Development of global temperature and pH calibrations based on bacterial 3-hydroxy fatty acids in soils

Pierre Véquaud[1], Sylvie Derenne[1], Alexandre Thibault[2], Christelle Anquetil[1], Giuliano Bonanomi[3], Sylvie Collin[1], Sergio Contreras[4], Andrew Nottingham[5], Pierre Sabatier[6], Norma Salinas[7], Wesley Phillip Scott[8], Josef P. Werne[8], Arnaud Huguet[1]

[1]Sorbonne Université, CNRS, EPHE, PSL, UMR METIS, Paris, 75005, France
[2]Antea Group, Innovation Hub, 803 boulevard Duhamel du Monceau, Olivet, 45160, France
[3]Dipartimento di Agraria, Università di Napoli Federico II, via Università 100, Portici, NA, 80055, Italy
[4]Laboratorio de Ciencias Ambientales (LACA), 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
[5]School of Geosciences, University of Edinburgh, Crew Building, Kings Buildings, Edinburgh EH9 3FF United Kingdom
[6]Univ. Savoie Mont Blanc, CNRS, EDYTEM, Le Bourget du Lac, 73776, France
[7]Instituto de Ciencias de la Naturaleza, Territorio y Energías Renovables, Pontificia Universidad Catolica del Peru, Av. Universitaria 1801, San Miguel, Lima 32, Peru
[8]Department of Geology and Environmental Science, University of Pittsburgh, Pittsburgh, PA 15260, USA

Correspondence to: Arnaud Huguet (arnaud.huguet@sorbonne-universite.fr)

Abstract. 3-hydroxy fatty acids (3-OH FAs) with 10 to 18 C atoms are membrane lipids mainly produced by Gram-negative bacteria. They have been recently proposed as temperature and pH proxies in terrestrial settings. Nevertheless, the existing correlations between pH/temperature and indices derived from 3-OH FA distribution (RIAN, RAN_{15} and RAN_{17}) are based on a small soil dataset (ca. 70 samples) and only applicable regionally. The aim of this study was to investigate the applicability of 3-OH FAs as mean annual air temperature (MAAT) and pH proxies at the global level. This was achieved using an extended soil dataset of 168 topsoils distributed worldwide, covering a wide range of temperatures (5°C to 30°C) and pH (3 to 8). The response of 3-OH FAs to temperature and pH was compared to that of established branched GDGT-based proxies (MBT'_{5Me}/CBT). Strong linear relationships between 3-OH FA-derived indices (RAN_{15}, RAN_{17} and RIAN) and MAAT/pH could only be obtained locally, for some of the individual transects. This suggests that these indices cannot be used as paleoproxies at the global scale using simple linear regression models, in contrast with the MBT'_{5Me} and CBT. However, strong global correlations between 3-OH FA relative abundances and MAAT/pH were shown by using other algorithms (multiple linear regression, k-NN and random forest models). The applicability of the three aforementioned models for paleotemperature reconstruction was tested and compared with the MAAT record from a Chinese speleothem. The calibration based on the random forest model appeared to be the most robust. It generally

48 showed similar trends with previously available records and highlighted known climatic events

49 poorly visible when using local 3-OH FA calibrations. Altogether, these results demonstrate

50 the potential of 3-OH FAs as paleoproxies in terrestrial settings.

51

52 **Keywords:** 3-hydroxy fatty acids; branched GDGTs; soils; global calibration; temperature and

53 pH proxy

54

55

1. Introduction

Investigating past climate variations is essential to understand and predict future environmental changes, especially in the context of global anthropogenic change. Direct records of environmental parameters are available for the last decades, the so-called "instrumental" period. Beyond this period, proxies can be used to obtain indirect information on environmental parameters. A major challenge is to develop reliable proxies which can be applied to continental environments in addition to marine ones. Indeed, available proxies have been mainly developed and used in marine settings, as the composition and mechanism of formation of marine sedimentary cores is less complex than in continental settings, which are highly heterogeneous. Several environmental proxies based on organic (e.g. the alkenone unsaturation index ($U^{k'}_{37}$; Brassell et al., 1986) and inorganic (Mg/Ca ratio and $^{18}O/^{16}O$ ratio of foraminifera; Emiliani, 1955; Erez and Luz, 1983) fossil remains were notably developed for the reconstruction of sea surface temperatures.

Some of the existing proxies are based on membrane lipids synthesized by certain microorganisms (Eglinton and Eglinton, 2008; Schouten et al., 2013). These microorganisms are able to adjust the composition of their membrane lipids in response to the prevailing environmental conditions in order to maintain an appropriate fluidity and to ensure the optimal state of the cellular membrane (Singer and Nicolson, 1972; Sinensky, 1974; Hazel and Williams, 1990; Denich et al., 2003). The structure of glycerol dialkyl glycerol tetraethers (GDGTs), which are membrane lipids biosynthesized by archaea and some bacteria, is especially known to be related to environmental conditions. Archaeal GDGTs are constituted of isoprenoid alkyl chains ether-linked to glycerol, whereas bacterial GDGTs are characterized by branched alkyl chains instead of isoprenoid ones. The latter compounds are ubiquitous in terrestrial (Weijers et al., 2007; Peterse et al., 2012; De Jonge et al., 2014; Naafs et al., 2017) and aquatic environments (Peterse et al., 2009; Tierney and Russell, 2009; Sinninghe Damsté et al., 2009; Loomis et al., 2012; Peterse et al., 2015; Weber et al., 2015). These branched GDGTs (brGDGTs) are produced by still unidentified bacteria, although some of them may belong to the phylum *Acidobacteria* (Sinninghe Damsté et al., 2011, 2014, 2018). The analysis of brGDGTs in 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). Despite improvements in brGDGT analytical methods and development of refined calibration models (De Jonge et al., 2014; Dearing Crampton-Flood et al., 2020), the Root Mean Square Error (RMSE) associated

89   with MAAT reconstruction using the global brGDGT calibrations in soils remains high (>4°C).

90   Thus, development of new molecular proxies, independent of and complementary to brGDGTs,

91   is essential to improve the reliability of temperature reconstructions in such settings.

92   Recent studies have unveiled the potential of another family of bacterial lipids − 3-

93   hydroxy fatty acids (3-OH FAs) − for temperature and pH reconstructions in terrestrial (Wang

94   et al., 2016, 2018; Huguet et al., 2019) and marine (Yang et al., 2020) settings. 3-OH FAs with

95   10 to 18 carbon atoms are specifically produced by Gram-negative bacteria and are bound to

96   the lipopolysaccharide (LPS) by ester or amide bonds (Wollenweber et al., 1982; Wollenweber

97   and Rietschel, 1990). Three types of 3-OH FAs can be distinguished, with either *normal* chains

98   or branched chains, *iso* or *anteiso*.

99   The analysis of 3-OH FAs in soils showed that the ratio of C_{15} or C_{17} *anteiso* 3-OH

100   FA to *normal* C_{15} or C_{17} 3-OH FA (RAN_{15} and RAN_{17} indices, respectively) were negatively

101   correlated with MAAT along the three mountains investigated so far: Mts. Shennongjia (China;

102   Wang et al., 2016), Rungwe and Majella (Tanzania and Italy, respectively; Huguet et al., 2019).

103   This suggests that Gram-negative bacteria producing these fatty acids respond to colder

104   temperatures with an increase in *anteiso*-C_{15}/C_{17} vs. *n*-C_{15}/C_{17} 3-OH FAs, in order to maintain

105   a proper fluidity and optimal state of the bacterial membrane, the so-called homeoviscous

106   adaptation mechanism (Sinensky, 1974; Hazel and Eugene Williams, 1990). Nevertheless, the

107   relationships between RAN_{15} and MAAT along the three mountain transects showed the same

108   slopes but different intercepts (Wang et al., 2016; Huguet et al., 2019), suggesting that regional

109   relations may be more adapted to apply RAN_{15} as a temperature proxy in soils. In contrast, a

110   significant calibration between RAN_{17} and MAAT could be established using combined data

111   from the three mountain regions (Wang et al., 2016; Huguet et al., 2019).

112   Another index, defined as the cologarithm of the sum of *anteiso* and *iso* 3-OH FAs

113   divided by the sum of *normal* homologues (RIAN index), was shown to be strongly negatively

114   correlated with soil pH along the three aforementioned mountains (Wang et al., 2016; Huguet

115   et al., 2020), reflecting a general relative increase in normal homologues compared to branched

116   (*iso* and *anteiso*) ones with increasing pH. This mechanism was suggested to reduce the

117   permeability and fluidity of the membrane for the cell to cope with lower pH (Russell et al.,

118   1995; Denich et al., 2003; Beales, 2004).

119   3-OH FA indices were recently applied for the first time to the reconstruction of the

120   temperature and hydrological changes over the last 10,000 years in a speleothem from China

121   (Wang et al., 2018), showing the potential of 3-OH FAs as independent tools for environmental

122   reconstruction in terrestrial settings. A very recent study based on marine sediments from the

North Pacific Ocean suggested that the distribution of 3-OH FAs could also be used to reconstruct sea surface temperature (Yang et al., 2020).

