FROG: A global machine-learning temperature calibration for branched GDGTs in soils and peats

Pierre Véquaud, Alexandre Thibault, Sylvie Derenne, Christelle Anquetil, Sylvie Collin, Sergio Contreras, Andrew T Nottingham, Pierre Sabatier, Josef P Werne, Arnaud Huguet

To cite this version:

Pierre Véquaud, Alexandre Thibault, Sylvie Derenne, Christelle Anquetil, Sylvie Collin, et al.. FROG: A global machine-learning temperature calibration for branched GDGTs in soils and peats. Geochimica et Cosmochimica Acta, 2022, 318, pp.468-494. 10.1016/j.gca.2021.12.007 . hal-03552122

HAL Id: hal-03552122
https://hal.sorbonne-universite.fr/hal-03552122

Submitted on 2 Feb 2022

Click here to view linked References

FROG: A GLOBAL MACHINE-LEARNING TEMPERATURE CALIBRATION FOR BRANCHED GDGTS IN SOILS AND PEATS

Pierre Véquaud[a], Alexandre Thibault[b], Sylvie Derenne[a], Christelle Anquetil[a], Sylvie Collin[a], Sergio Contreras[c], Andrew T. Nottingham[d,e], Pierre Sabatier[f], Josef P. Werne[g], Arnaud Huguet[a*]

[a]Sorbonne Université, CNRS, EPHE, PSL, UMR METIS, Paris, 75005, France
[b]Antea Group, Innovation Hub, 803 boulevard Duhamel du Monceau, Olivet, 45160, France
[c] Departamento de Química Ambiental, Facultad de Ciencias & Centro de Investigación en Biodiversidad y Ambientes Sustentables (CIBAS), Universidad Católica de la Santísima Concepción, Casilla 297, Concepción, Chile
[d]School of Geosciences, University of Edinburgh, Crew Building, Kings Buildings, Edinburgh EH9 3FF United Kingdom
[e]School of Geography, University of Leeds, Leeds, United Kingdom
[f]Univ. Savoie Mont Blanc, CNRS, EDYTEM, Le Bourget du Lac, 73776, France
[g]Department of Geology and Environmental Science, University of Pittsburgh, Pittsburgh, PA 15260, USA

Abstract

Branched glycerol dialkyl glycerol tetraethers (brGDGTs) are a family of bacterial lipids which have emerged over time as robust temperature and pH paleoproxies in continental settings. Nevertheless, it was previously shown that other parameters than temperature and pH, such as soil moisture, thermal regime or vegetation can also influence the relative distribution of brGDGTs in soils and peats. This can explain a large part of the residual scatter in the global brGDGT calibrations with mean annual air temperature (MAAT) and pH in these settings. Despite improvements in brGDGT analytical methods and development of refined models, the root-mean-square error (RMSE) associated with global calibrations between brGDGT distribution and MAAT in soils and peats remains high (~ 5 °C). The aim of the present study was to develop a new global terrestrial brGDGT temperature calibration from a worldwide extended dataset (i.e. 775 soil and peat samples, i.e. 112 samples added to the previously available global calibration) using a machine learning algorithm. Statistical analyses highlighted five clusters with different effects of potential confounding factors in addition to MAAT on the relative abundances of brGDGTs. The results also revealed the limitations of using a single index and a simple linear regression model to capture the response of brGDGTs to temperature changes. A new improved calibration based on a random forest algorithm was

---

[*] Corresponding author. Tel: + 33-144-275-172; fax: +33-144-275-150.

E-mail address: arnaud.huguet@sorbonne-universite.fr (A. Huguet).

thus proposed, the so-called random **F**orest **R**egression for Pale**O**MAAT using br**G**DGTs (**FROG).** This multi-factorial and non-parametric model allows to overcome the use of a single index, and to be more representative of the environmental complexity by taking into account the non-linear relationships between MAAT and the relative abundances of the individual brGDGTs. The FROG model represents a refined brGDGT temperature calibration ($R^2 = 0.8$; RMSE = 4.01°C) for soils and peats, more robust and accurate than previous global soil calibrations while being proposed on an extended dataset. This novel improved calibration was further applied and validated on two paleo archives covering the last 110 kyr and the Pliocene, respectively.

Keywords: branched GDGTs; global temperature calibration; soil; peat; machine learning

1. Introduction

Investigating past climate variations is essential to understand and predict future environmental changes, especially in the context of global anthropogenic changes. To this aim, "indirect" indicators of past climates – so-called proxies – have been developed and used regularly since the last century, including those based on microbial lipids. Microorganisms are able to modify the lipid composition of their membranes to maintain a functional fluidity and permeability of the latter. The temperature and pH of the microorganism living environment are considered to be the predominant factors influencing the membrane lipid distribution (Lauber et al., 2009; Siles and Margesin, 2016; Hofmann et al., 2016; Shen et al., 2019).

Among microbial lipids, branched glycerol dialkyl glycerol tetraethers (brGDGTs) have been increasingly used as temperature and pH proxies in continental settings over the last 15 years. These membrane lipids are produced by still unidentified bacteria, although some of them may belong to the phylum *Acidobacteria* (Sinninghe Damsté et al., 2011, 2014, 2018). They are ubiquitous in terrestrial (Weijers et al., 2007; Peterse et al., 2012; De Jonge et al., 2014; Naafs et al., 2017a) and aquatic environments (Blaga et al., 2009; Peterse et al., 2009, 2015; Damsté et al., 2009; Tierney and Russell, 2009; Loomis et al., 2012; Weber et al., 2015). The analysis of brGDGTs, based on a large number of soils distributed worldwide showed that the relative distribution of these compounds is mainly related to mean annual air temperature (MAAT) and soil pH (Weijers et al., 2007; Peterse et al., 2012; De Jonge et al., 2014). The average number of pentane rings, reflected in the Cyclisation of Branched Tetraethers (CBT) index (Weijers et al., 2007; Peterse et al., 2012), has been correlated with soil pH, while the average number of methyl groups, referred to as the Methylation of Branched Tetraethers (MBT) index, has been initially correlated with mean annual mean air temperature (MAAT) and, to a lesser extent, soil pH (Weijers et al., 2007 (r^2=0.77; RMSE =4.8°C, *n*=134); Peterse et al., 2012 (r^2=0.59; RMSE=5.0°C; *n*=176). More recently, new brGDGT isomers have been detected trough improved analytical methods, with methyl groups being present in either 5th , 6th, 7th or 8th position (De Jonge et al., 2013, 2014; Hopmans et al., 2016; Ding et al., 2016). It was observed that 6-methyl isomers were strongly and predominantly dependent on soil pH and 5-methyl brGDGTs on temperature (De Jonge et al., 2014). This led to the development of a new MBT index excluding 6-methyl isomers - the $\mathrm{MBT'_{5Me}}$ index - which correlates preferentially with MAAT (r^2 = 0,64; RMSE= 4.8°C; *n* = 231).

Hence, brGDGTs have emerged over time as robust temperature and pH paleoproxies in multiple types of settings – lakes (Blaga et al., 2009; Powers et al., 2010; Fawcett et al., 2011;

Harning et al., 2020), peatlands (Weijers et al., 2011b; Coffinet et al., 2018; Wu et al., 2020), soil/paleosols (Ding et al., 2015 ; Lu et al., 2016 ; Feng et al., 2019 ; Wang et al., 2020) and speleothems (Baker et al., 2019). Nevertheless, numerous studies showed that additional parameters other than temperature and pH, such as soil moisture, thermal regime or vegetation cover, may also influence the relative distribution of brGDGTs in peat and soils (Weijers et al., 2011a ; Dirghangi et al., 2013; Huguet et al., 2010, 2013 ; Menges et al., 2014; Davtian et al., 2016 ; Liang et al., 2019). This can explain a large part of the residual scatter in the global brGDGT soil/peat calibrations with MAAT and pH (De Jonge et al., 2014; Naafs et al., 2017a, b; Dearing Crampton-Flood et al., 2020).

To overcome these limitations, refinements in the global brGDGT calibrations were proposed over the years. De Jonge et al (2014) developed a global temperature calibration ($R^2 = 0.64$; RMSE = 4.8°C; $n = 231$) based on the MBT'$_{5ME}$ index, excluding the 6-methyl brGDGTs isomers. Naafs et al. (2017a) showed that a stronger correlation between MBT'$_{5Me}$ and MAAT in soils could be obtained by excluding samples dominated by 6-methyl brGDGTs (i.e. ratio of 5- vs. 6-methyl brGDGTs (IR$_{6Me}$) higher than 0.5; $n = 177$; $R^2 = 0.76$; RMSE = 4.1°C). More recently, Dearing Crampton-Flood et al. (2020) used Bayesian statistics instead of more classically applied single linear regressions to investigate the relationship between MBT'$_{5Me}$ and MAAT in soils ($n = 353$; $R^2 = 0.64$; RMSE = 6 °C). Samples with IR$_{6Me}$>0.5 were included in this calibration, as excluding them did not significantly change the strength of the correlation with MAAT. The robustness of the Bayesian approach relies on the fact that (i) it considers a given index (e.g., the MBT'$_{5Me}$ for brGDGTs) as the variable dependent on environmental parameters, consistent with the fact that bacterial lipids are produced in response to the variations of environmental parameters and that (ii) it avoids regression dilution phenomena, in contrast with most of the models based on linear regressions.

Despite improvements in brGDGT analytical methods and development of refined models, the RMSE associated with global calibrations between brGDGT distribution and MAAT in soils and peat remains high (> 4 °C). Part of this uncertainty may be related to our lack of understanding of the mechanism behind the relationship between MAAT and brGDGT distribution. This relationship has initially been explained by a biophysiological mechanism, i.e. the adjustment of the membrane lipid composition by the brGDGT-producing bacteria in response to changes in environmental conditions (homeoviscous adaptation; Weijers et al., 2007). Nevertheless, changes in bacterial community composition may also explain changes in brGDGT distribution, as recently shown by the lipid characterization of Acidobacterial cultures (Sinninghe Damsté et al., 2018) and the concomitant study of brGDGTs and bacterial community composition in soils from well-documented experimental sites (De Jonge et al., 2019, 2021).

Moreover, most of the previous global brGDGT calibrations in soils were based on a correlation between MAAT and a single index (i.e. MBT'$_{5Me}$; De Jonge et al., 2014; Naafs et al., 2017a, b; Dearing Crampton-Flood et al., 2020), even though the relative distribution of brGDGTs is likely to be concomitantly influenced by several environmental parameters. In contrast, using relative abundances of bacterial lipids rather than a single index in models appears more representative of the environmental complexity (Wang et al., 2020; Véquaud et al., 2020; Dunkley Jones et al., 2020). In this way, multiple regression models were also used to describe the relationships between brGDGT distribution and given environmental variables (MAAT, pH) in soils (e.g. Peterse et al., 2012; De Jonge et al., 2014) or lakes (e.g. Pearson et al., 2011; Russell et al., 2018). It was previously shown that the uncertainty in brGDGT calibrations can be improved through the use of multiple regression methods vs. single predictor methods (e.g. Loomis et al., 2012; Wang et al., 2020). Nevertheless, as other linear models, the multiple regression ones cannot take into account non-linear influences, which may occur in complex environmental settings. Such a limitation can be overcome using non-parametric models such as machine-learning algorithms. Machine-learning models were very recently used to develop global calibrations between the relative abundance of isoprenoid GDGTs and sea surface temperature (SST) in marine settings (Dunkley Jones et al., 2020) and between the relative abundance of bacterial 3-hydroxy fatty acids and MAAT in soils (Véquaud et al., 2020; Wang et al., 2021). These models allow overcoming the use of a single index as they are based on the whole suite of microbial lipids. They are built on a proportion of the total dataset (randomly defined) and then tested on the rest of the dataset, considered as independent. Such an approach improves the robustness of the model and avoids the phenomenon of regression dilution. As they are non-parametric, they also capture non-linear environmental influences, in line with the intrinsic complexity of the environmental settings.

In the present study, a machine-learning algorithm (random forest) was developed with the aim of proposing a new global brGDGT calibration for MAAT reconstruction in soils and peats with a reduced RMSE. It was based on an extended global dataset comprising 775 peat and soil samples (with 112 samples added to the previous global brGDGT calibration by Dearing Crampton-Flood et al., 2020). This dataset was statistically separated into clusters to better understand the parameters affecting brGDGT distribution in soils at the global scale. The clusters differed by the influence of environmental parameters – MAAT, mean annual precipitation (MAP), soil pH and the number of frozen days during the year (FRS) on the relative abundance of brGDGTs. This mechanistic approach highlighted the limitations of the MBT'$_{5Me}$-MAAT relationship at the global scale and then led to the development of a refined brGDGT temperature calibration (so-called FROG model) based on a random forest machine-learning algorithm and the whole suite of individual brGDGTs. Alternative models were also proposed to test the influence of confounding variables on the FROG calibration and potentially further improve its accuracy.

2. Materials and methods

2.1. Global soil dataset and environmental parameters

The dataset of the present study is comprised of the globally distributed surface peat and soil samples (*n*=663) used in previous brGDGT global calibrations (Weijers et al., 2007; Peterse et al., 2012; De Jonge et al., 2014; Yang et al., 2015; Ding et al., 2015; Xiao et al., 2015; Lei et al., 2016; Wang et al., 2016; Naafs et al., 2017b; Dearing Crampton-Flood et al., 2020). This dataset was extended with 112 soil samples from 6 altitudinal transects located in France, Italy, Tibet, Chile and Peru and for which brGDGT data were recently published (Huguet et al., 2019; Véquaud et al., 2020, 2021). The details of the dataset (*n*=775) are provided in Table 1.

Actual MAAT, pH and MAP values measured from the nearest weather stations, when available, were used to better determine the environmental reality, diversity and complexity. Such values were available for most of the samples, i.e. 598 of the 775 samples. Nevertheless, for the other samples, MAAT and pH values were extracted from the 0.5 gridded CRU TS v. 3.26 dataset (Harris et al., 2014), using the same approach as Dearing-Crampton Flood et al. (2020). This approach would have been inappropriate for the 6 aforementioned altitudinal transects, where large temperature variations derive from differences in elevation that can vary across short distances, as noticed by Pérez-Angel et al. (2020).

To constrain the applicability of the MBT'$_{5Me}$ as a temperature proxy in peat and soils, Naafs et al. (2017a) used a thermal regime indicator, the Growing Degree Days (GDD). This index is calculated by summing the daily temperatures above 0 °C over a year within a soil (Choler 2018) and interpreted as a proxy of heat accumulation within the latter (McMaster and Wilhelm, 1997; Choler, 2018). The GDD better reflects the growth temperatures encountered by bacterial communities in soils and peats. Unfortunately, as the daily temperatures were not available for the whole dataset of the present study, the GDD index could not be calculated. Instead, another thermal regime indicator was used, the number of frozen days during the year (FRS) for one location, proposed by Harris et al., (2014). The FRS was obtained for most of the samples (i.e. those with site coordinates available, Table 1).

The MBT'$_{5Me}$ index, reflecting the methylation level in 5-methyl isomers of GDGTs and considered as related to MAAT, was calculated according to De Jonge et al. (2014; Eq. 1):

$$\text{MBT'}_{5\text{Me}} = \frac{[Ia+Ib+Ic]}{[Ia+Ib+Ic]+[IIa+IIb+IIc]+[IIIa]} \qquad (1)$$

The CBT' index was calculated as follows (De Jonge et al., 214; Eq. 2):

$$\text{CBT'} = \log\left(\frac{[Ic]+[IIa']+[IIb']+[IIc']+[IIIa']+[IIIb']+[IIIc']}{[Ia]+[IIa+IIIa]}\right) \quad (2)$$

The IR_{6Me} reflects the relative abundance of 6- vs. 5-methyl brGDGTs, as proposed by Dang et al. (2016; Eq. 3):

$$IR_{6Me} = \log\left(\frac{[IIa']+[IIb']+[IIc']+[IIIa']+[IIIb']+[IIIc']}{[IIa']+[IIb'+IIc']+[IIIa']+[IIIb']+[IIIc']+[IIa]+[IIb]+[IIc]+[IIIa]+[IIIb]+[IIIc]}\right) \quad (3)$$

The Roman numerals correspond to the different GDGT structures presented in De Jonge et al. (2014). The 6-methyl brGDGTs are denoted by an apostrophe after the Roman numerals for their corresponding 5-methyl isomers.

2.2. Statistical analyses

A Principal Component Analysis (PCA) was performed on the entire dataset with R software (version 4.0.3; R Core Team, 2020) to observe the distribution of the different samples based on their brGDGT relative abundances. A cluster classification of the samples based on the k-means method was proposed. In order to choose the optimal number of clusters, the ratio of Within-Cluster-Sum-of-Squares (WCSS) over the total sum of squares was calculated. The WCSS is the sum of squares of the distances of each data point in all clusters to their respective centroids. The optimal number of clusters corresponds to the minimum value of the ratio of the WCSS over the total sum of squares (a WCSS = 0 means one sample corresponds to one cluster). In order to choose the threshold for the optimal WCSS value, and so the optimal number of clusters, the elbow method was used. It consists in plotting the WCSS values against the number of clusters, then allowing to derive the optimal number of clusters.

Redundancy analysis (RDA) was first performed on the global dataset and then carried out on each cluster derived from the PCA analysis to evaluate and compare the influence of the environmental parameters on brGDGT distribution (i) at the global scale and (ii) in each cluster. RDA is a "constrained" analysis, used to directly visualize the variation in the lipid data as a function of the environmental variables. It allows not only assessing but also quantifying the influence of each explanatory variable (i.e. environmental variables) on the distribution of bacterial lipids. RDA yields the influence of each variable, with regard to the statistical variance, on the pool of bacterial lipids, and allowed a quantification in percent of the influence of each parameter (i.e. conditional effect). Conditional effects summarize the effects of each variable taking into account the effect of variables with the greatest influence (Braak and Smilauer, 2002). RDA analyses were performed on centered and standardized data using the

CANOCO v. 5.04 software (Braak and Smilauer, 2002). The relationships between each variable and the dimensions of the RDA were investigated using the corresponding r-values and the percentages of variance.

In order to refine the threshold of the community index (CI) proposed by De Jonge et al. (2019), all linear regressions between MAAT and MBT'$_{5Me}$ on the global dataset were tested by successive iteration of CI values from 0 to 1 (0.001 step) using the R software, version 4.0.3 (R Core Team, 2014).

2.3. Machine learning: Random forest model

The random forest algorithm was used to develop a global calibration between brGDGT relative abundances and MAAT. The random forest algorithm is a supervised learning method notably used for regressions (e.g. Ho, 1995; Breiman, 2001; Denisko and Hoffman, 2018). This model works by building a multitude of decision trees from a training dataset and producing the mean prediction of the individual trees. Decision tree learning is one of the predictive modeling approaches used to move from observations to conclusions about the target value of an item.

In order to calculate the model based on the random forest algorithm, the global dataset was divided into two subsets: a training dataset and a test dataset. The training dataset corresponds to the samples used to fit the model. The test dataset corresponds to the samples used to provide an unbiased evaluation of the model previously fit on the training dataset. The training phase required for the random forest regression was performed on 75% of the sample set (which allow to neglect the overfit of the model), with 500 trees and an iteration of ten-fold cross-validations per model. The cross validation allows the optimization of the hyperparameters (number of variables in each node and minimal node size) of the models. Data selection was performed randomly on the dataset, but with a stratification modality according to the MAAT to limit the impact of extreme values. Then, the robustness and precision of the different models, developed from the random forest algorithm, were tested on the remaining 25 % of samples, considered as an independent dataset. Random forest models were performed with R software, version 4.0.3 (R Core Team, 2014) using the packages tidymodels (version 0.1.02)- ranger (version 0.12.1).

A R package with a web-application is available on a GITHUB repository (paleoFROG) for the reconstruction of brGDGT-derived MAAT using the FROG models proposed in the present study.

The performances of the random forest model were compared with those of the Bayesian models, BayMBT and BayMBT$_0$, proposed by Dearing Crampton-Flood et al. (2020). The latter were performed with the MATLAB code available from the GITHUB repository of Jessica Tierney (https://github.com/jesstierney; Dearing Crampton-Flood et al., 2020), using MATLAB, version 9.8. The *prior* mean was set to the MAAT mean for all soil samples (10°C), with a *prior* standard deviation of 30°C.

