM does not produce true probabilities by default and needs to be calibrated for these results (Kull et al., 2017b). 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 K-NN, CART, and RF calibration improved the models on both datasets.

The Random Forest model utilizing the SMOTE dataset achieves an F1 score of 89% (Table 2) in classification, which is lower than the 95% F1 score reported for the BIGMaC model by Martínez-Sosa et al., (2023). The difference in F1 scores is anticipated as a result of differing training datasets and methodologies, including the incorporation of isoGDGTs by Martínez-Sosa et al., (2023), and their establishment 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 F1 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 on 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 (Supplement 1, Fig. S6~~3~~ and S7~~4~~). 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, ~~and~~ XRF, and water depth estimates derived from these proxies to check for quantitative similarities (Fig. 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 peat depositional environments (Fig. 5). Down core records indicate distinct associations between pollen from algae, specifically *Pediastrum* and *Botryococcus*, which are typically associated with open lakes, and the brGDGT based ML probability estimates associated with lake depositional 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 soil depositional environments. In the Padul record, these spores were likely introduced through human activity (Ramos-Roman 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 (Fig. 6 and 7).

The comparison of reconstructed water depth results with the probability estimates from the brGDGT-trained ML model for both cores indicates that the two proxies exhibit ~~similar~~ have similar quantitative similarly patterns of increase and decrease. Similarly, PCA data derived from XRF datasets from the same core, follow these trends. Robles et al., (2022) associates higher lake levels with the PCA-1 associated with positive loadings (P, K, Al, Mg, Si, Ti, Fe) and negative (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, ~~and~~ 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 changepoint 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).


Padul
Phase 1
Phase 2
Phase 3
Phase 4
A
Deeper
Shallower
0.5000000
0.0625000
0.0078125
Pollen Water Depth
logarithmic scale
B
Geochemical data
PCA-1
-4
-2
0
2
Deeper
Shallower
C
0.6
0.5
0.4
Padul MBT'5me
D
Padul CBT'
0.2
0.0
-0.2
-0.4
E
100
80
60
40
20
Padul Macrophytes
F
Padul Algae
100
80
60
40
20
G
0.100
0.075
0.050
0.025
0.000
Padul brGDGT IIIa''ratio
H
Padul Random Forest
1.0
0.8
0.6
0.4
0.2
0.0
Lake
Soil
Peat
0
5000
10000
15000
20000
25000
30000
35000
Age Cal. BP