Even though these results are promising, the linear regressions between pH/MAAT and 3-OH FA indices in terrestrial environments are still based on a rather small dataset (ca. 70 soil samples; Wang et al., 2016; Huguet et al., 2019). The aim of this study was to investigate the applicability of 3-OH FAs as MAAT and pH proxies at the global level using an extended soil dataset and refined statistical tools. 3-OH FA distribution from 54 soils was determined in four globally distributed altitudinal transects (Tibet, Italy, Peruvian Andes and Chile) and was combined with data previously published by Wang et al. (2016; Mt Shennongjia, China), Huguet et al. (2019; Mt. Rungwe, Tanzania and Mt. Majella, Italy) and Véquaud et al. (2021; Mts. Lautaret-Bauges, France), leading to a total of 168 samples. In addition to linear regressions, non-parametric, machine learning models were used to improve the global relationships between 3-OH FA distribution and MAAT/pH and present the advantage of taking into account non-linear environmental influences, in line with the intrinsic complexity of the environmental settings. Finally, these new models were tested and compared by applying them to a speleothem archive (Wang et al., 2018) representing to date the only available MAAT record derived from 3-OH FA proxies in continental setting. As brGDGTs are the only microbial organic proxies which can be used for temperature and pH reconstructions in terrestrial settings so far, they can serve as a reference proxy to understand the temperature and pH dependency of 3-OH FAs analyzed in the same dataset, taking into account the large uncertainties persisting in the global temperature/pH brGDGT calibrations (De Jonge et al., 2014; Dearing Crampton-Flood et al., 2020). 3-OH FAs and brGDGTs have thus been concomitantly analyzed to assess their reliability and complementarity as independent temperature and pH proxies.

2. Material and methods

2.1. Soil dataset

2.1.1. Study sites

The dataset of the present study is comprised of the globally distributed surface soils previously analyzed for brGDGTs and 3-OH FAs and collected along 4 altitudinal transects: Mts. Shennongjia (China; Yang et al., 2015; Wang et al., 2016), Rungwe (Tanzania ; Coffinet

et al., 2017; Huguet et al., 2019), Majella (Italy; Huguet et al., 2019) and Lautaret-Bauges (France; Véquaud et al., 2021). This set was extended with surficial soils (0-10 cm) from 4 additional altitudinal transects described below, located in Italy, Tibet, Peru and Chile (Table 1).

Soil samples were collected from 13 sites along Mount Pollino in the Calabria region (Italy) between 0 and 2,200 m above sea level (a.s.l.) (Table 1). Mt. Pollino is located in the calcareous Apennine range and is 2,248 m a.s.l. It is framed to the northwest by the Sierra de Prete (2,181 m high) and to the south by the Pollino Abyss. The alpine to subalpine area (above 2,100 m a.s.l.) is characterized by the presence of Mediterranean grasslands (*Festuca bosniaca*, *Carex kitaibeliana*) and the presence of sinkholes (Todaro et al., 2007; Scalercio et al., 2014). The mountainous vegetation (over 1,200 m a.s.l.) is dominated by *Fagus sylvatica* forests and, at the treeline, by scattered *Pinus leucodermis* (Bonanomi et al., 2020). The soil is poorly developed and dominated by calcareous soils. Between 0 to 1,200 m a.s.l (Scalercio et al., 2014 and reference therein), Mt. Pollino is characterized by the presence of *Q. ilex* forests or shrubs. Climate along this mountain is humid Mediterranean, with high summer temperatures and an irregular distribution of rainfall throughout the year with pronounced summer drought (39.5% in winter, 23.7% in spring, 29.2% in autumn, 7.6% in summer; average annual precipitation: 1,570 mm; see Todaro et al., 2007). MAAT is comprised between 7 °C (2,200 m a.s.l) and 18 °C (0 m a.s.l; Scalercio et al., 2014). MAAT along Mt. Pollino was estimated using a linear regression between two MAAT (16°C at 400 m a.s.l and 10°C at 1,600 m a.s.l.) from the meteorological data (Castrovillari station) recorded by Scalercio et al. (2014). The pH of the soils analyzed in the present study ranges between 4.5 and 6.8 (Table 1).

Soil samples were collected from 17 sites along along Mount Shegyla between 3,106 and 4,474 m a.s.l. (southeastern Tibet, China), as previously described by Wang et al. (2015). Different climatic zonations are observed along this high-altitude site (2,700 to 4,500 m a.s.l): (i) a mountainous temperate zone between 2,700 and 3,400 m, (ii) a subalpine cold temperate zone between 3,400 and 4,300 m and (iii) a cold alpine zone above 4,300 m. Plant species, such as brown oak (*Q. semecarpifolia*) or common fir (*Abies alba*) are abundant within the mountainous and subalpine levels. In the cold subalpine zone, the Forrest's fir (*Abies georgei var. smithii)* is endemic to western China. In the cold alpine zone, coniferous species (*Sabina saltuaria*) as well as species typical of mountainous regions such as *Rhododendron* are observed. MAAT was estimated using a linear regression between 7 measured MAAT from the data recorded by Wang et al. (2015). The average MAAT along the transect is 4.6°C, with a

188   minimum of 1.1 °C at ca. 4,500 m a.s.l. and a maximum of 8.9 °C at ca. 3,100 m a.s.l. (Table

189   1). Soil pH ranges between 4.6 and 6.4 (Table 1).

190   Soils were sampled from 14 sites in the Peruvian Andes along the Kosñipata transect,

191   located in south-eastern Peru, in the upper part of the Madre de Dios/Madeira watershed, east

192   of the Andes Cordillera (Nottingham et al., 2015). This transect (190 m to 3,700 m a.s.l) is well-

193   documented and is the object of numerous ecological studies (Malhi et al., 2010; Nottingham

194   et al., 2015). There is a shift in vegetation zonation with increasing elevation, from tropical

195   lowland forest to montane cloud forest and high-elevation 'Puna' grassland. The tree line lies

196   between 3,200 and 3,600 m a.s.l. For the 14 sites sampled in this study, the lower 13 sites are

197   forest and the highest site is grassland. The 14 sites are part of a network of 1 ha forest plots

198   (Nottingham et al., 2015); for each 1 ha plot, 0-10 cm surface soil was sampled from 5

199   systematically distributed locations within each 1 ha plot. Mean annual precipitation does not

200   vary significantly with altitude (mean = 2448 mm.y^{-1}, SD = 503 mm.y^{-1}; Rapp and Silman, 2012;

201   Nottingham et al., 2015). MAAT is comprised between 26.4 °C at 194 m altitude and 6.5°C at

202   3644 m altitude (Table 1). The pH is characteristic of acidic soils (3.4 - 4.7; Table 1). Further

203   information on these sites and soils is available in Nottingham et al. (2015).

204   Soil samples were collected from 10 sites between 690 m and 1,385 m a.s.l. from the

205   lake shore (20 to 50 m offshore) of 10 Andean lakes located in Chile (38–39°S) within the

206   temperate forest (Table 1). High-frequency measurements of MAAT over a period of one year

207   are available for the different sampling sites. MAAT is comprised between 5.75°C and 9.2°C.

208   Soil pH ranges between 4.4 and 6.8 (Table 1).

209

210   *2.1.2. pH measurement*

211   Following sampling, soils were immediately transported to the laboratory and stored at

212   -20 °C. Soil samples from the Peruvian Andes, Mt. Pollino and Mt. Shegyla were then freeze-

213   dried, ground and sieved at 2 mm. The pH of the freeze-dried samples was measured in

214   ultrapure water with a 1:2.5 soil water ratio. Typically, 10 ml of ultrapure water were added to

215   4 g of dry soil. The soil solution was stirred for 30 min, before decantation for 1 hand pH

216   measurement (Carter et al., 2007).

217

218   **2.2. Lipid analyses**

219   BrGDGTs and 3-OH FAs were analyzed in all samples from the Peruvian Andes,

220   Chilean Andes, Mt. Pollino and Mt. Shegyla.

2.2.1. 3-OH FA analysis

Sample preparation for 3-OH FA analysis was identical to that reported by Huguet et al. (2019) and Véquaud et al. (2021). Soil samples were subjected to acid hydrolysis (3 M HCl) and extracted with organic solvents. This organic fraction was then rotary-evaporated, methylated in a 1M HCl-MeOH solution at 80 °C for 1 h and separated into three fractions over an activated silica column: (i) 30 ml of heptane/EtOAc (98: 2), (ii) 30 ml of EtOAc and (iii) 30 ml of MeOH. 3-OH FAs contained in the second fraction were derivatized at 70°C for 30 min with a solution of *N,O*- bis(trimethylsilyl)trifluoroacetamide (BSTFA) – Trimethylchlorosilane (TMCS) 99:1 (Grace Davison Discovery Science, USA) before gas chromatography-mass spectrometry (GC-MS) analysis.

3-OH FAs were analyzed with an Agilent 6890N GC-5973N using a Restek RXI-5 Sil MS silica column (60 m × 0.25 mm, i.d. 0.25 µm film thickness), as previously described (Huguet et al., 2019). 3-OH FAs were quantified by integrating the appropriate peak on the ion chromatogram and comparing the area with an internal standard (3-hydroxytetradecanoic acid, 2,2,3,4,4-d5; Sigma-Aldrich, France). The internal standard (0.5 mg/ml) was added just before injection as a proportion of 3 µl of standard to 100 µl of sample, as detailed by Huguet et al. (2019). The different 3-OH FAs were identified based on their retention time, after extraction of the characteristic *m/z* 175 fragment (*m/z* 178 for the deuterated internal standard; cf. Huguet et al., 2019).

The RIAN index was calculated as follows (Wang et al., 2016 ; Eq. 1) in the range C_{10}-C_{18} :

$$RIAN = -\log[(I + A)/ N] \tag{1}$$

where I, A, N represent the sum of all *iso*, *anteiso* and *normal* 3-OH FAs, respectively.

RAN_{15} and RAN_{17} indices are defined as follows (Wang et al., 2016; Eq. 2 and 3):

$$RAN_{15} = [\text{ } anteiso \text{ } C_{15}] / [normal \text{ } C_{15}] \tag{2}$$

$$RAN_{17} = [\text{ } anteiso \text{ } C_{17}] / [normal \text{ } C_{17}] \tag{3}$$

Analytical errors associated with the calculation of RIAN, RAN_{15} and RAN_{17} indices are respectively 0.006, 0.3 and 0.2 based on the analysis of one sample injected nine times during the analysis and five samples injected in triplicates.