3. Results

3.1. Principal component analysis and clustering on the global dataset

In order to explore the global dataset and understand which samples could explain the scattering on the global MBT'$_{5Me}$-MAAT calibration, we performed a statistical clustering of the extended peat and soil dataset without any *a priori* assumptions on the basis of their brGDGT fractional abundances. With this aim, a Principal Component Analysis (PCA) was performed on the entire brGDGT dataset (Fig. 1). The first 3 axes of the PCA carry most of the variance (70.3%; Figs. 1a, b, c). Consequently, the description of the analysis will be restricted to these axes. A cluster classification of the samples based on the k-means method was performed, yielding 5 clusters (Fig. 1d), based on the Within Cluster Sum of squares and the elbow method. The distribution of the samples between the different clusters is heterogeneous (between 76 and 230 samples), with various proportions of soil and peat samples (Table 2). The clusters are well-differentiated, with different means and amplitudes for MAAT, FRS and pH (Fig. 2), and also based on their geographical locations (Fig. 3). Clusters B and D contain a larger proportion of peat samples representative of acidic environments, which can explain the lower pH values by ca. 1 to 2 units compared to the other clusters (Gorham, 1991; Killops, 2005; Dedysh et al., 2006; Comont et al., 2006) (Table 2, Fig. 2). Cluster A shows samples mainly distributed over tropical and subtropical latitudes (Fig. 3) associated with high MAAT (22.4 ± 6.0 °C) and rather high MAP (1069 ± 385 mm/yr; Table 2, Fig. 2). Cluster B samples are distributed over temperate to subtropical latitudes, with precipitation amounts (1237 ± 643 mm/yr) as high as for samples from cluster A, but lower MAAT (16.0 ± 7.5 °C; Fig. 3; Table 2). Clusters A and B are characterized by comparable and higher MAATs than those from the other clusters, and conversely lower FRS (Fig. 2, Table 2). Samples of cluster C are mostly distributed in China and correspond to loess samples (Fig. 3). Within this cluster, MAP (453 ± 643 mm/yr) is the lowest of all clusters and MAAT (6.7 ± 5.1 °C) is on average lower than in clusters A and B, associated with a higher FRS (Table 2; Fig. 2). The samples from clusters D and E show similar geographical distributions, mostly in the northern hemisphere, at temperate latitudes, and even polar latitudes (Fig. 3). This results in lower MAP and MAAT especially for cluster D (784 ± 457mm/yr and 3.9 ± 5.9 °C, respectively), and a high FRS, similar to cluster C (Table 2; Fig. 2). Thus, the statistical differentiation of the brGDGT dataset into different clusters is reflected through various descriptive environmental parameters.

3.2. BrGDGT distribution in the global dataset and associated clusters

The fractional abundances of the individual brGDGTs were determined in the global dataset and in the five clusters statistically derived from the latter (Fig. 4). In the global dataset, the acyclic brGDGTs *Ia, IIa* and *IIa'* were predominant. Distinct brGDGT distributions were observed in each cluster. In cluster A, the tetra-methylated brGDGTs *Ia* and *Ib* as well as penta-methylated brGDGT *IIa* are the most abundant. Acyclic brGDGT *Ia* is largely predominant (ca. 75% of total brGDGT relative abundance) in cluster B. In cluster C, 6-methyl acyclic isomers of the penta- and hexa-methylated brGDGTs (*IIa'* and *IIIa'*) and brGDGT *Ia* represent altogether ca. 65% of the total brGDGT relative abundance. The brGDGT distribution of cluster D is dominated by acyclic compounds *Ia* and *IIa*. In cluster E, the 6-methyl brGDGTs are slightly more abundant than the 5-methyl isomers, with acyclic brGDGTs *Ia, IIa, IIa', IIIa, IIIa'* and monocyclic brGDGTs *Ib, IIb, IIb'* representing each between ca. 10 and 20% of total brGDGT relative abundance. The obvious differences in brGDGT distribution between the 5 clusters are also reflected in the indices derived from these compounds. Thus, the $\text{MBT'}_{5\text{Me}}$ is higher in clusters A and B (mean 0.88 ± 0.09 and 0.82 ± 0.13, respectively) than in clusters C (0.56 ± 0.14) as well as D and E (0.47 ± 0.09 and 0.49 ± 0.12, respectively, Fig. 5a). Regarding the CBT', it is much lower in clusters B and D (mean -1.29 ± 0.55 and -0.95 ± 0.63, respectively) than in cluster C (0.40 ± 0.24), E (0.13 ± 0.25; Fig. 5b) and A (-0.16 ± 0.33),. Similarly, the relative abundance of 6-methyl vs. 5-methyl brGDGTs ($\text{IR}_{6\text{Me}}$ ratio; Eq. 3) is much lower in clusters B and D (mean ~ 0.2) than in the other three clusters (mean comprised between 0.62 and 0.80; Fig. 5c).

3.3. Relationships between $\text{MBT'}_{5\text{Me}}$ and MAAT

A strong and significant correlation between MAAT and $\text{MBT'}_{5\text{Me}}$ is observed when considering the total soil dataset (Supp. Fig. 1; Eq. 4):

$$\text{MAAT (°C)} = 35.98 \times \text{MBT'}_{5Me} - 12.74 \ (n\text{=775; R}^2\text{= 0.65, RMSE= 5.2°C}) \quad (4)$$

Nevertheless, this global calibration shows a considerable scatter. The linear regressions between the MBT'$_{5Me}$ and MAAT were further explored for each cluster derived from the PCA analysis (Fig. 1). Clusters A and B show strong significant linear relationships (R^2 = 0.61 and 0.77, respectively; p <0.0001) between MAAT and MBT'$_{5Me}$ (Fig. 6a, b;) associated with improved RMSE (3.8°C and 3.6°C, respectively) compared to the global calibration (Supp. Fig. 1; Eq. 4). In contrast, for the other clusters (Figs. 4c, d, e), especially D and C, significant (p < 0.0001) but weak relationships between MBT'$_{5Me}$ and MAAT are observed (Fig. 6).

To further investigate the influence of the proposed environmental variables (MAAT, MAP, pH, FRS) on the brGDGT relative abundance in (i) the global dataset and (ii) the different clusters, RDA was performed (Fig. 7). Regarding clusters A and B, the first two axes explain 63.9% and 74.6% of the total variance of the dataset, with an explained fitted variation of 97.3% and 99.9%, respectively (i.e. explained fitted variation; relationship between the fractional abundances of brGDGTs and the selected environmental variables, calculated as the sum of all constrained eigenvalues) (Fig. 7a, b; Table 3). The first axis of the RDA for clusters A and B is well correlated with MAAT (r=0.80; r=0.90 respectively), FRS (r=-0.80; r=-0.81), MAP (r=0.77; r=0.78) and pH (r=-0.75; r=-0.60) (Fig. 5a, b; Table 3). Axis 2 of cluster A is mainly correlated with the FRS (r=0.56) and to a lesser extent pH (r=-0.34) and MAAT (r=-0.26), while axis 2 of cluster B is predominantly negatively correlated with pH (r=-0.76) and to a lesser extent with MAAT (r=-0.41; Fig. 7a, b; Table 3). The quantification of the combined influence of the different environmental variables on the brGDGT distribution shows a predominant effect of the FRS (39.6%) and to a lesser extent pH (18.6%) for cluster A and MAAT (57.3%) and to a lesser extent pH (16.7%) for cluster B (Table 3). The predominant influence of the thermal regime (MAAT and FRS) on the relative distribution of brGDGTs in these two clusters, despite large variation in pH range, especially in cluster B (Fig. 2), explains why the linear regressions between MBT'$_{5Me}$ and MAAT are stronger (Fig. 6) than that observed for the global dataset (Supp. Fig. 1; Eq. 4).

The first two axes of the RDAs for clusters C and D explain 34.9% and 53.4% of the total inertia of the dataset, with an explained fitted variation of 95.7% and 99.4%, respectively (Fig. 7c, d; Table 3). For these two clusters, axis 1 is strongly negatively correlated with pH (r=-0.91 and -0.96, respectively) and positively correlated with MAP values (r=0.59 and 0.48, respectively). Axis 2 is controlled by the thermal regime, being mainly correlated with FRS (r=0.74 and -0.68 for clusters C and D, respectively) and MAAT (r=-0.78 and 0.39, respectively). The quantification of the combined influence of the environmental variables on brGDGT distribution in clusters C and D shows a predominant effect of soil pH (25.6% and 47.4%, respectively) and, only to a much lesser extent, MAP and MAAT (<6%) (Table 3). This is consistent with the weak correlation between MBT'$_{5Me}$ and MAAT for cluster C, and absence of correlation for cluster D as well as the high scattering of the corresponding values (Fig. 6c, d).

The first two axes of the RDA for cluster E explain 28.4% of the total inertia of the dataset and the explained fitted variation is 94.40% (Fig. 7e; Table 3). Axis 1 is mainly correlated with MAAT (r=-0.76), FRS (r=0.87), pH (r=0.71) and to a lesser extent, MAP (r=-0.41; Table 3). Axis 2, on the other hand, is mainly influenced by MAAT (r=0.62), and to a lesser extent pH (r=0.41; Table 3). When examining the combined effect of the different environmental variables in this cluster, it appears that brGDGT distribution is mainly and significantly controlled by FRS (15.2%) and to a lesser extent by MAAT (6.6%) and pH (7.7%; Table 3). The major influence of the thermal regime (FRS, MAAT) on brGDGT distribution in cluster E is consistent with the relationship (R^2=0.44) observed between MBT'$_{5Me}$ and MAAT (Fig. 6e). Nevertheless, in contrast with cluster A, the additional influence of pH may explain the moderate determination coefficient of this correlation.

Regarding the global dataset, the first two axes explain 69.5% of the total variance of the dataset (Fig. 7f; Table 3) and the selected environmental variables explain 99.0% of the variance of the brGDGT relative abundances (Table 3). Axis 1 is strongly controlled by pH (r=-0.92) and MAP (r=0.72) and to a lesser extent by FRS (r=-0.54) and MAAT (r=0.48). Axis 2 is strongly correlated with MAAT (r=0.83), and to a lesser extent FRS (r=-0.58) followed by MAP (r=0.31) and pH (r=0.34). The quantification of the combined influence of the environmental variables on brGDGT distribution in the global dataset shows a predominant effect of soil pH (51.8%) and to a lesser extent MAAT (15.7%), with only a minor influence of FRS and MAP (<2%; Table 3).

4. Discussion

4.1. Constraints on the MBT'$_{5Me}$-MAAT relationship in soils

4.1.1. Global level

The MBT'$_{5Me}$ was shown to be linearly and strongly correlated with MAAT in the present extended soil dataset (Supp. Fig. 1; Eq. 2), as previously observed at the global level (e.g., De Jonge et al., 2014; Dearing Crampton-Flood et al., 2020). Nevertheless, in line with these previous studies, the RMSE remains high (5.2 °C). This scatter may have multiple sources, such as the fact that the brGDGT calibrations are achieved against MAAT, whereas brGDGT-producing bacteria live in soils. Soil temperature is not necessarily equivalent to MAAT and also depends on the vegetation cover (e.g. Wang et al., 2020), which may explain part of the scatter. Moreover, another source of uncertainty may be related to the fact that climatic data derived from the nearest weather stations (gridded datasets) are often used to develop brGDGT calibrations, while they may not appropriately reflect the local air temperatures nor the soils ones, as recently reported by Pérez-Angel et al. (2020). Last, part of the remaining uncertainty in the brGDGT-MAAT calibrations may be due to the influence of other environmental parameters than MAAT on brGDGT distribution, as discussed below.

The RDA analysis performed on our global dataset showed that soil pH was the main environmental control on brGDGT distribution besides MAAT (Fig. 7), as also previously reported (e.g. Weijers et al., 2007; Peterse et al., 2012; De Jonge et al., 2014; Naafs et al., 2017a). As expected, 6-methyl brGDGTs were all located in the left quadrant along axis 2, mainly controlled by pH, in line with the positive correlations previously observed between these compounds and pH (De Jonge et al., 2014; Dang et al., 2016). These isomers were purposefully excluded from the calculation of the MBT so that it is no more related to pH and only to MAAT (De Jonge et al., 2014). Nevertheless, the relative abundances of two of the main brGDGTs involved in the MBT'$_{5Me}$ (*Ia* and *IIa*) were shown to be significantly correlated with pH (R = 0.52 and 0.32, respectively; $p < 0.001$; Sup. Table 1) in the present dataset, as also previously observed (De Jonge et al., 2014). Such correlations with pH were even higher than those observed with MAAT ($R^2 = 0.28$ and 0.24 for compounds *Ia* and *IIa*, respectively; $p < 0.001$; Sup. Table 1), which may explain part of the remaining uncertainty in the MBT'$_{5Me}$ relationship. Very recently, De Jonge et al. (2021) highlighted the importance of taking into account the effect of soil pH on MBT'$_{5Me}$ values and associated temperature reconstructions, as soil pH was shown to be the main factor responsible for concomitant changes in brGDGT distribution and bacterial community composition in mid- and high-latitude experimental sites and hypothesized that such conclusions were also valid at the global scale.

Soil moisture has also been suggested to have an effect on the relative abundance of brGDGT distribution (e.g. Loomis et al., 2010; Dirghanghi et al., 2013; Menges et al., 2014; Dang et al., 2016; Naafs et al., 2017a), with weak or no linear relationships between MBT/MBT'$_{5Me}$ and MAAT in arid soils (MAP < 500 mm/yr). The relative soil moisture is related to pH variations, with arid soils mainly being alkaline (Naafs et al., 2017a). This may have a role on the diversity of bacterial communities (Lauber et al., 2009; Shen et al., 2019). Alternatively and/or complementarily, it was suggested that brGDGT producers may change their membrane composition in response to soil moisture changes (Loomis et al., 2010; Dang et al., 2016). The relative soil moisture may also impact the capacity of a soil to retain heat (Idso et al., 1975; Davidson et al., 1998; Balleza et al., 2014; Dang et al., 2016), indirectly influencing the methylation degree of brGDGTs. Dang et al (2016) and Naafs et al. (2017a) especially showed that the MBT'$_{5Me}$ was only significantly correlated with MAAT when the ratio of 5- vs. 6-methyl brGDGTs (IR$_{6Me;}$ Eq. 3) was lower than 0.5. Therefore, to potentially improve the accuracy of the global MBT'$_{5Me}$-MAAT calibration (Supp. Fig. 1), the total dataset was divided into two subgroups based on a threshold value of 0.5 for the IR$_{6Me}$ ratio as proposed by Dang et al. (2016) and Naafs et al. (2017a). Two subgroups with similar number of samples (*n*=389 for IR$_{6Me}$>0.5; *n*=384 for IR$_{6Me}$<0.5) were thus obtained. The linear regressions between MBT'$_{5Me}$ and MAAT in the two subgroups were statistically similar (Supp. Fig. 2), even though a slightly higher determination coefficient and lower RMSE were observed when IR$_{6Me}$<0.5. Moreover, the regressions obtained for the two subgroups did not show obvious improvements with the one derived from the total dataset. Therefore, a separation of the present dataset based on the IR$_{6Me}$ values does not appear necessary, as previously observed by Dearing Crampton-Flood (2020) for their dataset, showing that the relative abundance of 6- vs. 5-methyl brGDGTs only has a limited influence on the MBT'$_{5Me}$-MAAT relationship at the global scale.

More recently, De Jonge et al. (2019) suggested that the response of brGDGTs to temperature changes is strongly dependent on the nature of bacterial communities present in soils. These authors initially showed that the distribution of brGDGTs in a set of geothermally warmed soils from Iceland changed when the average annual soil temperature was above 14°C. This sudden change in brGDGT distribution coincided with an abrupt shift in the bacterial community composition. A relative increase in brGDGT *Ia* vs. homologues *IIa* and *IIIa* was observed in the soils with annual soil temperature > 14 °C (warm soil cluster). This was reflected in a change in the community index (CI) proposed by De Jonge et al. (2019; Eq. 5):

$$\mathrm{CI} = [Ia]/([Ia] + [IIa] + [IIIa]) \quad (5)$$

The CI index is similar to the MBT'$_{5Me}$ (Eq. 1), except that it excludes the compounds containing cyclopentyl moieties, i.e. those suspected to be pH-sensitive compounds. De Jonge et al. (2019) proposed a CI threshold of 0.64 to separate the geothermal soils with an annual soil temperature higher than 14°C (CI >0.64) and those with a lower temperature (CI <0.64).This observation was extended to the global scale, revealing two distinct clusters over the entire peat and soil dataset compiled by De Jonge et al. (2019) – one considered as a “cold” subgroup (*n*=251 soils and peats; MAAT between -8.3°C and 18.2°C) and another one as a “warm” subgroup (*n*= 195 soils and peats; MAAT between 0.4°C and 27.1°C) – with different responses to temperature and pH. The slopes and determination coefficients of the MBT’$_{5Me}$-MAAT relationship were significantly different in the two subgroups, which may explain part of the uncertainty in the MBT’$_{5Me}$-MAAT correlation at the global scale. The extended dataset proposed in the present study (*n*=775) was thus divided into two subgroups based on the CI threshold of 0.64 proposed by De Jonge et al. (2019). A strong linear relationship between MBT'$_{5Me}$ and MAAT was observed for the warm cluster (R^2=0.71), while it was much weaker for the cold subgroup (R^2=0.20; Fig. 8), as previously observed by De Jonge et al. (2019) on their global dataset. The discrimination of samples in clusters based on their different brGDGT signature and response to environmental changes allows better understanding the MBT’$_{5Me}$-MAAT relationship.

The CI threshold (0.64) defined by De Jonge et al. (2019) was based on a smaller number of soils (*n*=446) than available in the present study. To refine this value based on the present extended sample set (*n*=775), all linear regressions between MAAT and MBT'$_{5Me}$ were tested by successive iteration for CI values from 0 to 1 (0.001 step: Supp. Fig. 3a). The refined threshold is 0.69 (corresponding to the best adjusted R^2), still in agreement with the clusters presented in this study (Fig. 5d) and those of De Jonge et al. (2019). It leads to only limited changes in the slopes and intercepts of the MBT’$_{5Me}$-MAAT relationships in the warm and cold clusters (Sup. Fig.3b) in comparison with those obtained for the previously defined threshold of 0.64 (Fig. 8). The CI might be considered first and foremost in any use of the MBT'$_{5Me}$ index to reconstruct paleo-MAATs, as proposed by De Jonge et al. (2019).

4.1.2. Sample clustering

Complementarily to the empirical approach of De Jonge et al. (2019) described above, we used statistical tools to (i) classify the samples of the present peat and soil dataset based on their brGDGT distribution and (ii) investigate and compare the influence of the environmental factors on the brGDGT distribution in each cluster (Fig. 1).

Clusters A and B, which encompassed soils from temperate and (sub)tropical areas, were characterized by similarly high MBT'$_{5Me}$ values (mean > 0.8; Fig. 5a) and CI > 0.64 (Fig. 5d) and can be related to "warm" groups as defined by De Jonge et al. (2019). These two clusters differed by the more acidic nature of samples from cluster B than from cluster A, reflected in the much lower CBT' values in cluster B, consistent with the positive relationship usually observed between CBT' and pH (De Jonge et al., 2014). In line with the increase in the fractional abundance of 6-methyl brGDGTs with pH previously observed in soils (De Jonge et al., 2014), the samples from cluster A were also characterized by higher IR$_{6Me}$ ratio than those from cluster B. Despite these differences related to pH, the brGDGT distributions of these clusters were mainly impacted by the thermal regime (Table 3). Thus, moderate ($R^2 > 0.25$) to strong correlations ($R^2 > 0.50$, $p < 0.001$) between acyclic 5-methyl brGDGTs (*Ia, IIa* and *IIIa*) and MAAT were obtained in clusters A and B, respectively, as previously observed at the global level (De Jonge et al., 2014; Naafs et al., 2017a). This was reflected in the strong positive correlations observed between MBT'$_{5Me}$ and MAAT in clusters A and B (Fig. 6a,b), with non-significant differences between slopes, intercepts and RMSE, highlighting the overall similar response of brGDGT source microorganisms to temperature changes in soil and peat samples from the two "warm" subgroups.

