Quaternary Science Reviews 369 (2025) 109615

Contents lists available at ScienceDirect

Quaternary Science Reviews

journal homepage: www.elsevier.com/locate/quascirev

Development of new global lake brGDGT-temperature calibrations: advances, applications, challenges, and recommendations

Emma J. Pearson [a,*], Steve Juggins [a], Harry Allbrook [b,1], Louise C. Foster [a,c], Dominic A. Hodgson [c], B. David A. Naafs [b], Tony Phillips [c], Stephen J. Roberts [c]

[a] School of Geography, Politics and Sociology, Newcastle University, Newcastle-upon-Tyne, NE1 7RU, UK
[b] School of Chemistry, University of Bristol, Cantock's Close, Bristol, BS8 1TS, UK
[c] British Antarctic Survey, High Cross, Madingley Road, Cambridge, CB3 0ET, UK

ARTICLE INFO

Handling editor: P Rioual

Keywords:
Branched GDGTs
Calibration models
Temperature reconstruction
Lake sediments
Quaternary

ABSTRACT

Branched glycerol dialkyl glycerol tetraethers (brGDGTs) are a group of temperature-sensitive membrane lipids found in bacteria that have been widely used in palaeo-temperature reconstruction. Despite recent advances in analytical methods, calibration datasets and statistical modelling approaches, one of the current challenges in Quaternary science remains in determining the most appropriate calibration model for reconstructing past changes in climate. We address this challenge by expanding existing calibration datasets, and by evaluating calibration models constructed using a range of statistical modelling approaches. We further evaluate model performance by applying the calibrations to published downcore records from contrasting environments and across different Quaternary timescales.

Our study expands existing calibrations and includes new data from Antarctic lakes, providing greater confidence and applicability across a wider range of global environments. Results show robust brGDGT-temperature relationships on a global scale within the temperature range of approximately −2 °C to +31 °C covered in this study, with the random forest (RF) models performing the best (highest R^2_{cv} and lowest RMSEP) to estimate mean temperature of Months Above Freezing (MAF) and Mean Summer (air) Temperature (MST). Examination of uncertainties suggests the best models are accurately modelling all the features of the brGDGT-temperature relationships.

To evaluate model performance downcore we apply and recommend a suite of exploratory statistical analyses to help identify core-samples that have unusual, no-analogue compositions, and use measures of correlation and concordance to summarise the similarity in trends and absolute values among reconstructions as a tool to suggest which reconstructions may be more reliable and where to use caution. Our results demonstrate that, although cross-validated calibration R^2_{cv} and RMSEP may indicate good model performance for the calibration data, a thorough assessment is required to assess reconstruction reliability when a model is applied downcore at a specific site. Our findings highlight the complexities and caveats of different methods for global temperature calibrations. The implications of our work are also relevant to other calibration studies in Quaternary science.

1. Introduction

Quantitative palaeoclimate reconstructions are fundamental to understand long-term trends in natural climate variability and to test climate models used to make projections of future anthropogenic climate change. Advances in molecular organic geochemistry and related fields have led to the successful calibration and application of branched glycerol dialkyl glycerol tetraethers (brGDGTs) as temperature indicators in lakes. BrGDGTs are cell membrane lipids found in bacteria (Sinninghe Damsté et al., 2000) from diverse depositional settings and which have a structural variability that depends strongly on growth temperature (Weijers et al., 2006a,b), with the number of methyl groups (and cyclopentane rings) in the brGDGT structure being a key factor in the adaptation to temperature change, and a decrease occurring in the

* Corresponding author.
E-mail address: emma.pearson@ncl.ac.uk (E.J. Pearson).
[1] (present address) Department of Geological Sciences and Institute of Arctic and Alpine Research, University of Colorado, Boulder, 4001 Discovery Drive 80303, USA.

https://doi.org/10.1016/j.quascirev.2025.109615
Received 13 February 2025; Received in revised form 5 September 2025; Accepted 6 September 2025
Available online 19 September 2025

degree of methylation with increasing temperature (Supplementary Fig. 1). While the extent of brGDGT sources still remains largely unknown, advances in culture studies and bioinformatics have improved our understanding of the biosynthetic pathways of brGDGTs with implications for their application (e.g. Tetraether synthase, Tes; Chen et al., 2022; Lloyd et al., 2022). Results from molecular dynamics simulations of brGDGT membranes have shown that the empirically observed correlation between the degree of methylation and temperature allows brGDGT-producing bacteria to maintain adequate membrane fluidity via homeoviscous adaptation (Naafs et al., 2021). The physiological basis for the empirical relationship between brGDGT methylation number and temperature has also been supported by laboratory and environmental incubation studies (Martínez-Sosa et al., 2020) and using cultured *Solibacter usitatus* from the globally abundant bacterial phylum Acidobacteria, which may be an important brGDGT producer in nature (Chen et al., 2022; Halamka et al., 2023).

The development and application of temperature proxies based on brGDGTs led to the construction of global (Pearson et al., 2011) and regional lacustrine brGDGT-temperature calibration models which compare present day lake or air (mean) temperature with the brGDGT distributions in surface sediments. These regional calibrations include for eastern Africa (Tierney et al., 2010; Loomis et al., 2012) , Baffin Island, Canada (Shanahan et al., 2013), Tibet (Günther et al., 2014), Arctic Canada and Siberia (Peterse et al., 2014), Chile (Kaiser et al., 2015), Antarctica and the sub-Antarctic islands (Foster et al., 2016), and New Zealand (Zink et al., 2016). Calibrations have been applied down core to reconstruct past temperatures on a range of Quaternary timescales using lake sediments from a range of contrasting environments, including in eastern Africa (e.g. Sinninghe Damsté et al., 2012), Antarctica (Foster et al., 2016; Roberts et al., 2017), Siberia (Keisling et al., 2017), the USA (Krause et al., 2018), and Australia (Thomas et al., 2022).

The original analytical method for analysing brGDGTs used high performance liquid chromatography (HPLC) coupled to mass spectrometry (LCMS) with normal phase separation and a single cyano (CN) column (Hopmans et al., 2000). Accurate quantification of GDGTs using this method is challenging as the imperfect separation of some GDGT isomers can result in late eluting shoulders for one or more of the brGDGTs leading to an increase in analytical error for these compounds. De Jonge et al. (2013) determined that the late eluting shoulders comprise 6-methyl rather than 5-methyl brGDGTs and that improved chromatographic separation using a silica rather than cyano column had an impact on GDGT-derived proxies. De Jonge et al. (2014) subsequently redefined the MBT soil temperature index (methylation index of branched tetraethers; Weijers et al., 2007) to include both the 5-methyl and 6-methyl isomers, and found that removal of the 6-methyl isomers improved temperature calibrations, thus defining a new MBT'_{5ME} index using only the 5-methyl isomers. Based on these advances, the single cyano column (SC) analytical method of Hopmans et al. (2000) has been refined to use two silica columns in tandem (i.e. a dual column (DC) method) to improve chromatography and separation of the GDGT isomers (Hopmans et al., 2016). Further work has resulted in the separation and identification of both 5- and 6-methyl isomers of penta-methylated (GDGT-II) and hexa-methylated (GDGT-III) brGDGTs (Supplementary Fig. S1) in lakes (e.g., Dang et al., 2016; Li et al., 2017; Russell et al., 2018; Weber et al., 2018; Ning et al., 2019; Qian et al., 2019; Cao et al., 2020; Stefanescu et al., 2021; Martínez-Sosa et al., 2021; Raberg et al., 2021), and has led to DC-derived global (Martínez-Sosa et al., 2021; Raberg et al., 2021) and regional (Russell et al., 2018, eastern Africa; Lei et al., 2023, North American subtropics; Bauersachs et al., 2024, central Europe) lake calibration datasets.

Despite these advances there is clear variation in the robustness of SC and DC temperature reconstructions at different sites. In order to improve the accuracy and reliability of GDGT-based temperature reconstructions we consider two main routes of advancement: (i) expansion of the calibration dataset to capture a wider range of natural variability, and (ii) improvement of reconstructions via robust numerical methods. Here, we address these key issues by expanding and maximising the temperature gradient of both the single column (SC) and dual column (DC) surface sample calibration datasets, and by evaluating a range of different numerical approaches to calibration and reconstruction. We examine both DC and older SC methods because an updated global SC calibration can be applied to refine existing SC-derived core data (e.g., Sinninghe Damsté et al., 2012; Keisling et al., 2017; Krause et al., 2018; Foster et al., 2016; Roberts et al., 2017; Thomas et al., 2022; Heredia-Barión et al., 2023a, 2023b), and the existing global SC calibration (Pearson et al., 2011) has also been applied to modified DC-derived core data (Baxter et al., 2023) and was considered the most reliable method for reconstructing past temperature history from the sediments of Lake Challa (Baxter et al., 2024).

Our expanded datasets include Antarctic and sub-Antarctic sites to improve current quantitative brGDGT-temperature reconstructions in cold regions. To date the global calibration of Pearson et al. (2011) and the Antarctic calibration of Foster et al. (2016) are the only published studies reporting GDGTs in Antarctic lakes but no studies have been published investigating 5- and 6-methyl brGDGTs in Antarctic lake environments. This work therefore expands the Antarctic dataset used in the SC global calibration of Pearson et al. (2011), while providing the first addition of Antarctic samples in a DC global calibration, thus providing the first truly global brGDGT dataset comprising both 5- and 6-methyl brGDGTs. The Antarctic lakes span a range of Mean Annual Air Temperatures (MAAT) from −11.8 to 6.1C and Mean Summer Air Temperatures (MSAT) from −2.2 to 10.3C, cover a depth range of 0.5–55m and a range of pH and conductivities (see Foster et al., 2016).

Published brGDGT calibrations using lake sediments have used a range of statistical modelling techniques, including multiple linear regression of individual compounds (Pearson et al., 2011; Loomis et al., 2012; Foster et al., 2016; Bauersachs et al., 2024) , multiple regression using quadratic terms (Raberg et al., 2021), linear regression of a single index (e.g. MBT'_{5ME}) in a Bayesian framework (Martínez-Sosa et al., 2021), and machine learning methods such as regression trees (Véquaud et al., 2022) and deep learning artificial neural networks (Häggi et al., 2023). These methods each have advantages and disadvantages. For example, methods that collapse compounds into a single index (e.g., MBT'_{5ME}, e.g. Martínez-Sosa et al., 2021) potentially reduce noise in the non-linear temperature relationships of individual compounds but may not extract or model all features of the data. Similarly, linear regression with quadratic terms (e.g. Raberg et al., 2021) may better model non-linearities in the data but may be prone to overfitting and spurious extrapolation, particularly at the ends of the temperature gradient where non-linearities are less well constrained (Hahn, 1977). Here we apply these different statistical approaches to the same calibration datasets to provide a baseline assessment of method performance.

The performance of a calibration is usually assessed by measures of prediction error derived from the calibration dataset under cross-validation, as this guards against overfitting (Yates et al., 2023). However, performance of calibration models assessed purely on modern surface samples is not necessarily a good guide to the robustness of down-core reconstructions (Juggins, 2013; Sun et al., 2024), especially in cases where core data lie outside the geochemical composition space of the calibration data. Thus, in addition to evaluating different numerical approaches in terms of their prediction errors, we also evaluate the reconstructions from each method by applying them to a range of existing published sediment cores from contrasting environments.

Our specific aims and objectives were to: (i) develop new global lacustrine brGDGT-temperature calibrations based on single column (SC) and dual column (DC) LCMS analytical methods by expanding the current SC and DC datasets and including samples at the cold Antarctic and sub-Antarctic end of the temperature gradient; (ii) compare and assess a range of different statistical modelling methods that can be used for calibrations, with a focus on Mean Summer Temperature (MST) and Mean temperature of months Above Freezing (MAF), to identify the

most appropriate, robust, and realistic calibration method(s); (iii) apply new SC and DC calibrations to previously published down-core data from different lake environments and Quaternary timescales to evaluate the models and resulting downcore temperature reconstructions; (iv) assess the implications of our study for Quaternary science and provide recommendations to consider for future calibrations and reconstructions.

2. Materials and methods

2.1. Lake locations, datasets and brGDGT analyses

Locations of the lakes included in this study are shown in Fig. 1. To create new global brGDGT-temperature calibrations for lakes using the single column (SC) LCMS analytical method (Hopmans et al., 2000) we expand the global brGDGT-temperature calibration dataset of Pearson et al. (2011) by including the Antarctic and sub-Antarctic (Foster et al., 2016) datasets, along with other previously published regional lacustrine brGDGT-temperature datasets published between 2010 and 2018 from eastern Africa (Tierney et al., 2010; Loomis et al., 2012), Baffin Island, Canada (Shanahan et al., 2013), Arctic Canada and Siberia (Peterse et al., 2014), Chile (Kaiser et al., 2015), and New Zealand (Zink et al., 2016). Our new SC global lake calibration dataset (GLC-SC) comprises a total of 349 lakes. While this uses the original LCMS analytical method, which comes with its own caveats, it nonetheless enables refinement of reconstructions that have previously used this method (e.g., Sinninghe Damsté et al., 2012; Keisling et al., 2017; Krause et al., 2018; Foster et al., 2016; Roberts et al., 2017), and can be used by research groups that still use this analytical method (e.g., Thomas et al., 2022).

We also created a new global DC brGDGT-temperature calibration (GLC-DC) dataset comprising surface samples from a total of 378 lakes. Following the recent analytical advances which enables separation of 5- and 6-methyl compounds (Hopmans et al., 2016), we re-analysed Antarctic and sub-Antarctic samples from Pearson et al. (2011) and Foster et al. (2016) using the DC LCMS method and merged these data with existing published DC derived data published between 2016 and 2021 from Dang et al. (2016) (China), Li et al. (2017) (Inner Mongolia), Russell et al. (2018) (eastern Africa), Weber et al. (2018) (Switzerland and Italy), Ning et al. (2019), Qian et al. (2019), Cao et al. (2020) (China), Stefanescu et al. (2021) (USA), Martínez-Sosa et al. (2021) (global), Raberg et al. (2021) (Iceland and Canada).