 2.2.2. brGDGT analysis

256   Sample preparation for brGDGT analysis was similar to that reported by Coffinet et

257   al. (2014). Briefly, ca. 5-10 g of soil was extracted using an accelerated solvent extractor (ASE

258   100, Dionex-ThermoScientific, USA) with a dichloromethane (DCM) / methanol (MeOH)

259   mixture (9: 1) for 3×5 min at 100 °C and a pressure of 100 bars in 34 ml cells. The total lipid

260   extract was rotary evaporated and separated into two fractions of increasing polarity on a

261   column of activated alumina: (i) 30 ml of heptane: DCM (9: 1, v:v) ; (ii) 30 ml of DCM: MeOH

262   (1: 1, v:v). GDGTs are contained in the second fraction, which was rotary evaporated. An

263   aliquot (300 μL) was re-dissolved in heptane and centrifuged using an Eppendorf MiniSpin

264   centrifuge (Eppendorf AG, Hamberg, Germany) at 7000 rpm for 1 min.

265   GDGTs were then analyzed by high pressure liquid chromatography coupled with

266   mass spectrometry with an atmospheric pressure chemical ionisation source (HPLC-APCI-MS)

267   using a Shimadzu LCMS 2020. GDGT analysis was performed using two Hypersil Gold silica

268   columns in tandem (150 mm × 2.1 mm, 1.9 μm; Thermo Finnigan, USA) thermally controlled

269   at 40 °C, as described by Huguet et al. (2019). This methodology enables the separation of 5-

270   and 6-methyl brGDGTs. Semi-quantification of brGDGTs was performed by comparing the

271   integrated signal of the respective compound with the signal of a C_{46} synthesized internal

272   standard (Huguet et al., 2006) assuming their response factors to be identical.

273   The MBT'$_{5Me}$ index, reflecting the average number of methyl groups in 5-methyl

274   isomers of GDGTs and considered as related to MAAT, was calculated according to De Jonge

275   et al. (2014; Eq. 4):

276

277    $$\text{MBT'}_{5Me} = \frac{[Ia+Ib+Ic]}{[Ia+Ib+Ic]+[IIa+IIb+II]+[IIIa]} \tag{4}$$

278

279   The CBT' index, reflecting the average number of cyclopentyl rings in GDGTs and

280   considered as related to pH, was calculated as follows (De Jonge et al., 2014; Eq. 5):

281

282    $$CBT' = \log \left(\frac{[Ic]+[IIa\prime]+[IIb\prime]+[IIc\prime]+[IIIa\prime]+[IIIb\prime]+[IIIc\prime]}{[Ia]+[IIa+IIIa]} \right) \tag{5}$$

283

284   The Roman numerals correspond to the different GDGT structures presented in De

285   Jonge et al. (2014). The 6-methyl brGDGTs are denoted by an apostrophe after the Roman

286   numerals for their corresponding 5-methyl isomers. Analytical errors associated with the

287 calculation of MBT'$_{5Me}$ and CBT' indices are 0.015 and 0.02 respectively, based on the analysis

288 of three samples in triplicate among the 44 soil samples.

289 **2.3. Statistical analysis**

290 In order to investigate the correlations between environmental variables (pH, MAAT)

291 and the relative abundances of bacterial lipids (brGDGTs and 3-OH FAs) or the indices based

292 on these compounds, pairwise correlation matrices were performed in addition to single or

293 multiple linear regressions. As the dataset is not normally distributed, Spearman correlation was

294 used with a confidence level of 5%.

295 Principal component analyses (PCA) were performed on the different soil samples to

296 statistically compare the 3-OH FA/brGDGT distributions along the different altitudinal

297 transects. The fractional abundances of the bacterial lipids (3-OH FAs and brGDGTs) were

298 used for these PCAs, with MAAT, pH and location of the sampling site representing

299 supplementary variables (i.e. not influencing the principal components of the analysis).

300 Independent models should be used for the development of environmental calibrations,

301 as each of them has its own advantages and limits. Linear regression methods are simple to use

302 but many of them suffer from the phenomenon of regression dilution, as previously noted

303 (Naafs et al., 2017; Dearing Crampton-Flood et al., 2020). That is why other models than

304 ordinary least squares or single/multiple regression were also proposed in this study (cf. section

305 4.2. for discussion of the models) : the the k-nearest neighbor (k-NN) and random forest models.

306 These models are based on machine-learning algorithms, which are built on a proportion of the

307 total dataset (randomly defined, i.e. training dataset) and then tested on the rest of the dataset,

308 considered as independent (test dataset).

309 The k-NN model is based on the estimation of the mean distances between the different

310 samples. This is a supervised learning method (e.g. Gangopadhyay et al., 2009). A training

311 database composed of N "input-output" pairs is initially constituted to estimate the output

312 associated with a new input x. The method of the k-neighbors takes into account the k training

313 samples whose input is the closest to the new input x, according to a distance to be defined.

314 This method is non-parametric and is used for classification and regression. In k-NN regression,

315 the result is the value for this object, which is the average of the values of the k nearest

316 neighbors. Its constraints lie in the fact that, by definition, if a range of values is more frequent

317 than the others, then it will be statistically predominant among the k closest neighbors. But, in

318 the present case, data selection was performed randomly on the dataset with a stratification

modality according to the MAAT or the pH to limit the impact of extreme values as detailed below. This allows to overcome this limitation of the k-NN method.

The random forest algorithm is also a supervised learning method used, among other things, for regressions (e.g. Ho, 1995; Breiman, 2001; Denisko and Hoffman, 2018;). This model works by constructing a multitude of decision trees at training time 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. Decision trees where variables are continuous values are called regression trees.

The training phase required for the random forests, k-NN and multiple linear regression was performed on 75% of the sample set with an iteration of ten cross-validations per model. Data selection was performed randomly on the dataset (with no pre-processing of the individual 3-OH FAs) but with a stratification modality according to the MAAT or the pH to limit the impact of extreme values. Then, the robustness and precision of the different models were tested on the remaining 25 % of samples, considered as an independent dataset. Simple and Multiple linear regressions, PCA, k-NN and random forest models were performed with R software, version 3.6.1 (R Core Team, 2014) using the packages - tidymodels (version 0.1.0)- kknn (version 1.3.1), ranger (version 0.11.2). A web application is available online (https://athibault.shinyapps.io/paleotools) for the reconstruction of 3-OH FA-derived MAAT using the machine learning models proposed in the present study.

3. Results

3.1. Distribution of bacterial lipids

3.1.1. 3-OH FAs

3-OH FAs were identified in the whole dataset, representing eight elevation transects and 168 samples (Supplementary table 1; Yang et al., 2015; Wang et al., 2016; Coffinet et al., 2017; Huguet et al., 2019; Véquaud et al., 2021). Their chain lengths range between 8 and 26 C atoms, indicating that these compounds have various origins (bacteria, plants, and fungi; Zelles, 1999; Wang et al., 2016 and reference therein). The homologues of 3-OH FAs with 10 to 18 C atoms are considered to be produced exclusively by Gram-negative bacteria (Wollenweber and Rietschel, 1990; Szponar et al., 2003) and will be the only ones considered in the following. Compounds with an even carbon number and *normal* chains were the most

abundant 3-OH FAs in all samples (mean 67.9 % of the total 3-OH FAs, Standard Deviation (SD) 6.8%), with a predominance of the n-C_{14} homologue (21.9%, SD 3.23%; Fig. 1). *Iso* (mean 22.9%, SD 5.01%) and *anteiso* (mean 6.33 %, SD 1.79%) isomers were also present. It must be noted that *anteiso* isomers were only detected for odd carbon-numbered 3-OH FAs (Yang et al., 2015; Wang et al., 2016; Coffinet et al., 2017; Huguet et al., 2019).

The distribution of 3-OH FAs in the soils of the different altitudinal transects did not show a large variability (Fig. 1). Thus, there was no major difference in the relative abundances of most of the 3-OH FAs (i-C_{11}, a-C_{11}, n-C_{11}, i-C_{12}, a-C_{13}, n-C_{13}, i-C_{14}, n-C_{15}, i-C_{16}, a-C_{17} and n-C_{17}) between the 8 study sites, even though slight differences could be observed for some compounds as detailed below. For example, the Peruvian samples were characterized by higher average proportions of n-C_{18} 3-OH FA and lower contribution of the n-C_{10} and n-C_{12} homologues than those from the other transects. Soils from Mt. Shegyla were characterized by lower average proportions of n-C_{14} 3-OH FAs and higher abundances of i-C_{17} compounds compared to the other transects (Fig. 1).

3.1.2. brGDGTs

The relative abundances of brGDGTs were compared between the same transects as for 3-OH FAs, representing a total of 168 samples. The 5- and 6-methyl isomers were separated in most of the samples (Fig. 2; Sup. Table 2), except in older dataset, i.e. soils from Mt. Rungwe (Coffinet et al., 2014, 2017). BrGDGT data from Mt. Rungwe will not be further considered in this study.

The brGDGT distribution was dominated by acyclic compounds (Ia, IIa, IIa', IIIa, IIIa') which represent on average ca. 83.4% of total brGDGTs (SD = 14.5%; Fig. 2). The tetramethylated (Ia-c; mean 39.3%, SD of 20.5%) and the pentamethylated (IIa-c; 44.8%, SD 12.8%) brGDGTs were predominant over the hexamethylated ones (IIIa-c; Fig. 2). The 5-methyl isomers were on average present in a higher proportion (mean 71.9%, SD 23.4%) than the 6-methyl compounds (Fig. 2).