In contrast with clusters A and B, the three other clusters were characterized by CI < 0.64 (Fig. 5d), corresponding to "cold" groups (De Jonge et al., 2019). They encompassed soil and peat samples from cold to (sub)temperate zones (Fig. 3), with similar range of MBT'$_{5Me}$ values, which were much lower than those of clusters A and B, consistent with the increase of the methylation degree of brGDGTs at lower temperatures (Weijers et al., 2007). Samples from cluster D were more acidic than those from cluster C and E, leading to distinct brGDGT distributions, and especially lower average CBT' values in cluster D. The differences in brGDGT distribution between the three "cold" clusters were also reflected in the IR$_{6Me}$ ratio, the highest values of the latter in cluster C being consistent with the higher relative abundance of 6-methyl vs. 5-methyl brGDGTs generally observed in arid/alkaline soils (e.g. De Jonge et al., 2014; Naafs et al., 2017a), as those from cluster C (Table 2). In addition, the three clusters largely differed in their dependence to environmental parameters. Thus, brGDGT distributions in clusters C and D were predominantly influenced by pH and to a lesser extent by the thermal regime (Table 3). This led to weak (or no) correlations ($R^2 < 0.1$) between individual brGDGTs and MAAT (or FRS) (Supp. Tables 4, 5), explaining in turn the weak relationships between the MBT'_{5Me} and MAAT in these two clusters (Fig. 6c, d). In contrast with clusters C and D, the influence of the thermal regime (MAAT/FRS) on brGDGT distribution in samples from cluster E was higher than the one of pH, with weak to moderate correlations (R^2 0.2-0.45) only between the relative abundance of tetramethylated brGDGTs *Ib* and *Ic* and hexamethylated brGDGT IIIa and MAAT (Sup. Table 6; Fig. 7e), as also observed at the global level (De Jonge et al., 2014). Nevertheless, no correlations between brGDGTs *Ia/IIa* and MAAT were observed in cluster E. This contrasts with observations made at the global level in the previous (De Jonge et al., 2014; Naafs et al., 2017a) or present soil datasets (Sup. Table 1), where these compounds were considered as temperature-sensitive. This explains the more moderate correlation between the MBT'_{5Me} and MAAT in cluster E (R^2 0.44) than in clusters A and B or the total dataset ($R^2 > 0.6$; Fig. 6). It should be noted that in the cold soil cluster of De Jonge et al. (2019), the MAAT was similarly only correlated with the relative abundances of brGDGTs *Ia* and *IIIc*, leading to a weak correlation between MBT'_{5Me} and MAAT (R^2 0.28). Nevertheless, in the global cluster cold defined by De Jonge et al. (2019), brGDGT Ia was significantly negatively correlated with pH (R^2 0.69; $p< 0.001$), in contrast with cluster E (R^2 0.16; $p = 0.0003$). Such a difference may be related to the smaller size ($n = 77$) and different samples constituting cluster E vs. the global cold cluster ($n = 251$) of De Jonge et al. (2019).

Overall, the brGDGT distribution was differently impacted by environmental variables in each of the clusters of the present study (Fig. 7; Table 3), the effect of the thermal regime being predominant only in the two warm clusters (A and B) and to a lesser extent cold cluster E. We also observed a different dependency of the brGDGT distribution to temperature in the two warm clusters vs. cold cluster E, leading to distinct correlations between MBT'_{5Me} and MAAT for the two types of clusters, consistent with previous observations by de Jonge et al. (2019). Despite the smaller size of the cold cluster E (n =77) vs. the two warm clusters A (n=76) and B (n =174) altogether, the relationship between MBT'_{5Me} and MAAT for the global dataset ($n = 775$) was observed to be driven by the one of cluster E, as revealed by the similar slopes and intercepts (Fig. 6). This implies that the MBT'_{5Me} proxy is much more influenced by the temperature changes encountered in cold cluster E than in the other warm clusters. Such a difference in sensitivity between warm and cold groups has also been previously reported by De Jonge et al. (2019), who suggested that different bacterial communities, with different brGDGT fingerprints, may be associated with the warm and cold groups. Similarly, this shift in bacterial communities could at least partly explain the different MBT'$_{5Me}$-MAAT relationships in clusters A/B vs. E. Additionally, the thermal regime may influence the activity of brGDGT-producing microorganisms and associated biosynthesis, as several previous studies suggested that brGDGTs may be preferentially produced during warm seasons (Weijers et al., 2011b; Huguet et al., 2013; Deng et al., 2016). As the thermal regime is much higher in the cold clusters (C, D and E), they should be the most impacted by the seasonal production of brGDGTs, thus weakening the relationships between MAAT and brGDGT distribution. Raberg et al. (2021) very recently proposed to replace MAAT by a warm season index (mean temperature of months above freezing) to take into account the thermal regime effect (especially large at high-latitude), thus improving brGDGT calibrations in lake sediments. Similarly, Dearing Crampton-Flood et al. (2020) showed that the MBT'$_{5Me}$ in soils and peats was better correlated with the average temperature of months above 0 °C (BayMBT0 model, R^2 0.70) than with MAAT (BayMBT, R^2 0.64), with a reduced RMSE (3.8 vs. 6.0 °C for the BayMBT$_0$ and BayMBT, respectively).

In any case, the fact that (i) a strong relationship between MBT'$_{5Me}$ and MAAT is only observed for the warmer clusters (Fig. 6), (ii) the MBT'$_{5Me}$-MAAT relationship at the global level is driven by the moderate correlation of cold subgroup E containing only ca. 10% of the samples from the total dataset and (iii) weak correlations between MBT'$_{5Me}$ and MAAT are observed for ca. 55% of the dataset (clusters C and D) highlights the limitations of using a single index and a simple linear regression model to capture the response of brGDGTs to MAAT changes. Thus, very recently, Pérez-Angel et al. (2020) showed that the fractional abundance of brGDGT *Ia* was non-linearly related to MAAT in a dataset comprised of tropical soils (*n*=175). Other models taking into account such non-linear trends should be complementarily developed to better reflect this complex response to MAAT changes. Machine-learning models can be used to this aim and will be tested in the following.

4.2. A novel global calibration between MAAT and brGDGT distribution using the random forest algorithm

4.2.1. Model development

A machine-learning model – the random forest algorithm – was tested to potentially derive stronger and more accurate global MAAT calibration from brGDGT distributions than single linear regressions based on the MBT'$_{5Me}$ index.

The random forest algorithm was first "trained" during the learning phase to estimate MAAT from brGDGT relative abundances. During this phase, the model produces decision trees that will automatically discriminate the compounds whose relative abundances are not influenced by temperature, and thus selects those to be used to estimate MAAT, while taking in account the interdependency of environmental parameters. The developed model was called "random **F**orest **R**egression for Pale**O**MAAT using br**G**DGTs" (FROG).

As the 5-methyl brGDGTs are considered to be mainly correlated with MAAT in contrast with 6-methyl isomers (De Jonge et al., 2014), the random forest model was initially applied to the fractional abundances of the 5-methyl brGDGTs only (FROG$_{5Me}$). After training on the sampling dataset (75% of the global dataset; *n*=583), a strong (R^2 = 0.80) global calibration between 5-methyl brGDGT relative abundance and MAAT was obtained (Fig. 9a), with a RMSE of 4.12 °C. The random forest algorithm was then separately trained and tested using the relative abundances of all brGDGTs including the 6-methyl isomers (FROG). This "alternative" global calibration (R^2=0.81 and RMSE =4.09 °C; Fig. 9b) appeared similar to the FROG$_{5Me}$ model (Fig. 9a). The RMSE of both random forest calibrations (FROG$_{5Me}$ and FROG) are lower than those of the simple linear regression between MBT'$_{5Me}$ and MAAT (Fig. 9c, d; Table 4), with no clear trends in the residuals of the models. As observed by Naafs et al. (2017a) and Dearing Crampton-Flood et al. (2020) in their global MBT'$_{5Me}$-MAAT calibration, a major part of the uncertainty in the FROG model is likely due to the wide dispersion of predicted temperatures for MAAT < 10°C (Fig. 9c, d), corresponding to samples from high latitudes and/or high elevations. The characteristics of the FROG and FROG$_{5Me}$ model for the whole dataset and test dataset (R^2, RMSE, variance in residuals) were provided in Table 4 and are comparable, which shows that the FROG models do not overfit the data or, if so, to an extremely moderate extent.

Even though the random forest algorithm agnostically selected the set of brGDGTs which best describes the variability of MAAT, it presents the advantage of not being a blackbox, as the mechanism behind it can be described. First, the independent variations in the individual brGDGT variations with estimated MAAT for the FROG model can be compared, with non-linear trends for all the compounds (Sup. Fig. 4), showing the interest of using a non-linear model to describe the relationship between brGDGT distribution and MAAT. For example, the relative abundance of tetramethylated brGDGT *Ia* was observed to non-linearly increase with predicted MAAT, in contrast with compounds *IIa* and *IIIa* (Sup. Fig. 4). This is consistent with the trends obtained from the linear regression models both on the present extended dataset (Sup. Table 1) and on previous soil datasets (e.g. De Jonge et al., 2014; Naafs et al., 2017a), reflecting the decrease in the methylation degree and thus increase in the MBT'$_{5Me}$ with MAAT (Supp. Fig. 1; Eq. 2).

Second, the weight of the different variables used to define the random forest model with MAAT could be quantified using the permutation importance method (Fig 9e, f; Breiman, 2001). This method consists in considering each compound separately and determining how much model performance decreases if this compound is removed from the model. The brGDGTs *Ia, IIa, IIIa* were the homologues predominantly used by the FROG$_{5Me}$ model (Fig. 9e), consistent with the major influence of these compounds in the MBT'$_{5Me}$ (De Jonge et al., 2014) and the existing linear relationship between this index and MAAT (Sup. Fig. 1). Other brGDGT homologues, considered as being less temperature sensitive than the aforementioned acyclic brGDGTs (e.g. De Jonge et al., 2014), were also taken into account by the FROG$_{5Me}$ model, especially compounds *Ib* and *IIb,* which contain one cyclopentyl moiety. These compounds were similarly selected by a stepwise forward selection method to develop brGDGT calibration with MAAT, e.g. for East African lacustrine sediments (Russell et al., 2018) or Chinese soils (Wang et al., 2020). Therefore, different statistical models suggest that homologues *Ib* and *IIb* can improve the brGDGT-MAAT relationship in various environments, as it might take into account the covariations between MAAT and other environmental parameters. In contrast, brGDGTs *Ic* and *IIc*, which are low-abundant in the peat and soil dataset, have only a minor weight in the FROG$_{5Me}$ model (Fig. 9e).

Regarding the random forest model established with all brGDGTs (FROG model), the predominant homologues were also the acyclic brGDGTs *Ia*, *IIa*, and *IIIa,* and to a lesser extent, compound *Ib* as well as compounds *IIIa'*, *IIa'*, i.e. 6-methyl isomers, pointing to their contribution to MAAT reconstruction. 6-methyl brGDGTs were usually considered as being predominantly influenced by pH variations (De Jonge et al., 2014b). Nevertheless, they were also included in some recent local soil calibrations between brGDGT fractional abundances and MAAT in China (Wang et al., 2020) and Mongolia (Dugerdil et al., 2021) and in a regional calibration based on tropical soils (*n*=175; Pérez-Angel et al., 2020). These compounds were shown to be part of those sensitive to temperature and improved the correlation with MAAT (R^2 and RMSE) in comparison with linear MBT'$_{5Me}$-MAAT regression. It should also be noted that both linear and non-linear regression models were tested for the development of the aforementioned pan-tropical soil brGDGT calibration with MAAT (Pérez-Angel et al., 2020). The non-linear multiple regression showed slight improvements over the linear multiple regression based on brGDGT fractional abundances, which was interpreted as showing the importance of taking into account non-linear influences when establishing brGDGT calibrations in warm environments. In agreement with these recent results, the FROG model integrates both types of brGDGT isomers, including 6-methyl ones, thus reflecting the environmental reality, as the whole distribution of brGDGTs, and not only 5-methyl ones, can be concomitantly and non-linearly influenced by environmental parameters including MAAT (Sup. Fig. 4).

The respective weights of the different brGDGTs were observed to differ between the FROG and FROG$_{5Me}$ calibrations (Fig. 9). Nevertheless, the comparison of the statistical characteristics of these models do not allow favoring one vs. another, as both of them showed similar determination coefficients, RMSE and estimation ranges (Table 4). Therefore, both models will be tested on modern soil samples and sedimentary archives and compared in section 4.4.

4.2.2. Comparison of the FROG model with previous global soil calibrations

The FROG/FROG$_{5Me}$ model were compared with previously published global soil/peat linear calibrations based on Ordinary Least Square (De Jonge et al., 2014), Deming (Naafs et al., 2017a) and Bayesian (Dearing Crampton-Flood et al., 2020) regressions. The FROG and FROG$_{5Me}$ models are characterized by a higher determination coefficient and lower RMSE (Fig. 9; R^2=0.81 vs. 0.8 and RMSE=4.09 °C vs. 4.12 °C, respectively) than available calibrations (De Jonge et al., 2014: R^2=0.61, RMSE=4.8°C; *n* =231; Naafs et al., 2017a: R^2= 0.71, RMSE= 4.1°C, *n*=177; Dearing Crampton-Flood et al., 2020: R^2=0.64, RMSE=6.0 °C, *n*=343; Table 4), while being obtained from a larger dataset (*n*=775).

In order to make a direct statistical comparison of the different models, the calibrations by De Jonge et al. (2014), Naafs et al. (2017a) and Dearing Crampton-Flood et al. (2020; BayMBT model) were applied to the same dataset as the one used for the random forest model (Fig. 10; Table 4). This confirms that the FROG model performs better in terms of robustness and accuracy than the previously published global soil and peat calibrations (De Jonge et al., 2014; Naafs et al., 2017a; Dearing-Crampton Flood et al., 2020). The FROG model and the BayMBT both increases the upper limit of MAAT estimation by > 3 °C and 5 °C, respectively, in comparison with the calibrations by De Jonge et al. (2014) and Naafs et al. (2017a; Table 4). In contrast, the lower limit of MAAT estimates for the random forest model is higher than that of the other linear calibrations by De Jonge et al. (2014), Naafs et al. (2017a) and especially Dearing Crampton-Flood et al. (2020). This difference is related to the fact that the extended peat and soil dataset contains a lower number of samples collected under cold climates (12% with MAAT < 0°C). The low representation of such type of samples is a limitation for the training phase of the random forest model. In contrast, by definition, the linear models are able to reach higher or lower limits, even without the presence of « extreme » samples in the dataset. Nevertheless, machine-learning algorithms are flexible and the lower limit of temperature prediction of the FROG model could be decreased by analyzing brGDGTs in a larger number of soil/peat samples from cold settings and adding them to the model.

The higher robustness and accuracy of the FROG model compared to those based on the MBT'_{5Me} (Table 4) could be explained by its non-parametric nature and the fact that it takes into account non-linear influences on the brGDGT distribution (Supp. Fig. 4), unlike the linear models (De Jonge et al., 2014; Naafs et al., 2017a; Dearing Crampton-Flood et al., 2020). In addition, one of the prerequisites of the Bayesian model, as recently proposed by Dearing Crampton-Flood et al. (2020), is to determine a *prior*, before analysis. Nevertheless, as shown in this study, the influence of environmental parameters on brGDGT distribution in soils/peat is sample-dependent (e.g. Fig. 7; Table 3). Therefore, it seems difficult to determine a *prior* adapted to all the samples in such a large dataset. Bayesian models could be more efficient when applied to local/regional calibrations, where the *prior* can be determined more precisely. In any case, this study shows that efficient statistical approaches are useful to improve brGDGT calibrations.

4.3. Development of alternative MAAT calibrations

Although the FROG MAAT calibration presents a strong determination coefficient, the associated RMSE is still ca. 4 °C (Table 4). To test the influence of the thermal regime and MAP on the FROG model and potentially improve its accuracy, alternative submodels based on different subsets of the extended dataset were developed.

4.3.1. Alternative model considering the influence of thermal regime

The FRS was shown to have a significant impact on the brGDGT distribution of the extended soil dataset based on RDA analyses, especially on the 5-methyl isomers (Fig. 7; Table

3). The FRS can be considered as an indirect indicator of the thermal regime which may influence the growth of brGDGT source microorganisms and the production of these lipids. Nevertheless, the impact of the thermal regime on brGDGT distribution and concentration was the object of contrasting observations. The latter was not observed to be affected by seasonal temperature variations in mid-latitude soils (Weijers et al., 2011b; Lei et al., 2016). In contrast, several studies suggested that brGDGTs may be preferentially produced during summer in mid- to high latitude peats (Weijers et al., 2011a; Huguet et al., 2013) or soils (e.g. Deng et al., 2016), reflecting an effect of the seasonality on the brGDGT production, with an enhancement of the latter during the seasonal optimum, for example at the warm season when the soil is not frozen. An increase in the microbial biomass production was reported in unfrozen soils (Schimel and Clein, 1996; Nedwell, 1999; Schimel et al., 2007).

To take into account the potential influence of the thermal regime on brGDGT distribution, an alternative Bayesian calibration was proposed by Dearing Crampton-Flood et al. (2020). This model ($BayMBT_0$) estimates the average temperature of all months that have an average temperature above 0 °C using the MBT'_{5Me} index. The $BayMBT_0$ model ($R^2 = 0.70$; RMSE = 3.8°C; lower limit = 0.9°C, upper limit = 27.1°C; Dearing Crampton-Flood et al., 2020) improves the strength and accuracy of the BayMBT model but does not allow reconstructing negative temperatures. To evaluate the influence of the thermal regime on the FROG model, the same approach as that proposed by Dearing Crampton-Flood et al. (2020) was applied.

In the present extended dataset, monthly temperatures are available for 661 out of the 775 soil samples. This excludes some of the samples, collected in the French Alps, Peruvian Andes, Mts Pollino, Shegyla and Italy (Véquaud et al., 2020). The alternative calibration based on the average temperature of all months that have a temperature above 0 °C ($FROG_0$) is as strong as the global random forest calibration (FROG; Fig. 11), with an $R^2 = 0.84$, and is more accurate (RMSE = 2.5 °C) than the latter (Table 4, Supp. Fig. 6), even though it is based on a slightly reduced dataset (n=661). The $FROG_0$ calibration also performs better than the $BayMBT_0$ model ($R^2 = 0.56$; RMSE = 4.1°C), with a higher R^2, lower RMSE but with a slightly lower range of MAAT estimation (Table 4). Overall, the present study, through RDA analyses (Fig. 7) and the $FROG_0$ model, highlights the effect of the thermal regime on brGDGT distribution. In particular, it can be assumed that the activity of brGDGT source organisms will be reduced within frozen soils. In addition, snowfall can isolate the soil from the atmospheric compartment, decoupling atmospheric temperatures from soil temperatures during the cold season. This confirms the interest of using alternative models including this effect to improve the accuracy of MAAT reconstruction, as also suggested by Dearing Crampton-Flood et al. (2020).

4.3.2. Alternative model considering the influence of MAP

As previously discussed, the brGDGT distribution may be largely impacted by soil moisture (e.g., Dirghangi et al., 2013; Naafs et al., 2017a). In the present study, MAP was used as a proxy of the soil water content (SWC), which varies as a first approximation according to the precipitation regime, although other factors can play a role, such as relief (topography), evapotranspiration, grain size or vegetation cover (Crave and Gascuel-Odoux, 1997; Gómez-Plaza et al., 2001).

The influence of MAP was tested with the FROG_{500} random forest model. This alternative to the FROG model excludes all the samples with MAP < 500 mm/year, corresponding to soils previously defined as dry, and generally described as alkaline soils from arid regions, poor in organic matter (Peterse et al., 2012; Naafs et al., 2017a; Dearing Crampton-Flood et al., 2020). The FROG_{500} model contains 442 samples and only shows a slight improvement in MAAT reconstruction in comparison with the FROG calibration, with a slightly higher determination coefficient (R^2 =0.85) and lower RMSE (3.5°C; Table 4). The slight decrease in the RMSE of the FROG_{500} vs. the FROG model (by ca. 0.5 °C) is likely related to the large reduction of the dataset size by more than 300 samples. Therefore, the use of the FROG_{500} model, as the BayMBT_{500} proposed by Dearing Crampton-Flood et al. (2020), does not seem preferable, as its performance is only slightly better than the original FROG calibration, even though it contains a much lower number of samples.

4.4. Paleo application of the FROG global calibration

The statistical characteristics including determination coefficients and RMSE are not sufficient enough to discriminate between calibrations for paleotemperature reconstructions, as recently noticed by Dugerdil et al. (2021). Therefore, the performance and validity of the FROG models were tested and compared with the temperature record from Pliocene sediments from the North Sea basin (Dearing Crampton-Flood et al., 2018, 2020) and from a Chinese loess-paleosol sequence covering the last 110 kyr (Gao et al., 2012; Lu et al., 2016; Wang et al., 2020). These archives were the object of previous paleostudies, providing a context for the interpretation of the MAAT data from the FROG models.

4.4.1. Application of the FROG models to sediments from the Pliocene

The performance and reliability of the FROG models were tested using a paleorecord based on Pliocene sediments from the North Sea basin (Dearing Crampton-Flood et al., 2018, 2020). BrGDGTs were previously analyzed in this archive, with subsequent reconstruction of past temperature variations during the Pliocene using the MBT'$_{5Me}$ (Dearing Crampton-Flood et al., 2018). More recently, the validity of the BayMBT$_0$ model proposed by Dearing Crampton-Flood et al. (2020) was tested on this archive and compared with the MAAT records derived from the soil brGDGT calibrations by De Jonge et al. (2014) and Naafs et al. (2017a).