The datasets used in this study thus comprise (i) published brGDGT data from lake calibrations published over past decades using the single column LCMS method (SC), and (ii) published brGDGT data from lake studies using the dual column (DC) LCMS method, in combination with our new Antarctic DC derived brGDGT dataset. Method details for the original data for each of the lakes can be found in the corresponding original publications, and DC analysis details for the Antarctic samples are provided in Supplementary SI1. Information on methods used to verify the lake locations, elevations and temperature data is provided in Supplementary SI2. The lake calibration datasets are not restricted to a particular lake chemistry type (i.e. they encompass fresh, brackish and saline lakes), and we apply our new calibration models to published data from contrasting environments spanning different Quaternary time slices and timescales (see Section 2.3).

2.2. Temperature data

While some early published lake-temperature calibrations used Mean Annual (Air) Temperature (MAT or MAAT) (Tierney et al., 2010; Zink et al., 2016), the global brGDGT-temperature calibration (Pearson et al., 2011) and Antarctic regional calibration (Foster et al., 2016) use Mean Summer (air) Temperature (MST) because brGDGT production is expected to occur primarily during the warmest season when biological productivity is greatest (also see Sun et al., 2011; Shanahan et al., 2013). The influence of summer temperature (as opposed to annual) is particularly relevant in polar lakes as they undergo large seasonal temperature fluctuations between two seasons (summer and winter) and are ice-covered for a significant part of the year. More recent studies have found strong relationships between brGDGT compositions in lacustrine sediments and mean temperature of Months Above Freezing, MAF (e.g. Martínez-Sosa et al., 2021; Raberg et al., 2021) which account more realistically for more than a single season of brGDGT production. Note that in the low latitudes, where temperatures do not reach below 0 °C, MAF = MAT.

In our study we examine the relationships of both MST and MAF with lake sedimentary brGDGT distributions. For MST and MAF temperature data, we use originally reported observed published values where available. Gaps in the observed data were filled with values extracted from the ERA5 reanalysis dataset (Hersbach et al., 2019, 2020) using Copernicus Climate Change Service Information [2020, 2021] (see

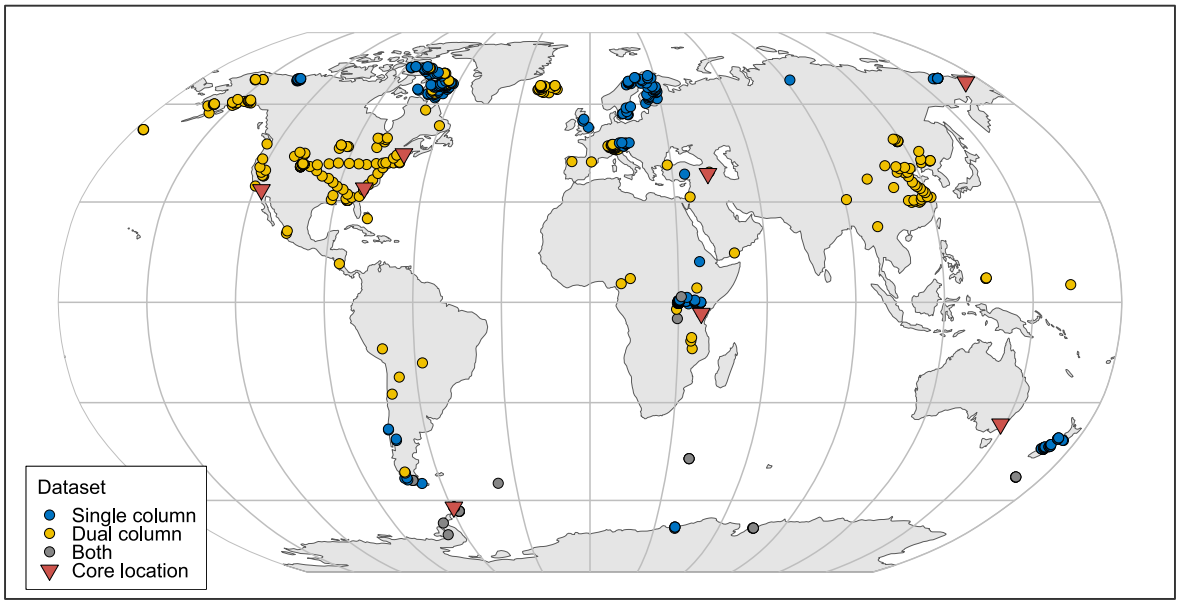

Fig. 1. World map showing locations of the core sites and lakes used in the single and dual column calibration datasets.

Supplementary SI2). For the latter we used the mean MST (average temperature of June, July, August (JJA) or, for austral summer, December, January, February (DJF)) or MAF, both averaged over the 30 years prior to the date of sampling to capture the approximate period covered by the surface sample. For our SC dataset the temperature range spans −2.2 °C–31.2 °C (MST) and 0.44 °C–25.7 °C (MAF). For our DC dataset the temperature range spans −2.2 °C–30.8 °C (MST) and 0.44 °C–28.1 °C (MAF).

2.3. Numerical calibration methods, cross validation, and application downcore

Inferring past temperatures from sedimentary GDGT compositions is a problem of multivariate calibration. This is a well-established area of statistics (e.g. Varmuza and Filzmoser, 2009) but GDGT data possess a number of properties that make calibrations using traditional methods difficult, including non-linear relationships, constant sum constraints and collinearity between predictors, the influence of other environmental variables and measurement error in both GDGT and temperature data. A useful calibration should be complex enough to model non-linear trends in multivariate compound abundances but not overly complex so that it over-fits the calibration dataset and performs poorly with core data. It should also use an appropriate statistical method that takes account of the specific properties of, in this case, brGDGT data. A range of numerical calibration techniques have been used in the literature, and each addresses the specific challenges of brGDGT calibration in different ways. Since there is no accepted and demonstrated "best" method we explore a range of numerical calibration techniques and evaluate their performance on our modern calibration datasets and downcore reconstructions.

For both our SC and DC datasets we investigated and assessed a range of different calibration statistical modelling approaches including linear regression (LR), multi-model averaging (MMA), generalised additive modelling (GAM), Dirichlet regression (DR), random forests (RF) and deep-learning neural networks (NN) to construct new global calibrations (see Supplementary SI3). In addition to these methods, we also compared the performance of the original (global) regression model of Pearson et al. (2011) and recent calibrations based on the methylation index of branched tetraethers (MBT'$_{5ME}$) from Martínez-Sosa et al. (2021) and using linear regression with quadratic terms from Raberg et al. (2021, eqn (11), R11). We also investigated converting DC to SC formatted data for application of the Pearson et al. (2011) model by summing the fractional abundances of the 5- and 6-methyl brGDGT compounds which was considered by Baxter et al. (2023, 2024) to be a robust method for African Lake Challa. However, in our expanded global dataset this approach (not shown) suggests this finding is site dependent and not globally applicable. See Supplementary SI3 for more information on the numerical calibration methods explored.

Following model construction, cross-validation should be performed to properly test and validate the model. We assess the performance of each of our models using the squared correlation between observed and predicted temperatures (R^2_{cv}) and root mean squared error of prediction (RMSEP) calculated under 10-fold cross-validation. This approach provides a guard against overfitting and hence is a better and more reliable basis for comparison among models, and also gives a more realistic idea of prediction errors when the model is applied to core data.

We applied our single column global lake calibrations (GLC-SC) and dual column global lake calibrations (GLC-DC) to previously published lake sediment core data from contrasting environments that span different timescales and time slices within the Quaternary period to assess how well the different models performed when applied to core data. Specifically, we applied our updated GLC-SC to published core data from Antarctica (Yanou Lake; Roberts et al., 2017), eastern Africa (Lake Challa; Sinninghe Damsté et al., 2012), Australia (Club Lake; Thomas et al., 2022), Siberia (El'gygytgyn; Keisling et al., 2017) and the USA (White Pond; Krause et al., 2018). We applied our GLC-DC to published core data from the USA (Basin Pond; Miller et al., 2018; Lake Elsinore; Feakins et al., 2019) and Turkey (Lake Van; Stockhecke et al., 2021). A brief summary of lake information is provided in Supplementary SI4 and their locations are indicated on Fig. 1.

When applying calibrations down core it is important to also examine the extent to which the calibration dataset encompasses the range of brGDGT values found in the core. Core samples with brGDGT compositions that are markedly different from the calibration dataset represent non-analogue situations, that can be difficult to reconstruct and might not be useable for palaeoclimate reconstructions. The position of our core samples within the surface calibration space was visualised using principal components analysis (PCA). We also calculated Mahalanobis distance between each core sample and the calibration dataset and converted these distances to probabilities using a chi-squared cumulative probability distribution (Aggarwal, 2017). Core samples with probabilities of greater than 0.99 are considered outliers in their brGDGT composition. Finally, we compare different reconstructions using Pearson's product-moment correlation coefficient and Lin's concordance correlation coefficient (Lin, 1989). The former assesses the extent to which different reconstructions follow similar trends (but may have different absolute values) while the latter expresses deviations from 1:1 concordance between reconstructions and reflects differences in both trends and absolute values. Like Pearson's correlation, Lin's concordance ranges from −1 to 1, with 1 indicating perfect agreement and −1 indicating strong discordance. The overall similarity among reconstructions was then summarised by cluster analysis of the resulting correlation/concordance matrices.

BrGDGTs were expressed as fractional abundances of the sum of all brGDGTs identified using the SC and DC methods prior to all analyses because we were interested in compositional differences in brGDGTs between sites. Our GLC-SC dataset comprises the nine major brGDGTs including the 5-methyl isomers originally identified using the SC method (noting that some 6-methyl isomers co-elute), and our GLC-DC dataset comprises fifteen brGDGTs including the separated 5- and 6-methyl isomers (see Supplementary Fig. S1). An additional consideration with multivariate calibrations is minor compounds, which are subject to large integration errors and can have undue weight in the analysis. We therefore excluded from all calibrations the following compounds which had an abundance of less than 1.0 % in 75 % of the calibration samples: fIc, fIIc, fIIIb, fIIIc (SC dataset) and fIc, fIIc, fIIc', fIIIb, fIIIb', fIIIc, fIIIc' (DC dataset).

All numerical analyses were performed using R software for statistical computing and graphics (R Core Team, 2024) with the following additional packages for ordination (vegan: Oksanen, 2024), GAMs (mgcv: Wood, 2017), MMA (MuMIn: Bartoń, 2024), Dirichlet regression (DirichletReg: Maier, 2014), random forests (ranger: Wright and Ziegler, 2017), and neural networks (keras: Allaire and Chollet, 2024; tensorflow: Allaire and Tang, 2024).

3. Results

3.1. Development and assessment of new global lake brGDGT-temperature calibration models

The performance of each single column (SC) and dual column (DC) calibration method is summarised in Table 1 which includes both the R^2 and RMSE values and also the cross-validated R^2_{cv} and RMSEP values, the latter which are the focus for discussion since these have undergone rigorous cross-validation and thus provide a more realistic indicator of how well the model will perform with new data. The relationships between observed and predicted MST and MAF are shown in Fig. 2. A key observation is that the brGDGT distribution in lacustrine sediments is highly correlated with temperature across the globe.

For the SC dataset and MST as the target variables the Pearson et al. (2011) calibration model is the worst performing method overall (R^2_{cv} = 0.77, RMSEP = 5.06 °C), with a marked over-estimation at the lower end

Table 1
Performance of the SC and DC datasets using different numerical calibration methods.