High variability of the brGDGT distribution was observed among the different transects. The relative abundance of brGDGT Ia was much higher in the Peruvian soils (mean 83%, SD 12.6%) than in the other transects (mean between 17.3% and 61.7%; Fig. 2). The 5-methyl isomers were more abundant than the 6-methyl isomers for all sites except for Mt. Pollino (mean 5-methyl = 44%, SD=11.7%) and Mt. Majella (mean 5-methyl = 33.7 %, SD = 5.5%; Fig. 2).

3.2. 3-OH FA and brGDGT-derived indices

3.2.1. 3-OH FA

The RIAN index roughly varied by an order of magnitude among the eight elevation transects (Table 1). The RIAN index ranged from 0.37 to 0.67 for the Peruvian Andes, 0.23 to 0.56 for Mt. Shegyla, 0.15 to 0.34 for Mt. Pollino, 0.21 to 0.53 for the Chilean Andes, 0.26 to 0.80 for Mt. Rungwe (Huguet et al., 2019), 0.16 to 0.46 for Mt. Majella (Huguet et al., 2019), 0.20 to 0.69 for Mt. Shennongjia (Wang et al., 2016) and 0.13 to 0.56 for the French Alps (Véquaud et al., 2021).

The RAN_{15} varied greatly among the different sites (Table 1). It was in the same range along Mts. Rungwe (1.04-5.73) and Majella (0.68-6.43; Huguet et al., 2019). In contrast, its upper limit was higher for Mts. Shennongjia (0.67-10.77; Wang et al., 2016), Shegyla (4.07-12.17), Pollino (2.41-10.26), the Peruvian Andes (2.45-13.77) and the French Alps (1.44-12.26). The range of variation in RAN_{15} was narrower for the Chilean Andes (3.82-6.40).

The RAN_{17} values were similar among the different altitudinal transects (Table 1), ranging from 1.72 to 3.90 along Mt. Shegyla, 0.68 to 6.43 along Mt. Majella (Huguet et al., 2019), 1.19 to 4.54 along Mt. Pollino, 1.91 to 4.25 for the Chilean Andes and 1.12 to 3.57 along Mt. Shennongjia (Wang et al., 2016). The range of RAN_{17} values was narrower for Mt. Rungwe (0.33-1.62; Huguet et al., 2019) and the Peruvian Andes (0.61-2.39) and wider for the French Alps (0.89-6.42; Véquaud et al., 2021) compared to the other sites.

3.2.2. brGDGT

The range of variation in the MBT'$_{5Me}$ index was homogeneous along most transects (0.32-0.63; Table 1), except the Peruvian Andes, with higher values (0.58-0.98; Table 1). Regarding the CBT' index, it showed similar ranges along Chilean Andes (-2.28 to -0.32) and Mt. Shegyla (-2.39 to -0.35; Table 1). This index showed different ranges of variations along the other altitudinal transects: Mts. Shennongjia (-1.18 to 0.50; Yang et al., 2015), Pollino (-0.24 to 0.43) and Peruvian Andes (-1.91 to -1.09). Finally, The CBT' values varied within a narrow range along Mt.Majella (0.23-0.59; Huguet et al., 2019) and within a wide range along the French Alps (-2.29 to 0.52 ; Véquaud et al., 2021).

3.3. Principal component analysis and clustering of 3-OH FA and brGDGT distribution

Principal component analyses were performed to refine the comparison of bacterial lipid distribution (3-OH FAs and brGDGTs) among the different altitudinal transects. *3.3.1. 3-OH FA*

The first two axes of the 3-OH FA PCA explained 39.1% of the total variance in the dataset (Fig. 3a). Dimension 1 (23.9%) opposed samples from Mt. Pollino in the right quadrant to Peruvian soils and samples from Mt. Shennongjia. Dimension 2 (15.2%) especially separated individuals from Chile and Mt. Rungwe. The Wilks' test showed that the location of the sampling sites was the best variable discriminating the distribution of the individuals in the PCA.

Principal component analysis performed on the temperature (RAN_{15}, RAN_{17}) and pH (RIAN) indices derived from 3-OH FAs showed that most of the variance was carried by the first two axes of the PCA (Axis 1 = 56.09%; Axis 2 = 35.29%; Supp. Fig. 2). The first axis was highly correlated with the RAN_{15} ($r = 0.87$) and RAN_{17} ($r = 0.93$) as well as with MAAT ($r = -0.67$), while Axis 2 showed strong correlations with the RIAN ($r = 0.96$) and pH ($r = -0.61$). The PCA allowed visualizing relationships at the scale of the whole dataset, between MAAT and RAN_{15} and RAN_{17} ($r = -0.61$; $r = -0.64$ respectively) and between pH and RIAN ($r = -0.53$).

3.3.2. brGDGT

The first two axes of the brGDGT PCA explained 57.7% of the total variance in the dataset (Fig. 3b). Dimension 1 (42.6%) strongly discriminated soils from Mt. Majella and, to a lesser extent, Mt. Pollino, in the right quadrant from those from Mt. Shegyla, Peruvian Andes and Chilean Andes in the left quadrant. Mts Majella and Pollino were also discriminated negatively along dimension 2 (15.1%). Samples from Mts. Shennongjia and Lautaret-Galibier were distributed over the entire PCA. As for the 3-OH FAs, Wilks' test showed that the location of the sampling sites was the best variable discriminating the distribution of the brGDGTs in the PCA.

4. Discussion

4.1. 3-OH FA and brGDGT-derived proxies

Previous studies conducted on soils from individual altitudinal transects revealed (1) local linear relationships between MAAT/pH and 3-OH FA indices and (2) the potential for

447 combined calibrations using simple linear regressions (Wang et al., 2016; Huguet et al., 2019;
448 Véquaud et al., 2021). In the present study, the existence of linear relationships between 3-OH
449 FA-derived indices and environmental variables was further investigated using an extended soil
450 dataset and the corresponding results were compared with those derived from the brGDGTs,
451 used as an established reference proxy.
452
4.1.1. Relationships between pH and bacterial lipid-derived proxies

454   The relationship between RIAN and pH was investigated along each of the altitudinal
455 transects (Fig. 4a; Sup. Table 3). No significant linear relationship was obtained for the
456 Peruvian Andes, Mts. Rungwe, Pollino and Majella (Huguet et al., 2019) and weak to moderate
457 correlations were observed along Mts. Shegyla and Lautaret-Bauges (R^2 = 0.29-0.46; Sup.
458 Table 3). In contrast, strong regressions between RIAN and pH were observed along Mt.
459 Shennongjia (R^2 = 0.71) and in Chilean Andes (R^2 = 0.66). A weak linear relationship between
460 RIAN and pH (R^2=0.34; RMSE = 0.99; $p = 7.39 \times 10^{-17}$) was also obtained when considering
461 the 168 samples for the eight elevation transects altogether. Therefore, our results confirm the
462 general influence of pH on the relative abundance of 3-OH FAs (Huguet et al., 2019) but
463 suggest that strong linear correlations between RIAN and pH can only be obtained (i) at a local
464 level and (ii) only for some of the sites.
465   As previously suggested (Huguet et al., 2019), the absence or weakness of linear
466 correlations between RIAN and pH may be at least partly due to the small range of variation of
467 pH (<2 units) along some mountains, such as Mts. Rungwe, Majella, and the Peruvian Andes
468 (Fig. 4a; Table 1, Huguet et al., 2019). Transects for the Peruvian Andes and Mt. Majella were
469 also characterized by the absence of relationships between pH and the brGDGT-derived CBT'
470 index, supporting the hypothesis that low pH ranges limit the potential of obtaining linear
471 relationships between indices based on bacterial lipids and pH. Nevertheless, the existence of a
472 narrow pH range was not the only limiting factor in obtaining a strong linear regression between
473 RIAN and pH. Indeed, MAAT rather than soil pH was the dominant driver of soil bacterial
474 diversity and community composition for the Peruvian transect (using 16S rRNA sequencing
475 (Nottingham et al., 2018); and using phospholipid fatty acids (Whitaker et al., 2014)), consistent
476 with the weak correlation between soil pH and bacterial lipids. The weakness of the RIAN-pH
477 relationship may also be partly due to the heterogeneity of soils encountered along a given
478 altitudinal transect, representing specific microenvironments and to the large diversity of
479 bacterial communities in soils from different elevations (Siles and Margesin, 2016). The
480 distribution of 3-OH FAs varies greatly among Gram-negative bacterial species (Bhat and

481 Carlson, 1992) which may account for the significant variability in RIAN values observed in
482 soils from a given transect. Altogether, these results suggest that linear models are not the most
483 suitable for establishing a global calibration between RIAN and pH in soils.

484   Concerning GDGTs, moderate to strong relationships between brGDGT-derived CBT'
485 index and pH were observed along 5 of the 7 altitudinal transects investigated (Fig. 4b; Sup.
486 Table 3). All the individual linear relationships between CBT' and pH, where present, had
487 similar slopes and ordinates and share (for most of the samples) the same 95% confidence
488 intervals (p-value <0.5). This resulted in a strong linear relationship between CBT' index and
489 pH values for the dataset (R^2 = 0.68; RMSE = 0.71; n = 140), which is weaker than the global
490 calibration (R^2 = 0.85; RMSE = 0.52; n = 221) proposed by De Jonge et al. (2014).