The aforementioned calibrations, as well as the FROG/FROG$_{5Me}$ models (Fig.11a), showed an overall decrease in MAAT during the Pliocene, consistent with the global cooling that is documented in the literature (Lisiecki and Raymo, 2005; Dearing Crampton-Flood et al., 2018, 2020). The MAAT records derived from the different models showed similar qualitative trends over the reconstructed period, even though the range of variation appeared slightly smaller (3-4°C) for the FROG model than for the other calibrations (Fig. 11a). The FROG and FROG$_{5Me}$ calibrations provided similar results, implying that the inclusion of the 6-methyl isomers in the models did not significantly change the paleotemperature reconstructions in the present case.

The absolute temperatures reconstructed by the different brGDGT models (Fig. 11a) were generally lower than those expected in NW Europe (13-14 °C) over the Pliocene and derived from pollen assemblages and model outputs (Dearing Crampton-Flood et al., 2020 and references therein). The BayMBT record showed the lowest MAAT estimates, with very large oscillations and negative temperatures during the late Pliocene, in contrast with the other global calibrations (Fig. 11a). Nevertheless, it is unlikely that temperatures went below 0°C during the Pliocene based on the presence of pollen assemblage of warm-adapted species found in Dutch sediments of this period, as specified by Dearing Crampton-Flood et al. (2020). That is why Dearing Crampton-Flood et al. (2020) favored the use of the BayMBT$_0$ model, based on the mean temperature of all months above 0 °C, for MAAT reconstruction, as it is more accurate (RMSE = 3.8°C) than the BayMBT model (RMSE = 6°C). The BayMBT$_0$ model indeed provided higher absolute MAAT estimates than the BayMBT and much more consistent with temperatures estimated from other proxies (i.e. 13-14°C; Fig. 11b).

Although the FROG model did not result in such negative reconstructed MAAT, the FROG$_0$ model was also applied to the Pliocene archive to compare with the BayMBT$_0$ record (Fig. 11b). Both models showed a decreasing trend in temperature over the Pliocene, although the BayMBT$_0$ model displayed larger oscillations (between 3.5 °C and 16.2°C; mean 10.7°C) than the $FROG_0$ one (between 10 °C and 13.1°C; mean 11.7°C). As the MAAT estimates from the $FROG_0$ model are associated with a smaller error (2.5 °C; Table 4) than the other global brGDGT calibrations including the $BayMBT_0$ one (3.8 °C; Dearing Crampton-Flood et al., 2020), an improvement in the accuracy of the paleoreconstruction over this period can be anticipated when using the $FROG_0$ model. This model, unlike the FROG model, estimates the mean temperature of all months above 0°C, which can be considered as more reflecting warm season temperature. Unlike the FROG model, the $FROG_0$ model can neglect the thermal and nival influences on the samples, and so on the source organisms of brGDGTs. We suggest that the $FROG_0$ model should be used in addition to the FROG model rather than alone, in order to obtain complementary information on annual and seasonal temperatures dynamics in paleoclimate studies.

4.4.2. Application of the FROG model to a Chinese loess-paleosol sequence

The FROG calibration was also applied to the Lantian loess-paleosol sequence (LPS), located in the southern Chinese Loess Plateau (Fig. 1 in Gao et al., 2012) and covering the last 110 kyr (Gao et al., 2012). BrGDGT data from this 8.5 m sequence and associated paleotemperature reconstructions were previously published (Gao et al., 2012; Lu et al., 2016; Wang et al., 2020). Recently, Wang et al. (2020) analysed brGDGTs in 149 modern soils covering a large climate gradient in China and calibrated brGDGT distribution against both mean annual soil temperature (MAST) and MAAT. They applied these local MAST and MAAT brGDGT calibrations as well as the global MAAT calibration (MAT_{mr}) by De Jonge et al. (2014) to the LPS sequence over the last 60 kyr and showed that the local MAST calibration provided more reasonable variations in the past continental temperatures than the local (Wang et al., 2020) and global (Peterse et al., 2012; De Jonge et al., 2014) MAAT calibrations. This was notably related to past changes in vegetation coverage which may affect the relationship between MAST and MAAT.

Regarding the modern period, Wang et al. (2020) collected six surface soil samples adjacent to the Lantian LPS to serve as a reference for the present time. The reconstructed MAAT based on the FROG (11.9 ± 0.8 °C) and $FROG_{5Me}$ (11.9 ± 0.9 °C) for these soils are in agreement with the recorded MAAT at this site (12.6 °C; Wang et al., 2020), suggesting that these calibrations can at least be applied for modern MAAT reconstruction in the region. At this site, $MAAT_0$ can be considered as close to MAAT, as a winter (December; January; February) mean air temperature of ~1 °C was reported (Gao et al., 2012). The $FROG_0$ model provided temperature estimates (14.3 ± 0.3 °C) consistent with the expected one. In contrast with the different FROG calibrations, the temperature estimates derived from the BayMBT$_0$ model (16.8 ± 2.0 °C) were much higher than expected, with a large variability between the 6 soil samples. As for the local MAST calibration by Wang et al. (2020), it provided temperatures (12.8 ± 1.4°C) consistent with those measured in soils nearby as previously reported (Wang et al., 2020) (Fig. 12a).

So as to assess the reliability of the calibration, the local MAST calibration by Wang et al. (2020) was applied to the whole LPS covering the last 110 kyr and compared with the MAAT records derived from the different FROG models (FROG, FROG$_{5Me}$ and FROG$_0$) as well as the BayMBT$_0$ (Dearing-Crampton Flood et al., 2020). The BayMBT$_0$ model was chosen rather than the BayMBT one, as it provides most accurate temperature reconstructions, as reported by Dearing Crampton-Flood et al. (2020) and specified above. Nevertheless, it should be noted that the BayMBT$_0$ model, as the FROG$_0$ ones, allow the reconstruction of the mean temperature of all months > 0 °C (MAAT$_0$) instead of the MAAT.

When applied to the Lantian LPS, the different calibrations showed the same qualitative trends (Fig. 12b). Thus, the temperature oscillated between 110 and 60 kyr (with different amplitudes depending on the calibration) and then showed a continuous cooling trend between 60 kyr and 30 kyr, the lowest values being reached between ca. 22 and 27 kyr BP, corresponding to the local Last Glacial Maximum (LGM), as previously observed by Gao et al. (2012) and Lu et al. (2016). Then, the temperature increased rapidly and peaked at the Early Holocene before decreasing toward the present-day values. Such general trends are similar to those previously reported for the Lantian LPS (Gao et al., 2012; Lu et al., 2016; Wang et al., 2020). Local insolation (35 °N) was previously shown to be the dominant factor impacting the temperature records of this site, as similarly observed in other LPSs of the southern plateau (Peterse et al., 2011, 2014; Jia et al., 2013). Thus, the temperature maxima at ca. 12-15 kyr, 63 kyr and 82 kyr BP were consistent with insolation maxima, while the low temperature observed at ca. 90 kyr, 75 kyr, 25 kyr and the late Holocene coincided with insolation minima (Wang et al., 2008, Fig. 1; Gao et al., 2012, Fig. 3; Lu et al., 2016). The brGDGT-derived temperature records and the δ^{18}O record from Chinese speleothems (Wang et al., 2008), related to monsoon intensity, displayed roughly similar patterns (Fig. 12b), showing the relationship existing between the temperature and precipitation intensity, even though the temperature record was generally observed to precede the δ^{18}O record, as already noticed in the Lantian LPS (Gao et al., 2012; Wang et al., 2020) and other sites of the southern plateau (Peterse et al., 2011, 2014; Jia et al., 2013). The FROG, FROG$_{5Me}$ and FROG$_0$ calibrations showed especially well-defined peaks over the period between 75 kyr and 110 kyr BP, corresponding to the glacial and interglacial substages of Marine Isotope Stage 5. These extrema, also apparent when using the local MAST calibration by Wang et al. (2020) and, with a more reduced amplitude, the $BayMBT_0$ model, were similarly observed in the Chinese speleothem $\delta^{18}O$ record (Wang et al., 2008; Fig. 12b), although the exact timing differed probably in relation to age modelling uncertainties (Wang et al., 2008).

The results derived from the brGDGT calibrations were in general agreement with the previously published records and climate models. Thus, the temperature estimates for the LGM derived from the FROG/$FROG_{5Me}$/$FROG_0$ and local MAST (Wang et al., 2020) calibrations were, respectively, ca. 2 °C and 4 °C lower than those of the present-day surficial sediments (Fig. 12a), consistent with the difference of ca. 2-4 °C derived from East Asian climate models (Ju et al., 2007). As for the $BayMBT_0$ calibration, the difference in temperature estimates between LGM and present-day (ca. 3 °C, respectively) should be interpreted with care, as the temperature derived from the top-core sediment (ca. 17 °C) is abnormally high compared to the recorded MAAT (ca. 12 °C), as discussed above for soils surrounding the Lantian LPS, hence $BayMBT_0$ seems to be less relevant than FROG models.

After the local LGM (at ca. 21-24 kyr BP), the temperature was shown to increase by ca. 8 to 11°C (depending on the calibration) between the LGM and the peak at the Early Holocene (Fig. 12b), consistent with the increase of ca. 10 °C observed during this deglaciation based on lacustrine records from central eastern Europe (Sanchi et al., 2014) and western North America (Feakins et al., 2019). The FROG model showed a higher warming trend (by ca. 3 °C) than the $FROG_{5Me}$ one. Such a difference is likely related to the inclusion of the 6-methyl isomers in the FROG model. Loess deposits are developed under arid conditions, which favors the domination of 6-methyl vs. 5-methyl brGDGTs (De Jonge et al., 2014; Naafs et al., 2017a), as it is the case in the Lantian LPS (average IR_{6Me} over the whole sequence: 0.58 ± 0.07). The high abundance of 6-methyl brGDGTs in arid/alkaline soils (i.e. $IR_{6Me} > 0.5$) was shown to make complex the applicability of the MBT'_{5Me} (Naafs et al., 2017a). In the Lantian LPS, this led to the development of specific local calibrations, such as the one proposed by Wang et al. (2020) to reconstruct MAST, based on stepwise regression method and including several 6-methyl brGDGTs (*IIa', IIb', IIIa'*) to improve temperature reconstruction. Similarly, it may not be excluded that the presence of 6-methyl brGDGTs in the FROG model could help in improving paleotemperature reconstructions in comparison with the $FROG_{5Me}$ calibration containing only 5-methyl brGDGTs. Nevertheless, except in the early Holocene, the difference in temperature estimates between the FROG and $FROG5_{Me}$ models was generally <1.5 °C, which is much lower than the RMSE of these calibrations (ca. 4 °C; Table 4). As both of them provided similar qualitative trends, this does not favor one calibration vs. another for paleotemperature reconstruction of this archive.

To conclude, the FROG, FROG$_{5Me}$ and FROG$_0$ calibrations were able to accurately reconstruct present-day temperatures at the Lantian LPS, in contrast with the BayMBT$_0$ model. The reliability of the FROG model was further demonstrated when applied to the whole sedimentary record. It showed documented climatic variations, with a reduced error (4°C) compared to previous global soil calibrations (Peterse et a., 2012; De Jonge et al., 2014), and consistent with the trends derived from a local MAST calibration (Wang et al., 2020).

5. Conclusions

Several global brGDGT calibrations for MAAT reconstruction in soils and peats have been proposed over the last years. Nevertheless, the uncertainty in brGDGT-based temperature estimates is still substantial, largely due to the influence of the various environmental variables in addition to MAAT on brGDGT distribution. A statistical clustering and analysis of the globally distributed brGDGT dataset allowed hierarchizing the parameters affecting brGDGT distribution in soils and thus the MBT'$_{5Me}$-MAAT relationship at the global scale. pH was shown to be the main environmental control on brGDGT distribution, followed by MAAT, over the whole dataset. The five statistical clusters were well-differentiated based on environmental parameters (MAAT, FRS, MAP, pH) and geographical locations and were characterized by distinct brGDGT distributions. A strong relationship between MBT'$_{5Me}$ and MAAT was only observed for the warmer clusters while the MBT'$_{5Me}$-MAAT relationship at the global level was shown to be driven by the moderate correlation corresponding to a cold subgroup containing only ca. 10% of the samples from the total dataset. This highlighted the limitations of using a single index and a simple linear regression model to capture the response of brGDGTs to temperature changes.

A new improved MAAT calibration based on random forest algorithm was then proposed, the so-called Random **F**orest **R**egression for Pale**O**MAAT using br**G**DGTs (**FROG).** The FROG model, which is multi-factorial and non-parametric, appears to be more robust and accurate than previous global calibrations while being proposed on an extended soil and peat dataset. This is related to the fact that it takes into account the non-linear influences on the relationships between MAAT and the relative abundances of individual brGDGTs. Finally, the FROG model was applied to two existing paleorecords and compared with available calibrations, showing its suitability for paleoreconstructions. Application of this new model should improve the accuracy of brGDGT-based MAAT reconstructions in soils and peats, especially in environments where the MBT'$_{5Me}$ shows some limitations because of potential confounding factors. As the random forest algorithm is adaptative and flexible, the FROG model, freely available to the community through an R package, could be easily further improved by the implementation of additional samples in the dataset. The machine-learning approach proposed in this study for calibrating the brGDGT-MAAT relationship could be applied to other settings, such as lacustrine ones.

Research data. FROG models presented in this study are freely available using a R package with a web-application on a GITHUB repository (paleoFROG). The soil dataset used in this study will be added on Pangaea.

Acknowledgments. We thank Sorbonne Université for a PhD scholarship to Pierre Véquaud and the Labex MATISSE (Sorbonne Université) for financial support. The EC2CO programme (CNRS/INSU – BIOHEFECT/MICROBIEN) is thanked for funding of the SHAPE project. Arnaud Huguet and Sergio Contreras are grateful for funding of the ECOS SUD/ECOS ANID #C19U01/190011 project. Andrew T. Nottingham was supported by the UK Natural Environment Research Council (NERC), grant NE/T012226.

References

Baker A., Blith A.J., Jex C.N., Mcdonald J.A., Woltering M., Khan S.J., 2019. Glycerol dialkyl glycerol tetraethers (GDGT) distributions from soil to cave: Refining the speleothem paleothermometer. *Organic Geochemistry* **136**, 103890.

Balleza D., Garcia-Arribas A. B., Sot J., Ruiz-Mirazo K. and Goñi F. M. (2014) Ether- versus Ester-Linked Phospholipid Bilayers Containing either Linear or Branched Apolar Chains. *Biophysical Journal* **107**, 1364–1374.

Blaga C. I., Reichart G.-J., Heiri O. and Sinninghe Damsté J. S. (2009) Tetraether membrane lipid distributions in water-column particulate matter and sediments: a study of 47 European lakes along a north–south transect. *J Paleolimnol* **41**, 523–540.

Braak C. J. F. ter and Smilauer P. (2002) *CANOCO Reference Manual and CanoDraw for Windows User's Guide: Software for Canonical Community Ordination (version 4.5).*, www.canoco.com, Ithaca NY, USA.

Breiman L. (2001) Random Forests. *Machine Learning* **45**, 5–32.

Choler P. (2018) Winter soil temperature dependence of alpine plant distribution: Implications for anticipating vegetation changes under a warming climate. *Perspectives in Plant Ecology, Evolution and Systematics* **30**, 6–15.

Coffinet S., Huguet A., Bergonzini L., Pedentchouk N., Williamson D., Anquetil C., Gałka M., Kołaczek P., Karpińska-Kołaczek M., Majule A., Laggoun-Défarge F., Wagner T. and Derenne S. (2018) Impact of climate change on the ecology of the Kyambangunguru crater marsh in southwestern Tanzania during the Late Holocene. *Quaternary Science Reviews* **196**, 100–117.

Comont L., Laggoun-Défarge F. and Disnar J.-R. (2006) Evolution of organic matter indicators in response to major environmental changes: The case of a formerly cut-over peat bog (Le Russey, Jura Mountains, France). *Organic Geochemistry* **37**, 1736–1751.

Crave A. and Gascuel- Odoux C. (1997) The Influence of Topography on Time and Space Distribution of Soil Surface Water Content. *Hydrological Processes* **11**, 203–210.

Dang XinYue, Xue J., Yang H. and 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.

Dang Xinyue, Yang H., Naafs B. D. A., Pancost R. D. and Xie S. (2016) Evidence of moisture control on the methylation of branched glycerol dialkyl glycerol tetraethers in semi-arid and arid soils. *Geochimica et Cosmochimica Acta* **189**, 24–36.

Davidson E. A., Belk E. and Boone R. D. (1998) Soil water content and temperature as independent or confounded factors controlling soil respiration in a temperate mixed hardwood forest. *Global Change Biology* **4**, 217–227.

Davtian N., Ménot G., Bard E., Poulenard J. and Podwojewski P. (2016) Consideration of soil types for the calibration of molecular proxies for soil pH and temperature using global soil datasets and Vietnamese soil profiles. *Organic Geochemistry* **101**, 140–153.

De Jonge C., Hopmans E. C., Stadnitskaia A., Rijpstra W. I. C., Hofland R., Tegelaar E. and 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. *Organic Geochemistry* **54**, 78–82.

De Jonge C., Hopmans E. C., Zell C. I., Kim J.-H., Schouten S. and Sinninghe Damsté J. S. (2014) Occurrence and abundance of 6-methyl branched glycerol dialkyl glycerol tetraethers in soils: Implications for palaeoclimate reconstruction. *Geochimica et Cosmochimica Acta* **141**, 97–112.

De Jonge C., Radujković D., Sigurdsson B. D., Weedon J. T., Janssens I. and Peterse F. (2019) Lipid biomarker temperature proxy responds to abrupt shift in the bacterial community composition in geothermally heated soils. *Organic Geochemistry* **137**, 103897.

De Jonge C., Kuramae E.E., Radujković D., Weedon J.T., Janssens I.A., Peterse F. (2021). The influence of soil chemistry on branched tetraether lipids in mid- and high latitude soils: Implications for brGDGT-based paleothermometry. *Geochimica et Cosmochimica Acta* **310**, 95-112.

Dearing Crampton-Flood E., Peterse F., Munsterman D. and Sinninghe Damsté J. S. (2018) Using tetraether lipids archived in North Sea Basin sediments to extract North Western European Pliocene continental air temperatures. *Earth and Planetary Science Letters* **490**, 193–205.

Dearing Crampton-Flood E., Tierney J. E., Peterse F., Kirkels F. M. S. A. and Sinninghe Damsté J. S. (2020) BayMBT: A Bayesian calibration model for branched glycerol dialkyl glycerol tetraethers in soils and peats. *Geochimica et Cosmochimica Acta* **268**, 142–159.

Dedysh S. N., Pankratov T. A., Belova S. E., Kulichevskaya I. S. and Liesack W. (2006) Phylogenetic Analysis and In Situ Identification of Bacteria Community Composition in an Acidic Sphagnum Peat Bog. *Appl. Environ. Microbiol.* **72**, 2110–2117.

Denisko D. and Hoffman M. M. (2018) Classification and interaction in random forests. *Proc Natl Acad Sci USA* **115**, 1690–1692.

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

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

Dirghangi S. S., Pagani M., Hren M. T. and Tipple B. J. (2013) Distribution of glycerol dialkyl glycerol tetraethers in soils from two environmental transects in the USA. *Organic Geochemistry* **59**, 49–60.

Dugerdil L., Joanin S., Peyron O., Jouffroy-Bapicot I., Vannière B., Boldgib B., Unkelbach J., Behling H., Ménot G. (2021). Climate reconstructions based on GDGT and pollen surface datasets from Mongolia and Baikal area: calibrations and applicability to extremely cold–dry environments over the Late Holocene. *Climate of the Past* **17**, 1199-1226.

Dunkley Jones T., Eley Y. L., Thomson W., Greene S. E., Mandel I., Edgar K. and Bendle J. A. (2020) OPTiMAL: a new machine learning approach for GDGT-based palaeothermometry. *Climate of the Past* **16**, 2599–2617.

Fawcett P. J., Werne J. P., Anderson R. S., Heikoop J. M., Brown E. T., Berke M. A., Smith S. J., Goff F., Donohoo-Hurley L., Cisneros-Dozal L. M., Schouten S., Sinninghe Damsté J. S., Huang Y., Toney J., Fessenden J., WoldeGabriel G., Atudorei V., Geissman J. W. and Allen C. D. (2011) Extended megadroughts in the southwestern United States during Pleistocene interglacials. Nature 470, 518–521.

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 and Planetary Science Letters* **515**, 164-172.

Gao L., Nie J., Clemens S., Liu W., Sun J., Zech R., Huang Y. (2012). The importance of solar insolation on the temperature variations for the past 110 kyr. *Palaeogeography, Palaeoclimatology, Palaeoecology* **317-318**, 128-133.