Variable	Method	R^2	RMSE	R^2_{cv}	RMSEP
Single column					
MST	Pearson et al. (2011)	0.77	5.06	0.77	5.06
MST	Linear regression	0.79	3.72	0.78	3.78
MST	Multi-model averaging	0.79	3.74	0.78	3.80
MST	Generalised additive model	0.83	3.36	0.81	3.53
MST	Dirichlet regression	0.76	4.57	0.75	4.51
MST	Random forests	0.96	1.75	0.83	3.29
MST	Neural network	0.87	2.90	0.81	3.50
MAF	Linear regression	0.86	2.58	0.86	2.61
MAF	Multi-model averaging	0.87	2.60	0.86	2.63
MAF	Generalised additive model	0.91	2.15	0.90	2.24
MAF	Dirichlet regression	0.83	3.36	0.82	3.23
MAF	Random forests	0.97	1.67	0.91	2.17
MAF	Neural network	0.94	1.80	0.89	2.33
Dual column					
MST	Pearson et al. (2011)	0.78	3.93	0.78	3.93
MST	Linear regression	0.82	3.45	0.82	3.51
MST	Multi-model averaging	0.82	3.48	0.81	3.57
MST	Generalised additive model	0.84	3.31	0.82	3.45
MST	Dirichlet regression	0.79	4.56	0.79	4.55
MST	Random forests	0.96	1.66	0.85	3.19
MST	Neural network	0.94	2.08	0.84	3.26
MAF	MBT M-S	0.79	3.32	0.79	3.32
MAF	Raberg et al. eqn 11	0.82	2.92	0.82	2.92
MAF	Linear regression	0.84	2.61	0.84	2.68
MAF	Multi-model averaging	0.84	2.68	0.83	2.75
MAF	Generalised additive model	0.88	2.32	0.85	2.61
MAF	Dirichlet regression	0.82	3.34	0.81	3.34
MAF	Random forests	0.97	1.26	0.86	2.46
MAF	Neural network	0.95	1.46	0.85	2.58

Note: Pearson et al. (2011)= *original regression model of* Pearson et al. (2011)*; Linear Regression (LR) = linear regression modelling using* Pearson et al. (2011) *approach but calibrated with the new expanded dataset; MBT M-S= MBT eqn. (7) after* Martínez-Sosa et al. (2021)*; Raberg et al. eqn. 11= calibration equation (11) after* Raberg et al. (2021). See Supplementary SI3 for details of the statistical modelling approaches used.

of the temperature gradient (Fig. 2). This was also a feature, albeit less pronounced, of the original Pearson et al. (2011) calibration (also see Foster et al., 2016; Roberts et al., 2017; Heredia Barión et al., 2023a; Heredia Barión et al., 2023b). Improved results are obtained when using the same linear regression (LR) approach as Pearson et al. (2011) but calibrated with the new expanded dataset (R^2_{cv} = 0.78, RMSEP = 3.78 °C). This model has almost identical model performance to multi-model averaging (MMA; $R^2_{cv} = 0.78$, RMSEP = 3.80 °C). Replacing the linear fits of the LR and BMA models with smooth functions in the generalised additive model (GAM) improves the performance slightly (R^2_{cv} = 0.81, RMSEP = 3.53 °C). Dirichlet regression is the worst performing method for R^2_{cv} ($R^2_{cv} = 0.75$, RMSEP = 4.51 °C). Random forests

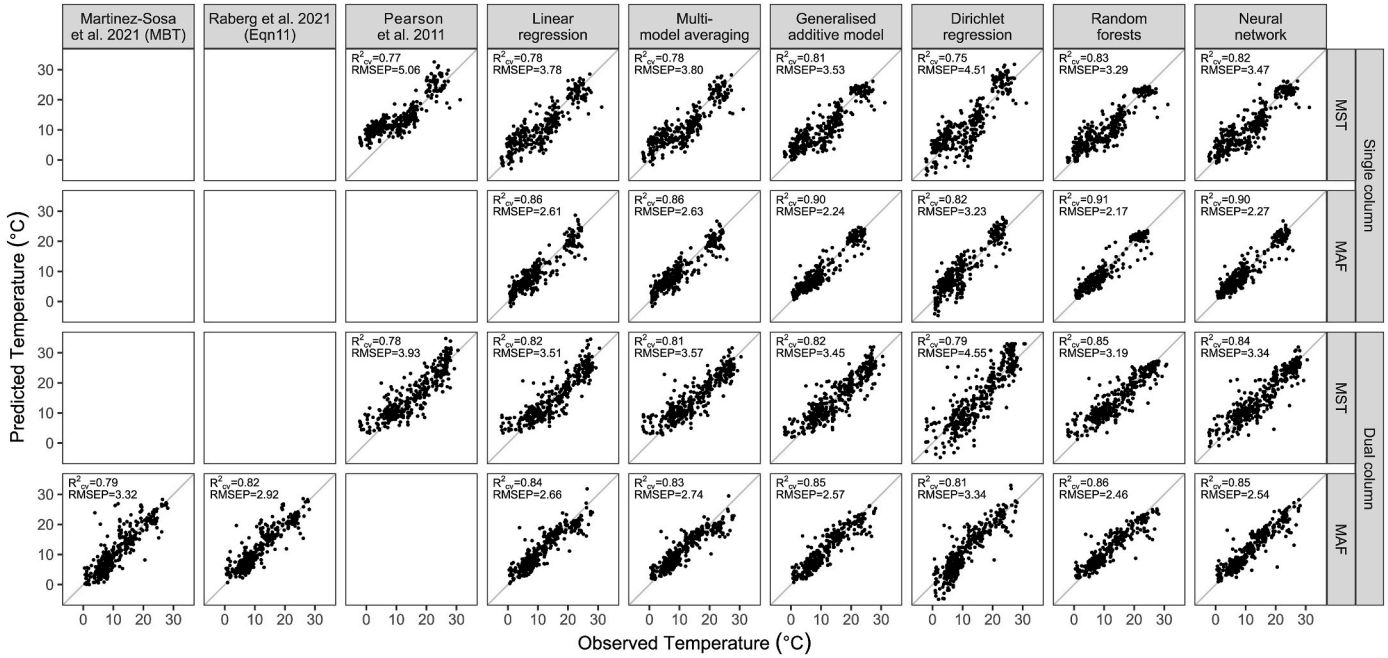

Fig. 2. Relationships between observed and predicted (under 10-fold cross-validation) MST and MAF for the Single and Dual column datasets for each numerical method.

is the best performing method overall (R^2_{cv} = 0.83, RMSEP = 3.29 °C) and slightly outperforms neural network, the other machine learning method (R^2_{cv} = 0.81, RMSEP = 3.50 °C).

The pattern of model performance among methods for MAF is similar to that for MST except that the squared correlations are slightly higher and RMSEP lower for any given method (e.g. for random forests R^2_{cv} and RMSEP values are 0.83/3.29 °C and 0.91/2.17 °C for MST and MAF respectively).

Model performance among methods applied to the DC dataset also follows a broadly similar pattern to that in the SC dataset, except that for MST the squared correlations between observed and predicted MST are slightly higher for the DC dataset for a given method (e.g. for LR, R^2_{cv} = 0.78 and 0.82 for SC and DC datasets, respectively), and for MAF the model performance is slightly lower for the DC dataset for a given method (e.g., for LR, R^2_{cv} = 0.86 and 0.84 for SC and DC datasets respectively). Overall, Dirichlet regression is again the worst performing and tends to under-predict at the low end of the temperature gradient, while random forests (RF) and neural networks (NN) perform best (RF: MST R^2_{cv} = 0.85, RMSEP = 3.19 °C and MAF R^2_{cv} = 0.86, RMSEP = 2.46 °C; NN: MST R^2_{cv} = 0.84, RMSEP = 3.26 °C and MAF = R^2_{cv} 0.85, RMSEP = 2.58 °C).

In addition to the methods described above, we also applied two additional previously published MAF calibrations to the DC dataset. The first, the MBT'$_{5ME}$ index (Martinez-Sosa et al. (2021, Eqn. (7)), performed relatively poorly with the second highest RMSEP (3.32 °C). The other, linear regression using quadratic terms, which was the best performing full set calibration for MAF proposed by Raberg et al. (2021; Eqn. (11)), performs slightly worse than the regression approaches of LR, MMA and GAM (R^2_{cv} = 0.82, RMSEP = 2.92 °C).

Overall Dirichlet regression is the worst performing method in both datasets. The published SC calibration of Pearson et al. (2011) and DC calibration based on MBT'$_{5ME}$ of Martínez-Sosa et al. (2021) also have higher RMSEP for MST and MAF respectively than the new calibrations presented in Table 1. For the remaining methods the differences in RMSEP are relatively small, with machine learning methods (random forests and neural networks) slightly outperforming regression-based approaches (linear regression, GAMs, multi-model averaging, and quadratic regression). Overall, random forests is the best performing method (defined as highest R^2_{cv} and lowest RMSEP) for both for both MST and MAF with the single column (SC) and dual column (DC) datasets. Of the two temperature variables, random forest models for MAF have the lowest overall RMSEP (SC = 2.17 °C; DC = 2.46 °C).

3.2. Temperature reconstructions and comparison of methods

Temperature reconstructions for each downcore record using our new models are shown in Fig. 3 and Supplementary Figs. S4. For each core we compare reconstructions for the calibration methods described above and also show a number of associated diagnostic plots to help interpretation. These include plots of the standard deviation (SD) of the different reconstructions to quantify variability among methods, a PCA (Principal Components Analysis) of the combined core and calibration datasets to visualise where core samples plot outside or on the periphery of the calibration data and represent no-analogue conditions, and boxplots summarising the abundances of each brGDGT compound in the core and calibration samples. Core samples with Mahalanobis probabilities greater than 0.99 and 0.999 are considered outliers and extreme outliers respectively (a brGDGT distribution unlike the modern

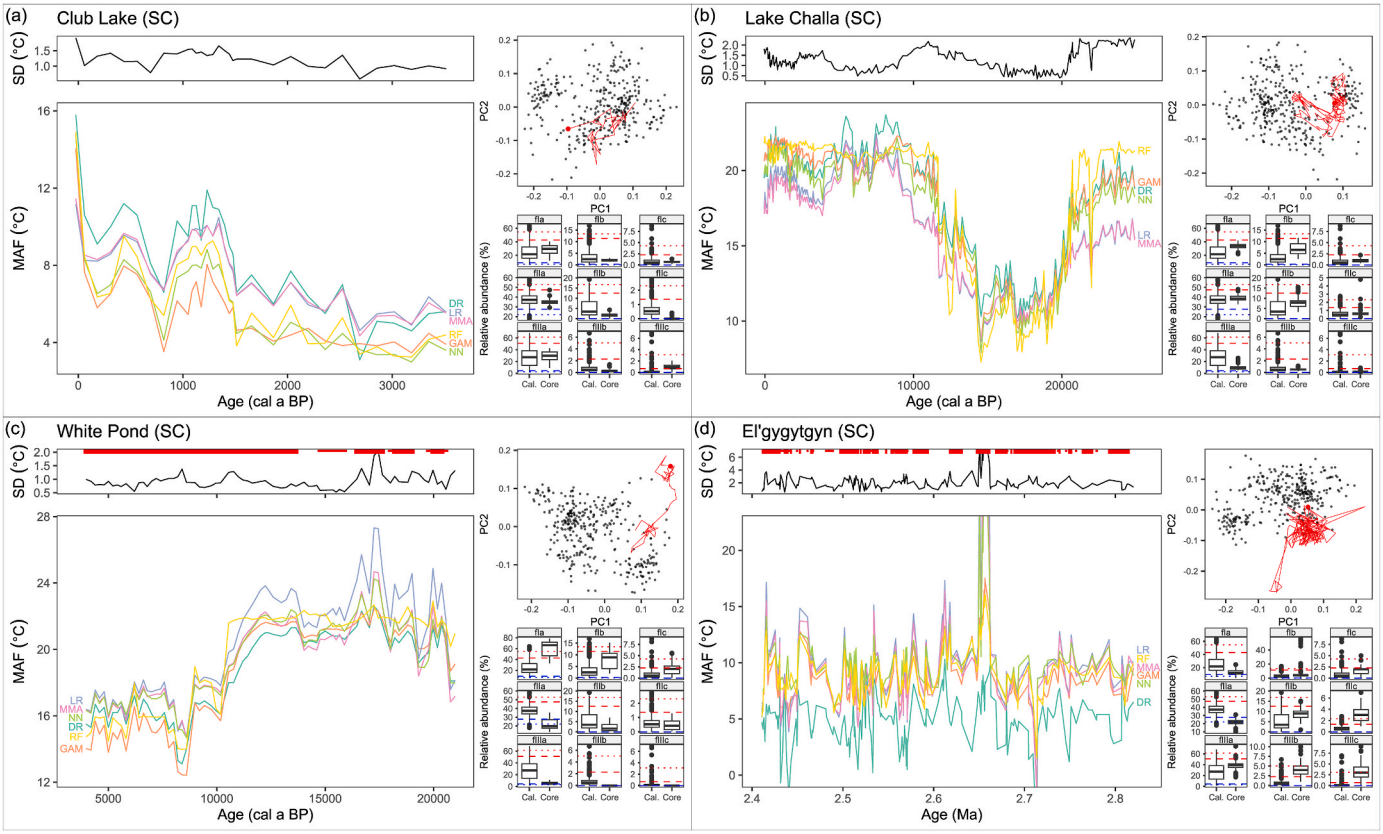

Fig. 3. Temperature reconstructions and diagnostic plots for each core using the numerical methods described in Section 2.3. Plots include temperature reconstructions, standard deviation (SD) of the different reconstructions, PCA of the combined core and calibration datasets, and boxplots summarising brGDGT compound abundance in core and calibration samples. Red symbol on PCA plots indicates top of the core. Blue and red dotted and dashed lines on boxplots indicate 1st, 5th, 95th, 99th percentiles of values in calibration dataset. Thin and thick red lines in the SD plot indicate outliers and extreme outliers respectively. Note different x and y axis scales. See text for details and Supplementary SI5 for enlarged plots.