435

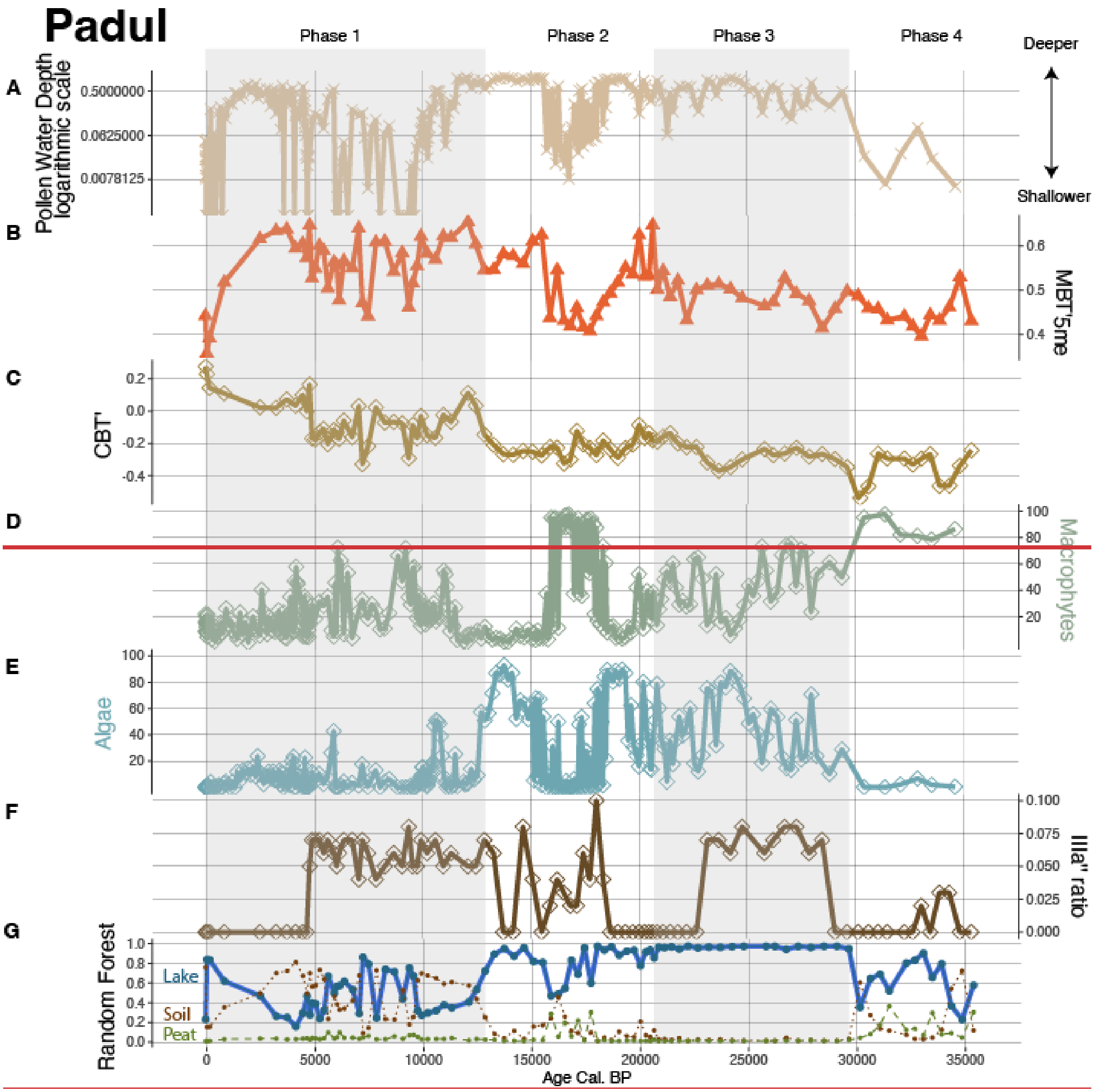


Figure 6: Comparison of probability estimates from the Padul record with aquatic pollen, ~~and~~ NPPs, XRF, ~~and~~ brGDGT indexes (data from Ramos-Román et al., 2018; Camuera et al., 2018;2019; Rodrigo-Gámiz et al., 2022). (A) Pollen and NPP based water-depth reconstructions (B) Output from Principle component analysis (PCA-1) from the XRF and geochemical data (Camuera et al., 2018) (C~~B~~) MBT'$_{5ME}$ brGDGT index (D~~C~~) CBT'~~'~~ brGDGT index (E~~D~~) Selected aquatic plants including Cyperaceae and *Typha*. (F~~E~~) Selected algae *Pediastrum*, *Botryococcus*, and *Mougeota* as well as aquatic plants

Cyperaceae and *Typha*. (G~~F~~) IIIa″~~’’~~ brGDGT ratio (H~~G~~) Probability estimates for the lake depositional environments (this study).

PCA analysis was also conducted on the XRF dataset on both the Padul core, and the authors used it as a proxy for lake-level change. For Padul Camuera et al., (2018) attributed negative loadings of PCA-1 with higher lake levels (Ca, Sr, Si, A, MS) and positive loadings with lower (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 for 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 ~~from ~~(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