491   The discrepancy in relationships between temperature and brGDGTs and 3-OH FAs
492 might partly be due to differences in the relative abundance of these lipids among bacterial
493 communities. The brGDGTs are produced by a more restricted and less diverse number of
494 bacterial species than 3-OH FAs, which are arguably biosynthesized by a large diversity of
495 Gram-negative bacteria species (e.g. Wakeham et al., 2003, Zelles et al., 1995; Zelles, 1999).
496 So far, only bacteria from the *Acidobacteria* phylum were identified as putative brGDGT
497 producers in soils (Sinninghe Damsté et al., 2018). The hypothetical lower diversity of brGDGT
498 producers, in contrast with 3-OH FAs might explain the more homogenous response and lower
499 scatter of the relationships between pH and CBT' index. Moreover, the CBT' index is a ratio
500 based on a restricted number of compounds, representing the direct dependence of the degree
501 of cyclisation of bacterial GDGTs on pH. Conversely, the RIAN index is calculated from the
502 relative abundances of all the individual 3-OH FAs between C_{10} and C_{18} (Wang et al., 2016). It
503 cannot be ruled out that some of the compounds used to calculate the RIAN index are
504 preferentially synthesized, as part of the homeoviscous mechanism, in response to
505 environmental variables other than pH. This calls for a better understanding of the ecology of
506 3-OH FA-producing bacteria and their adaptation mechanisms.

507
508

    4.1.2 Relationships between MAAT and bacterial lipid-derived proxies

510   RAN_{15} was previously shown to be correlated with MAAT along Mts. Rungwe,

511   Majella and Shennongjia (Wang et al., 2016; Huguet et al., 2019). Moderate to strong linear

512   correlations (R^2 =0.49-0.79) between RAN_{15} and MAAT were also observed along most of the

513   individual transects investigated (Fig. 5a; Sup. Table 3, except along the Chilean and Lautaret-

514   Bauges transects. The individual correlations do not share the same 95% confidence intervals

515   and even when some of them present similar slopes, the regression lines display significantly

516   different intercepts (*p*-value > 0.05) (Fig. 5a). This supports the hypothesis of a site-dependent

517   effect of the linear RAN_{15}-MAAT relationship previously made by Huguet et al. (2019).

518   Similarly to RAN_{15}, RAN_{17} was moderately to strongly correlated (R^2 =0.53-0.81)

519   with MAAT along 5 out of 8 individual transects (Fig. 5b; Sup. Table 3). The small range of

520   variation in MAAT along the Chilean transect (6.0-9.2 °C) (Table 1), associated with that of

521   the RAN_{15} /RAN_{17}, could explain the lack of a linear relationship between the MAAT and these

522   indices. As for the French Alps (Mts Lautaret-Bauges), the influence of local environmental

523   parameters (pH and to a lesser extent soil moisture and grain size, related to vegetation and soil

524   types, or thermal regimes associated with the snow cover) on 3-OH FA distribution was shown

525   to be predominant over that of MAAT (Véquaud et al., 2021). In contrast with RAN_{15}, the linear

526   regressions between RAN_{17} and MAAT along Mts. Shegyla, Shennongjia, Rungwe and the

527   Peruvian Andes transects share confidence intervals at 95% and have similar slope and intercept

528   values (*p*-value <0.05; Fig. 5b; Sup. Table 3), suggesting that RAN_{17} could be a more effective

529   global proxy for MAAT reconstructions than RAN_{15}.

530   In order to test the hypothesis that RAN_{17}, rather than RAN_{15}, is a more effective

531   global proxy for MAAT, the global calibrations between RAN_{15}/RAN_{17} and MAAT based on

532   the entire soil dataset (*n* = 168) were compared. The two linear regressions had similar moderate

533   determination coefficients (R^2 = 0.37 and 0.41 for RAN_{15} and RAN_{17}, respectively) and similar

534   high RMSE (RMSE = 5.46°C and 5.28°C for RAN_{15} and RAN_{17}, respectively; Sup. Table 4).

535   For all transects (except for the Mt Majella RAN_{17}/MAAT relationship), the individual local

536   regressions between RAN_{15}/RAN_{17} and MAAT outperformed the proposed global linear

537   calibrations in terms of determination coefficients (0.49-0.81) and RMSE (1.98-3.57 °C; Sup.

538   Table 4), suggesting that local rather than global linear transfer functions based on RAN_{15} or

539   RAN_{17} may be more appropriate for paleotemperature reconstructions in soils.

540   The difficulties in establishing global linear RAN_{15}/RAN_{17}-MAAT calibrations may

541   partly be due to the fact that microbial diversity, especially for 3-OH FA-producing Gram-

542   negative bacteria (Margesin et al., 2009; Siles and Margesin, 2016), can vary greatly from one

soil to another, resulting in variation of the RAN_{15}/RAN_{17} indices, as also assumed for the RIAN. The strong regional dependence of the 3-OH FA distribution may thus explain the weak correlation between 3-OH FA-derived indices (RAN_{15}, RAN_{17} and RIAN) and environmental variables (MAAT/pH) at a global level. This regional dependency was further supported by the PCA of the relative abundance of 3-OH FAs across the global dataset, which showed that the individuals were grouped based on the sampling location (Fig. 3a).

In addition to 3-OH FAs, the relationships between brGDGT distribution and MAAT were investigated along the seven transects for which the 5- and 6-methyl brGDGT isomers were separated (Mts Shegyla, Pollino Majella, Lautaret-Bauges, Shennongjia, Peruvian Andes and Chilean Andes). These individual transects showed moderate to strong relationships between MAAT and MBT'$_{5Me}$ (R^2 0.35-0.89; Fig. 6 and Sup. Table 3), with similar slopes and ordinates (except for the Peruvian Andes) and shared 95% confidence intervals for most of the samples. A distinct relationship between MBT'$_{5Me}$ and MAAT was observed along the Peruvian Andes and Mt Majella transects (Fig. 6a), as also observed for the RIAN and RAN_{15} indices (Figs 4a and 5a). The singularity of the Peruvian soils is also visible on the PCA performed on the brGDGT distribution (Fig. 3b), where the samples from this region are pooled very narrowly from the rest of the dataset. This specific trend is difficult to explain, even though the Peruvian Andes are subjected to warmer climatic conditions (Table 1) than the other temperate transects, which may in turn affect the nature of the microbial communities encountered in the soils and the bacteria lipid distribution (Siles and Margesin, 2016; Hofmann et al., 2016; De Jonge et al., 2019).

A moderate linear relationship between MAAT and MBT'$_{5Me}$ (MAAT = 24.5 × MBT'$_{5Me}$ -4.78; R^2 = 0.57, RMSE = 3.39 °C, n = 140; Sup. Table 3) was observed after combining the data for the seven aforementioned altitudinal transects. This global relationship follows a similar trend as the calibration proposed by De Jonge et al. 2014 (MAAT = 31.45 × MBT'$_{5Me}$ - 8.57) and is more robust and accurate than those obtained between the RAN_{15}/RAN_{17} and MAAT (Sup. Table 3). This confirms that the MBT'$_{5Me}$ index can be applied at a global scale using a simple linear regression model as previously shown (De Jonge et al., 2014; Naafs et al., 2017), in contrast with the RAN_{15} and RAN_{17} proxies, for which only strong local calibrations with MAAT were found.

As a similar conclusion was obtained for the RIAN-pH proxy, it appears necessary to use more complex models to develop global calibrations between 3-OH FA-derived proxies and MAAT/pH. This novel method allows taking into account the complexity and specificity of each environmental site.

4.2. Development of new models for the reconstruction of MAAT and pH from 3-OH FA

Several complementary methods were recently used to derive calibrations with environmental parameters from organic proxies. Most calibrations between lipid distribution and environmental variables were based on simple linear regression models, most often the ordinary least square regression (e.g. for brGDGTs: De Jonge et al., 2014; Wang et al., 2016), as it is simple and easy to implement and understand. Other linear models, such as Deming regression (Naafs et al., 2017) or Bayesian regression (Tierney and Tingley, 2014; Dearing Crampton-Flood et al., 2020) were also used. Nevertheless, these single linear regression methods rely on a given index (e.g. MBT'$_{5Me}$ or CBT' for brGDGTs) which is correlated with environmental parameters. This represents a limitation, as the relative distribution of bacterial lipids can be concomitantly influenced by several environmental parameters (e.g. Véquaud et al., 2021) and can also depend on the microbial diversity and sources of these compounds (Parker et al., 1982; Bhat and Carlson, 1992; Zelles, 1999). In contrast, using bacterial relative abundances rather than a single index in the relationships with environmental variables appears less restrictive, and more representative of the environmental complexity. Other models can be used in this way, such as those based on multiple regressions (e.g. Peterse et al., 2012; De Jonge et al., 2014; Russell et al., 2018), describing the relationships between one or several explained variables (e.g. bacterial lipid abundances) and one or several explanatory variables (e.g. MAAT, pH). Multiple regressions can reveal the presence of linear relationships among several known variables but cannot take into account non-linear influences, which may occur in complex environmental settings. This limitation, common to all linear models, can be overcome using non-parametric methods such as some of the machine-learning algorithms (e.g. nearest neighbours or random forest; Dunkley Jones et al., 2020). The reliability of the latter models lies in the fact that they are non-linear, which helps capturing the intrinsic complexity of the environmental setting, and that they avoid the regression dilution phenomenon observed in most linear models. Moreover, their robustness is improved by the fact that they are built on a randomly defined proportion of the total dataset and then tested on the rest of the dataset, considered as independent. Last, these machine-learning algorithms are flexible and are continuously evolving when adding new samples.