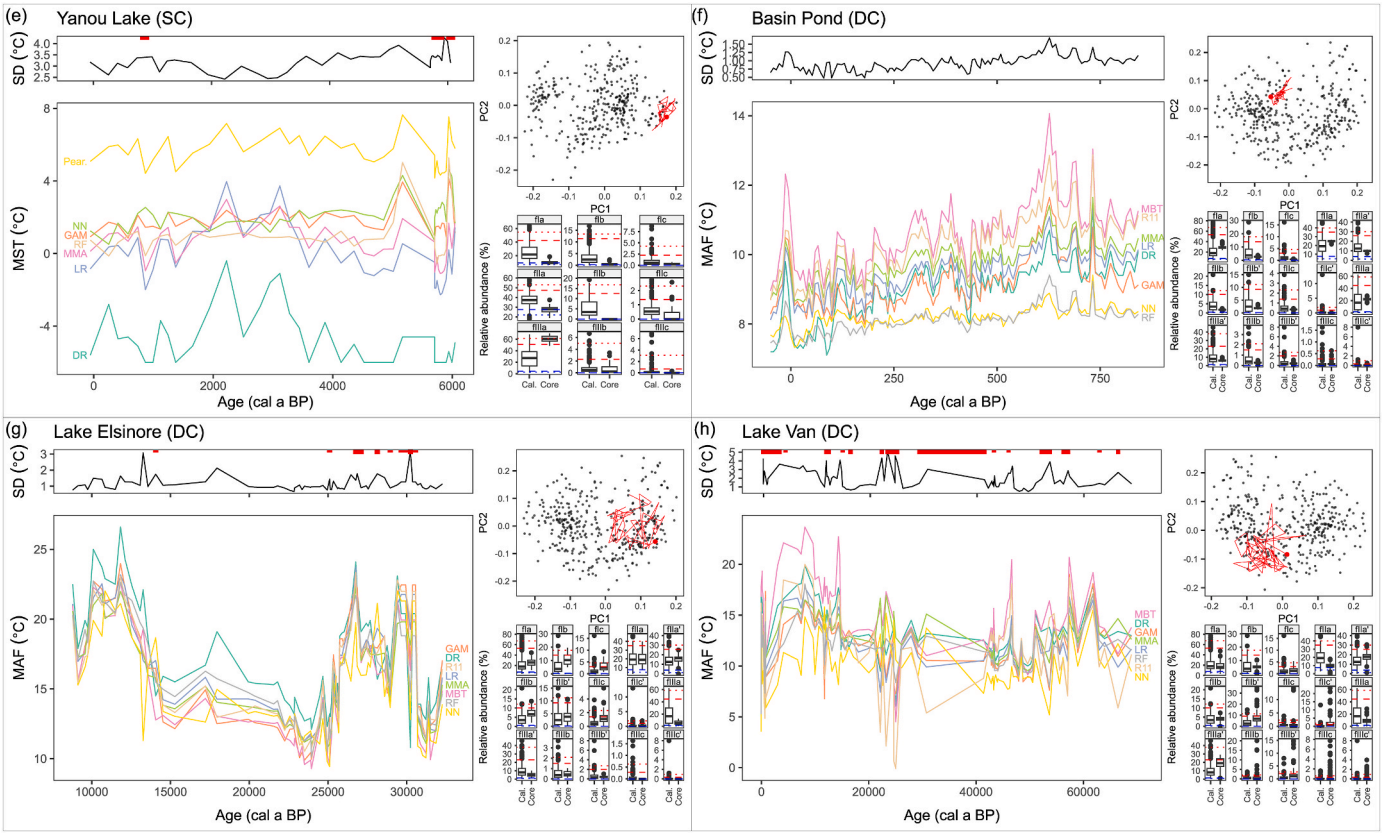

Fig. 3. (*continued*).

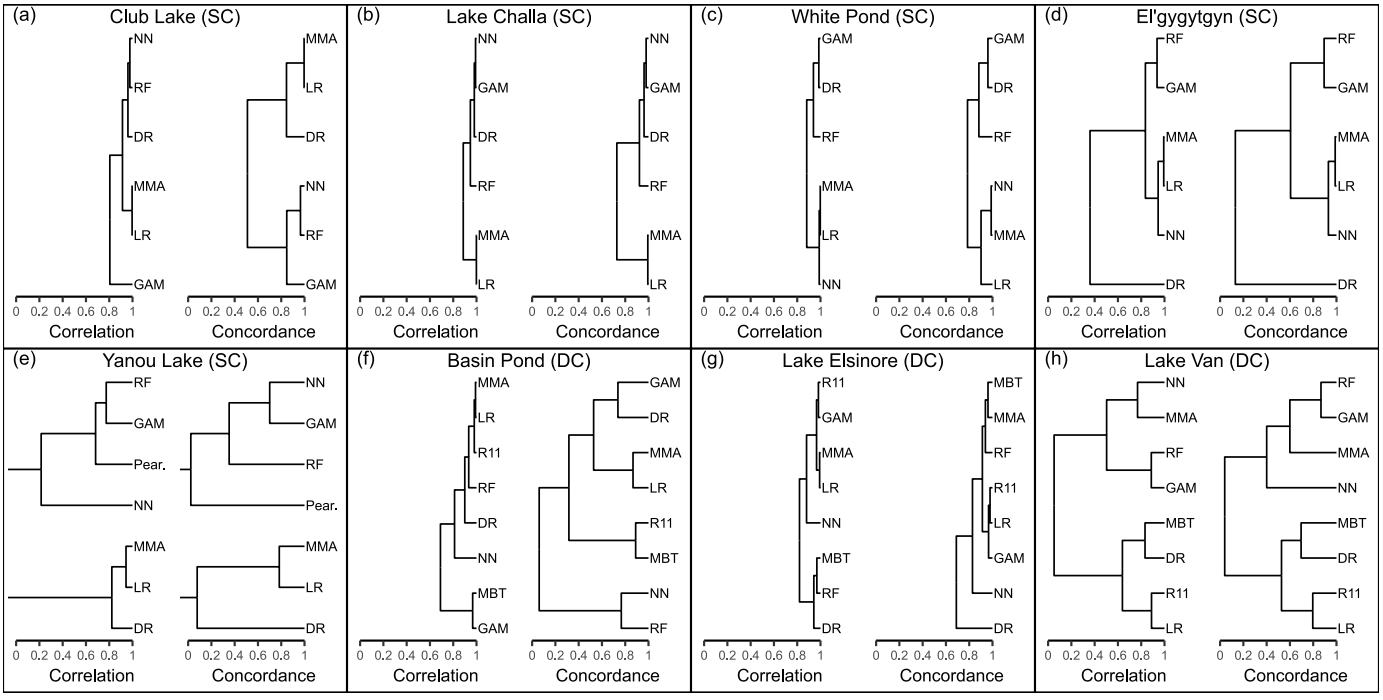

Fig. 4. Dendrograms summarising correlation and concordance among different reconstructions for each core.

calibration dataset) and are indicated as thin and thick red bars on the SD plot. Reconstructions for MAF and MST show similar trends for all cores, so we only show MAF reconstructions as this is currently the most commonly reconstructed brGDGT temperature variable, except for Yanou Lake (Antarctica) where we show the reconstruction for MST since this is a cold polar lake which experiences long periods of ice cover and two seasons (summer and winter) that means MST values are more appropriate and meaningful than MAF.

We also quantify the overall similarity in trends and absolute values among reconstructions using Pearson's product-moment correlation coefficient and Lin's concordance coefficient respectively.

3.2.1. Single column (SC) reconstructions

The Club Lake record (Australia, Fig. 3a) spans the last ~3.5 kyr and most of the core plots within the first two principal components of the brGDGT calibration space. Only the surface sample is somewhat atypical with unusually low relative abundance of fIIb, although this is not considered an outlier. Reconstructions all follow similar trends throughout the core and concordance between methods is generally high (Fig. 4a) and with low standard deviation (c. 1–1.52 °C), although reconstructions for LR, MMA and DR are consistently c. 2 °C higher than those for RF, GAM and NN.

Reconstructions for Lake Challa (eastern Africa; Fig. 3b), which spans the last ~ 25 kyr, also plot well within the calibration data and has no outliers. Correlation and concordance among methods is generally high (Fig. 4b) although there is greater variability between methods in the early and later parts of the core, with SD values of up to c. 2.5 °C, with temperatures predicted by LR and MMA c. 3–4 °C lower than other methods and RF that are c. 2 °C higher.

Reconstructions for White Pond (USA; Fig. 3c), spanning from 21 to 4 cal ka BP, all follow similar trends and show high correlation and high concordance (Fig. 4c) with generally low SD (c. 1.0 °C), although LR is unusual in predicting higher MAF in the early, older, part of the core, and, importantly, with higher variance. Most of the core samples plot outside or on the periphery of the calibration data and most are considered extreme outliers, with unusually high abundance of compounds fIa and fIb.

The El'gygytgyn (Siberia; Fig. 3d) record spans the late Pliocene between 2.82 and 2.41 Ma. Most core samples plot at the periphery or outside of the calibration dataset space, and most are considered extreme outliers, with unusually high values of fIb and fIIIb, and unusually low values of fIa and fIIa. However, with the exception of DR, there is high correlation and moderate concordance among reconstructions although spuriously high and low MAF are reconstructed by all methods around 2.65 and 2.72 Ma respectively.

Yanou Lake (Antarctica; Fig. 3e) is a cold polar site. The late Holocene record spans the last ~6 kyr, and although the whole core has unusually high abundance of fIIIa and plots at the edge of the calibration dataset only a few samples in the early, and one in the later, part of the core are considered outliers using a Mahalanobis distance criterion. Samples plot in a tight cluster in calibration space indicating little variation in brGDGT composition that is reflected in a narrow range of MST reconstructions for individual methods. Except for LR and MMA there is generally low correlation and low concordance among methods: DR and Pearson et al. (2011) show markedly different values, with DR predicting sub-zero temperatures for the whole core, and Pearson et al. temperatures of over 4 °C. SD values are correspondingly high and range from c. 2.5–4 °C.

3.2.2. Dual column (DC) reconstructions

Fig. 3f shows MAF reconstructions for Basin Pond (USA) spanning the last ~900 years. All samples fall within the variability of the calibration dataset and there are no outliers. Core samples plot very closely together on the PCA highlighting little difference in sample composition downcore, which is reflected in the narrow range of MAF reconstructions and the low standard deviations among reconstructions. All methods reconstruct a gradual cooling over the last 850 years and although correlations between different reconstructions are high, their concordance is low, with methods showing a consistent offset from NN and RF that predict low MAF with low variance, to MBT'_{5ME} and R11 that reconstruct higher MAF with much higher variance.

Reconstructions for Lake Elsinore (USA; Fig. 3g), spanning the interval ~8–32 cal ka BP, exhibit high correlation and high concordance, with all methods showing very similar trends and, except for DR in the middle and later parts of the core, broadly similar values. The core samples lie within the calibration set but a few have unusually high values of compounds fIb and fIIb, and a number of samples towards the base of the core are outliers.

For Lake Van (Turkey; Fig. 3h), spanning the last ~68 kyr, most of the core is positioned within or towards the edge of the calibration dataset and many core samples from 60 cal ka BP onwards are marked as outliers, primarily as a result of unusually high relative abundance of the isomers fIIb' and fIIIa' and low abundance of fIIa. Variation among reconstructions is large (SD 2–4 °C), there is a general pattern of very low correlation and concordance throughout most of the record with reconstructions showing high variance and spurious high and low values in several sections of the core.

4. Discussion

Here we first examine and compare the performance of our new SC and DC calibration models (Section 4.1), highlight the importance of considering sources of errors in calibration datasets and reconstructions, and explore sources of model uncertainty (Section 4.2). We then discuss applicability of our models for down-core reconstructions, highlighting by example those sites which demonstrate calibration application robustness and those that require calibration application caution (Section 4.3). We also provide some recommendations from our findings for use in brGDGT calibrations and applications in Quaternary science (also see Section 5).

4.1. Comparison of calibration models

Our new global single-column (SC) and dual-column (DC) brGDGT calibration datasets were examined using a range of numerical modelling methods to assess how the different approaches impact palaeotemperature reconstructions. We applied several existing and new regression methods that model the potentially non-linear relationship between brGDGTs and temperature with varying levels of complexity, from simple least squares regression (LR) to essentially black-box machine learning methods of random forests (RF) and deep-learning neural networks (NN).

As shown in Table 1 and Fig. 2 and the Results Section 3.1, for both our SC and DC datasets the Dirichlet regression was the worst performing method with a tendency to under-predict at the low end of the temperature gradient. This is surprising, given that Dirichlet regression is a classical regression approach recommended in ecological and environmental applications using count based fractional data and statistically the most appropriate method for modelling the brGDGT compositional data (Douma et al., 2019; see Supplementary SI3). Exploratory plots (Supplementary Fig. S2) indicate large residuals are associated with unusually high or low values of fIa and fIIIa, suggesting that predictions from this method are more sensitive to unusual values of these compounds than other techniques. While fIIIa has a strong correlation with temperature across a wide range of lakes (Pearson et al., 2011), changes in redox and bacterial communities may also influence the abundance of fIIIa at some sites (e.g. by some bacteria with lower oxygen requirements producing more fIIIa; Yao et al., 2020), and brGDGT fIa is thought to be derived from catchment soil bacteria (e.g. Sinninghe-Damsté et al., 2000; Hopmans et al., 2004; Weijers et al., 2006a,b). Unusual values of these compounds may therefore reflect overprinting of changes in redox and/or bacterial communities on the brGDGT composition which are accentuated when using this approach.

Improved performance was obtained using best subsets linear regression (LR), which gave a very similar model performance and pattern of residuals to multi-modelling averaging (MMA). The potential increase in model robustness and prediction accuracy obtained by averaging over several best models is not observed in practice with these data. Replacing linear fits with smooth functions using GAMs improved the performance slightly, though the machine learning approaches of

random forests (RF) and neural networks performed slightly better with RF returning the lowest prediction errors for SC and DC datasets (SC: RMSEP = 2.17 °C for MAF; 3.29 °C for MST; DC: RMSEP = 2.46 °C for MAF; 3.19 °C for MST).

While model performance using the DC dataset follows a broadly similar pattern to the SC dataset, overall, the squared correlations between observed and predicted MST are slightly higher in the DC dataset for any given method and for MAF the model performance is slightly lower for a given method (Table 1 and Fig. 2). The poor performance when applying the MBT'$_{5ME}$ index (Martínez-Sosa et al., 2021) to the DC dataset is, in part, a result of large residuals with unusually high or low abundance of fIa, as with the Dirichlet results, perhaps indicative of unusual bias of soil bacterial inputs (Supplementary Fig. S3). Testing linear regression using quadratic terms (Eqn. (11); Raberg et al., 2021) did not improve on the simpler multiple regression models of LR and MMA. GAM improved over LR, MMA and LR with quadratic terms, suggesting that there are some non-linear features in the data that can be modelled but that these are better modelled using data-driven smoothers rather than quadratic terms.