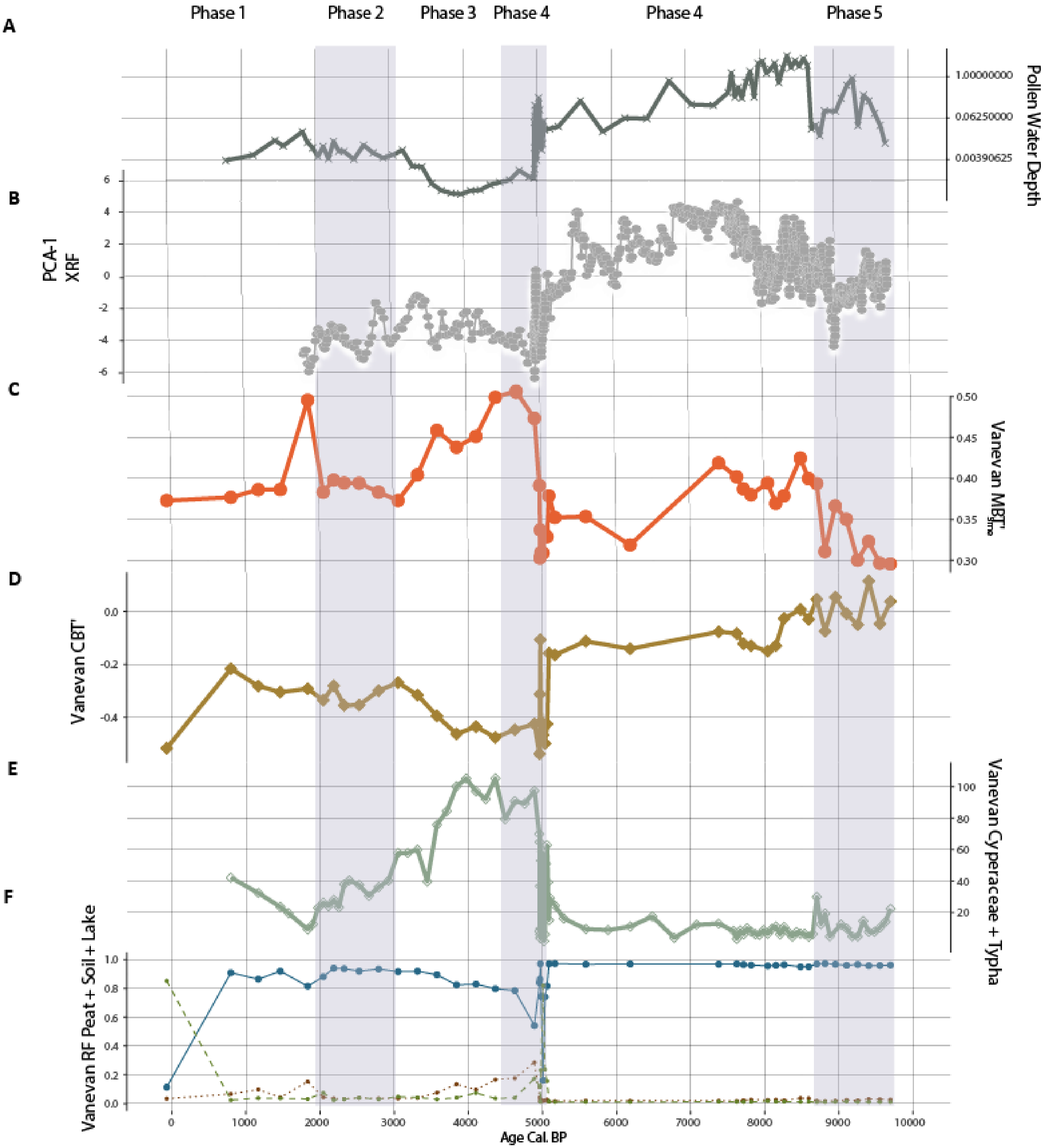
4.2.1 Identifying provenance changes in downcore records (and their impact on the MBT′~~’~~$_{5ME}$).

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 machine learning model from Martínez-Sosa et al., (2023), which classified most samples as a lake depositional 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 35000 – 30000 cal. BP and from 10000 cal. BP – to the present differences between models include samples classified as soil (our model) rather than peat or lake (Supplement 1, 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 soil or peat depositional environments.

Secondly, our findings indicate that the identification of *in-situ* produced lacustrine brGDGTs from lake depositional 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 (Figure 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; Ding et al., 2018). However, the lack of discussion regarding the absence of 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 peat and soil depositional 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 peat depositional environments compared to other tetramethylated brGDGTs linked to lake environments, highlighting the need to advance beyond ternary diagrams for provenance identification.

Vanevan


Phase 1
Phase 2
Phase 3
Phase 4
Phase 4
Phase 5
A
Pollen Water Depth
1.00000000
0.06250000
0.00390625
B
PCA-1 XRF
6
4
2
0
-2
-4
-6
C
Vanevan MBT'me
0.50
0.45
0.40
0.35
0.30
D
Vanevan CBT'
0.0
-0.2
-0.4
E
Vanevan Cyperaceae + Typha
100
80
60
40
20
F
Vanevan RF Peat + Soil + Lake
1.0
0.8
0.6
0.4
0.2
0.0
0
1000
2000
3000
4000
5000
6000
7000
8000
9000
10000
Age Cal. BP