As shown in section 4.1., robust global calibrations between 3-OH FA-derived indices (RIAN, RAN$_{15}$ and RAN$_{17}$) and MAAT/pH could not be established using a simple linear regression model, contrary to what was observed with brGDGT-derived indices. Therefore,

610 three different independent and complementary models were tested to potentially establish
611 stronger statistical relationships between 3-OH FA distributions and pH/MAAT at the global
612 level : (i) a parametric model – multiple linear regression; (ii) two non-parametric models –
613 random forest (e.g. Ho, 1995; Denisko and Hoffman, 2018) and k-NN algorithms (e.g.
614 Gangopadhyay et al., 2009). As discussed above, the multiple linear regression model allows
615 the determination of linear relationships between MAAT/pH and the individual relative
616 abundances of 3-OH FAs, instead of indices derived from the latter. As for the two non-
617 parametric models, they present among other things the advantage of taking into account non-
618 linear environmental influences.

619 The three models, based on a supervised machine learning approach, were applied to
620 the total soil dataset (n=168). All the 3-OH FA homologues, whatever their abundance, were
621 included in the models to keep the maximum variability and take into account the specificity
622 and complexity of each altitudinal transect. Indeed, the nature of the individual 3-OH FAs
623 whose fractional abundance is mainly influenced by MAAT/pH may be site-dependent, as
624 previously observed (Véquaud et al., 2021). The performances of these three models were
625 compared with those of the linear calibrations between 3-OH FA-derived indices (RAN$_{15}$,
626 RAN$_{17}$, RIAN) and MAAT/pH (Table 2).

627

628 *4.2.1. Temperature calibrations*

629 The multiple linear regression model yielded a strong relationship between 3-OH FA
630 relative abundances and MAAT (Fig. 7a; Eq.6):
631 *MAAT (°C) = -59.02 × [nC$_{10}$] + 102.1 × [iC$_{11}$] + 2628.49 × [aC$_{11}$] – 165.58× [nC$_{11}$] – 79.799*
632 *× [nC$_{12}$] + 89.93 × [iC$_{13}$] + 205.06 × [aC$_{13}$] – 136.25 × [nC$_{13}$] – 309.71 × [iC$_{14}$] – 43.16 ×*
633 *[nC$_{14}$] – 9.27 × [iC$_{15}$] – 308.53 × [aC$_{15}$]+ 66.06 × [nC$_{15}$] – 60.57 × [iC$_{16}$] + 15.53 × [nC$_{16}$]+*
634 *13.52 × [iC$_{17}$] – 228.76 × [aC$_{17}$] – 91.12 × [nC$_{17}$]+ 42.16 × [nC$_{18}$]+43.71*
635 *(n = 168; R^2 = 0.79; RMSE = 3.0 °C)* (6)
636 This model, which takes into account the whole suite of 3-OH FAs (C$_{10}$-C$_{18}$), presents a higher
637 strength than the global linear relationships between 3-OH FA derived indices and MAAT
638 (R²=0.37 and 0.41; RMSE =5.5°C and 5.3°C for RAN$_{15}$ and RAN$_{17}$, respectively; Table 2). The
639 multiple linear regression also improves the accuracy and robustness of MAAT prediction in
640 comparison with single linear relationships, with lower RMSE (3.0 °C), variance of the
641 residuals (9.2 °C; Fig. 7d) and mean absolute error (MAE; 2.3 °C) than with the RAN$_{15}$ and
642 RAN$_{17}$ calibrations (RMSE of 5.5 and 5.3 °C; variance of 29.8 and 27.9 °C; MAE of 4.0 and
643 3.9 °C for RAN$_{15}$ and RAN$_{17}$, respectively; Table 2).

644   Similarly to the multiple linear regression model (Fig. 7a), the random forest (Fig. 7b)

645 and k-NN (Fig. 7c) calibrations are characterized by strong determination coefficients (R^2 0.83

646 and 0.77, respectively). The variance in residuals, MAE and RMSE of the random forest

647 calibration are slightly lower than those of the multiple linear regression and k-NN models

648 (Table 2). An advantage of the random forest algorithm lies in the fact that the weight of the

649 different variables used to define the model can be quantified using the permutation importance

650 method (Breiman, 2001). The a-C_{15}, i-C_{14}, a-C_{17}, n-C_{12}, n-C_{15}, and to a lesser extent n-C_{17}, n-

651 C_{16} and i-C_{13} 3-OH FAs were observed to be the homologues predominantly used by the model

652 to estimate MAAT values (Fig. 9a). They include all the 3-OH FAs involved in the calculation

653 of the RAN_{15} and RAN_{17} indices, especially the a-C_{15} homologue. This may explain why linear

654 relationships between the RAN_{15}/RAN_{17} and MAAT could be established along some of the

655 altitudinal transects investigated until now (Wang et al., 2016; Huguet al., 2019; Véquaud et

656 al., 2021; this study). Nevertheless, other individual 3-OH FAs than those appearing in the

657 calculation of the RAN_{15} and RAN_{17} have also a major weight in the random forest model and

658 seem to be influenced by temperature changes, explaining the moderate determination

659 coefficients of the global RAN_{15}/RAN_{17}-MAAT linear relationships observed in this study.

660   On the whole, the strength and accuracy of the multiple linear regression, k-NN and

661 random forest models are much higher than those based on the RAN_{15} and RAN_{17} indices

662 (Table 2). This is likely related to the fact that the three aforementioned models integrate the

663 whole suite of 3-OH FAs homologues (C_{10} to C_{18}) and thus better capture the complexity of the

664 response of soil Gram-negative bacteria and their lipid distribution to temperature changes than

665 the RAN_{15} and RAN_{17} indices. They also present the advantage of increasing the range of

666 temperature which may be predicted by more than 4 °C in comparison with the RAN_{15} and

667 RAN_{17} calibrations (Table 2). Indeed, even though the lower limit of MAAT estimates for the

668 three models tested in the present study is slightly higher than those based on the RAN_{15} and

669 RAN_{17} indices, the upper limit of the MAAT which can be estimated using the multiple linear

670 regression, random forest and k-NN models is substantially higher (ca. 25 °C) than that based

671 on the RAN_{15} or RAN_{17} indices (ca. 17 °C; Table 2).

672   The three proposed models show the potential of 3-OH FAs as MAAT proxies at the

673 global level, which was not visible using RAN_{15} and RAN_{17} indices. The non-parametric

674 models (random forest and k-NN) may benefit from the fact that they take into account the

675 complex, non-linear relationships between environmental parameters and bacterial lipid

676 abundance. This is highlighted when comparing the independent variations of the individual 3-

677 OH FA relative abundances with estimated MAAT for the three proposed models, with non-

678 linear trends for the k-NN and random forest models, in contrast with the multiple linear

679 regression (Supp. Fig. 2).

680

681   *4.2.2. pH calibrations*

682   A robust linear relationship between the RIAN and pH could not be obtained from the

683 whole soil dataset (Fig. 4a; Table 2). In contrast, the multiple regression model provided a

684 strong correlation between the 3-OH FA fractional abundances and pH (Fig. 8a; Eq. 7):

685 *pH = -1.45 × [nC₁₀] − 31.70 × [iC₁₁] − 162.09 × [aC₁₁] − 53.22 × [nC₁₁] − 6.21× [nC₁₂] +*

686 *56.24× [iC₁₃] − 2.02 × [aC₁₃] + 15.10 × [nC₁₃] + 23.99 × [iC₁₄] − 4.54× [nC₁₄] − 13.79 ×*

687 *[iC₁₅] − 15.74 × [aC₁₅] + 1.93 × [nC₁₅] − 46.29 × [iC₁₆] − 3.20 × [nC₁₆] − 1.80 × [iC₁₇] −*

688 *8.90 × [aC₁₇] + 11.46 ×[nC₁₇] − 3.63×[nC₁₈] + 7.84*   *(n* = 168; R² = 0.64; RMSE = 0.8)   (7)

689   The random forest (Fig. 8b) and k-NN pH models (Fig. 8c) appeared to be slightly more

690 robust and accurate than the multiple linear regression (Fig. 8a), as the former two models

691 presented slightly higher determination coefficients (R² = 0.68 and 0.70 for k-NN and random

692 forest, respectively) and slightly lower RMSE (0.7), variance in residuals (0.5) and MAE (0.5)

693 than the multiple linear regression (Table 2).

694   As for the MAAT random forest model, the weight of the individual 3-OH FAs in the

695 pH random forest calibration was determined (Fig. 9b). Three homologues − i-C₁₃, n-C₁₅, i-C₁₆

696 − had a larger weight in the global pH model than the others (Fig. 9b). This is consistent with a

697 detailed study of 3-OH FA distribution in soils from the French Alps (Véquaud et al., 2021),

698 where the i-C₁₃ and i-C₁₆ 3-OH FAs were observed to be predominantly influenced by pH.

699 Nevertheless, in addition to the three aforementioned homologues, most of the C₁₀ to C₁₈ 3-OH

700 FAs have a non-negligible influence in the random forest pH model, except the a-C₁₅ and i-C₁₄

701 compounds (Fig. 9b). This is in line with the definition of the 3-OH FA-based pH index (RIAN)

702 defined by Wang et al. (2016) which includes the whole suite of 3-OH FAs. These results

703 suggest that soil Gram-negative bacteria may respond to pH variations by modifying the whole

704 distribution of 3-OH FAs, even though this needs to be further confirmed by working at the

705 microbial level.

706   In any case, in contrast with the RIAN index, the multiple linear regression, k-NN and

707 random forest models provided strong global calibrations with pH (Fig. 8), as robust as the

708 global CBT'-pH relationship (Fig. 4b). The three proposed models also increase the range of

709 pH which can be estimated (~ 4 pH units) in comparison with the RIAN global calibration (~ 3

710 pH units), further strengthening the potential of these models for soil pH reconstruction. As

711 MAAT models, the independent variations of the individual 3-OH FA relative abundances with

712 estimated pH highlight non-linear trends for the k-NN and random forest models, in contrast

713 with the multiple linear regression (Supp. Fig. 3), which might favor the use of the two non-

714 parametric models in order to take into account such non-linear influences. The machine-

715 learning MAAT and pH models proposed in this paper are flexible and could be further

716 improved by increasing the number of soil samples analyzed and the representativeness of the

717 different MAAT and pH values within the dataset.