Gómez-Plaza A., Martínez-Mena M., Albaladejo J. and Castillo V. M. (2001) Factors regulating spatial distribution of soil water content in small semiarid catchments. *Journal of Hydrology* **253**, 211–226.

Gorham E. (1991) Northern Peatlands: Role in the Carbon Cycle and Probable Responses to Climatic Warming. *Ecological Applications* **1**, 182–195.

Harning D.J., Curtin L., Geirsdóttir Á., D'Andrea W.J., Miller G.H., Sepúlveda J. (2020). Lipid Biomarkers Quantify Holocene Summer Temperature and Ice Cap Sensitivity in Icelandic Lakes. *Geophysical Research Letters* **47**, e2019GL085728.

Harris I., Jones P. D., Osborn T. J. and Lister D. H. (2014) Updated high-resolution grids of monthly climatic observations - the CRU TS3.10 Dataset: UPDATED HIGH-RESOLUTION GRIDS OF MONTHLY CLIMATIC OBSERVATIONS. *Int. J. Climatol.* **34**, 623–642.

Hofmann K., Lamprecht A., Pauli H. and Illmer P. (2016) Distribution of Prokaryotic Abundance and Microbial Nutrient Cycling Across a High-Alpine Altitudinal Gradient in the Austrian Central Alps is Affected by Vegetation, Temperature, and Soil Nutrients. *Microb Ecol* **72**, 704–716.

Hopmans E. C., Schouten S. and Sinninghe Damsté J. S. (2016) The effect of improved chromatography on GDGT-based palaeoproxies. *Organic Geochemistry* **93**, 1–6.

Huguet A., Coffinet S., Roussel A., Gayraud F., Anquetil C., Bergonzini L., Bonanomi G., Williamson D., Majule A. and Derenne S. (2019) Evaluation of 3-hydroxy fatty acids as a pH and temperature proxy in soils from temperate and tropical altitudinal gradients. *Organic Geochemistry* **129**, 1–13.

Huguet A., Fosse C., Laggoun-Défarge F., Delarue F. and Derenne S. (2013) Effects of a short-term experimental microclimate warming on the abundance and distribution of branched GDGTs in a French peatland. *Geochimica et Cosmochimica Acta* **105**, 294–315.

Huguet A., Fosse C., Laggoun-Défarge F., Toussaint M.-L. and Derenne S. (2010) Occurrence and distribution of glycerol dialkyl glycerol tetraethers in a French peat bog. *Organic Geochemistry* **41**, 559–572.

Huguet A., Francez A.-J., Jusselme M. D., Fosse C. and Derenne S. (2014) A climatic chamber experiment to test the short term effect of increasing temperature on branched GDGT distribution in Sphagnum peat. *Organic Geochemistry* **73**, 109–112.

Idso S. B., Schmugge T. J., Jackson R. D. and Reginato R. J. (1975) The utility of surface temperature measurements for the remote sensing of surface soil water status. *Journal of Geophysical Research (1896-1977)* **80**, 3044–3049.

Jia G.D., Rao Z.G., Zhang J., Li Z.Y., Chen F.H. (2013). Tetraether biomarker records from a loess-paleosol sequence in the western Chinese Loess Plateau. *Frontiers in Microbiology* **4** https://doi.org/10.3389/fmicb.2013.0019.

Ju L.,Wang H., Jiang D. (2007). Simulation of the Last Glacial Maximum climate over East Asia with a regional climate model nested in a general circulation model. *Palaeogeography, Palaeoclimatology, Palaeoecology* **248**, 376-390.

Killops S. (2005) Introduction to Organic Geochemistry, 2nd edn (paperback) ed. V. Killops. *Geofluids* **5**, 236–237.

Lauber C. L., Hamady M., Knight R. and Fierer N. (2009) Pyrosequencing-Based Assessment of Soil pH as a Predictor of Soil Bacterial Community Structure at the Continental Scale. *Appl. Environ. Microbiol.* **75**, 5111–5120.

Lei Y., Yang H., Dang X., Zhao S. and Xie S. (2016) Absence of a significant bias towards summer temperature in branched tetraether-based paleothermometer at two soil sites with contrasting temperature seasonality. *Organic Geochemistry* **94**, 83–94.

Liang J., Russell J. M., Xie H., Lupien R. L., Si G., Wang J., Hou J. and Zhang G. (2019) Vegetation effects on temperature calibrations of branched glycerol dialkyl glycerol tetraether (brGDGTs) in soils. *Organic Geochemistry* **127**, 1–11.

Lisiecki L. E. and Raymo M. E. (2005) A Pliocene-Pleistocene stack of 57 globally distributed benthic δ18O records. *Paleoceanography* **20**.

Loomis S.E., Russell J.M., Sinninghe Damsté J.S. (2010). Distributions of branched GDGTs in soils and lake sediments from western Uganda: Implications for a lacustrine paleothermometer. Organic Geochemistry 42, 739-751.

Loomis S. E., Russell J. M., Ladd B., Street-Perrott F. A. and Sinninghe Damsté J. S. (2012) Calibration and application of the branched GDGT temperature proxy on East African lake sediments. *Earth and Planetary Science Letters* **357–358**, 277–288.

Lu H., Liu W., Wang H. and Wang Z. (2016) Variation in 6-methyl branched glycerol dialkyl glycerol tetraethers in Lantian loess–paleosol sequence and effect on paleotemperature reconstruction. *Organic Geochemistry* **100**, 10–17.

McMaster G. and Wilhelm W. (1997) Growing degree-days: one equation, two interpretations. *Publications from USDA-ARS / UNL Faculty*.

Menges J., Huguet C., Alcañiz J. M., Fietz S., Sachse D. and Rosell-Melé A. (2014) Influence of water availability in the distributions of branched glycerol dialkyl glycerol tetraether in soils of the Iberian Peninsula. *Biogeosciences* **11**, 2571–2581.

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

Naafs B. D. A., Inglis G. N., Zheng Y., Amesbury M. J., Biester H., Bindler R., Blewett J., Burrows M. A., del Castillo Torres D., Chambers F. M., Cohen A. D., Evershed R. P., Feakins S. J., Gałka M., Gallego-Sala A., Gandois L., Gray D. M., Hatcher P. G., Honorio Coronado E. N., Hughes P. D. M., Huguet A., Könönen M., Laggoun-Défarge F., Lähteenoja O., Lamentowicz M., Marchant R., McClymont E., Pontevedra-Pombal X., Ponton C., Pourmand A., Rizzuti A. M., Rochefort L., Schellekens J., De Vleeschouwer F. and Pancost R. D. (2017b) Introducing global peat-specific temperature and pH calibrations based on brGDGT bacterial lipids. *Geochimica et Cosmochimica Acta* **208**, 285–301.

Nedwell D. B. (1999) Effect of low temperature on microbial growth: lowered affinity for substrates limits growth at low temperature. *FEMS Microbiol Ecol* **30**, 101–111.

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. *Geochimica et Cosmochimica Acta* 75, 6225-6238.

Pérez-Angel L.C., Sepúlveda J., Molnar P., Montes C., Rajagopalan B., Snell K., Gonzalez-Arango C., Dildar N. (2020). Soil and air temperature calibrations using branched GDGTs for the Tropical Andes of Colombia: Toward a Pan-tropical calibration. Geochemistry, Geophysics, Geosystems 21, e2020GC008941

Peterse F., Kim J.-H., Schouten S., Kristensen D. K., Koç N. and Sinninghe Damsté J. S. (2009) Constraints on the application of the MBT/CBT palaeothermometer at high latitude environments (Svalbard, Norway). *Organic Geochemistry* **40**, 692–699.

Peterse F., Prins M.A., Beets C.J., Troelstra S.R., Zheng H., Gu Z., Schouten S., Sinninghe Damsté J.S. (2011). Decoupled warming and monsoon precipitation in East Asia over the last deglaciation. *Earth and Planetary Science Letters* **301**, 256-264.

Peterse F., van der Meer J., Schouten S., Weijers J. W. H., Fierer N., Jackson R. B., Kim J.-H. and Sinninghe Damsté J. S. (2012) Revised calibration of the MBT–CBT paleotemperature proxy based on branched tetraether membrane lipids in surface soils. *Geochimica et Cosmochimica Acta* **96**, 215–229.

Peterse F., Martínez-García A., Zhou B., Beets C.J., Prins M.A., Zheng H., Eglinton T.I. (2014). Molecular records of continental air temperature and monsoon precipitation variability in East Asia spanning the past 130,000 years. *Quaternary Science Reviews* **83**, 76-82.

Peterse F., Moy C. M. and Eglinton T. I. (2015) A laboratory experiment on the behaviour of soil-derived core and intact polar GDGTs in aquatic environments. *Biogeosciences* **12**, 933–943.

Powers L., Werne J. P., Vanderwoude A. J., Sinninghe Damsté J. S., Hopmans E. C. and Schouten S. (2010) Applicability and calibration of the TEX86 paleothermometer in lakes. *Organic Geochemistry* **41**, 404–413.

Raberg J.H., Harning D.J., Crump S.E., de Wet G., Blumm A., Kopf S., Geirsdóttir A., Miller G.H., Sepúlveda J. (2021). Revised fractional abundances and warm-season temperatures substantially improve brGDGT calibrations in lake sediments. Biogeosciences 18, 3579-3603.

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 sediments: Effects of temperature, pH, and new lacustrine paleotemperature calibrations. *Organic Geochemistry* **17**, 56-69.

Sanchi L., Ménot G., Bard E. (2014). Insights into continental temperatures in the northwestern Black Sea area during the Last Glacial period using branched tetraether lipids. *Quaternary Science Reviews* **84**, 98-108.

Schimel J., Balser T. C. and Wallenstein M. (2007) Microbial Stress-Response Physiology and Its Implications for Ecosystem Function. *Ecology* **88**, 1386–1394.

Schimel J. P. and Clein J. S. (1996) Microbial response to freeze-thaw cycles in tundra and taiga soils. *Soil Biology and Biochemistry* **28**, 1061–1066.

Shen C., Shi Y., Fan K., He J.-S., Adams J. M., Ge Y. and Chu H. (2019) Soil pH dominates elevational diversity pattern for bacteria in high elevation alkaline soils on the Tibetan Plateau. *FEMS Microbiol Ecol* **95**.

Siles J. A. and Margesin R. (2016) Abundance and Diversity of Bacterial, Archaeal, and Fungal Communities Along an Altitudinal Gradient in Alpine Forest Soils: What Are the Driving Factors? *Microb Ecol* **72**, 207–220.

Sinninghé Damsté J. S., Ossebaar J., Abbas B., Schouten S. and Verschuren D. (2009) Fluxes and distribution of tetraether lipids in an equatorial African lake: Constraints on the application of the TEX86 palaeothermometer and BIT index in lacustrine settings. *Geochimica et Cosmochimica Acta* **73**, 4232–4249.

Sinninghé Damsté J. S., Rijpstra W. I. C., Foesel B. U., Huber K. J., Overmann J., Nakagawa S., Kim J. J., Dunfield P. F., Dedysh S. N. and Villanueva L. (2018) An overview of the occurrence of ether- and ester-linked iso-diabolic acid membrane lipids in microbial cultures of the Acidobacteria: Implications for brGDGT paleoproxies for temperature and pH. *Org. Geochem.* **124**, 63–76.

Sinninghé Damsté J. S., Rijpstra W. I. C., Hopmans E. C., Foesel B. U., Wüst P. K., Overmann J., Tank M., Bryant D. A., Dunfield P. F., Houghton K. and Stott M. B. (2014) Ether- and Ester-Bound iso-Diabolic Acid and Other Lipids in Members of Acidobacteria Subdivision 4. *Appl. Environ. Microbiol.* **80**, 5207–5218.

Sinninghé Damsté J. S., Rijpstra W. I. C., Hopmans E. C., Weijers J. W. H., Foesel B. U., Overmann J. and Dedysh S. N. (2011) 13,16-Dimethyl Octacosanedioic Acid (iso-Diabolic Acid), a Common Membrane-Spanning Lipid of Acidobacteria Subdivisions 1 and 3. *Appl. Environ. Microbiol.* **77**, 4147–4154.

Tierney J. E. and Russell J. M. (2009) Distributions of branched GDGTs in a tropical lake system: Implications for lacustrine application of the MBT/CBT paleoproxy. *Organic Geochemistry* **40**, 1032–1036.

Véquaud P., Derenne S., Anquetil C., Collin S., Poulenard J., Sabatier P. and Huguet A. (2021) Influence of environmental parameters on the distribution of bacterial lipids in soils from the French Alps: Implications for paleo-reconstructions. *Organic Geochemistry* **153**, 104194.

Véquaud P., Derenne S., Thibault A., Anquetil C., Bonanomi G., Collin S., Contreras S., Nottingham A., Sabatier P., Salinas N., Scott W. P., Werne J. P. and Huguet A. (2020) Development of global temperature and pH calibrations based on bacterial 3-hydroxy fatty acids in soils. *Biogeosciences Discussions*, 1–40.

Wang Y.J., Cheng H., Edwards R.L., Kong X.G., Shao X.H., Chen S.T., Wu J.Y., Jiang X.Y., Wang X.F., An Z.S. (2008). Millennial- and orbital-scale changes in the East Asian monsoon over the past 224,000 years. *Nature* **451**, 1090-1093.

Wang H., An Z., Lu H., Zhao Z. and Liu W. (2020) Calibrating bacterial tetraether distributions towards in situ soil temperature and application to a loess-paleosol sequence. *Quaternary Science Reviews* **231**, 106172.

Wang H., Liu W. and Lu H. (2016) Appraisal of branched glycerol dialkyl glycerol tetraether-based indices for North China. *Organic Geochemistry* **98**, 118–130.

Weber Y., De Jonge C., Rijpstra W. I. C., Hopmans E. C., Stadnitskaia A., Schubert C. J., Lehmann M. F., Sinninghe Damsté J. S. and Niemann H. (2015) Identification and carbon isotope composition of a novel branched GDGT isomer in lake sediments: Evidence for lacustrine branched GDGT production. *Geochimica et Cosmochimica Acta* **154**, 118–129.

Weijers J. W. H., Schouten S., van den Donker J. C., Hopmans E. C. and Sinninghe Damsté J. S. (2007) Environmental controls on bacterial tetraether membrane lipid distribution in soils. *Geochimica et Cosmochimica Acta* **71**, 703–713.

Weijers J. W. H., Steinmann P., Hopmans E. C., Schouten S. and Sinninghe Damsté J. S. (2011a) Bacterial tetraether membrane lipids in peat and coal: Testing the MBT–CBT temperature proxy for climate reconstruction. *Organic Geochemistry* **42**, 477–486.

Weijers J.W., Bernhardt B., Peterse F., Werne J.P., Dungait J.A., Schouten S., Sinninghe Damsté J.S. (2011b). Absence of seasonal patterns in MBT-CBT indices in mid-latitude soils. *Geochimica et Cosmochimica Acta* **75**, 3179-3190.

Xiao W., Xu Y., Ding S., Wang Y., Zhang X., Yang H., Wang G. and Hou J. (2015) Global calibration of a novel, branched GDGT-based soil pH proxy. *Organic Geochemistry* **89–90**, 56–60.

Yang H., Lü X., Ding W., Lei Y., Dang X. and Xie S. (2015) The 6-methyl branched tetraethers significantly affect the performance of the methylation index (MBT′) in soils from an altitudinal transect at Mount Shennongjia. *Organic Geochemistry* **82**, 42–53.

Figure and table captions

Figure 1. Principal Component Analysis performed on the global dataset (*n*=767) with the k-means clustering. (a) PC2 vs PC1, (b) PC3 vs PC1, (c) Variance explained (%) for each component, (d) Optimal number of clusters according to the elbow method, based on the observation of the Within Cluster Sum of Squares (WCSS).

Figure 2. Boxplot showing the distribution of the 4 environmental variables considered: (a) pH, (b) MAAT (°C), (c) MAP (mm/year), (d) FRS. Interquartile range (IQR) = $Q_3 – Q_1$ where Q_3 is the 75th percentile and Q_1 is the 25th percentile. Outliers are defined with a 1.5 coefficient on the IQR. Letters on the panel show the differences between each cluster according to Kruskall-Wallis and Dunn post-hoc tests.

Figure 3. Spatial distribution of samples in the global dataset. The colors correspond to the different clusters.

Figure 4. Fractional abundances of the individual brGDGTs determined in the (a) Cluster A, (b) Cluster B, (c) Cluster C, (d) Cluster D, (e) Cluster E and (f) in the global dataset.

Figure 5. Boxplot showing the distribution of the (a) MBT'_{5Me}, (b) CBT', (c) IR_{6ME}, (d) Community Index (CI; Eq. 5; defined by De Jonge et al. (2019) for the 5 clusters defined after PCA analysis (Fig. 2). The CI thresholds of 0.64 and 0.69 separating "warm" and "cold" groups as proposed by De Jonge et al. (2019) and in the present study, respectively, are represented in panel (d). Letters on the panels show the differences between each cluster according to Kruskall-Wallis and Dunn post-hoc tests.

Figure 6. Linear regressions between MBT'_{5Me} and MAAT (°C) for (a) cluster A, (b) cluster B, (c) cluster C, (d) cluster D, (e) cluster E, (f) the whole dataset. Grey dots correspond to soils and green dots to peats samples.

Figure 7. Redundancy analysis between brGDGT distribution and environmental variables for (a) cluster A, (b) cluster B, (c) cluster C, (d) cluster D and (e) cluster E.

Figure 8. Comparison of the linear regressions between MBT'_{5Me} and MAAT (°C) in the two subgroups derived from the extended dataset: "cold" cluster (CI < 0.64) and "warm" cluster (CI > 0.64). The community index (CI; Eq. 3) was defined by De Jonge et al. (2019). The dashed line corresponds to the linear relationship between MAAT (°C) and MBT'_{5Me} in the global dataset used in this study (*n*=775).

Figure 9. MAAT predicted by the random forest models using the relative abundances of (a) 5-methyl brGDGTs ($FROG_{5Me}$) and (b) all brGDGTs (FROG model). Residuals of the (c) $FROG_{5Me}$ model and (d) FROG model with all brGDGTs plotted against predicted MAAT. Importance of the individual brGDGTs in (e) the $FROG_{5Me}$ and (f) FROG model with all brGDGTs, according to the permutation importance method (Breiman, 2001). These results were obtained from the test dataset. Grey dots correspond to soils and green dots to peats.

Figure 10. Comparison of the global MBT'_{5Me} calibration proposed by (a) De Jonge et al. (2014), (b) Naafs et al. (2017a) and (c) Dearing Crampton-Flood et al. (2020; BayMBT model) using the same dataset as the FROG model (*n* = 192). Grey dots correspond to soils and green dots to peats.

Figure 11. Reconstructed MAAT for the Pliocene marine sediment sequence from the Hank core located in the Netherlands (Dearing Crampton-Flood et al., 2018, 2020) (a) derived from the calibrations by De Jonge et al. (2014), Naafs et al. (2017a), Dearing Crampton-Flood et al. (2020; BayMBT) and from the FROG/$FROG_{5Me}$ models and (b) derived from the calibrations by Dearing Crampton-Flood et al. (2020; $BayMBT_0$) and from the $FROG_0$ model. Grey zones are the 95% intervals for the FROG and $FROG_0$ models

Figure 12. (a) Estimated temperatures from the 6 surface soils collected adjacent to the Lantian LPS, serving as a reference for the present time (Wang et al., 2020), derived from the $BayMBT_0$ model (Dearing Crampton-Flood et al. 2020) and FROG/$FROG_{5Me}$/$FROG_0$ models (this study). (b) Comparison of the MAAT estimates from the Lantian LPS sequence covering the last 110 kyr, derived from local MAST calibration (Wang et al., 2020) the $BayMBT_0$ model (Dearing Crampton-Flood et al. 2020) and FROG/$FROG_{5Me}$/$FROG_0$ models (this study), with a $\delta^{18}O$ record from a Chinese speleothem (in green; Wang et al., 2008).

Table 1. Location, number, and references for soils and peat samples used to establish the new global brGDGT calibration proposed in this study. Available parameters for the different sampling sites are shown: MAAT (Mean Annual Air Temperature (°C), pH, MAP (Mean Annual Precipitation (mm/yr), FRS (Number of frost days per year). Asterisks represent the new samples added to the global dataset.

Table 2. Quantitative description of the 5 clusters obtained after k-means on PCA on the brGDGT distribution of the total dataset.

Table 3. RDA correlation coefficients of the selected environmental variables along axes 1 and 2 for each cluster, and quantification of the influence of the different environmental variables on brGDGT relative abundances. Statistical significance ($p<0.05$) is shown with an asterisk.