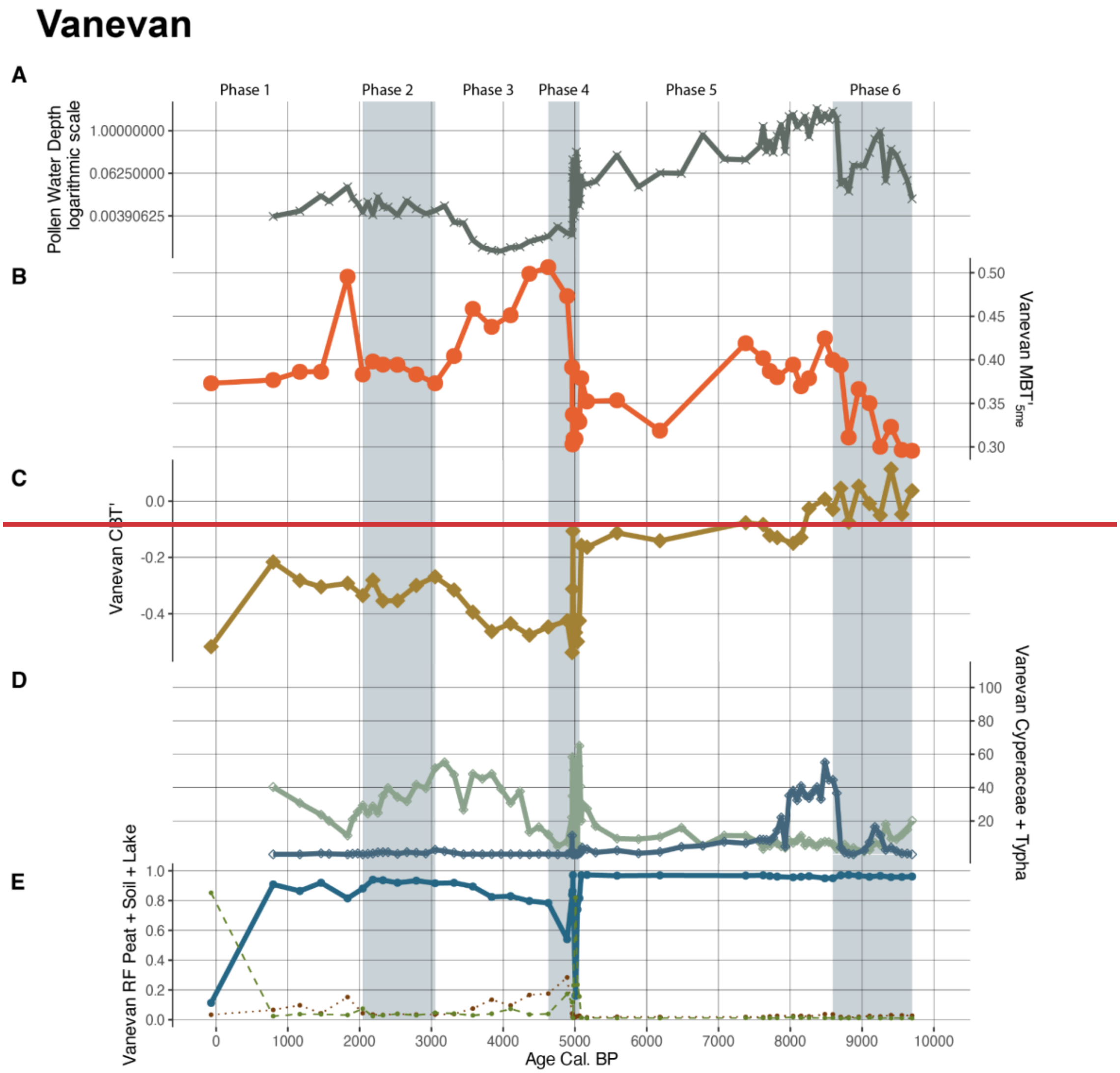


Figure 7: Comparison between the probability estimates on the Vanevan record and the aquatic pollen, ~~and~~ NPPs, XRF, and brGDGT indexes (data from Robles et al. 2022). (A) Pollen and NPP based water-depth reconstructions (B) MBT'$_{5ME}$ brGDGT index (C) CBT'~~'~~ brGDGT index (D) Select algae and aquatic plants for the Vanevan record algae is *Pediastrum*, *Botryococcus*, and *Mougeota* and aquatic plants are Cyperaceae and *Typha*. (E) Probability estimates for the lake depositional environments on both records.

The results of our models indicate that variations in brGDGT provenance, even within mixed sedimentary environments, significantly influence widely used indexes like the MBT'$_{5ME}$ (Fig. 6 and 7). Pollen and NPPs provide independent confirmation of the impact these changes have on the MBT'$_{5ME}$, particularly when analyzed alongside data from our new brGDGTs global database. In this database, soil (0.56) and peat (0.58) exhibit higher mean MBT'$_{5ME}$ values compared to the lower values observed in lakes (0.39) (Supplement 1, Fig. S5~~2~~). ~~However, it must be noted that analytical differences between laboratories (i.e., De Jonge et al., 2024) suggests the accuracy of these values may fluctuate, but the overall trends remain.~~ 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 the 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 ~~Camerua~~ 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 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 (Fig. 6 and 7). The Vanevan record indicates that the large shifts in the 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.