Importantly, while the RF model provides the best results overall (highest R^2_{cv} and lowest RMSEP) for both our SC and DC datasets, overall differences in prediction error between LR, MMA, GAM, RF and NN methods are small. Using cross-validation performance on the calibration set as a criterion it is therefore difficult to choose a single best method for downcore reconstruction. Furthermore, the worst performing methods (DR, MBT, R11) have cross-validation prediction errors of c 3.0–4.5 °C which compare favourably with some other previously reported terrestrial brGDGT temperature calibrations (e.g. non-cross-validated) RMSEs in the region of 4–5 °C for global soils (Naafs et al., 2017).

The temperature range of our SC dataset (−2.2 °C–31.2 °C for MST; 0.44–25.7 °C for MAF) slightly increases the range of our previous Pearson et al. (2011) global dataset (mean summer temperature range c. 1–31 °C) and also covers a wider range than other regional SC studies (e.g. Loomis et al. (2014) MAT range 1.5–26.8°C; Sun et al. (2011) warm months range 8.2–23.3 °C,MAT range −2.8 °C to 23.3°C; Tierney et al. (2010) MAT range 1–25 °C).

The cross-validated performance of our new SC RF model (R^2_{cv} = 0.91, RMSEP = 2.17 °C for MAF; R^2_{cv} = 0.83, RMSEP = 3.29 °C for MST, n = 349) gives improved or comparable results to previous published SC models, bearing in mind differences between reported raw R^2 and cross-validated R^2_{cv} values and differences in temperature parameters used e.g. Pearson et al. (2011) global calibration had MST R^2_{cv} = 0.88; RMSEP = 2.1 °C, n = 90, while the MAT calibration of Loomis et al. (2014) resulted in R^2 = 0.88, RMSEP = 2.1 °C, n = 111, and the MAT of Tierney et al. (2010) had R^2 = 0.94, RMSE = 2.2 °C, n = 46.

The temperature range of our DC dataset (−2.2 °C–30.8 °C for MST; 0.44°C–28.1 °C for MAF) improves on previous MST and MAF DC calibrations, e.g. Martínez-Sosa et al. (2021) MAF temperature ranges from 1.6 to 28.1°C; Raberg et al. (2021) MST temperature ranges from c. −1 to 29.5 °C, with a MAF range from 0.6 to 26.8 °C).

The performance of our new DC RF model (R^2_{cv} = RMSEP = 2.39 °C for MAF; R^2_{cv} = 0.85, RMSEP = 3.17 °C for MST, n = 378) improves on the DC modelled MAF calibration performance of Martínez-Sosa et al. (2021) using MBT'$_{5ME}$ (2021; R^2 = 0.85, RMSE = 2.8 °C, n = 272) and the full set MAF calibration of Raberg et al. (2021, eqn (11); R^2 = 0.91, RMSE = 1.97 °C, n = 182). As highlighted in Section 3.1, application of these calibrations also performed less well than several of the numerical calibrations derived using our new expanded dataset.

4.2. Sources of model uncertainty

Most of the calibration models listed in Table 1 have R^2_{cv} values of c. 0.8–0.9, indicating that c. 0.1–0.2 of the variance in temperature is not accounted for by the models. This unexplained variance emanates from a combination of model error (Var_{mod}: i.e. how well MST or MAF encapsulates a physiologically meaningful driver for GDGT composition, and how well the numerical methods accurately model the potentially complex and non-linear response to this variable), the measurement bias in estimates of compound abundance (Var_{gdgt}), and the measurement bias in MST or MAF (Var_{temp}: i.e. the accuracy of the estimate of temperature composite and how well this reflects the temperature of the lake water environment). The total variance in the temperature data (Var_{tot}) can thus be decomposed into components representing the variance explained by the calibration model (Var_{expl}) and the three components of unexplained variance described above:

$$Var_{tot} = Var_{expl} + Var_{mod} + Var_{gdgt} + Var_{temp}$$

The lack of fit (Var_{mod}) represents the systematic variation in the temperature that could potentially be modelled by a more complex model or with additional compounds. The combination of Var_{gdgt} and Var_{temp} sets the upper limit on the variance possible to model (Nilsson et al., 1996). Values of Var_{gdgt} and Var_{temp} are not well constrained, but Var_{gdgt} can be estimated from inter-laboratory comparisons which suggests a standard deviation of c. 1.25 °C for MAF estimates derived from MBT'$_{5ME}$ (De Jonge et al., 2024). For Var_{temp} a standard deviation of 1.5 °C has been suggested for a similar calibration dataset (Naafs et al., 2017). Using these values for Var_{gdgt} and Var_{temp}, the same uncertainty in the estimates of brGDGT fractional abundance data in both SC and DC datasets, and in estimates of MST and MAF, these values give RF calibrations shown in Table 2.

Table 2
Explained variance (Var_{expl}), and variance attributed to model error (Var_{mod}), compound measurement error (Var_{gdgt}), and MST/MAF measurement error (Var_{temp}) for each dataset and temperature variable as percentage of total variance.

Dataset	Variable	Var_{expl}	Var_{mod}	Var_{gdgt}	Var_{temp}
SC	MST	83.4	10.7	2.4	3.4
SC	MAF	90.5	1.8	3.2	4.5
DC	MST	85.0	9.3	2.3	3.3
DC	MAF	86.4	4.9	3.5	5.1

Several observations are immediately apparent from these decompositions. First, Var_{gdgt} + Var_{temp} values average 5.7 % for MST and 8.2 % for MAF, setting an upper limit for the maximum possible R^2 values of c. 0.94 and 0.92 for MST and MAF models respectively. This suggests that the RF calibration for MAF in the SC dataset (R^2_{cv} = 0.91) is indeed accurately modelling all the features of the brGDGT-temperature relationships and that any further improvements in model performance, as judged by an increase in R^2_{cv} or reduction in RMSEP, are likely to be modest.

Second, modelling errors (Var_{mod}) associated with MST are substantially larger than those for MAF, suggesting that MAF better summarises the temperature of the growing season of brGDGT producers than MST. Martínez-Sosa et al. (2021) also found that MAF was a better predictor of the MBT'$_{5ME}$ index than mean annual air temperature (MAAT) suggesting that brGDGT producers are less active below freezing.

Finally, Var_{mod} for MAF is substantially lower in the SC dataset: indeed, all of the calibration models perform slightly better for the SC dataset, despite the DC dataset being constructed using the new improved chromatography method. It is possible that improved analytical advances can expose further complexities between GDGT compositions and environmental relationships with individual 5- and 6-methyl isomers. For example, studies have found 5-methyl brGDGTs to have no relationship with temperature in some Chinese lakes (Dang et al., 2018; Qian et al., 2019; Wang et al., 2012, 2021; Wu et al., 2023) which is markedly different from findings from East African (Russell et al., 2018), North American (Martínez-Sosa et al., 2021), pantropical lakes (Zhao et al., 2023), and some freshwater lakes (Wang et al., 2021) which do show a correlation between 5-methyl brGDGTs and

temperature. Conversely, 6-methyl brGDGTs have shown a significant correlation with temperature in some Chinese lakes (Dang et al., 2018; Qian et al., 2019), while in some central European lakes there appears to be a correlation between both 5- and 6-methyl isomers and temperature, especially in high elevation lakes (Bauersachs et al., 2024). Such variation in the relationships between 5-methyl and 6-methyl isomers and temperature suggests that they don't necessarily or consistently contribute to improving temperature calibrations and, since these compounds also often occur as only a small fractional abundance, they have been excluded from some calibration studies (e.g., Russell et al., 2018), as has also been the case for some compounds here (see Section 2.3). The discrepancies in relationships with temperature could also be due to the 5- and/or 6-methyl compound abundances being influenced by environmental variable/s in addition to temperature (e.g. salinity, pH, oxygen) in some lakes (Wang et al., 2021; Kou et al., 2022; Halamka et al., 2023).

4.3. Assessment of downcore reconstructions

The calibration models evaluated in the previous section differ in the compounds used, the implicit weight given to each compound, and model the brGDGT-temperature response with varying degrees of complexity. Overall, RF calibrations produced the smallest RMSEP, and Dirichlet regression the largest, but differences between most methods are small. Furthermore, choosing a "best" method is not straightforward, as calibration model performance, even under cross-validation, is not always a good guide to the robustness of down-core reconstructions (Juggins, 2013). This point has been recently reiterated by Sun et al. (2024) who recommend that the significance test developed by Telford and Birks (2011) be applied to confirm that a palaeoenvironmental reconstruction derived from proxy data is robust and potentially reliable. Unfortunately, this test has a high Type II error when the number of predictor variables is low (<10) and so cannot be used with our GDGT calibrations. Instead, we evaluate the reconstructions using measures of no-analogue conditions and concordance between methods. PCA plots of the surface calibration dataset and core samples and plots of Mahalanobis distances between calibration and core samples (Fig. 3) help visualise differences between surface and core samples and highlight no-analogue conditions, and provide an indication of how robust we might expect the application of the models to be for each core site. Accompanying boxplots of brGDGT abundance in calibration and core samples helps identify which compounds are responsible for the no-analogue conditions.

Finally, cluster analyses of the Pearson's product-moment correlation and Lin's concordance correlation matrices summarise the overall similarity in trends (correlation) and absolute values (concordance) among reconstructions (Fig. 4). Consistency in trends and/or absolute values among different methods suggest reconstructions may be more reliable.

Considering all reconstructions, two can be immediately dismissed. These are Pearson et al. (2011) and DR at Yanou Lake. The first produces reconstructions that are c. 4 °C higher than other methods, and substantially higher than the contemporary mean summer temperature of 1.1 ± 0.8 °C (Heredia Barión et al., 2023a). The tendency of the Pearson et al. (2011) method to overestimate temperatures at the low end of the gradient was noted in Foster et al. (2016) and Roberts et al. (2017) and this method is not recommended for reconstructions at polar and sub-polar sites. DR is the worst performing method in terms of RMSEP in the calibration data and although it produces reconstructions generally consistent with other methods at most sites considered here, at Yanou Lake it substantially underestimates modern MST. This is most likely because of extrapolation due to unusually high abundance of isomer fIIIa. A similar extrapolation in reconstructing unrealistic low MAF values is also observed at the Arctic site El'gygytgyn. As mentioned above, DR seems to be particularly poorly constrained at the low temperature end of the gradient and our DR calibration also cannot be recommended for cold sites.

Excluding these outlier reconstructions there are few consistent patterns for each calibration method across different lakes. For example, excluding the Pearson et al. (2011) and DR reconstructions, LR and MMA predict higher temperatures than other methods for Club Lake, White Pond and Basin Pond but lower temperatures for Lake Challa (Fig. 5, Supplementary Table S1). Similarly, reconstructions using LR and MMA have much lower variance than other methods for Club Lake and Lake Challa but have implausibly high variance for White Pond and El'gygytgyn. Likewise, reconstructions using NN have much higher

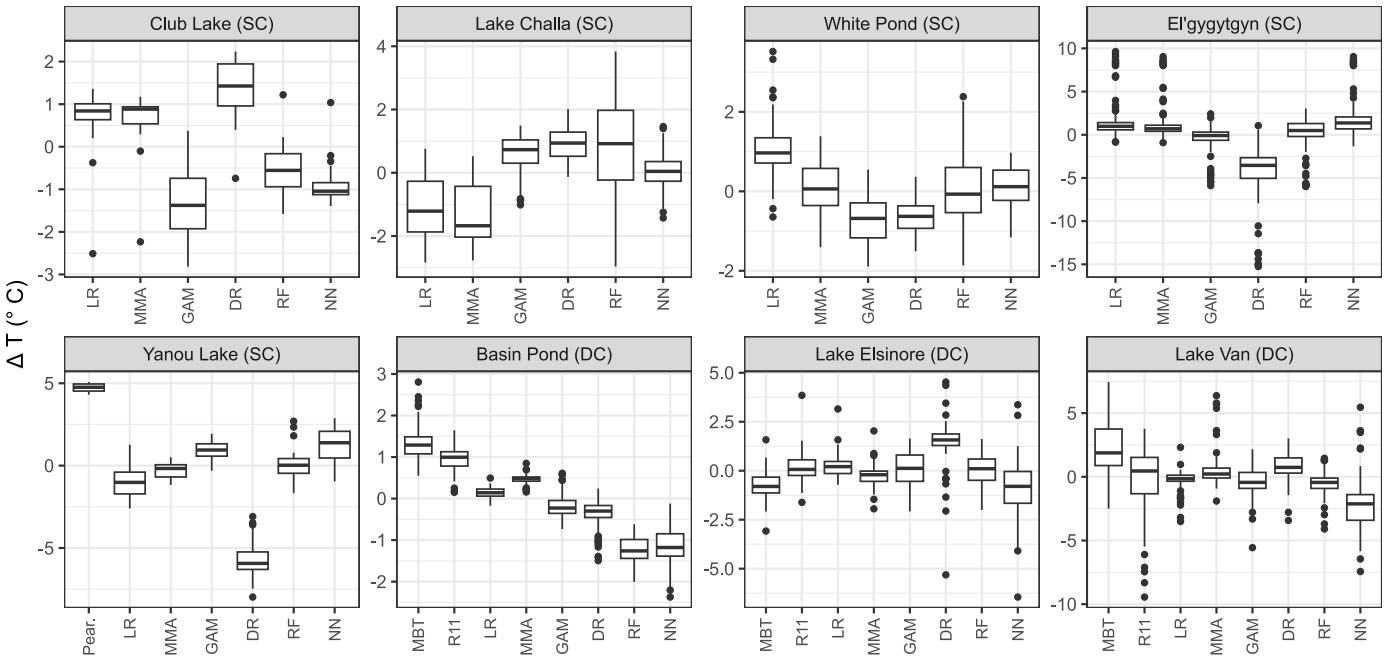

Fig. 5. Boxplots summarising the difference, ΔT (°C), between reconstructed temperature for each method and the consensus reconstruction (i.e. mean reconstructed temperature for all methods).

variance than other methods for El'gygytgyn and Lake Elsinore but have extremely low variance for Basin Pond and White Pond.