Table 4. Characteristics of the different brGDGT models compared in this study to estimate MAAT in terrestrial settings: R^2, RMSE (or RMSEP; i.e. Root Mean Square Error of Prediction, for the results on the test dataset), variance of the residuals and the upper and lower limits of estimation. The "training" samples (75%) were used to develop the different machine learning models, which were then tested on the remaining sample set. Characteristics are presented for the test dataset and all the dataset (under brackets, in italics). Previous calibrations from De Jonge et al. (2014), Naafs et al. (2017a) and Dearing Crampton-Flood et al. (2020) are indicated with asterisks.

Supplementary figure 1 Linear relationship between MBT'_{5Me} and MAAT (°C) in the global dataset used in this study (n=775).

Supplementary figure 2 Comparison of the linear regressions between MBT'_{5Me} and MAAT (°C) in the global dataset for samples with $IR_{6Me}>0.5$ and $IR_{6Me}<0.5$.

Supplementary figure 3 (a) Estimation of the community index thresholds, for each linear regression between MAAT and $MBT'_{5Me;}$ (b) Comparison of the linear regressions between MBT'_{5Me} and MAAT (°C) in the two subgroups derived from the extended dataset: “cold” cluster (CI < 0.69) and “warm” cluster (CI > 0.69). The dashed line corresponds to the linear relationship between MAAT (°C) and MBT'_{5Me} in the global dataset used in this study (n=775)

Supplementary figure 4. Partial plots of the individual brGDGT variations in the FROG model proposed to estimate MAAT.

Supplementary figure 5. Partial plots of the individual brGDGT variations in the $FROG_{5Me}$ model proposed to estimate MAAT.

Supplementary figure 6. (a) MAAT predicted by the $FROG_0$ model using the relative abundances of all brGDGTs. (b) Residuals of the $FROG_0$ model. (c) Importance of the individual brGDGTs in the $FROG_0$ model, according to the permutation importance method (Breiman, 2001). These results were obtained from the test dataset.

Supplementary Table 1. (a) Correlation matrix between the fractional abundances of the brGDGTs, and the MAAT, MAP, pH, FRS, CBT', MBT'5Me, the community index (CI) and the IR_{6Me} in the global dataset presented in this study (n=775). The values given are the R^2. (b) *p*-values of the correlations shown in the supplementary table 1a.

Supplementary Table 2. (a) Correlation matrix between the fractional abundances of the brGDGTs, and the MAAT, MAP, pH, FRS, CBT', MBT'_{5Me}, the community index (CI) and the IR_{6Me} in the Cluster A. The values given are the R^2. (b) *p*-values of the correlations shown in supplementary table 2a.

Supplementary Table 3. (a) Correlation matrix between the fractional abundances of the brGDGTs, and the MAAT, MAP, pH, FRS, CBT', MBT'_{5Me}, the community index (CI) and the IR_{6Me} in the Cluster B. The values given are the R^2. (b) *p*-values of the correlations shown in supplementary table 3a.

Supplementary Table 4. (a) Correlation matrix between the fractional abundances of the brGDGTs, and the MAAT, MAP, pH, FRS, CBT', MBT'_{5Me}, the community index (CI) and the IR_{6Me} in the Cluster C. The values given are the R^2. (b) *p*-values of the correlations shown in the supplementary table 4a

Supplementary Table 5. (a) Correlation matrix between the fractional abundances of the brGDGTs, and the MAAT, MAP, pH, FRS, CBT', MBT'_{5Me}, the community index (CI) and the IR_{6Me} in the Cluster D. The values given are the R^2. (b) *p*-values of the correlations shown in the supplementary table 5a.

Supplementary Table 6. (a) Correlation matrix between the fractional abundances of the brGDGTs, and the MAAT, MAP, pH, FRS, CBT', MBT'_{5Me}, the community index (CI) and the IR_{6Me} in the Cluster E. The values given are the R^2. (b) p-values of the correlations shown in the supplementary table 6a.

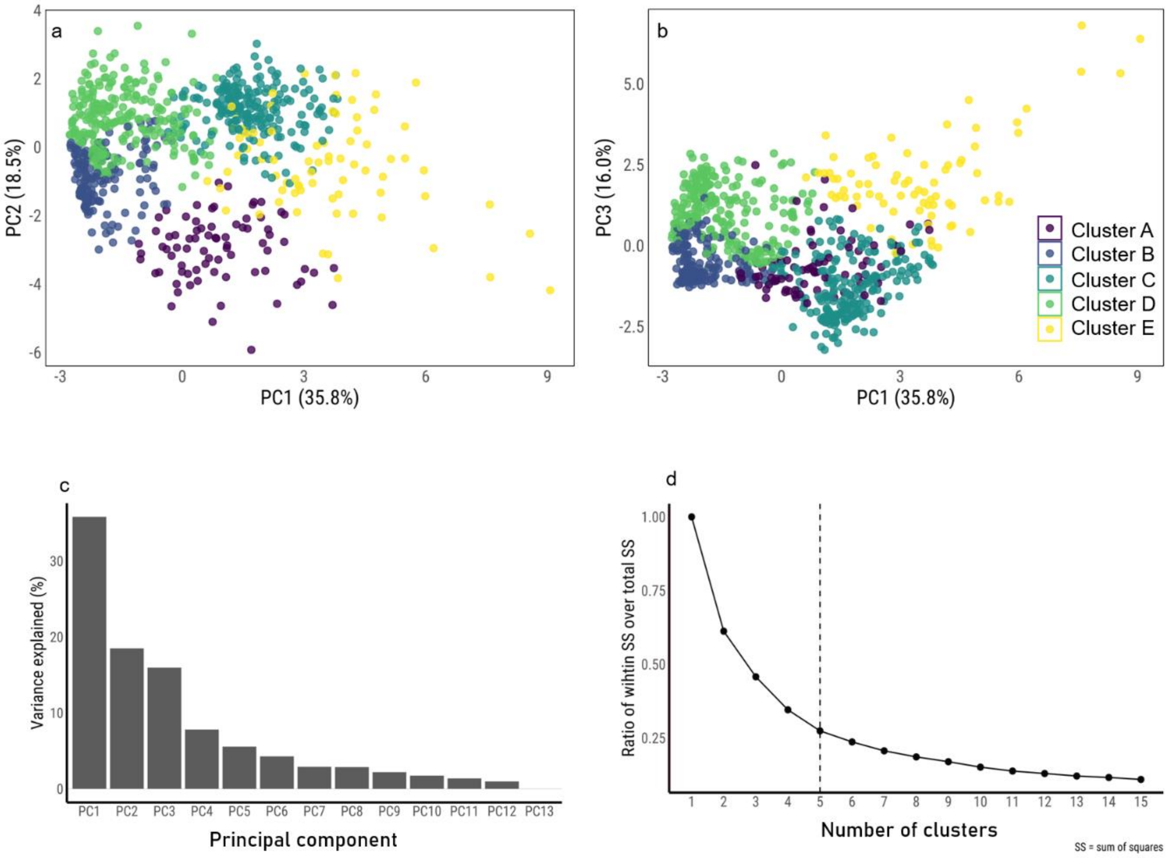

a
PC1 (35.8%)
PC2 (18.5%)
4
2
0
-2
-4
-6
-3
0
3
6
9
b
PC1 (35.8%)
PC3 (16.0%)
5.0
2.5
0.0
-2.5
Cluster A
Cluster B
Cluster C
Cluster D
Cluster E
c
Variance explained (%)
30
20
10
0
PC1
PC2
PC3
PC4
PC5
PC6
PC7
PC8
PC9
PC10
PC11
PC12
PC13
Principal component
d
Ratio of wihtin SS over total SS
1.00
0.75
0.50
0.25
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
Number of clusters
SS = sum of squares

Figure 1

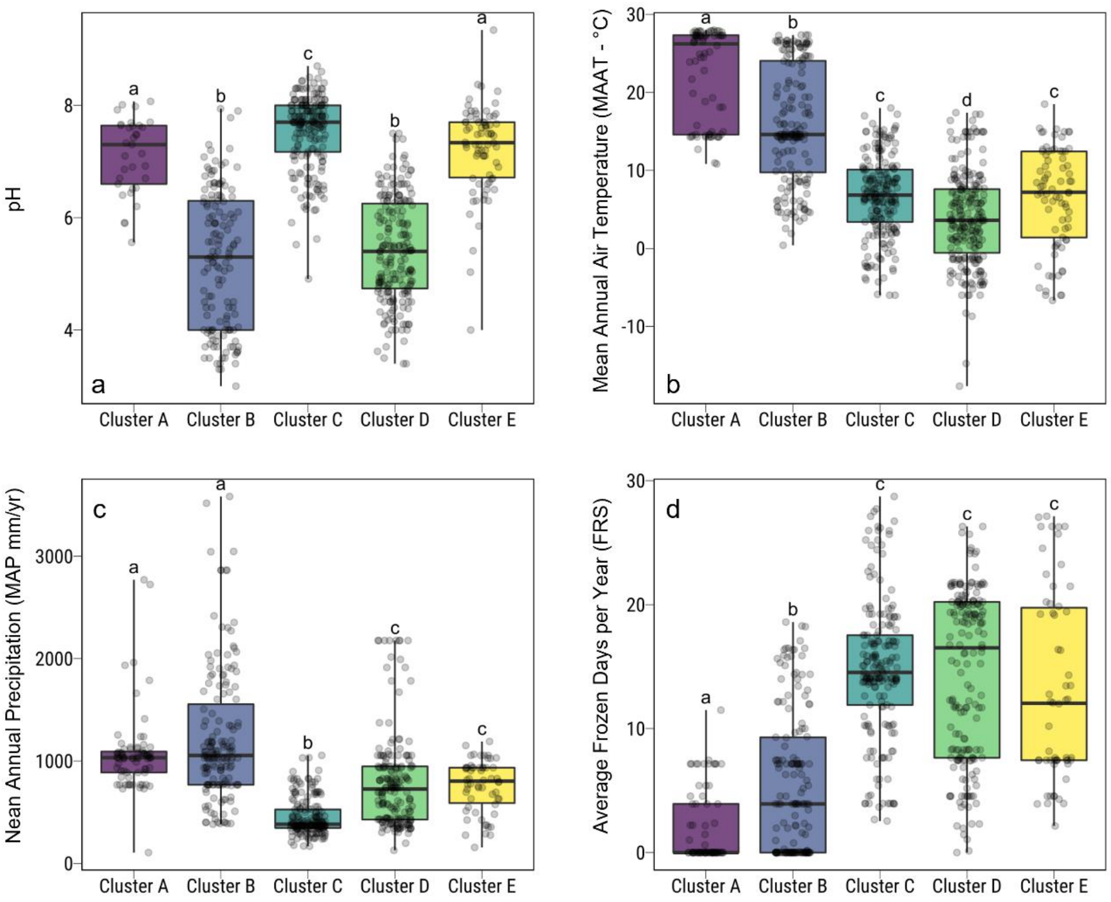

a
pH
4
6
8
b
Mean Annual Air Temperature (MAAT - °C)
-10
0
10
20
30
c
Nean Annual Precipitation (MAP mm/yr)
0
1000
2000
3000
d
Average Frozen Days per Year (FRS)
0
10
20
30
a
b
c
b
a
a
b
c
d
c
a
a
b
c
c
a
b
c
c
c
Cluster A
Cluster B
Cluster C
Cluster D
Cluster E

Figure 2

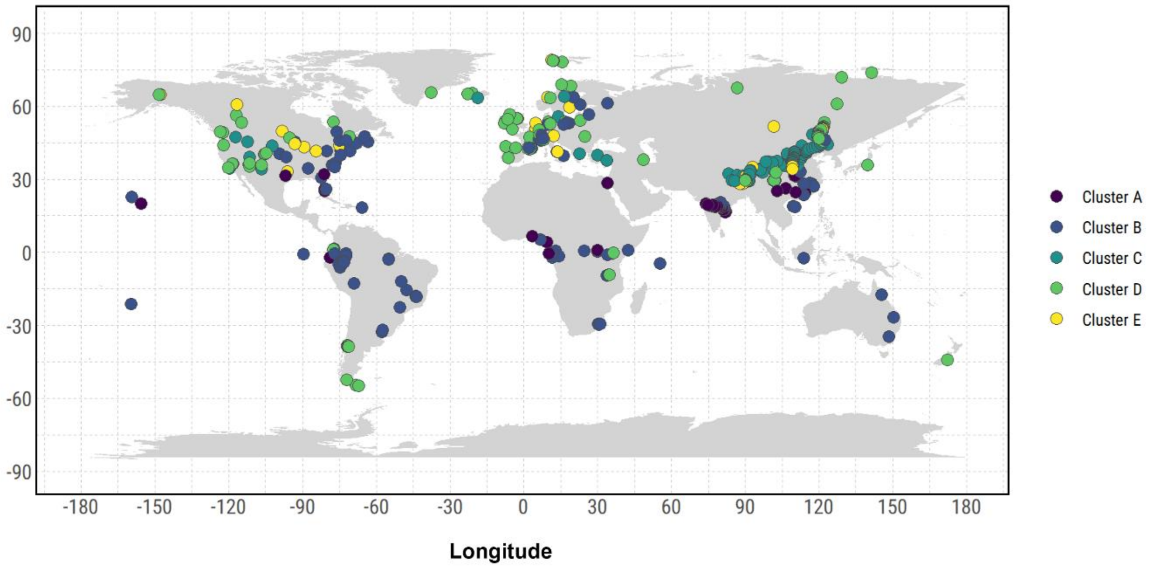

90
60
30
0
-30
-60
-90
-180
-150
-120
-90
-60
-30
0
30
60
90
120
150
180
Longitude
Cluster A
Cluster B
Cluster C
Cluster D
Cluster E

Figure 3

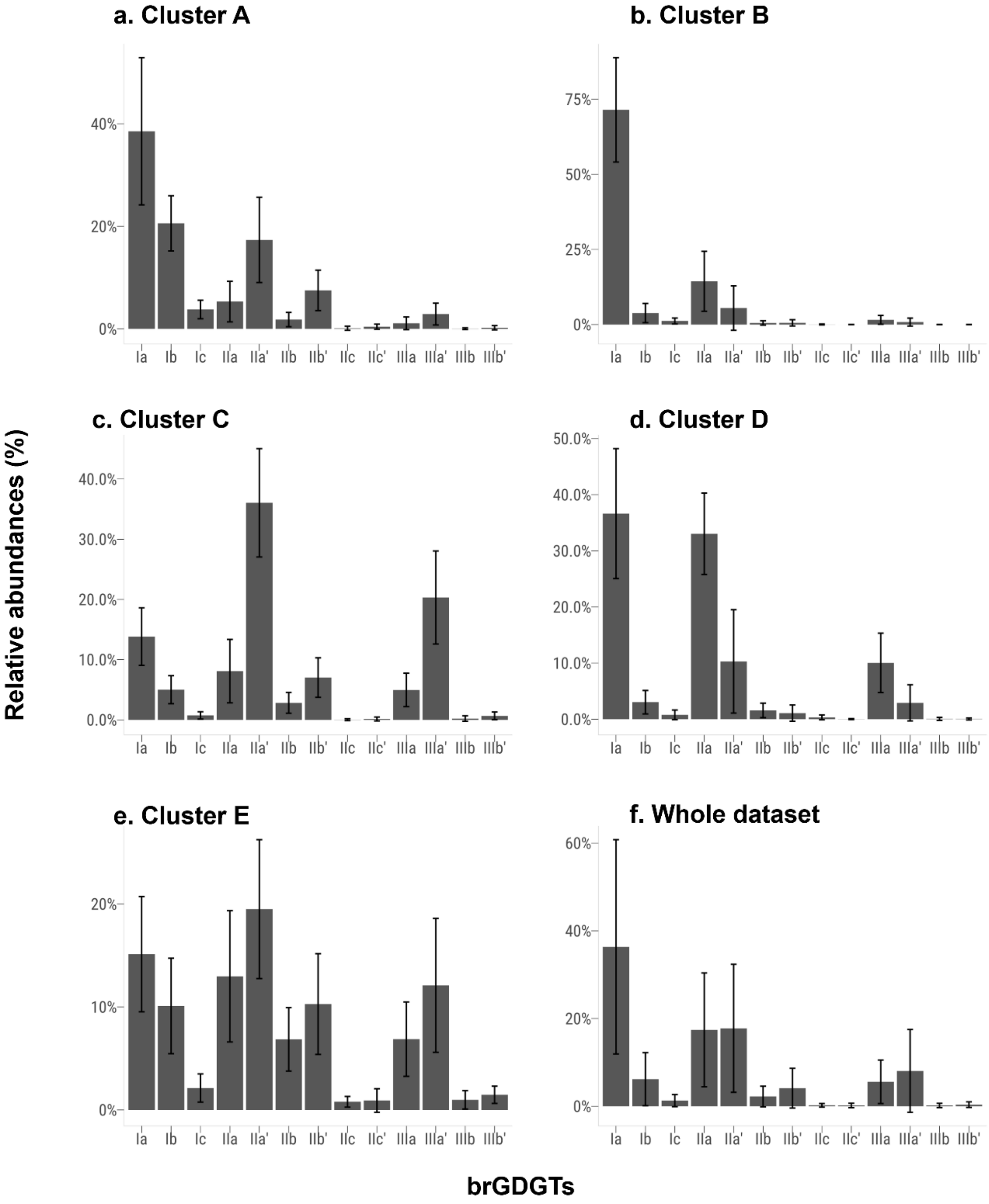

a. Cluster A
40%
20%
0%
b. Cluster B
75%
50%
25%
0%
c. Cluster C
40.0%
30.0%
20.0%
10.0%
0.0%
d. Cluster D
50.0%
40.0%
30.0%
20.0%
10.0%
0.0%
e. Cluster E
20%
10%
0%
f. Whole dataset
60%
40%
20%
0%
Ia
Ib
Ic
IIa
IIa'
IIb
IIb'
IIc
IIc'
IIIa
IIIa'
IIIb
IIIb'
Relative abundances (%)
brGDGTs

Figure 4

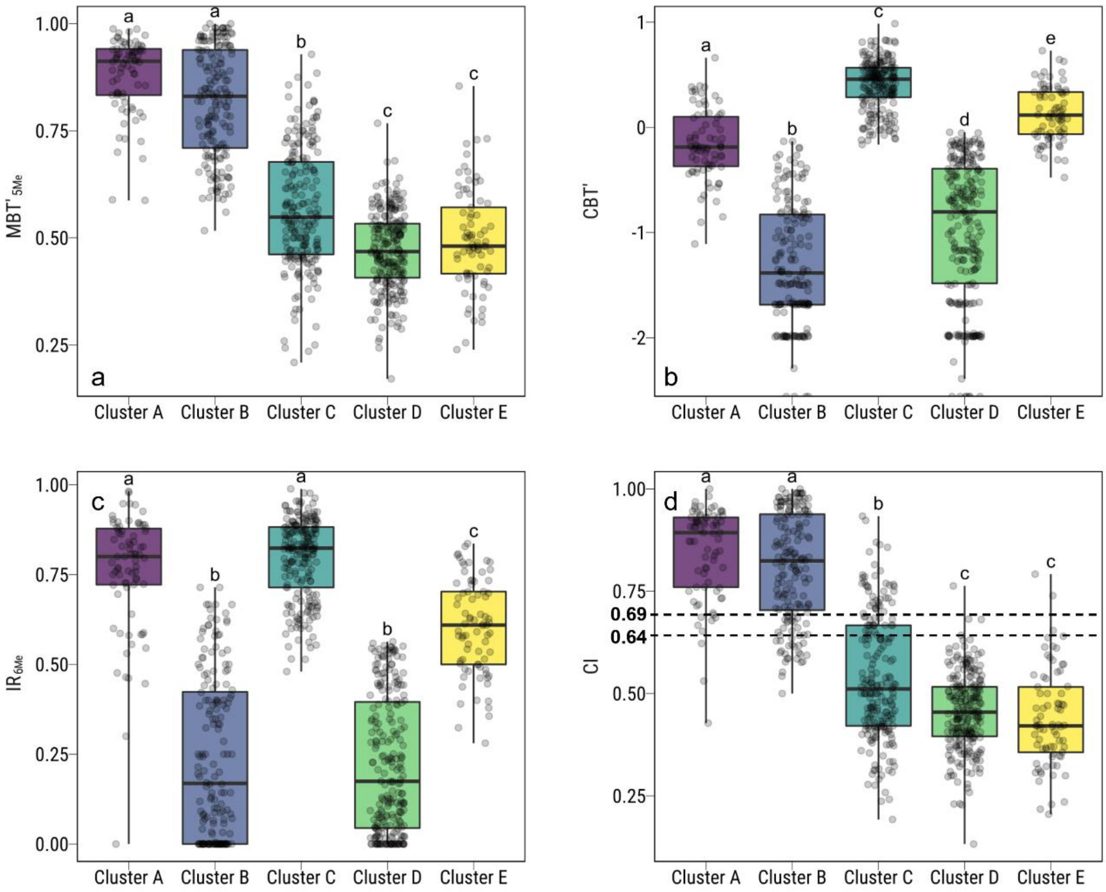

a
MBT'$_{5Me}$
1.00
0.75
0.50
0.25
a a b c c
b
CBT'
1
0
-1
-2
a b c d e
c
IR$_{6Me}$
1.00
0.75
0.50
0.25
0.00
a b a b c
d
CI
1.00
0.75
0.69
0.64
0.50
0.25
a a b c c
Cluster A
Cluster B
Cluster C
Cluster D
Cluster E