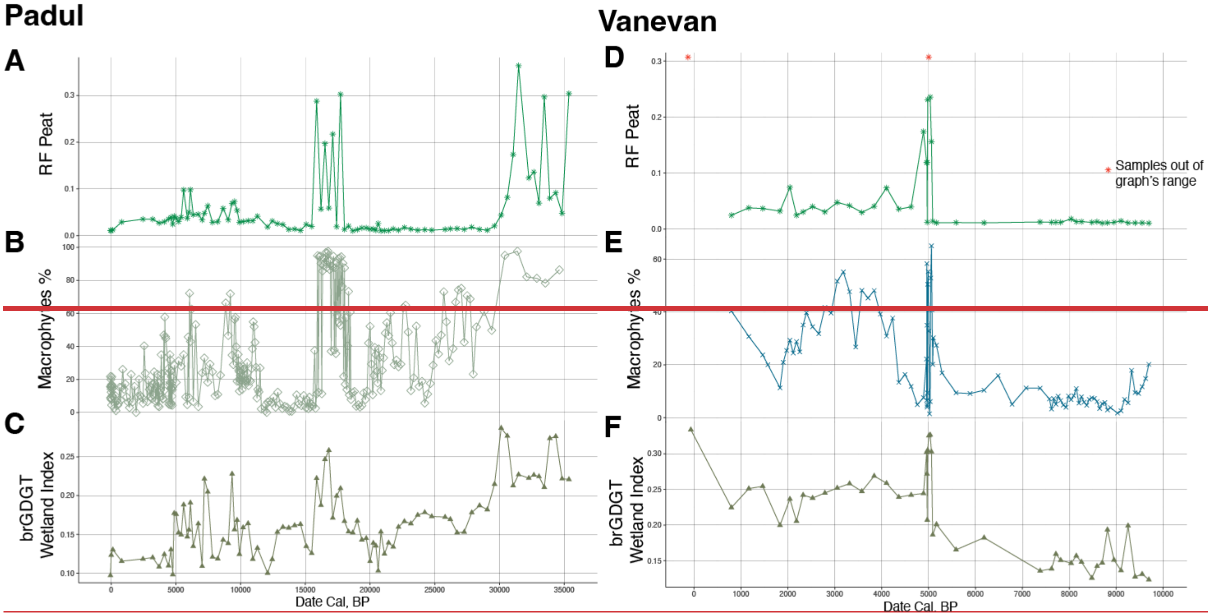
Rodrigo-Gámiz et al., (2022) identified a relationship in the Padul record between increases in reconstructed pH and MAAT variability within the upper 116 cm, approximately correlating to the last 5000 years. They associated this with a dried ephemeral lake and suggested potential bias in the MBT'$_{5ME}$ 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 on this variation (Fig. 7). Soil provenance changes appear to exert a greater influence on the 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 (Ofıti 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~~0.000003672~~). 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 the MBT′~~’~~$_{5ME}$ and the distribution of brGDGTs (Fig. 6 & 7).

~~4.2.4 Use of brGDGT indices to trace environmental changes~~

~~The shifts in the CBT’ which align with the increased soil probabilities in our brGDGT-based ML model suggests that the CBT’ can possibly be utilized as a screening tool to detect soil influences. In addition, a new wetland index was developed (Equation 1) to monitor wetland development over time, based on the strong correlation between Cyperaceae pollen and brGDGT [IIa], the association with peat depositional environments in the modern database, and higher median [IIa] values for peat in the global database distribution (Supplement 1 Fig. S1). BrGDGT [Ia] is linked to peatland depositional environments and exhibits a positive correlation with Cyperaceae in the Padul record; however, its inclusion in the equation resulted in a diminished correlation between Cyperaceae and the pollen-based water depth. Consequently, it has been excluded from the equation. In peatlands, incorporating the Ia into the equation may facilitate the monitoring of change.~~

~~Equation 1: Wetland Index: ([IIa]) / ([Ia] + [Ib] + [Ic] + [IIa] + [IIa’] +~~
~~[IIb] + [IIb’] + [IIc] + [IIc’] +~~
~~[IIIa] + [IIIb] + [IIIc] +~~
~~[IIIa’] + [IIIb’] + [IIIc’])~~