This lack of consistency makes it difficult to predict the behaviour of different methods at different sites and to identify those which may produce the most useful reconstructions. Despite this there are some overall patterns of correlation and concordance among reconstructions at the eight core sites which can be divided into three groups: those with *high correlation and high concordance* among reconstructions (Club Lake, Lake Chala, White Pond and Lake Elsinore), those with *high correlation but moderate concordance* among reconstructions (Basin Pond and El'gygytgyn), and those with *low correlation and low concordance* among reconstructions (Yanou Lake and Lake Van).

Three of the four sites in the first group (Club Lake, Lake Challa and Lake Elsinore) have brGDGT compositions that plot well within brGDGT calibration space, indicating that our expanded and improved surface calibration dataset should be appropriate for reconstructing past temperatures at these sites. Standard deviations among reconstructions are generally low (<2 °C) and only a few samples in Lake Elsinore have non-analogue compositions. There is generally high concordance among reconstructions for all sites in this group with the exception that LR and MMA show a consistent offset towards higher MAF in the Club Lake and White Pond records and lower MAF in the early and later parts of the record from Lake Challa, and RF reconstructions have higher variance than other methods for Lake Challa, driven by an oversensitivity to fluctuations in fIIIa. This compound is indicative of changes in redox (e.g. Yao et al., 2020) and in Lake Challa most of the brGDGT production has been found to occur in the anoxic zone (Van Bree et al., 2020). BrGDGT IIIa may also be responding to thermal stratification at this deep and permanently stratified site since more pronounced water-column stratification in turn affects the relative niche availability of different GDGT-producing microbes (Baxter et al., 2024). Our reconstructed MST and MAF at Club Lake show similar trends to those previously reported while the shifts in warming in the Lake Elsinore record correspond with multiproxy evidence reported by Feakins et al. (2019). The issues aside, the good fit of the core data to the calibration set and the general concordance between methods suggests that reconstructions for these sites are robust in terms of both trends and absolute values.

For the fourth site, White Pond, the Holocene part of the core plots outside the space of the calibration dataset, and large parts of the core are considered extreme outliers. Despite this, there is high concordance among methods, although LR and MMA exhibit much greater variance in the early part of the White Pond record than other methods. However, all methods reconstruct unfeasibly high temperatures for the pre-Holocene period and differ markedly from the original published reconstruction of Krause et al. (2018) that used the MBT calibration of Peterse et al. (2014). Krause et al. (2018) suggest that high BIT values at this site indicate a predominately terrestrial source for brGDGT inputs, which may account for the anomalously high abundances of isomers fIa and fIb which lead to all methods overestimating temperatures since brGDGT fIa is thought to be derived predominantly from catchment soil bacteria (Sinninghe-Damsté et al., 2000; Hopmans et al., 2004; Weijers et al., 2006a,b). Despite the high concordance among reconstructions the lack of fit of the core to calibration dataset suggests that reconstructions for this site are not reliable.

The second group of two sites have high correlation but only moderate concordance among reconstructions and differ in their fit to the calibration dataset. The first, Basin Pond has a good fit to the modern calibration dataset and there are no outliers. Like sites in the previous group, there is a high correlation between methods and all reconstruct similar trends, highlighting a gradual cooling from c. 600 cal a BP. However, concordance among methods is only moderate and there are substantial differences in the variability and magnitude of the reconstructed cooling trend. In their original publication Miller et al. (2018) applied Dang et al. (2018) and Russell et al. (2018) calibrations to White Pond and found that although they produced similar trends, they differed in variance and were offset by c. 4 °C for the whole record. They consequently advised caution when interpreting this record. Our results also suggest caution and while the overall reconstructed trends may be reliable, the absolute values and inferences about the magnitude of temperature changes are not.

For the second site in this group, El'gygytgyn, most of the core samples are considered extreme outliers and plot outside of the calibration data. With the exception of DR, reconstructions generally follow similar trends and there is moderate concordance among methods but this includes concordance in reconstructing spuriously high and low values, which are especially accentuated for LR and MMA. El'gygytgyn is located in Arctic Siberia and is an extremely large and deep lake. There are very few deep (>100m) lakes in our calibration dataset, and these are from Uganda and Tanzania and are not suitable analogues. In their original publication for this record, Keisling et al. (2017) used the MBT/CBT calibration of Sun et al. (2011) developed using lakes on the Tibetan Plateau. Their reconstruction shows similar trends and similar overall absolute values to ours (excluding DR), except our reconstructions have more variance and include several spuriously high and low values: our methods appear to be very sensitive to extrapolation under no-analogue conditions at this site and while the reconstructed trends may be reliable the absolute values should be treated with caution.

The final group of sites (Lake Van and Yanou Lake) exhibit both low correlation and low concordance among reconstructions and also differ in their fit to the calibration data. Yanou Lake samples plot on the periphery of the calibration dataset and a few samples from the middle of the record are considered outliers. The range of MST reconstructions for individual methods is relatively small, and different methods show different trends: LR and MMA suggest a warmer period between c. 3000-2000 cal a BP, while RF, GAM and NN suggest a relatively stable MAF from 4500 cal a BP. The variance in reconstructed MAF is noticeably different for different methods and concordance for Pearson et al. (2011) and DR is particularly low, with reconstructions c. 4 °C higher and c. 4 °C lower, respectively, than other methods. Discounting DR and Pearson et al. (2011) models at Yanou Lake, the trends in the remaining reconstructions differ in both magnitude and detail. Although only a few samples from this site are considered outliers, most of the core does have unusually high abundance of fIIIa. Notably, a high abundance of brGDGT fIIIa compound is a characteristic feature in Antarctic lakes (Pearson et al., 2011) and the higher abundance (mean c.60 %) in the Yanou Lake core samples than in the global surface sample training set (mean c. 30 %) reflects this. Redox influences on brGDGT fIIIa, with bacteria with lower oxygen requirements producing more abundant fIIIa (Weber et al., 2018; Yao et al., 2020), may play an important role in such lakes. Moreover, Antarctic lakes, such as Yanou Lake, are often characterised by subaquatic mosses which may play an important and unique role in brGDGT bacterial composition, and reflect the complex interplay between sources and processes in polar and cold-region lakes. Yanou Lake is a polar site that has likely undergone only small changes in temperature over the last 6000 years. The range of reconstructed temperatures is consequently relatively small, and reconstructions appear to be masked or confounded by the increased uncertainty caused by the unusual brGDGT compositions.

As with Yanou Lake, reconstructions for Lake Van display an overall similarity in gross trends but the detail reveals a pattern of very low correlation and concordance among methods with reconstructions showing high variance and some spurious high and low values. Although core samples from Lake Van plot within the first two PCA axes of the calibration dataset they have unusually high abundance of isomers fIIa', fIIb' and fIIIa' and are classified as extreme outliers using a Mahalanobis distance criterion. R11, MBT and NN in particular extrapolate poorly and reconstruct unfeasibly high or low temperatures under these conditions (Fig. 3, S4). Lake Van is a high elevation (1650m), seasonally stratified, endorheic lake. It is the largest soda lake in the world with alkaline waters reaching a pH of 9.8 and a salinity of 22 psμ (Stockhecke

et al., 2021). The unusual chemistry, limnology and stratification regime of this site may account for the no-analogue brGDGT compositions and high abundance of the specific 6-methyl isomers and consequent lack of coherence among methods.

In summary, as highlighted above, explanations for why some model reconstructions perform poorly include unusually high or low abundances of specific compounds, changes in source inputs, bacterial communities and associated environment (e.g. water depth) and chemistry (e.g. redox, salinity), limits of specific statistical methods or sources of model uncertainty. We evaluate the reconstructions using measures of no-analogue conditions and concordance between methods and, apart from the caveats about using DR at cold sites, there are no consistent patterns in the reconstructions that would allow us to recommend one or more methods that produce "better" or more reliable reconstructions. However, we find that a comparison of reconstructions at a site, and the concordance, along with measures of fit between the core and calibration brGDGT compositions can be used to identify reconstructions that may be reliable or not.

Sites with good fit to the calibration data and high concordance among methods (Club Lake, Lake Challa and Lake Elsinore) have reconstructions that are reliable and may be interpreted in terms of both trends and absolute values. Sites with a good fit but moderate or low concordance among methods (Basin Pond, Yanou Lake) have reconstructions that may be interpreted in terms of trends but not absolute values. Calibrations for sites with a poor fit to the calibration set (White Pond, El'gygytgyn and Lake Van) are problematic. For Lake Van different calibrations extrapolate in different ways under no-analogue conditions, leading to incoherent and spurious reconstructions. At White Pond and El'gygytgyn different methods extrapolate in a generally coherent fashion to produce similar trends but differ substantially in variance. At White Pond they produce unreliable reconstructions with implausibly high temperatures and trends but at El'gygytgyn the trends are plausible.

Our comparisons show that reconstructing palaeotemperature using branched GDGTs is challenging. To paraphrase the statistician George E. P. Box, "*all reconstructions are wrong, but some are useful*". We find that there is no single "best" numerical method and one chosen using criteria of highest cross-validated R^2 and lowest RMSEP may not give the most reliable reconstruction. Different methods emphasise different properties of the brGDGT data and extrapolate in different ways, especially under non-analogue conditions. Rather we show that it is important to consider both the fit of the core data to the calibration set to identify no-analogue problems (also noting that these relationships may be different for different parts of the core), and to inspect the concordance among different methods as a guide to identify which reconstructions may be useful.

5. Summary, conclusions and recommendations

We used new expanded calibration datasets developed using single column (SC) and dual column (DC) HPLC-MS analytical methods to construct global brGDGT-temperature calibrations for application to lakes downcore across a range of environments and Quaternary timescales. We evaluated a range of different statistical modelling approaches and applied them to a number of published downcore brGDGT records. Our study substantially expands on and improves existing SC and DC calibration datasets, including expanding the colder, Antarctic end of the SC method and including Antarctic sites in what is the first truly global DC brGDGT-temperature dataset. Our results show that brGDGT distributions in lacustrine sediments are correlated with temperature on a global scale, and show a robust brGDGT-temperature relationship across at least the range of c.-2 °C to c. + 31 °C covered in this study. By examining both SC and DC methods we have also contributed to the much-needed movement to standardise across methodologies in an interdisciplinary space and to keep older methods and data relevant, and demonstrate that both SC and DC calibrations can be used for temperature reconstructions. By comparing and assessing model performance and subsequent application downcore using a range of statistical tools we demonstrate a rigorous approach to assess the robustness and applicability of our calibration models and which is an approach advised for use in studies going forwards.

Of the statistical methods tested, Dirichlet regression was the worst performing method in both SC and DC datasets, despite it being statistically the most appropriate method for modelling the brGDGT compositional data (Douma et al., 2019). The application of the published SC calibration of Pearson et al. (2011) and DC calibration based on MBT'_{5ME} of Martínez-Sosa et al. (2021) also have higher RMSEP for MST and MAF respectively than the new calibrations. Machine learning methods (random forests and neural networks) slightly outperform the regression-based approaches (linear and quadratic regression, GAMs, and multi-model averaging). Overall, random forests (RF) gave the highest cross-validated R^2 and lowest RMSEP of all models constructed to estimate MAF and MST in global lakes when using both the SC and the DC global calibration datasets. Importantly, however, overall differences in prediction error between LR, MMA, GAM, RF and NN methods are small (0.02–0.5 °C), while the worst performing methods (DR, MBT, R11) have cross-validation prediction errors of c 3.0–4.5 °C which compare favourably with some other previously reported calibrations. Examination of model uncertainty was carried out by decomposing the total variance in the temperature data into the three components of unexplained variance (errors in model, errors in estimates of compound abundance, and errors in estimates in temperature data). Our results suggest that the best models are accurately modelling all the features of the brGDGT-temperature relationships and that any further improvements in model performance, as judged by an increase in R^2_{cv} or reduction in the RMSEP, are likely to be modest.

Our findings importantly demonstrate that, while a calibration R^2_{cv} and RMSEP may suggest good model performance, even under rigorous cross validation, a more thorough assessment of relationships between surface sample calibration dataset and core samples and an assessment of applicability to a specific given site is required to identify reliable reconstructions. To do this we recommend the use of exploratory statistical analyses including PCA, and boxplots of compound abundance in the calibration and core data, and the use of Mahalanobis distances between calibration and core data to identify core-samples that have unusual composition and lack analogues in the calibration data, and identify specific sections of the core that might be problematic, noting that a reconstruction may or may not be reliable throughout the whole of a specific core due to changes in, for example, source inputs or environment.

We also used the correlation and concordance between records to summarise the similarity in trends (correlation) and absolute values (concordance) among reconstructions, using consistency in trends and/or absolute values among different methods as a tool to suggest which reconstructions may be more reliable. Examples from different sites from contrasting Quaternary environments and timescales highlight reconstructions which demonstrate robustness and those which should be treated with caution. Cores that have good analogues and a high correlation and concordance between models likely produce the most useful reconstructions, and those with a poor fit to the calibration data and with low correlation and low concordance between different methods are more problematic and require caution.

Which compounds are driving the reconstructions can be indicative of changes in, for example, brGDGT source provenance/inputs, environmental conditions or brGDGT bacterial communities. Differences in individual brGDGT compound compositions between core and surface dataset samples, as highlighted, can (and should) therefore be examined to explore and highlight what might be driving discrepancies between model outputs downcore.