Figure 5

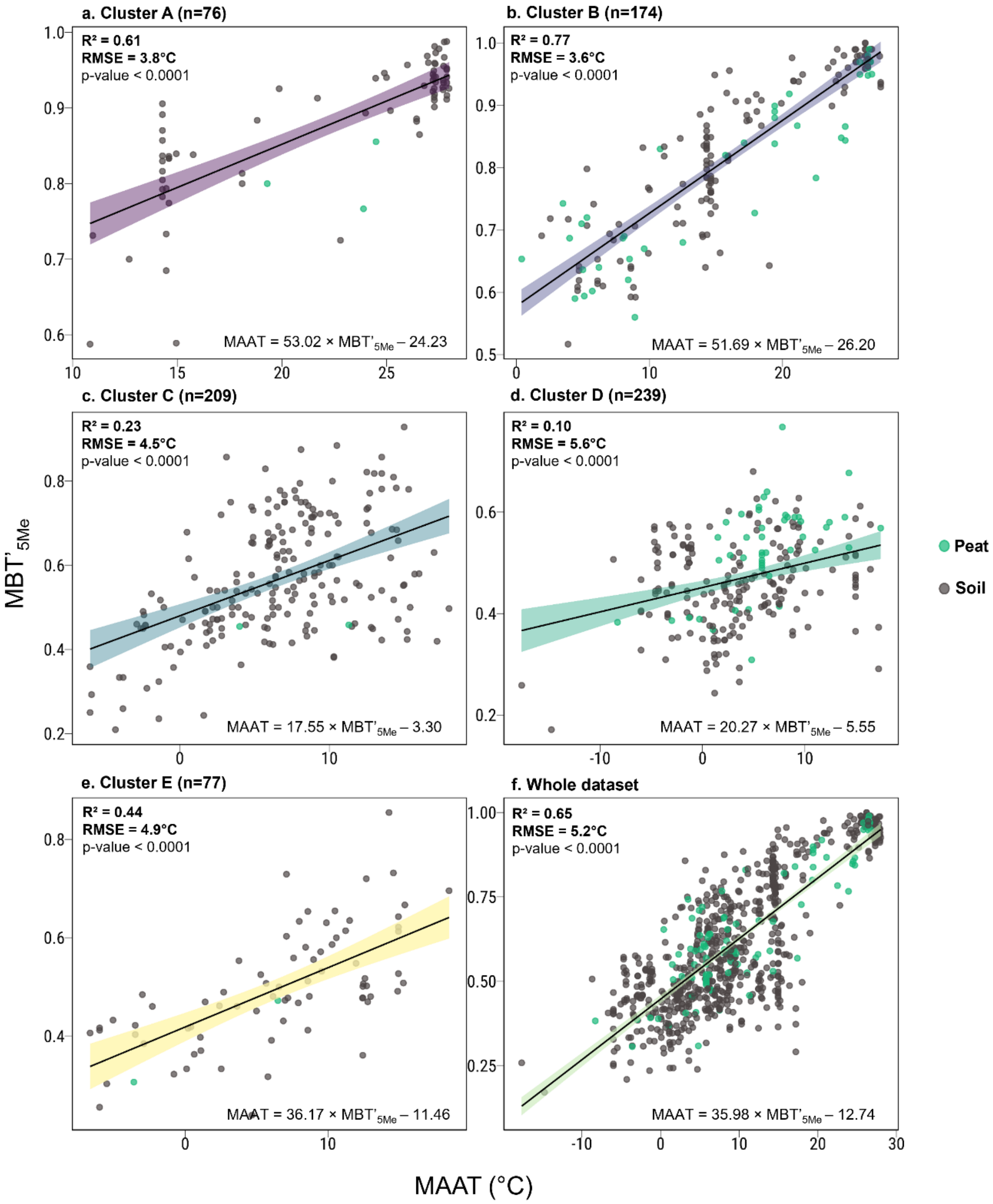

a. Cluster A (n=76)
R² = 0.61
RMSE = 3.8°C
p-value < 0.0001
MAAT = 53.02 × MBT'5Me – 24.23
b. Cluster B (n=174)
R² = 0.77
RMSE = 3.6°C
p-value < 0.0001
MAAT = 51.69 × MBT'5Me – 26.20
c. Cluster C (n=209)
R² = 0.23
RMSE = 4.5°C
p-value < 0.0001
MAAT = 17.55 × MBT'5Me – 3.30
d. Cluster D (n=239)
R² = 0.10
RMSE = 5.6°C
p-value < 0.0001
MAAT = 20.27 × MBT'5Me – 5.55
e. Cluster E (n=77)
R² = 0.44
RMSE = 4.9°C
p-value < 0.0001
MAAT = 36.17 × MBT'5Me – 11.46
f. Whole dataset
R² = 0.65
RMSE = 5.2°C
p-value < 0.0001
MAAT = 35.98 × MBT'5Me – 12.74
MBT'5Me
MAAT (°C)
Peat
Soil
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.2
1.00
0.75
0.50
0.25
10
15
20
25
0
-10
30

Figure 6

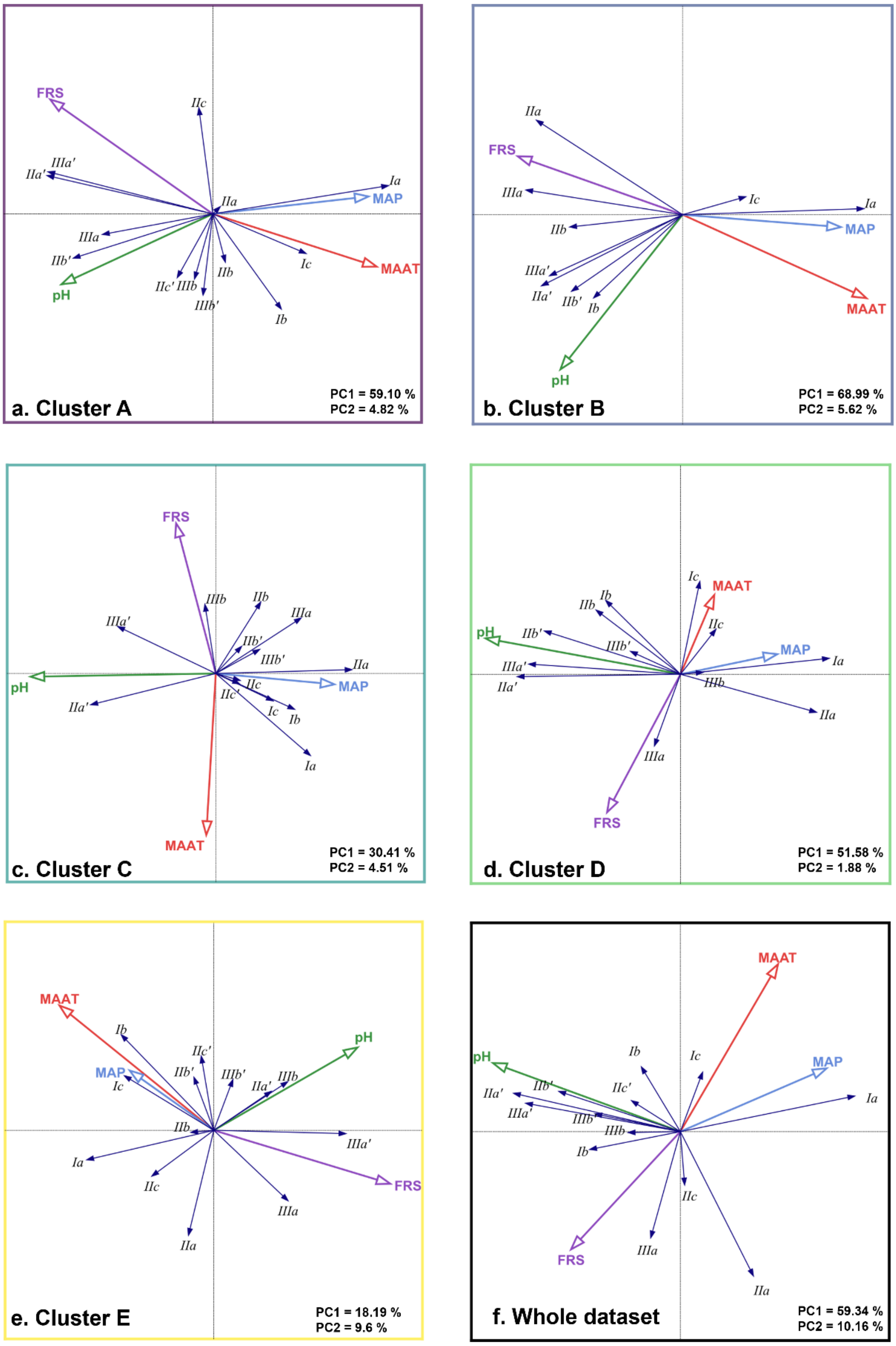

a. Cluster A
PC1 = 59.10 %
PC2 = 4.82 %
FRS
IIc
IIIa'
IIa'
IIa
Ia
MAP
IIIa
IIb'
pH
IIc'
IIIb
IIb
IIIb'
Ic
MAAT
Ib
b. Cluster B
PC1 = 68.99 %
PC2 = 5.62 %
IIa
FRS
IIIa
Ic
Ia
IIb
MAP
IIIa'
IIa'
IIb'
Ib
MAAT
pH
c. Cluster C
PC1 = 30.41 %
PC2 = 4.51 %
FRS
IIIb
IIb
IIIa
IIIa'
IIb'
IIIb'
IIa
pH
IIc
MAP
IIc'
IIa'
Ic
Ib
Ia
MAAT
d. Cluster D
PC1 = 51.58 %
PC2 = 1.88 %
Ic
MAAT
Ib
IIb
IIc
pH
IIb'
IIIb'
MAP
Ia
IIIa'
IIa'
IIIb
IIa
IIIa
FRS
e. Cluster E
PC1 = 18.19 %
PC2 = 9.6 %
MAAT
Ib
IIc'
pH
MAP
Ic
IIb'
IIIb'
IIb
IIa'
IIIb
IIIa'
Ia
IIc
FRS
IIIa
IIa
f. Whole dataset
PC1 = 59.34 %
PC2 = 10.16 %
MAAT
pH
Ib
Ic
MAP
IIa'
IIb'
IIc'
Ia
IIIa'
IIIb'
IIIb
Ib
IIc
FRS
IIIa
IIa

Figure 7

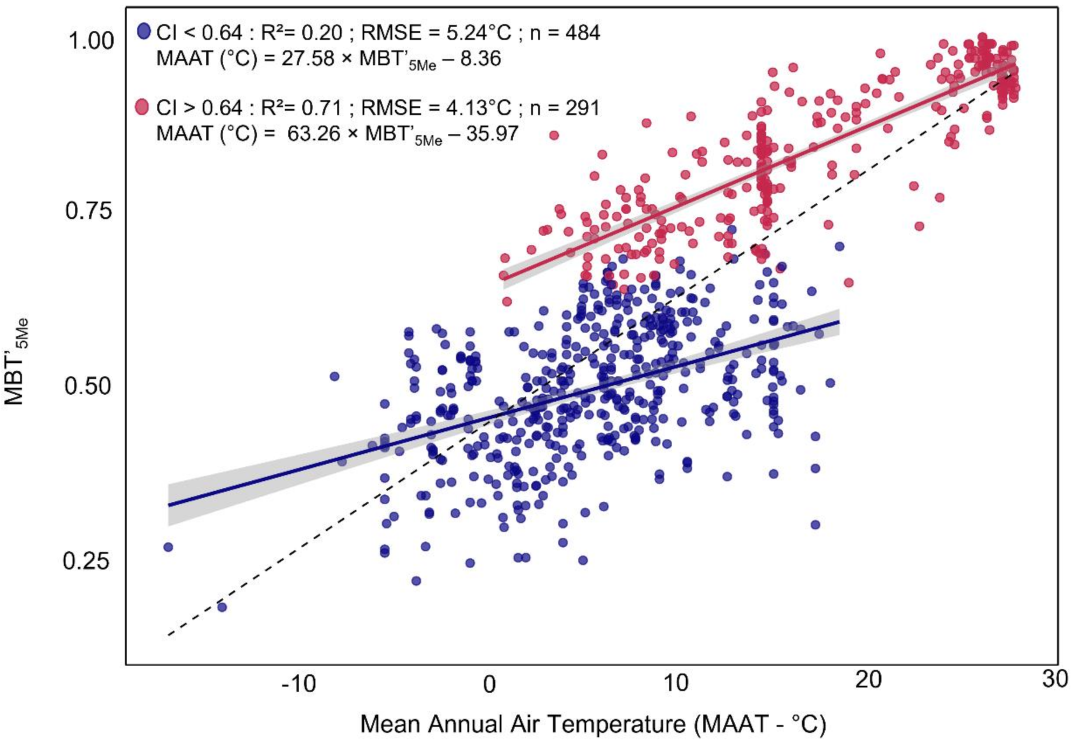

CI < 0.64 : R²= 0.20 ; RMSE = 5.24°C ; n = 484
MAAT (°C) = 27.58 × MBT'5Me – 8.36
CI > 0.64 : R²= 0.71 ; RMSE = 4.13°C ; n = 291
MAAT (°C) = 63.26 × MBT'5Me – 35.97
MBT'5Me
1.00
0.75
0.50
0.25
-10
0
10
20
30
Mean Annual Air Temperature (MAAT - °C)

Figure 8

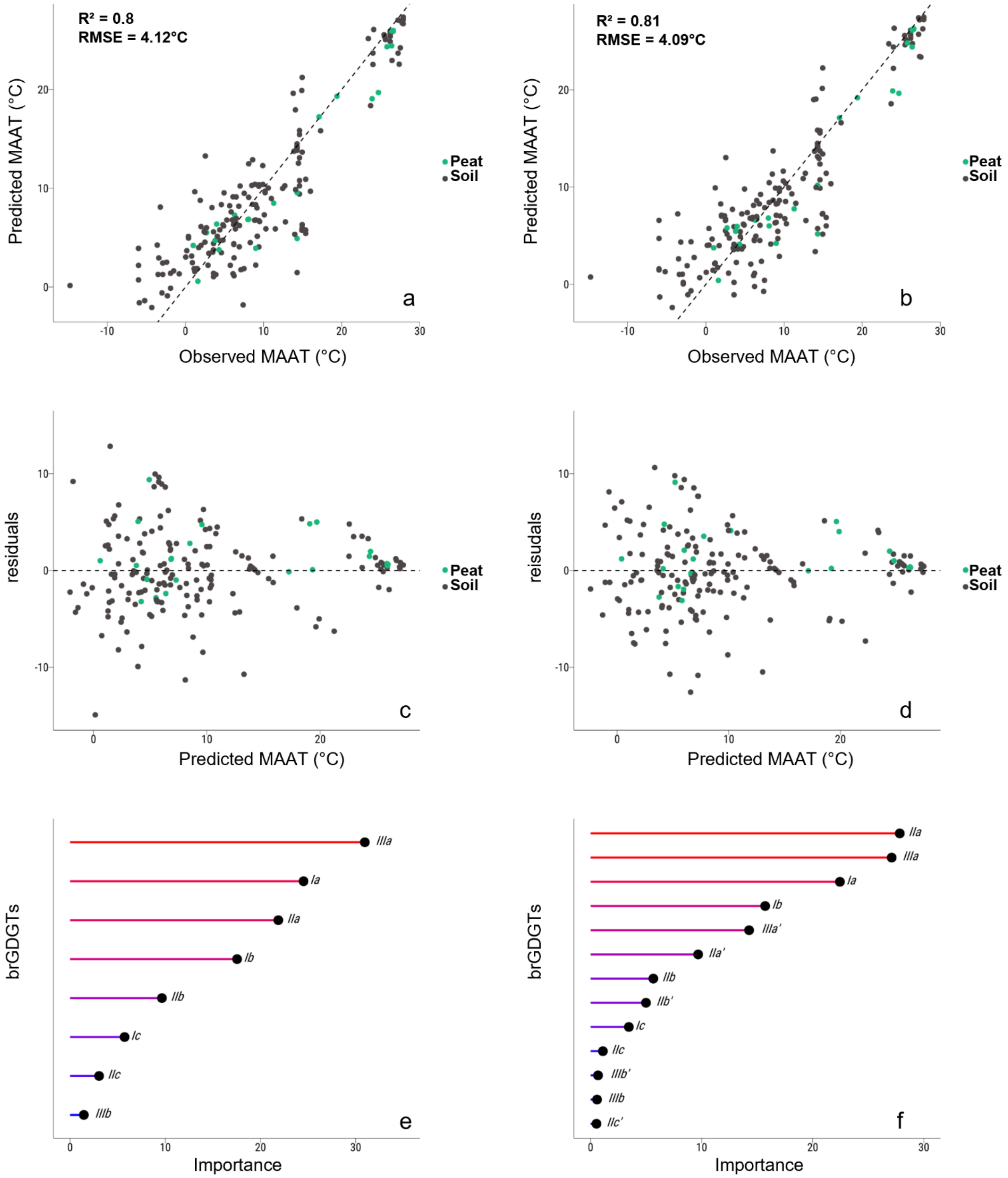

R² = 0.8
RMSE = 4.12°C
Predicted MAAT (°C)
Observed MAAT (°C)
Peat
Soil
a
R² = 0.81
RMSE = 4.09°C
b
residuals
Predicted MAAT (°C)
c
reisudals
d
brGDGTs
IIIa
Ia
IIa
Ib
IIb
Ic
IIc
IIIb
Importance
e
IIa
IIIa
Ia
Ib
IIIa'
IIa'
IIb
IIb'
Ic
IIc
IIIb'
IIIb
IIc'
f
-10
0
10
20
30

Figure 9

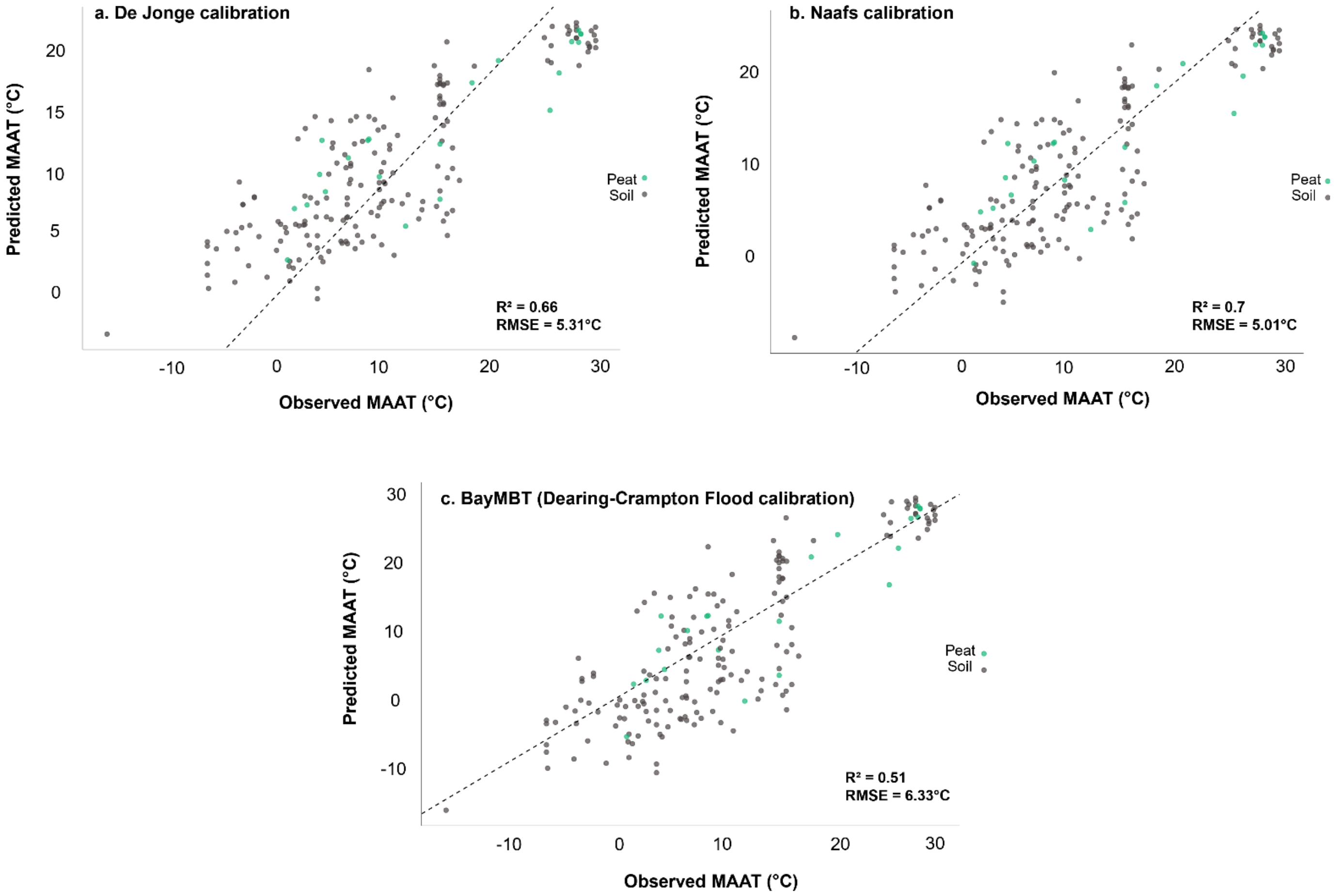

a. De Jonge calibration
Predicted MAAT (°C)
20
15
10
5
0
-10
0
10
20
30
Observed MAAT (°C)
Peat ●
Soil ●
R² = 0.66
RMSE = 5.31°C
b. Naafs calibration
Predicted MAAT (°C)
20
10
0
-10
0
10
20
30
Observed MAAT (°C)
Peat ●
Soil ●
R² = 0.7
RMSE = 5.01°C
c. BayMBT (Dearing-Crampton Flood calibration)
Predicted MAAT (°C)
30
20
10
0
-10
-10
0
10
20
30
Observed MAAT (°C)
Peat ●
Soil ●
R² = 0.51
RMSE = 6.33°C