~~**Figure 8:** Comparison between the Random Forest peat probabilities (A and D), pollen from Cyperaceae and *Typha* (B and E), and the brGDGT wetland index (Equation 1, C and F) on the Padul and Vanevan records.~~

~~The wetland index results align closely with the pollen percentages of Cyperaceae (Padul: 0.52, p-value: 0.000003, Vanevan: 0.70 p-value: 0.0000001) and with *Typha* (Padul: 0.50, p-value: 0.000007, Vanevan 0.65, p-value) for both records (Fig. 8). In testing the index on the modern sample database, an adequate division between depositional environments is observed, with median values of 0.135, 0.167, and 0.337 for soil, lake, and peat, respectively (Supplement 1, Fig. S2). Furthermore, while our emphasis was on the frequently employed $\mathrm{MBT'}_{5ME}$, the correlation between shifts and individual brGDGTs underscores the potential impact these modifications may have on any brGDGT index.~~

4.2.4 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. 8), 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); temperate environments are well represented, but tropical, arid, and arctic conditions are under-represented. 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 logloss score of .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 an 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 a shallow versus deep lakes, remain inadequately represented. To effectively utilize these tools, it is essential to collect additional information, including water depth, salinity, pH, redox conditions, and others, 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 are presently employing pollen-based climate reconstructions in conjunction with brGDGT reconstructions (e.g., Watson et al., 2018; Martin et al., 2019; Dugerdil et al., 2021b; Robles et al., 2022:2023; Stefanescu et al., 2021). The findings indicate that aquatic pollen, ~~and~~ 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., 2021b; 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 improved classification of wetland environments within the brGDGT literature.

5 Conclusion

This study demonstrated that multivariate methods enhance the understanding of how provenance changes affect brGDGT distributions and the 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 lake, soil, and peat depositional 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 the MBT'$_{5ME}$. ~~We developed a wetland index, utilizing the fractional abundance of IIa, to analyze shifts between these systems.~~

This study confirms that independent proxies, including aquatic pollen, and 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.

Funding: 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 a SNSF PRIMA grant number 179783.

Acknowledgements: ChatGPT was utilized to help streamline and improve performance for both the R and Python code utilized for this project.

Competing interests: The authors declare no competing interests

Code availability: Code and data for this project will be publicly available on https://github.com/amycromartie/ProbbrGDGT

Data availability: Additionally, database will be uploaded to Pangea

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

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, Organic Geochemistry, 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 Bioinformatics, 14, 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, Organic Geochemistry, 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 Science Reviews, 305, 108014, 2023.

Baxter, A. J., Verschuren, D., Peterse, F., Miralles, D. G., Martin-Jones, C. M., Maitituerdi, A., der Meeren, T., Van Daele, M., Lane, C. S., Haug, G. H., and others: Reversed Holocene temperature--moisture relationship in the Horn of Africa, Nature, 620, 336–343, 2023.

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, 2008.

Blaga, C. I., Reichart, G. J., Schouten, S., Lotter, A. F., Werne, J. P., Kosten, S., Mazzeo, N., Lacerot, G., and Sinninghe Damsté, J. S.: Branched glycerol dialkyl glycerol tetraethers in lake sediments: Can they be used as temperature and pH proxies?, Organic Geochemistry, 41, https://doi.org/10.1016/j.orggeochem.2010.07.002, 2010.

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, International Journal of Information Management Data Insights, 5, 100331, https://doi.org/10.1016/j.jjimei.2025.100331, 2025.

van Bree, L. G. J., Peterse, F., Baxter, A. J., De Crop, W., Van Grinsven, S., Villanueva, L., Verschuren, D., and Sinninghe Damsté, J. S.: Seasonal variability and sources of in situ brGDGT production in a permanently stratified African crater lake, Biogeosciences Discussions, 2020, 1–36, 2020.

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., Krzywinski, M., and Altman, N.: Machine learning: a primer, Nature methods, 14, 1119, 2017.

Bzdok, D., Altman, N., and Krzywinski, M.: Statistics versus machine learning, Nature Methods, 15, 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 ~200,000-year-long multiproxy sedimentary record from Padul, southern Iberian Peninsula, Quaternary Science Reviews, 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 Science Reviews, 205, https://doi.org/10.1016/j.quascirev.2018.12.013, 2019.