While we have focused here on temperature, with improved separation of 5- and 6-methyl isomers there are additional variables that could be considered in future studies such as salinity, pH, nutrients,

oxygen, or other confounding variables, where such data is available. Such data is usually lacking and is not consistently available in global datasets derived from multiple sources and this was not possible in this study. Studies to investigate relationships between brGDGTs and other environmental variables on a global scale to improve our understanding of the influence of confounding factors on brGDGT-temperature relationships in different environments still very much remains beyond the scope of large-scale global calibration studies.

We recommend the combination of tools and approaches we have used in our study to assess and address challenges in calibration studies in relation to identifying non-analogue conditions and the suitability of a particular calibration to a specific core site. Furthermore, reporting results from these approaches taken, and of the different calibrations, are recommended as a standard approach when reporting GDGT-based temperatures and to support their robustness, applicability and reliability. Such approaches to consider, while highlighted here for brGDGT-temperature calibrations applicable to lakes, are also applicable and significant for other proxy (e.g. pollen, chironomids, salinity, nutrients) calibration development studies and applications (e.g. to peat, marine sediments, soils) in Quaternary science.

Author contributions

EJP: Conceptualization; Formal analysis; Funding acquisition; Investigation; Methodology; Project administration; Resources; Supervision; Validation; Visualization; Writing - original draft; Writing - review & editing.

SJ: Conceptualization; Data curation; Software; Formal analysis; Funding acquisition; Investigation; Methodology; Resources; Supervision; Validation; Visualization; Writing - original draft; Writing - review & editing.

LCF: Investigation; Writing - review & editing.

HA: Investigation; Writing - review & editing.

DAH: Funding acquisition; Resources; Supervision; Writing - review & editing.

BDAN: Funding acquisition; Resources; Supervision; Writing - review & editing.

TP: Resources; Writing - review & editing.

SJR: Funding acquisition; Resources; Supervision; Writing - review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The Antarctic samples (SC method) were analysed as part of the British Antarctic Survey (BAS) Natural Environment Research Council (NERC) funded Science Program, with analyses funded by the British Antarctic Survey, and Newcastle University Faculty Research Fund (to EJP) and by NERC Studentship NE/J500173/1 to LCF (BAS and Newcastle University; supervisors SJR, EJP, DAH, SJ). The Antarctic samples were re-analysed (DC method) as part of NERC funded Life science Mass Spectrometry Facility (now National Environmental Isotope Facility) grant LSMSF BRIS/126/1518 (to SJR, EJP, SJ, DAH, BDAN) with analyses performed by HA. Results within this project contain modified Copernicus Climate Change Service information [2020, 2021]. Neither the European Commission nor the European Centre for Medium-range Weather Forecasts (ECMWF) is responsible for any use that may be made of the Copernicus information or data it contains.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.quascirev.2025.109615.

Data availability

Data and code associated with this article will be made available on GitHub (https://github.com/) and the Antarctic datasets at the NERC UK Polar Data Centre (https://www.bas.ac.uk/data/uk-pdc/).

References

Aggarwal, C., 2017. Outlier Analysis, second ed. Springer, Cham, p. 466.
Allaire, J., Chollet, F., 2024. Keras: R interface to 'keras'. R package version 2.15.0. https://CRAN.R-project.org/package=keras.
Allaire, J., Tang, Y., 2024. Tensorflow: R interface to 'TensorFlow'. R package version 2.16.0. https://CRAN.R-project.org/package=tensorflow.
Bartoń, K., 2024. MuMIn: multi-model inference. R package version 1.48.4. https://CRAN.R-project.org/package=MuMIn.
Bauersachs, T., Schubert, C.J., Mayr, C., Gilli, A., Schwark, L., 2024. Branched GDGT-based temperature calibrations from Central European lakes. Sci. Total Environ. 906, 167724. https://doi.org/10.1016/j.scitotenv.2023.167724.
Baxter, A.J., Peterse, F., Verschuren, D., Sinninghe Damsté, J.S., 2024. Assessment of branched glycerol monoalkyl glycerol tetraether (brGMGT)-based paleothermometry in the 250,000-year sediment record of Lake Chala, equatorial East Africa. Org. Geochem. 195. https://doi.org/10.1016/j.orggeochem.2024.104812.
Baxter, A.J., Verschuren, D., Peterse, F., Miralles, D.G., Martin-Jones, C.M., Maitituerdi, A., Van der Meeren, T., Van Daele, M., Lane, C.S., Haug, G.H., Olago, D. O., Sinninghe Damste, J.S., 2023. Reversed Holocene temperature-moisture relationship in the Horn of Africa. Nature 620, 336–343. https://doi.org/10.1038/s41586-023-06272-5.
Cao, J., Rao, Z., Shi, F., Jia, G., 2020. Ice formation on lake surfaces in winter causes warm-season bias of lacustrine brGDGT temperature estimates. Biogeosciences 17, 2521–2536. https://doi.org/10.5194/bg-17-2521-2020.
Chen, Y., Zheng, F., Yang, H., Yang, W., Wu, R., Liu, X., Liang, H., Chen, H., Pei, H., Zhang, C., Pancost, R.D., Zeng, Z., 2022. The production of diverse brGDGTs by an Acidobacterium providing a physiological basis for paleoclimate proxies. Geochem. Cosmochim. Acta 337, 155–165. https://doi.org/10.1101/2022.04.07.487437.
Dang, X., Ding, W., Yang, H., Pancost, R.D., Naafs, B.D.A., Xue, J., Lin, X., Lu, J., Xie, S., 2018. Different temperature dependence of the bacterial brGDGT isomers in 35 Chinese lake sediments compared to that in soils. Org. Geochem. 119, 72–79. https://doi.org/10.1016/j.orggeochem.2018.02.008.
Dang, X., Xue, J., Yang, H., Xie, S., 2016. Environmental impacts on the distribution of microbial tetraether lipids in Chinese lakes with contrasting pH: implications for lacustrine paleoenvironmental reconstructions. Sci. China Earth Sci. 59, 939–950. https://doi.org/10.1007/s11430-015-5234-z.
De Jonge, C., Hopmans, E.C., Stadnitskaia, A., Rijpstra, W.I.C., Hofland, R., Tegelaar, E., Sinninghe Damsté, J.S., 2013. Identification of novel penta- and hexamethylated branched glycerol dialkyl glycerol tetraethers in peat using HPLC–MS2, GC–MS and GC–SMB-MS. Org. Geochem. 54, 78–82. https://doi.org/10.1016/j.orggeochem.2012.10.004.
De Jonge, C., Peterse, F., Nierop, K.G.J., Blattmann, T.M., Alexandre, M., Ansanay-Alex, S., Austin, T., Babin, M., Bard, E., Bauersachs, T., Blewett, J., Boehman, B., Castañeda, I.S., Chen, J., Conti, M.L.G., Contreras, S., Cordes, J., Davtian, N., van Dongen, B., Duncan, B., Elling, F.J., Galy, V., Gao, S., Hefter, J., Hinrichs, K.U., Helling, M.R., Hoorweg, M., Hopmans, E., Hou, J., Huang, Y., Huguet, A., Jia, G., Karger, C., Keely, B.J., Kusch, S., Li, H., Liang, J., Lipp, J.S., Liu, W., Lu, H., Mangelsdorf, K., Manners, H., Martinez Garcia, A., Menot, G., Mollenhauer, G., Naafs, B.D.A., Naeher, S., O'Connor, L.K., Pearce, E.M., Pearson, A., Rao, Z., Rodrigo-Gámiz, M., Rosendahl, C., Rostek, F., Bao, R., Sanyal, P., Schubotz, F., Scott, W., Sen, R., Sluijs, A., Smittenberg, R., Stefanescu, I., Sun, J., Sutton, P., Tierney, J., Tejos, E., Villanueva, J., Wang, H., Werne, J., Yamamoto, M., Yang, H., Zhou, A., 2024. Interlaboratory comparison of branched GDGT temperature and pH proxies using soils and lipid extracts. G-cubed 25. https://doi.org/10.1029/2024gc011583.
De Jonge, C., Stadnitskaia, A., Hopmans, E.C., Cherkashov, G., Fedotov, A., Sinninghe Damsté, J.S., 2014. In situ produced branched glycerol dialkyl glycerol tetraethers in suspended particulate matter from the Yenisei River, Eastern Siberia. Geochem. Cosmochim. Acta 125, 476–491. https://doi.org/10.1016/j.gca.2013.10.031.
Douma, J.C., Weedon, J.T., Warton, D., 2019. Analysing continuous proportions in ecology and evolution: a practical introduction to beta and Dirichlet regression. Methods Ecol. Evol. 10, 1412–1430. https://doi.org/10.1111/2041-210x.13234.
Feakins, S.J., Wu, M.S., Ponton, C., Tierney, J.E., 2019. Biomarkers reveal abrupt switches in hydroclimate during the last glacial in southern California. Earth Planet Sci. Lett. 515, 164–172. https://doi.org/10.1016/j.epsl.2019.03.024.
Foster, L.C., Pearson, E.J., Juggins, S., Hodgson, D.A., Saunders, K.M., Verleyen, E., Roberts, S.J., 2016. Development of a regional glycerol dialkyl glycerol tetraether (GDGT)–temperature calibration for Antarctic and sub-Antarctic lakes. Earth Planet Sci. Lett. 433, 370–379. https://doi.org/10.1016/j.epsl.2015.11.018.

Günther, F., Thiele, A., Gleixner, G., Xu, B., Yao, T., Schouten, S., 2014. Distribution of bacterial and archaeal ether lipids in soils and surface sediments of Tibetan lakes: implications for GDGT-based proxies in saline high mountain lakes. Org. Geochem. 67, 19–30. https://doi.org/10.1016/j.orggeochem.2013.11.014.

Häggi, C., Naafs, B.D.A., Silvestro, D., Bertassoli, D.J., Akabane, T.K., Mendes, V.R., Sawakuchi, A.O., Chiessi, C.M., Jaramillo, C.A., Feakins, S.J., 2023. GDGT distribution in tropical soils and its potential as a terrestrial paleothermometer revealed by Bayesian deep-learning models. Geochem. Cosmochim. Acta 362, 41–64. https://doi.org/10.1016/j.gca.2023.09.014.

Hahn, G.J., 1977. The hazards of extrapolation in regression analysis. J. Qual. Technol. 9, 159–165. https://doi.org/10.1080/00224065.1977.11980791.

Halamka, T.A., Raberg, J.H., McFarlin, J.M., Younkin, A.D., Mulligan, C., Liu, X.L., Kopf, S.H., 2023. Production of diverse brGDGTs by Acidobacterium *Solibacter usitatus* in response to temperature, pH, and O(2) provides a culturing perspective on brGDGT proxies and biosynthesis. Geobiology 21, 102–118. https://doi.org/10.1111/gbi.12525.

Heredia Barión, P., Roberts, S.J., Spiegel, C., Binnie, S.A., Wacker, L., Davies, J., Gabriel, I., Jones, V.J., Blockley, S., Pearson, E.J., Foster, L., Davies, S.J., Roland, T. P., Hocking, E.P., Bentley, M.J., Hodgson, D.A., Hayward, C.L., McCulloch, R.D., Strelin, J.A., Kuhn, G., 2023a. Holocene deglaciation and glacier readvances on the Fildes Peninsula and King George Island (Isla 25 de Mayo), South Shetland Islands, NW Antarctic Peninsula. Holocene 33, 636–658. https://doi.org/10.1177/09596836231157059.

Heredia Barión, P.A., Strelin, J.A., Roberts, S.J., Spiegel, C., Wacker, L., Niedermann, S., Bentley, M.J., Pearson, E.J., Czalbowski, N.T.M., Davies, S.J., Schnetger, B., Grosjean, M., Arcusa, S., Perren, B., Hocking, E.P., Kuhn, G., 2023b. The impact of Holocene deglaciation and glacial dynamics on the landscapes and geomorphology of potter Peninsula, king George island (Isla 25 Mayo), NW Antarctic Peninsula. Front. Earth Sci. 10, 1073075. https://doi.org/10.3389/feart.2022.1073075.

Hersbach, H., Bell, B., Berrisford, P., Biavati, G., Horányi, A., Muñoz Sabater, J., Nicolas, J., Peubey, C., Radu, R., Rozum, I., Schepers, D., Simmons, A., Soci, C., Dee, D., Thépaut, J.-N., 2019. ERA5 monthly averaged data on single levels from 1979 to present. Copernicus Climate Change Service (C3S) Climate Data Store (CDS). https://doi.org/10.24381/cds.f17050d7 (Accessed on 08-sep-2020, 29-Mar-2021).

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., Nicolas, J., Peubey, C., Radu, R., Schepers, D., Simmons, A., Soci, C., Abdalla, S., Abellan, X., Balsamo, G., Bechtold, P., Biavati, G., Bidlot, J., Bonavita, M., De Chiara, G., Dahlgren, P., Dee, D., Diamantakis, M., Dragani, R., Flemming, J., Forbes, R., Fuentes, M., Geer, A., Haimberger, L., Healy, S., Hogan, R.J., Hólm, E., Janisková, M., Keeley, S., Laloyaux, P., Lopez, P., Lupu, C., Radnoti, G., de Rosnay, P., Rozum, I., Vamborg, F., Villaume, S., Thépaut, J.N., 2020. The ERA5 global reanalysis. Q. J. R. Meteorol. Soc. 146, 1999–2049. https://doi.org/10.1002/qj.3803.

Hopmans, E., Schouten, S., Pancost, R.D., van der Meer, M.T.J., Sinninghe Damste, J.S., 2000. Analysis of intact tetraether lipids in archaeal cell material and sediments by high performance liquid chromatography/atmospheric pressure chemical ionization mass spectrometry. Rapid Commun. Mass Spectrom. 14, 585–589. https://doi.org/10.1002/(SICI)1097-0231(20000415)14:7%3C585::AID-RCM913%3E3.0.CO;2-N.