Figure 10

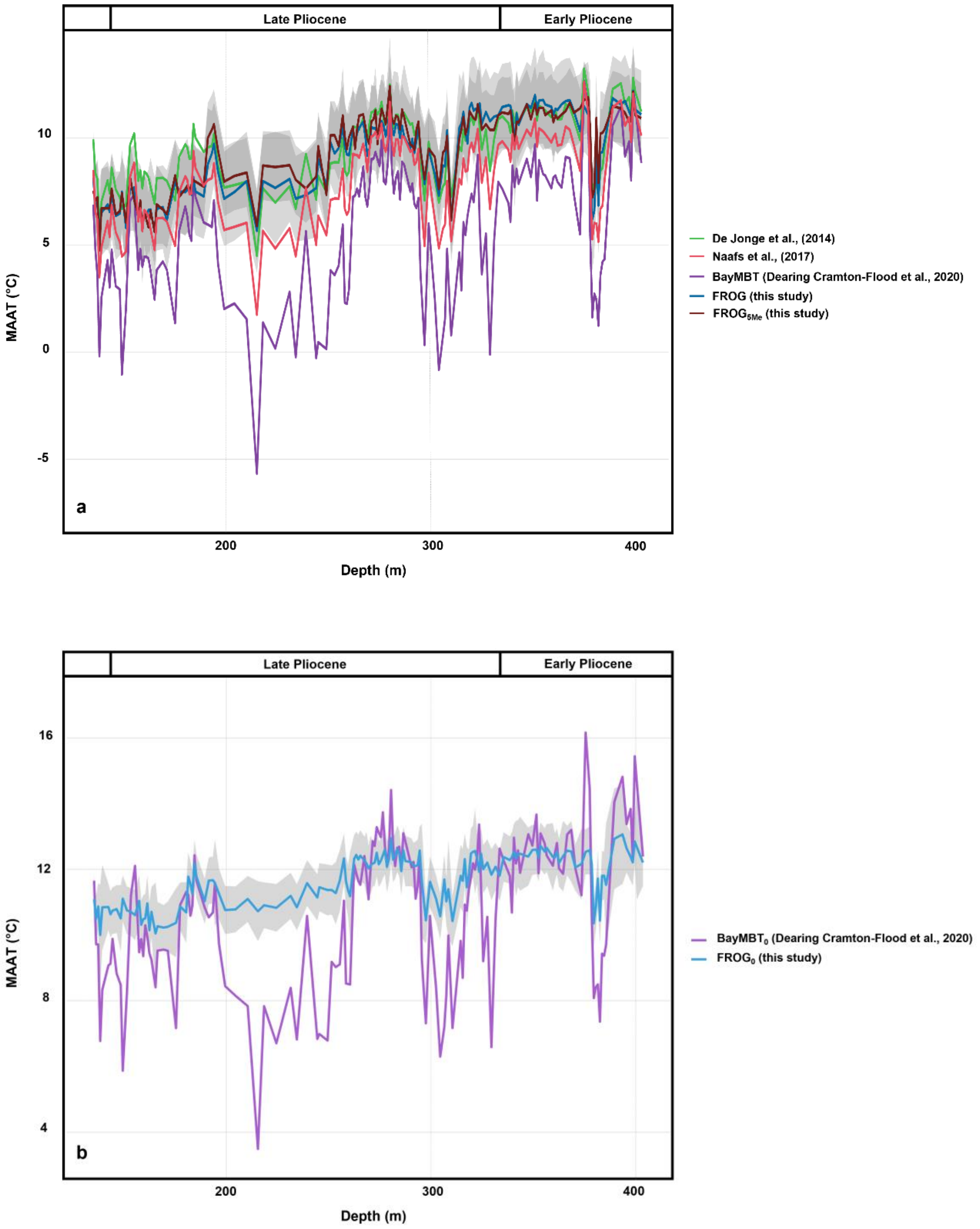

Late Pliocene
Early Pliocene
MAAT (°C)
10
5
0
-5
a
200
300
400
Depth (m)
De Jonge et al., (2014)
Naafs et al., (2017)
BayMBT (Dearing Cramton-Flood et al., 2020)
FROG (this study)
$FROG_{5Me}$ (this study)
Late Pliocene
Early Pliocene
MAAT (°C)
16
12
8
4
b
200
300
400
Depth (m)
$BayMBT_0$ (Dearing Cramton-Flood et al., 2020)
$FROG_0$ (this study)

Figure 11

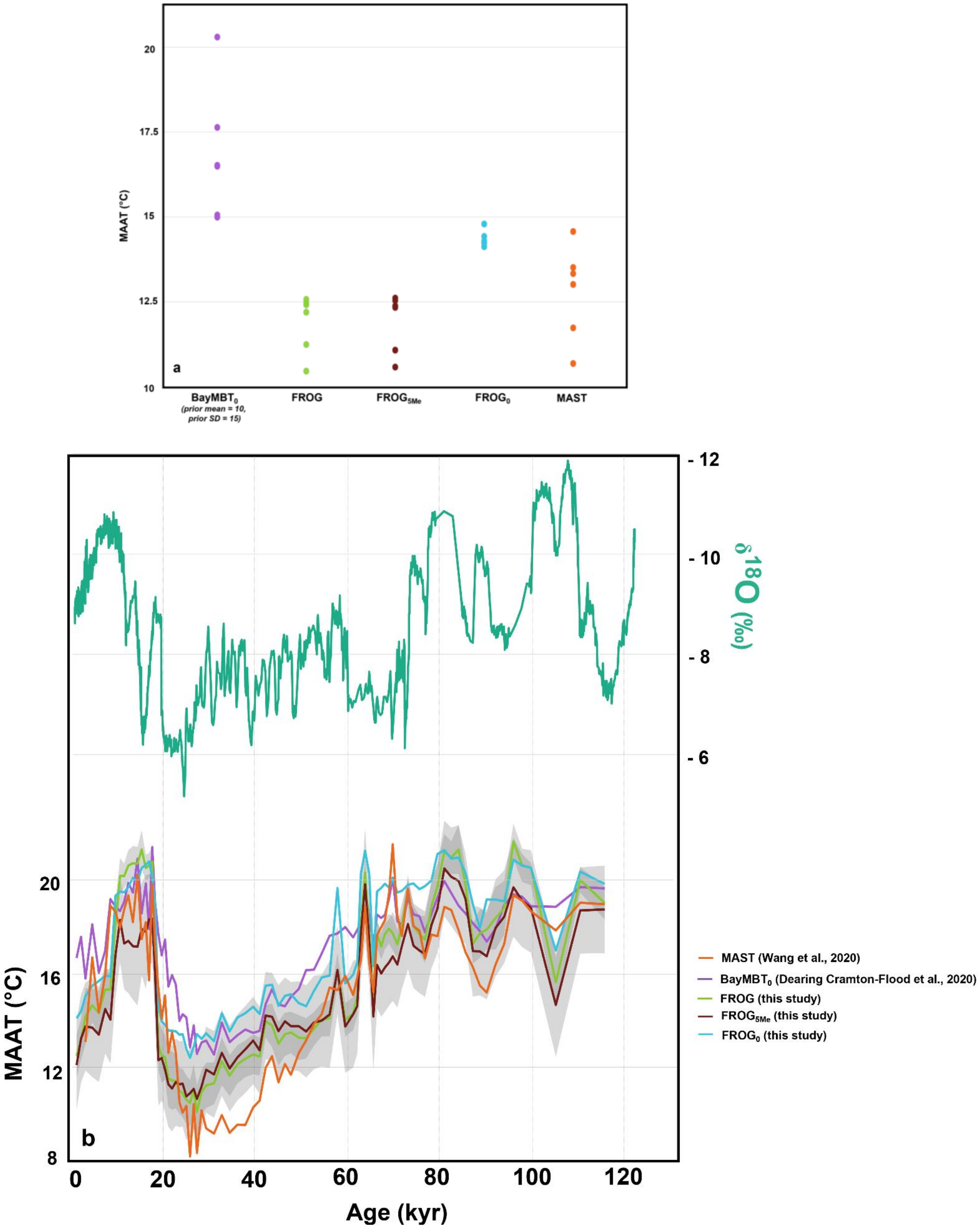
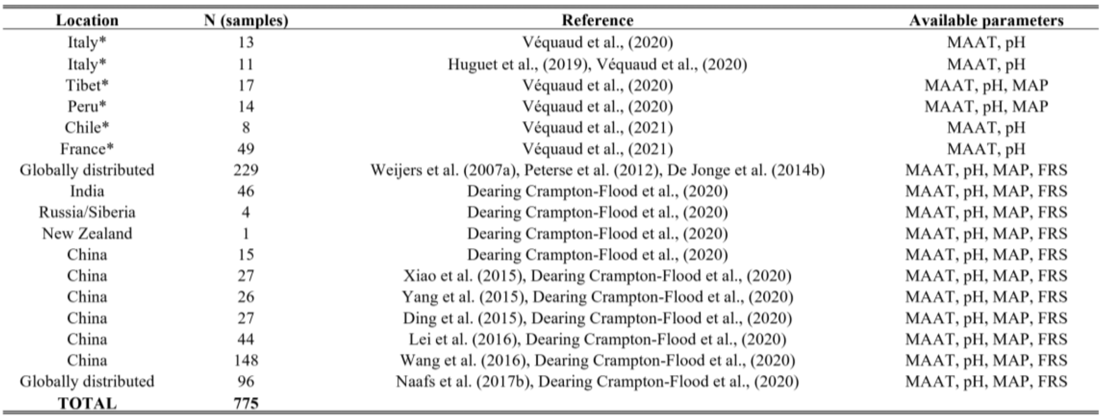

a
MAAT (°C)
20
17.5
15
12.5
10
BayMBT$_0$
(prior mean = 10, prior SD = 15)
FROG
FROG$_{5Me}$
FROG$_0$
MAST
b
MAAT (°C)
8
12
16
20
Age (kyr)
0
20
40
60
80
100
120
- 12
- 10
- 8
- 6
δ^{18}O (‰)
MAST (Wang et al., 2020)
BayMBT$_0$ (Dearing Cramton-Flood et al., 2020)
FROG (this study)
FROG$_{5Me}$ (this study)
FROG$_0$ (this study)

Figure 12



Location	N (samples)	Reference	Available parameters
Italy*	13	Véquaud et al., (2020)	MAAT, pH
Italy*	11	Huguet et al., (2019), Véquaud et al., (2020)	MAAT, pH
Tibet*	17	Véquaud et al., (2020)	MAAT, pH, MAP
Peru*	14	Véquaud et al., (2020)	MAAT, pH, MAP
Chile*	8	Véquaud et al., (2021)	MAAT, pH
France*	49	Véquaud et al., (2021)	MAAT, pH
Globally distributed	229	Weijers et al. (2007a), Peterse et al. (2012), De Jonge et al. (2014b)	MAAT, pH, MAP, FRS
India	46	Dearing Crampton-Flood et al., (2020)	MAAT, pH, MAP, FRS
Russia/Siberia	4	Dearing Crampton-Flood et al., (2020)	MAAT, pH, MAP, FRS
New Zealand	1	Dearing Crampton-Flood et al., (2020)	MAAT, pH, MAP, FRS
China	15	Dearing Crampton-Flood et al., (2020)	MAAT, pH, MAP, FRS
China	27	Xiao et al. (2015), Dearing Crampton-Flood et al., (2020)	MAAT, pH, MAP, FRS
China	26	Yang et al. (2015), Dearing Crampton-Flood et al., (2020)	MAAT, pH, MAP, FRS
China	27	Ding et al. (2015), Dearing Crampton-Flood et al., (2020)	MAAT, pH, MAP, FRS
China	44	Lei et al. (2016), Dearing Crampton-Flood et al., (2020)	MAAT, pH, MAP, FRS
China	148	Wang et al. (2016), Dearing Crampton-Flood et al., (2020)	MAAT, pH, MAP, FRS
Globally distributed	96	Naafs et al. (2017b), Dearing Crampton-Flood et al., (2020)	MAAT, pH, MAP, FRS
TOTAL	**775**		

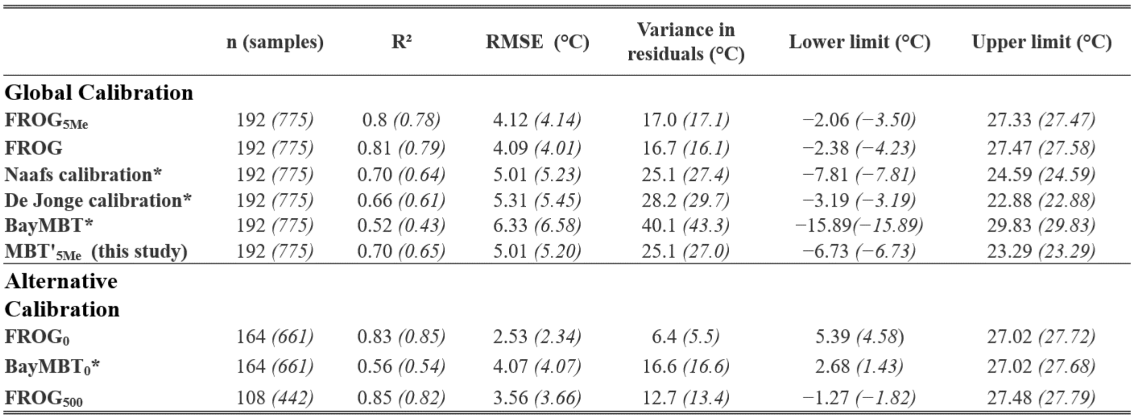
Table 1

	Cluster A	Cluster B	Cluster C	Cluster D	Cluster E	All
Soil	73	131	207	193	75	679
Peat	3	43	2	46	2	96
Samples	**76**	**174**	**209**	**239**	**77**	**775**
Min pH	5.6	3.0	4.9	3.4	4.0	3.0
Max pH	8.1	7.9	8.7	7.5	9.3	9.3
Mean pH	**7.1**	**5.2**	**7.5**	**5.4**	**7.2**	**6.3**
Min MAAT (°C)	10.8	0.4	-6.0	-17.7	-6.7	-17.7
Max MAAT (°C)	28.0	27.4	18.0	17.4	18.5	28.0
Mean MAAT (°C)	**22.4**	**16.0**	**6.7**	**3.9**	**6.5**	**9.5**
Min MAP (mm/yr)	106	383	168	128	156	106
Max MAP (mm/yr)	2770	3584	1055	2177	1191	3584
Mean MAP (mm/yr)	**1069**	**1237**	**453**	**784**	**756**	**855**
Min FRS	0	0	3	0	2	0
Max FRS	11	19	29	26	27	29
Mean FRS	**2**	**5**	**15**	**14**	**13**	**11**

Table 2

	Variables	**RDA correlation coefficients**		**Conditionnal effects (%)**
		Axis 1	**Axis 2**	
Cluster A	**pH**	-0.75	-0.34	18.6*
	MAAT (°C)	0.8	-0.26	5.3*
	MAP (mm/year)	0.77	0.08	2.2
	FRS	-0.8	0.56	39.6*
	Expl. variation (%)	59.10	4.82	
	Expl. fitted variation (%)	89.97	7.34	
Cluster B	**pH**	-0.6	-0.76	16.7*
	MAAT (°C)	0.9	-0.41	57.3*
	MAP (mm/year)	0.78	-0.06	0.2
	FRS	-0.81	0.29	0.5
	Expl. variation (%)	68.99	5.62	
	Expl. fitted variation (%)	92.42	7.52	
Cluster C	**pH**	-0.91	-0.02	25.6*
	MAAT (°C)	-0.05	-0.78	4.5*
	MAP (mm/year)	0.59	-0.05	5.7*
	FRS	-0.2	0.74	0.7
	Expl. variation (%)	30.41	4.51	
	Expl. fitted variation (%)	83.36	12.37	
Cluster D	**pH**	-0.96	0.18	47.4*
	MAAT (°C)	0.17	0.39	1.5*
	MAP (mm/year)	0.48	0.1	3.4*
	FRS	-0.36	-0.68	1.5*
	Expl. variation (%)	51.58	1.88	
	Expl. fitted variation (%)	95.9	3.5	
Cluster E	**pH**	0.71	0.41	7.7*
	MAAT (°C)	-0.76	0.62	6.6*
	MAP (mm/year)	-0.41	0.29	0.6
	FRS	0.87	-0.26	15.2*
	Expl. variation (%)	18.19	9.6	
	Expl. fitted variation (%)	62.5	31.9	
Whole dataset	**pH**	-0.92	0.34	51.8*
	MAAT (°C)	0.48	0.83	15.7*
	MAP (mm/year)	0.72	0.31	1.1*
	FRS	-0.54	-0.58	1.6*
	Expl. variation (%)	59.34	10.16	
	Expl. fitted variation (%)	84.52	14.46	

Table 3



	n (samples)	R²	RMSE (°C)	Variance in residuals (°C)	Lower limit (°C)	Upper limit (°C)
Global Calibration						
FROG$_{5Me}$	192 *(775)*	0.8 *(0.78)*	4.12 *(4.14)*	17.0 *(17.1)*	−2.06 *(−3.50)*	27.33 *(27.47)*
FROG	192 *(775)*	0.81 *(0.79)*	4.09 *(4.01)*	16.7 *(16.1)*	−2.38 *(−4.23)*	27.47 *(27.58)*
Naafs calibration*	192 *(775)*	0.70 *(0.64)*	5.01 *(5.23)*	25.1 *(27.4)*	−7.81 *(−7.81)*	24.59 *(24.59)*
De Jonge calibration*	192 *(775)*	0.66 *(0.61)*	5.31 *(5.45)*	28.2 *(29.7)*	−3.19 *(−3.19)*	22.88 *(22.88)*
BayMBT*	192 *(775)*	0.52 *(0.43)*	6.33 *(6.58)*	40.1 *(43.3)*	−15.89*(−15.89)*	29.83 *(29.83)*
MBT'$_{5Me}$ (this study)	192 *(775)*	0.70 *(0.65)*	5.01 *(5.20)*	25.1 *(27.0)*	−6.73 *(−6.73)*	23.29 *(23.29)*
Alternative Calibration						
FROG$_0$	164 *(661)*	0.83 *(0.85)*	2.53 *(2.34)*	6.4 *(5.5)*	5.39 *(4.58*)	27.02 *(27.72)*
BayMBT$_0$*	164 *(661)*	0.56 *(0.54)*	4.07 *(4.07)*	16.6 *(16.6)*	2.68 *(1.43)*	27.02 *(27.68)*
FROG$_{500}$	108 *(442)*	0.85 *(0.82)*	3.56 *(3.66)*	12.7 *(13.4)*	−1.27 *(−1.82)*	27.48 *(27.79)*

Table 4