Cearns, M., Hahn, T., Clark, S., and Baune, B. T.: Machine learning probability calibration for high-risk clinical decision-making, Australian \& New Zealand Journal of Psychiatry, 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, Chemical Geology, 559, 119908, https://doi.org/10.1016/j.chemgeo.2020.119908, 2021.

Crampton-Flood, E. D., 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, 2020.

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 history of a mountain steppe: A Holocene reconstruction from the South Caucasus, Shenkani, Armenia, Quaternary Science Reviews, 246, 106485, https://doi.org//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, Geochimica et Cosmochimica Acta, 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, Organic Geochemistry, 119, https://doi.org/10.1016/j.orggeochem.2018.02.008, 2018.

Dankowski, T. and Ziegler, A.: Calibrating random forests for probability estimation, Statistics in medicine, 35, 3949–3960, 2016a.

Dankowski, T. and Ziegler, A.: Calibrating random forests for probability estimation, Statistics in medicine, 35, 3949–3960, 2016b.

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

Dawid, A. P.: Rejoinder: Calibration-Based Empirical Probability, The Annals of Statistics, 13, https://doi.org/10.1214/aos/1176349738, 2007.

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, Geochimica et Cosmochimica Acta, 125, 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, Geochimica et Cosmochimica Acta, 141, 97–112, 2014b.

De Jonge, C., Radujković, D., Sigurdsson, B. D., Weedon, J. T., Janssens, I., and Peterse, F.: Lipid biomarker temperature proxy responds to abrupt shift in the bacterial community composition in geothermally heated soils, Organic Geochemistry, 137, 103897, https://doi.org/10.1016/j.orggeochem.2019.07.006, 2019.

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, Geochimica et Cosmochimica Acta, 310, 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, Organic Geochemistry, 187, https://doi.org/10.1016/j.orggeochem.2023.104706, 2024.

Dearing Crampton-Flood, E., Peterse, F., and Sinninghe Damsté, J. S.: Production of branched tetraethers in the marine realm: Svalbard fjord sediments revisited, Organic Geochemistry, 138, 103907, https://doi.org/10.1016/j.orggeochem.2019.103907, 2019.

Dembla, G.: Intuition behind Log-loss score, Towards Data Science, 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, Journal of Archaeological Science, 96, 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, Organic Geochemistry, 102, 52–58, 2016.

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, Climate of the Past, 17, 1199–1226, https://doi.org/10.5194/cp-17-1199-2021, 2021a.

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 Science Reviews, 273, https://doi.org/10.1016/j.quascirev.2021.107235, 2021b.

Genuer, R., Poggi, J.-M., Genuer, R., and Poggi, J.-M.: Random forests, Springer, 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, 2022.

Hastie, T., Tibshirani, R., Friedman, J., Hastie, T., Tibshirani, R., and Friedman, J.: Random forests, The elements of statistical learning: Data mining, inference, and prediction, 587–604, 2009.

Hilbe, J. M.: Practical guide to logistic regression, crc Press, 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 and Planetary Science Letters, 224, 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, Organic Geochemistry, 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, 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, Organic Geochemistry, 37, 1036–1041, https://doi.org/10.1016/j.orggeochem.2006.05.008, 2006.

Hunter, J. D.: Matplotlib: A 2D graphics environment, Computing in Science \& Engineering, 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, International Journal of Science and Research (IJSR), 5, 1842–1845, 2016.

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), Journal of Quaternary Science, 29, 70–82, https://doi.org/10.1002/jqs.2679, 2014.

Jung, Y.: Multiple predicting K-fold cross-validation for model selection, Journal of Nonparametric Statistics, 30, 197–215, https://doi.org/10.1080/10485252.2017.1404598, 2018.

Kalita, J.: Machine learning: Theory and practice, Chapman and Hall/CRC, 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, Biometrical Journal, 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, 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 (south-eastern shore of Lake Sevan, Armenia), Quaternary International, 395, 5–18, 2016.

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, Organic Geochemistry, 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, Organic Geochemistry, 127, 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, Palaeogeography, Palaeoclimatology, Palaeoecology, 611, https://doi.org/10.1016/j.palaeo.2022.111358, 2023.

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, Organic Geochemistry, 42, 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, Methods of Information in Medicine, 51, 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: 2017 International Conference on Wireless Technologies, Embedded and Intelligent Systems (WITS), Citation Key: manzali2017impure, 1–6, 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), Organic Geochemistry, 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 Science Reviews, 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, Geochimica et Cosmochimica Acta, 305, 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, Paleoceanography and Paleoclimatology, 38, 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, Journal of King Saud University - Computer and Information Sciences, 35, https://doi.org/10.1016/j.jksuci.2023.01.014, 2023.