718

719 **4.3. Paleoclimate application of the new 3-OH FA/MAAT models**

720

721 The multiple regression, random forest and k-NN models developed for MAAT

722 reconstruction using 3-OH FAs were similar in terms of robustness and precision (Figs. 7a, b,

723 c; Table 2). The performance and validity of these global terrestrial calibrations for

724 paleotemperature reconstructions was thus tested and compared with the MAAT record from a

725 Chinese speleothem (HS4 stalagmite) covering the last 9,000 years BP (Wang et al., 2018).

726 This terrestrial archive was the object of previous paleostudies, thus providing a context for the

727 interpretation of the MAAT data and, to the best of our knowledge, represents the only

728 published application of 3-OH FAs as a paleotemperature proxy in terrestrial settings (Wang et

729 al., 2018). The local comparison of 3-OH FA distributions in the overlying soils and stalagmites

730 and the analyses of bacterial diversity and transport pathways suggested that the 3-OH FAs in

731 the HS4 speleothem were mainly soil-derived (Wang et al., 2018), supporting the application

732 of soil calibrations for MAAT reconstruction from this archive, although not being a paleosoil

733 itself. The first paleoapplication of 3-OH FAs (Wang et al., 2018) on this speleothem relied on

734 a local calibration between the RAN_{15} index and MAAT proposed by Wang et al. (2016) using

735 soils from Mt. Shennogjia. The MAAT estimates derived from our global soil calibrations were

736 compared with those obtained from this local soil calibration (Wang et al., 2016).

737

738

739 *4.3.1 Comparison of the multiple linear regression, k-NN and random forest global*

740 *MAAT calibrations*

741 The multiple regression model (Eq. 6; Fig. 7a) yielded MAAT estimates ranging

742 between -35 and 22.8 °C over the last 9,000 years (Supp. Fig. 4). The temperature minimum (-

743 35°C) observed at 560 yrs BP can be considered as an outlier, with a significantly lower MAAT

744 estimate than those provided by the other samples. After having ignored this apparent outlier,

745 the MAAT range over the last 9,000 years was comprised between 3.2°C and 22.8°C, with

temperature shifts of up to 15 °C within very short periods of time. The observed range of MAAT and large variations in temperature over such short periods appear far too excessive, as the expected amplitude of MAAT during the Holocene is expected to be up to ca. 2-3 °C (Liu et al., 2014). This highly questions the reliability of the multiple linear regression model for MAAT reconstruction from this archive.

MAAT estimates derived from the k-NN calibration ranged between 6.5 and 19.7 °C over the last 9,000 years (Supp. Fig. 4). Abrupt shifts in MAAT of more than 10 °C were observed between 2,000 and 4,000 yrs BP. Such variations, higher than the RMSE of the calibration, appear excessive for the Holocene period, as previously discussed for the multiple regression model. The bias in MAAT estimates may be due to the intrinsic definition of the k-NN model, which is better suited for uniformly distributed datasets. This is not the case here, as the individual transects heterogeneously cover a wide range of temperatures. The application of a global calibration at the local scale – that of the HS4 stalagmite – using the k-NN method and based on the similarities among samples, thus does not appear appropriate. Such a calibration might be improved by extending the dataset with samples more equally distributed across a wider range of global climatic gradients.

Finally, the random forest model yielded MAAT estimates between 10.6 and 19.3°C, i.e. a much smaller range than the k-NN algorithm and multiple regression model (Supp. Fig. 4). The amplitude of the shifts observed between 2,000 and 4,000 yrs BP was ca. 4°C, which is climatically more consistent than the variations obtained with the k-NN method and multiple regression model, even though these large variations in MAAT over such short periods of time still appear too excessive. Furthermore, the application of the global random forest calibration roughly provided similar temperature trends as those derived from the local RAN$_{15}$ calibration by Wang et al. (2018; Fig. 10), despite some largest oscillations for the global model. These results suggest that the random forest calibration is more reliable than the multiple regression and k-NN ones. This can be explained by the intrinsic definition of the random forest algorithm, which averages the results of several independent models (so-called decision trees), thus reducing the variance and thus the forecast error on the final model. This is also in line with the slightly higher accuracy of the random forest calibration compared with the other two models (Table 2), as previously discussed. In contrast, the multiple regression calibration was the less performant of the three models on the investigated archive. This may be related to its parametric nature and the fact that it does not take into account the natural non-linear variations on 3-OH FA fractional abundances highlighted by the random forest and k-NN models (Supp. Figs. 2 and 3).

780   In conclusion, the three models proposed in this study, especially the random forest,

781 have potential for MAAT reconstruction, even though the application to a well-known

782 paleoclimate archive showed their limitations. This highlights the importance of testing new

783 calibrations on well-characterized archives to investigate their reliability.

784

785   *4.3.2. Comparison of the global random forest and local RAN$_{15}$ calibrations for MAAT*

786 *reconstruction*

787   The random forest model was observed to be the most reliable of the three proposed

788 global MAAT calibrations (Fig. 7). To go further, we compared the temperature record derived

789 from our global random forest calibration with that derived from the local MAAT/RAN$_{15}$

790 transfer function proposed by Wang et al. (2016; Fig. 10). The application of the local RAN$_{15}$

791 calibration to the HS4 stalagmite yielded an average MAAT of ca. 18.4 °C over the most recent

792 part of the record (last 800 yrs; Fig. 10), consistent with the MAAT of 18 °C recorded *in situ*

793 by a temperature logger (Hu et al., 2008; Wang et al., 2018). In contrast, absolute MAAT

794 estimates derived from the random forest model were on average 14.2 °C over the last 800 yrs

795 and were generally lower than those obtained from the local RAN$_{15}$ calibration over the whole

796 record. Altogether, these results suggest that the random forest model tends to underestimate

797 absolute MAAT, in contrast with the RAN$_{15}$ calibration proposed by Wang et al. (2016). This

798 discrepancy may be due the fact that the calibration proposed in the present study is based on a

799 global dataset, with samples subject to a large variety of environmental and climatic conditions,

800 whereas the RAN$_{15}$-MAAT transfer function by Wang et al. (2016) was constructed using soil

801 samples from a regional altitudinal transect, located at only 120 km distance from the stalagmite

802 site (Wang et al., 2018).

803   Even though the local calibration by Wang et al. (2016) provides more accurate

804 absolute MAAT values than the present global random forest model, as it could be expected,

805 both calibrations roughly generate similar qualitative MAAT trends over time. A regular slight

806 decrease in temperature of ca. 1 °C was observed between 9,000 and ca. 1,000 yrs BP based on

807 the local RAN$_{15}$ calibration (Fig. 10a; Wang et al., 2018). This general decreasing trend was

808 also visible when using the random forest model, but with larger oscillations and mainly

809 between 9,000 and 4,000 yrs BP, in agreement with the general trend recorded by the $\partial^{18}O$

810 record (mixture of temperature and hydrological signals, Wang et al., 2018) of the HS4

811 stalagmite (Fig. 10c,d; Hu et al., 2008). In addition, both the global random forest, local RAN$_{15}$

812 calibrations and the $\partial^{18}O$ record allowed the identification of several climatic events in the

813 Northern hemisphere, in agreement with the reconstructed total solar irradiance (TSI,

814 Steinhilber et al., 2009, Fig. 10e). Thus, both models highlighted, with slightly different
815 amplitudes, the Medieval Warm Period (800-1000 years BP) and Little Ice Age (LIA; 200-500
816 years BP) periods (Mann et al., 2008; Ljungqvist, 2010; Wang et al., 2018). The LIA event is
817 particularly well represented by the global random forest calibration, in line with the decrease
818 in the TSI (Fig. 10c,e) associated with a relative increase in the $\partial^{18}O$ of HS4 carbonates
819 (dry/cool event, Wang et al., 2018). Before the MWP, the global random forest calibration
820 shows slight oscillations, which can be assumed to be representative of TSI variations between
821 500 and 1,300 yrs BP. Similarly, an important cooling event, well correlated with a significant
822 decrease in the TSI (Fig. 10a, c, e), was recorded by the two calibrations at 1300 yr BP.

823   The global random forest calibration also highlighted two cooling events, poorly
824 represented by the local RAN$_{15}$ calibration: one at ca. 4,200 yrs BP ago and, to a lesser extent,
825 another one between 2,800 and 3,000 yrs BP (Bond et al., 2001; Mayewski et al., 2004). The
826 event at 4,200 yrs BP is consistent with the $\partial^{18}O$ and solar irradiance records and is referenced
827 in the literature as the "4.2 kiloyear event" (deMenocal, 2001). This intense drought event was
828 suggested to have had a major impact on different civilizations (collapses, migrations;
829 (Gibbons, 1993; Staubwasser et al., 2003; Li et al., 2018; Bini et al., 2019). Thus, in some parts
830 of China, the production of rice fields sharply decreased during this period, leading to a decrease
831 in population (Gao et al., 2007).