Hopmans, E.C., Schouten, S., Sinninghe Damsté, J.S., 2016. The effect of improved chromatography on GDGT-based palaeoproxies. Org. Geochem. 93, 1–6. https://doi.org/10.1016/j.orggeochem.2015.12.006.

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

Juggins, S., 2013. Quantitative reconstructions in palaeolimnology: new paradigm or sick science? Quat. Sci. Rev. 64, 20–32. https://doi.org/10.1016/j.quascirev.2012.12.014.

Kaiser, J., Schouten, S., Kilian, R., Arz, H.W., Lamy, F., Sinninghe Damsté, J.S., 2015. Isoprenoid and branched GDGT-based proxies for surface sediments from marine, fjord and lake environments in Chile. Org. Geochem. 89–90, 117–127. https://doi.org/10.1016/j.orggeochem.2015.10.007.

Keisling, B.A., Castañeda, I.S., Brigham-Grette, J., 2017. Hydrological and temperature change in Arctic Siberia during the intensification of Northern Hemisphere Glaciation. Earth Planet Sci. Lett. 457, 136–148. https://doi.org/10.1016/j.epsl.2016.09.058.

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

Krause, T.R., Russell, J.M., Zhang, R., Williams, J.W., Jackson, S.T., 2018. Late quaternary vegetation, climate, and fire history of the Southeast Atlantic Coastal plain based on a 30,000-yr multi-proxy record from white Pond, South Carolina, USA. Quat. Res. 91, 861–880. https://doi.org/10.1017/qua.2018.95.

Lei, Y., Strong, D.J., Caballero, M., Correa-Metrio, A., Pérez, L., Schwalb, A., Macario-González, L., Cohuo, S., Lozano-García, S., Ortega-Guerrero, B., Werne, J.P., 2023. Regional vs. global temperature calibrations for lacustrine BrGDGTs in the North American (sub)tropics: implications for their application in paleotemperature reconstructions. Org. Geochem. 184, 104660. https://doi.org/10.1016/j.orggeochem.2023.104660.

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

Lin, L.I.K., 1989. A concordance correlation coefficient to evaluate reproducibility. Biometrics 45, 255–268. https://doi.org/10.2307/2532051.

Lloyd, C.T., Iwig, D.F., Wang, B., Cossu, M., Metcalf, W.W., Boal, A.K., Booker, S.J., 2022. Discovery, structure and mechanism of a tetraether lipid synthase. Nature 609, 197–203. https://doi.org/10.1038/s41586-022-05120-2.

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

Loomis, S.E., Russell, J.M., Ladd, B., Street-Perrott, F.A., Sinninghe Damsté, J.S., 2012. Calibration and application of the branched GDGT temperature proxy on East African lake sediments. Earth Planet Sci. Lett. 357–358, 277–288. https://doi.org/10.1016/j.epsl.2012.09.031.

Maier, M.J., 2014. DirichletReg: Dirichlet Regression for Compositional Data in R., Research Report Series/Department of Statistics and Mathematic. University of Economics and Business, WU Vienna. https://doi.org/10.57938/ad3142d3-2fcd-4c37-aec6-8e0bd7d077e1. Vienna.

Martínez-Sosa, P., Tierney, J.E., Meredith, L.K., 2020. Controlled lacustrine microcosms show a brGDGT response to environmental perturbations. Org. Geochem. 145, 1–8. https://doi.org/10.1016/j.orggeochem.2020.104041.

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

Miller, D.R., Habicht, M.H., Keisling, B.A., Castañeda, I.S., Bradley, R.S., 2018. A 900-year New England temperature reconstruction from in situ seasonally produced branched glycerol dialkyl glycerol tetraethers (brGDGTs). Clim. Past 14, 1653–1667.

Naafs, B.D.A., Gallego-Sala, A.V., Inglis, G.N., Pancost, R.D., 2017. Refining the global branched glycerol dialkyl glycerol tetraether (brGDGT) soil temperature calibration. Org. Geochem. 106, 48–56. https://doi.org/10.1016/j.orggeochem.2017.01.009.

Naafs, B.D.A., Oliveira, A.S.F., Mulholland, A.J., 2021. Molecular dynamics simulations support the hypothesis that the brGDGT paleothermometer is based on homeoviscous adaptation. Geochem. Cosmochim. Acta 312, 44–56. https://doi.org/10.1016/j.gca.2021.07.034.

Nilsson, M.B., Dabakk, E., Korsman, T., Renberg, I., 1996. Quantifying relationships between near-infrared reflectance spectra of lake sediments and water chemistry. Environ. Sci. Technol. 30, 2586–2590. https://doi.org/10.1021/es950953a.

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

Oksanen, J., 2024. Vegan: community ecology package. R package version 2, 6–8. https://doi.org/10.32614/CRAN.package.vegan. http://CRAN.R-project.org/package=vegan.

Pearson, E.J., Juggins, S., Talbot, H.M., Weckström, J., Rosén, P., Ryves, D.B., Roberts, S. J., Schmidt, R., 2011. A lacustrine GDGT-temperature calibration from the Scandinavian Arctic to Antarctic: renewed potential for the application of GDGT-paleothermometry in lakes. Geochem. Cosmochim. Acta 75, 6225–6238. https://doi.org/10.1016/j.gca.2011.07.042.

Peterse, F., Vonk, J.E., Holmes, R.M., Giosan, L., Zimov, N., Eglinton, T.I., 2014. Branched glycerol dialkyl glycerol tetraethers in Arctic lake sediments: sources and implications for paleothermometry at high latitudes. J. Geophys. Res.: Biogeosciences 119, 1738–1754. https://doi.org/10.1002/2014jg002639.

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

R Core Team, 2024. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. Version 4.4.1. http://www.R-project.org/.

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

Roberts, S.J., Monien, P., Foster, L.C., Loftfield, J., Hocking, E.P., Schnetger, B., Pearson, E.J., Juggins, S., Fretwell, P., Ireland, L., Ochyra, R., Haworth, A.R., Allen, C.S., Moreton, S.G., Davies, S.J., Brumsack, H.J., Bentley, M.J., Hodgson, D.A., 2017. Past penguin colony responses to explosive volcanism on the Antarctic Peninsula. Nat. Commun. 8, 14914. https://doi.org/10.1038/ncomms14914.

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

Shanahan, T.M., Hughen, K.A., Van Mooy, B.A.S., 2013. Temperature sensitivity of branched and isoprenoid GDGTs in Arctic lakes. Org. Geochem. 64, 119–128. https://doi.org/10.1016/j.orggeochem.2013.09.010.

Sinninghe Damsté, J.S., Hopmans, E.C., Pancost, R.D., Schouten, S., Geenevasen, J.A.J., 2000. Newly discovered non-isoprenoid glycerol dialkyl glycerol tetraether lipids in sediments. Chem. Commun. 1683–1684. https://doi.org/10.1039/b004517i.

Sinninghe Damsté, J.S., Ossebaar, J., Schouten, S., Verschuren, D., 2012. Distribution of tetraether lipids in the 25-ka sedimentary record of Lake Challa: extracting reliable TEX86 and MBT/CBT palaeotemperatures from an equatorial African lake. Quat. Sci. Rev. 50, 43–54. https://doi.org/10.1016/j.quascirev.2012.07.001.

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

Stockhecke, M., Bechtel, A., Peterse, F., Guillemot, T., Schubert, C.J., 2021. Temperature, precipitation, and vegetation changes in the Eastern Mediterranean over the last deglaciation and Dansgaard-Oeschger events. Palaeogeogr. Palaeoclimatol. Palaeoecol. 577, 110535. https://doi.org/10.1016/j.palaeo.2021.110535.

Sun, P., Holden, P.B., Birks, H.J.B., 2024. Can machine learning algorithms improve upon classical palaeoenvironmental reconstruction models? Clim. Past 20, 2373–2398. https://doi.org/10.5194/cp-2023-69.

Sun, Q., Chu, G., Liu, M., Xie, M., Li, S., Ling, Y., Wang, X., Shi, L., Jia, G., Lü, H., 2011. Distributions and temperature dependence of branched glycerol dialkyl glycerol tetraethers in recent lacustrine sediments from China and Nepal. J. Geophys. Res. 116, G01008. https://doi.org/10.1029/2010jg001365.

Telford, R.J., Birks, H.J.B., 2011. A novel method for assessing the statistical significance of quantitative reconstructions inferred from biotic assemblages. Quat. Sci. Rev. 30, 1272–1278. https://doi.org/10.1016/j.quascirev.2011.03.002.

Thomas, Z.A., Mooney, S., Cadd, H., Baker, A., Turney, C., Schneider, L., Hogg, A., Haberle, S., Green, K., Weyrich, L.S., Perez, V., Moore, N.E., Zawadzki, A., Kelloway, S.J., Khan, S.J., 2022. Late Holocene climate anomaly concurrent with fire activity and ecosystem shifts in the eastern Australian Highlands. Sci. Total Environ. 802, 149542. https://doi.org/10.1016/j.scitotenv.2021.149542.

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

van Bree, L.G.J., Peterse, F., Baxter, A.J., De Crop, W., van Grinsven, S., Villanueva, L., Verschuren, D., Sinninghe Damsté, J.S., 2020. Seasonal variability and sources of in situ brGDGT production in a permanently stratified African crater lake. Biogeosciences 17, 5443–5463. https://doi.org/10.5194/bg-17-5443-2020.

Varmuza, K., Filzmoser, P., 2009. Introduction to Multivariate Statistical Analysis in Chemometrics. CRC Press, Boca Raton, p. 336. https://doi.org/10.1201/9781420059496.

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

Wang, H., Liu, W., He, Y., Zhou, A., Zhao, H., Liu, H., Cao, Y., Hu, J., Meng, B., Jiang, J., Kolpakova, M., Krivonogov, S., Liu, Z., 2021. Salinity-controlled isomerization of lacustrine brGDGTs impacts the associated $MBT_{5ME'}$ terrestrial temperature index. Geochem. Cosmochim. Acta 305, 33–48. https://doi.org/10.1016/j.gca.2021.05.004.

Wang, H., Liu, W., Zhang, C.L., Wang, Z., Wang, J., Liu, Z., Dong, H., 2012. Distribution of glycerol dialkyl glycerol tetraethers in surface sediments of Lake Qinghai and surrounding soil. Org. Geochem. 47, 78–87. https://doi.org/10.1016/j.orggeochem.2012.03.008.

Weber, Y., Sinninghe Damste, J.S., Zopfi, J., De Jonge, C., Gilli, A., Schubert, C.J., Lepori, F., Lehmann, M.F., Niemann, H., 2018. Redox-dependent niche differentiation provides evidence for multiple bacterial sources of glycerol tetraether lipids in lakes. Proceedings of the National Academy of Sciences USA 115, 10926–10931. https://doi.org/10.1073/pnas.1805186115.

Weijers, J.W., Schouten, S., Hopmans, E.C., Geenevasen, J.A., David, O.R., Coleman, J. M., Pancost, R.D., Sinninghe Damste, J.S., 2006a. Membrane lipids of mesophilic anaerobic bacteria thriving in peats have typical archaeal traits. Environ. Microbiol. 8, 648–657. https://doi.org/10.1111/j.1462-2920.2005.00941.x.

Weijers, J.W.H., Schouten, S., Spaargaren, O.C., Sinninghe Damsté, J.S., 2006b. Occurrence and distribution of tetraether membrane lipids in soils: implications for the use of the TEX86 proxy and the BIT index. Org. Geochem. 37, 1680–1693. https://doi.org/10.1016/j.orggeochem.2006.07.018.

Weijers, J.W.H., Schouten, S., van den Donker, J.C., Hopmans, E.C., Sinninghe Damsté, J. S., 2007. Environmental controls on bacterial tetraether membrane lipid distribution in soils. Geochem. Cosmochim. Acta 71, 703–713. https://doi.org/10.1016/j.gca.2006.10.003.

Wood, S.N., 2017. Generalized Additive Models: an Introduction with R, second ed. Chapman and Hall/CRC, p. 496.

Wright, M.N., Ziegler, A., 2017. Ranger: a fast implementation of random forests for high dimensional data in C++ and R. J. Stat. Software 77 (1), 1–17. https://doi.org/10.18637/jss.v077.i01.

Wu, J., Yang, H., Shen, C., Zhu, L., Pei, H., Dang, X., Huang, M., Xie, S., 2023. BrGDGT-based quantitative reconstructions of paleotemperature in lakes: regional vs. site-specific calibrations. Quat. Sci. Rev. 322, 108416. https://doi.org/10.1016/j.quascirev.2023.108416.

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

Yates, L.A., Aandahl, Z., Richards, S.A., Brook, B.W., 2023. Cross validation for model selection: a review with examples from ecology. Ecol. Monogr. 93 (1), e1557. https://doi.org/10.1002/ecm.1557.

Zhao, B., Russell, J.M., Tsai, V.C., Blaus, A., Parish, M.C., Liang, J., Wilk, A., Du, X., Bush, M.B., 2023. Evaluating global temperature calibrations for lacustrine branched GDGTs: seasonal variability, paleoclimate implications, and future directions. Quat. Sci. Rev. 310, 108124. https://doi.org/10.1016/j.quascirev.2023.108124.

Zink, K.G., Vandergoes, M.J., Bauersachs, T., Newnham, R.M., Rees, A.B.H., Schwark, L., 2016. A refined paleotemperature calibration for New Zealand limnic environments using differentiation of branched glycerol dialkyl glycerol tetraether (brGDGT) sources. J. Quat. Sci. 31, 823–835. https://doi.org/10.1002/jqs.2908.