Murphy, K. P.: Machine learning: a probabilistic perspective, MIT press, 2012.

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, Geochimica et Cosmochimica Acta, 208, 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, Organic Geochemistry, 106, 48–56, 2017b.

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, Organic Geochemistry, 106, 48–56, 2017c.

Nick, T. G. and Campbell, K. M.: Logistic regression, Topics in biostatistics, 273–301, 2007.

Niculescu-Mizil, A. and Caruana, R.: Predicting good probabilities with supervised learning, in: Proceedings of the 22nd international conference on Machine learning, Citation Key: niculescu2005predicting, 625–632, 2005.

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, Science of the Total Environment, 924, 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, 2019.

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), Climate of the Past, 19, 2127–2156, https://doi.org/10.5194/cp-19-2127-2023, 2023.

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, 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., and others: Scikit-learn: Machine learning in Python, Journal of machine learning research, 12, 2825–2830, 2011.

Peralta, A. L., Ludmer, S., Matthews, J. W., and Kent, A. D.: Bacterial community response to changes in soil redox potential along a moisture gradient in restored wetlands, Ecological Engineering, 73, https://doi.org/10.1016/j.ecoleng.2014.09.047, 2014.

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, Geochimica et Cosmochimica Acta, 96, 215–229, 2012.

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, https://doi.org/10.5194/bg-18-3579-2021, 2021.

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, Journal of Geophysical Research: Biogeosciences, 127, 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, Science Advances, 8, 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), Climate of the Past, 14, 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 and Planetary Change, 215, https://doi.org/10.1016/j.gloplacha.2022.103859, 2022.

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, Chemical Geology, 606, 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 Science Reviews, 277, 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., and others: Climate changes during the Late Glacial in southern Europe: new insights based on pollen and brGDGTs of Lake Matese in Italy, Climate of the Past, 19, 493–515, 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 Science Reviews, 281, 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, Organic Geochemistry, 117, 56–69, https://doi.org/10.1016/j.orggeochem.2017.12.003, 2018.

Simpson, G. L.: Analogue methods in palaeoecology: using the analogue package, Journal of Statistical Software, 22, 1–29, 2007.

Siriseriwan, W.: A collection of oversampling techniques for class imbalance problem based on SMOTE, 2019.

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, https://doi.org/10.1016/j.orggeochem.2020.104174, 2021.

Team, R. C.: R Core Team R: a language and environment for statistical computing, Foundation for Statistical Computing, 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, Organic Geochemistry, 40, https://doi.org/10.1016/j.orggeochem.2009.04.014, 2009.

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

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, Geochimica et Cosmochimica Acta, 318, 468–494, https://doi.org/10.1016/j.gca.2021.12.007, 2022.

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, Geochimica et Cosmochimica Acta, 305, https://doi.org/10.1016/j.gca.2021.05.004, 2021.

Wang, H., Chen, W., Zhao, H., Cao, Y., Hu, J., Zhao, Z., Cai, Z., Wu, S., Liu, Z., and Liu, W.: Biomarker-based quantitative constraints on maximal soil-derived brGDGTs in modern lake sediments, Earth and Planetary Science Letters, 602, https://doi.org/10.1016/j.epsl.2022.117947, 2023.

Wang, X., Zhang, H. H., and Wu, Y.: Multiclass Probability Estimation With Support Vector Machines, Journal of Computational and Graphical Statistics, 28, 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 Science Reviews, 182, 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, Geochimica et Cosmochimica Acta, 154, 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, Proceedings of the National Academy of Sciences of the United States of America, 115, 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, Environmental Microbiology, 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, Geochimica et Cosmochimica Acta, 71, 703–713, 2007.

Wickham, H.: ggplot2: Elegant Graphics for Data Analysis, Springer-Verlag New York, 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 O2 in a Chinese lake drive changes in microbial communities and impact sedimentary GDGT distributions, Chemical Geology, 579, 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, https://doi.org/10.5194/bg-13-5883-2016, 2016.

Yu, X., Ascencio, J., and French, R.: Open-Source Climate Classification Package: kgcPy, in: 2024 IEEE 52nd Photovoltaic Specialist Conference (PVSC), 1085–1085, 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, Organic Geochemistry, 41, 1060–1066, https://doi.org/10.1016/j.orggeochem.2010.03.004, 2010.