832   Both calibrations additionally shows a cooling period between 4,000 yrs and 3,200 yrs
833 BP, more pronounced based on the global random forest model, followed by another cooling
834 between 3,200 years BP and 3,000 yrs BP. This cooling period is consistent with the trends
835 derived from $\partial^{18}O$ and solar irradiance records. It culminates with a cold episode at 3000 yrs
836 BP, also known as Late Bronze Age Collapse (Kaniewski et al., 2013). Indeed, this cold
837 episode, combined with droughts, may have led to a decrease in agricultural production in
838 China, contributing to the degradation of trade routes and ultimately to the collapse of Bronze
839 Age civilizations (Weiss, 1982; Knapp and Manning, 2016). Last, the global random forest
840 calibration also highlights two additional cold events, between 5,600 and 5,900 yrs BP, as well
841 as around 7,100 yrs BP, corresponding to solar irradiance minima (Bond et al., 2001; Mayewski
842 et al., 2004) and which are not as clearly visible with the local RAN$_{15}$ calibration by Wang et
843 al. (2016).

844   The first application of the random forest calibration to a natural archive shows the
845 potential of 3-OH FAs as paleotemperature proxies at a global scale, as known and documented
846 climatic events were recorded, with a similar RMSE (2.8 °C; Table 2) as that of the local
847 calibration by Wang et al. (2.6 °C; 2016). This RMSE is also much lower than the one related

848   to the latest global MAAT-brGDGT calibrations (> 4 °C; De Jonge et al., 2014; Naafs et al.,

849   2017; Dearing Crampton-Flood et al., 2020), even though the latter are based on a larger number

850   of soil samples than the global 3-OH FA model proposed in the present study. In summary, we

851   demonstrate that 3-OH FAs are promising and effective temperature proxies for terrestrial

852   settings, complementary to, and independent of, the brGDGTs, and also highlight the usefulness

853   of non-parametric models using machine learning, especially the random forest algorithm, to

854   establish global MAAT calibrations. We expect that analyses of 3-OH FAs in a larger number

855   of globally distributed soils will further improve the accuracy and robustness of the global

856   random forest calibration for paleotemperature reconstruction.

857

858

859   **5. Conclusions**

860   3-OH FAs have been recently proposed as environmental proxies in terrestrial settings,

861   based on local studies. This study investigated for the first time the applicability of these

862   compounds as MAAT and pH proxies at the global scale using an extended soil dataset across

863   a series of globally distributed elevation transects ($n = 168$). Strong linear relationships between

864   3-OH FA-derived indices (RAN_{15}, RAN_{17} and RIAN) and MAAT/pH could only be obtained

865   locally, for some individual transects, suggesting that these indices cannot be used as

866   paleoproxies at the global scale through this kind of model. Other algorithms (multiple linear

867   regression, k-NN and random forest models) were tested and, in contrast with simple linear

868   regressions, provided strong global correlations between MAAT/pH and 3-OH FA relative

869   abundances. The applicability of these three models for paleotemperature reconstruction was

870   tested and compared with the MAAT record from the unique available record: a Chinese

871   speleothem. The calibration based on the random forest model appeared to be the most robust

872   and showed similar trends to previous reconstructions and known Holocene climate variations.

873   Furthermore, the global random forest model highlighted documented climatic events poorly

874   represented by the local RAN_{15} calibration. This new global model is promising for

875   paleotemperature reconstructions in terrestrial settings and could be further improved by

876   analyzing 3-OH FAs in a larger number of globally distributed soils. This study demonstrates

877   the major potential of 3-OH FAs as MAAT/pH proxies in terrestrial environments through the

878   different models presented and their application for paleoreconstruction.

879

880   **Data availability.** All data are available in the Supplementary tables.

881

Author contributions. P.V. performed the lipid and statistical analyses and wrote a first draft of the paper., A.H. and S.D. supervised the work of P.V. and corrected the first draft, P.V. and A.T. developed the different models, G.B., A.N., W.P.S., N.S., J.P.W. and S.C. provided samples and/or associated data, and all the co-authors reviewed and commented on the paper.

Competing interests. The authors declare that they have no conflict of interest.

Acknowledgments. We thank Sorbonne Université for a PhD scholarship to P.V. and the Labex MATISSE (Sorbonne Université) for financial support. The EC2CO program (CNRS/INSU – BIOHEFECT/MICROBIEN) is thanked for funding of the SHAPE project. A.H. and S.C. are grateful for funding of the ECOS SUD/ ECOS ANID #C19U01 project. We are grateful to Jérôme Poulenard for discussions on soil characteristics, and for comments on the manuscript. We thank Dr. Juntao Wang and Prof. Jinzheng He for having provided soils from Mt. Shegyla. We thank the reviewers for their comments which helped in improving the manuscript.

References

Beales, N.: Adaptation of Microorganisms to Cold Temperatures, Weak Acid Preservatives, Low pH, and Osmotic Stress: A Review, 3, 1–20, https://doi.org/10.1111/j.1541-4337.2004.tb00057.x, 2004.

Bhat, U. R. and Carlson, R. W.: A new method for the analysis of amide-linked hydroxy fatty acids in lipid-As from gram-negative bacteria, Glycobiology, 2, 535–539, https://doi.org/10.1093/glycob/2.6.535, 1992.

Bini, M., Zanchetta, G., Persoiu, A., Cartier, R., Catala, A., Cacho, I., Dean, J. R., Di Rita, F., Drysdale, R. N., Finné, M., Isola, I., Jalali, B., Lirer, F., Magri, D., Masi, A., Marks, L., Mercuri, A. M., Peyron, O., Sadori, L., Sicre, M.-A., Welc, F., Zielhofer, C., and Brisset, E.: The 4.2 ka BP Event in the Mediterranean region : an overview, 15, 555–577, 2019.

Bonanomi, G., Zotti, M., Mogavero, V., Cesarano, G., Saulino, L., Rita, A., Tesei, G., Allegrezza, M., Saracino, A., and Allevato, E.: Climatic and anthropogenic factors explain the variability of Fagus sylvatica treeline elevation in fifteen mountain groups across the Apennines, Forest Ecosystems, 7, 5, https://doi.org/10.1186/s40663-020-0217-8, 2020.

Bond, G., Kromer, B., Beer, J., Muscheler, R., Evans, M. N., Showers, W., Hoffmann, S., Lotti-Bond, R., Hajdas, I., and Bonani, G.: Persistent Solar Influence on North Atlantic Climate During the Holocene, 294, 2130–2136, https://doi.org/10.1126/science.1065680, 2001.

Brassell, S. C., Eglinton, G., Marlowe, I. T., Pflaumann, U., and Sarnthein, M.: Molecular stratigraphy: a new tool for climatic assessment, 320, 129–133, https://doi.org/10.1038/320129a0, 1986.

Breiman, L.: Random Forests, Machine Learning, 45, 5–32, https://doi.org/10.1023/A:1010933404324, 2001.

Carter, M. R., Gregorich, E. G., and Gregorich, E. G.: Soil Sampling and Methods of Analysis, CRC Press, https://doi.org/10.1201/9781420005271, 2007.

Coffinet, S., Huguet, A., Williamson, D., Fosse, C., and Derenne, S.: Potential of GDGTs as a temperature proxy along an altitudinal transect at Mount Rungwe (Tanzania), Organic Geochemistry, 68, 82–89, https://doi.org/10.1016/j.orggeochem.2014.01.004, 2014.

Coffinet, S., Huguet, A., Pedentchouk, N., Bergonzini, L., Omuombo, C., Williamson, D., Anquetil, C., Jones, M., Majule, A., Wagner, T., and Derenne, S.: Evaluation of branched GDGTs and leaf wax n-alkane δ2H as (paleo) environmental proxies in East Africa, Geochimica et Cosmochimica Acta, 198, 182–193, https://doi.org/10.1016/j.gca.2016.11.020, 2017.

De Jonge, C., Hopmans, E. C., Zell, C. I., Kim, J.-H., Schouten, S., and Sinninghe Damsté, J. S.: Occurrence and abundance of 6-methyl branched glycerol dialkyl glycerol tetraethers in soils: Implications for palaeoclimate reconstruction, Geochimica et Cosmochimica Acta, 141, 97–112, https://doi.org/10.1016/j.gca.2014.06.013, 2014.

De Jonge, C., Radujković, D., Sigurdsson, B. D., Weedon, J. T., Janssens, I., and Peterse, F.: Lipid biomarker temperature proxy responds to abrupt shift in the bacterial community

composition in geothermally heated soils, Organic Geochemistry, 137, 103897, https://doi.org/10.1016/j.orggeochem.2019.07.006, 2019.

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

deMenocal, P. B.: Cultural Responses to Climate Change During the Late Holocene, 292, 667–673, https://doi.org/10.1126/science.1059287, 2001.

Denich, T. J., Beaudette, L. A., Lee, H., and Trevors, J. T.: Effect of selected environmental and physico-chemical factors on bacterial cytoplasmic membranes, Journal of Microbiological Methods, 52, 149–182, https://doi.org/10.1016/S0167-7012(02)00155-0, 2003.

Denisko, D. and Hoffman, M. M.: Classification and interaction in random forests, Proc Natl Acad Sci USA, 115, 1690–1692, https://doi.org/10.1073/pnas.1800256115, 2018.

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

Eglinton, T. I. and Eglinton, G.: Molecular proxies for paleoclimatology, Earth and Planetary Science Letters, 275, 1–16, https://doi.org/10.1016/j.epsl.2008.07.012, 2008.

Emiliani, C.: Pleistocene Temperatures, The Journal of Geology, 63, 538–578, https://doi.org/10.1086/626295, 1955.

Erez, J. and Luz, B.: Experimental paleotemperature equation for planktonic foraminifera, Geochimica et Cosmochimica Acta, 47, 1025–1031, https://doi.org/10.1016/0016-7037(83)90232-6, 1983.

Gangopadhyay, S., Harding, B. L., Rajagopalan, B., Lukas, J. J., and Fulp, T. J.: A nonparametric approach for paleohydrologic reconstruction of annual streamflow ensembles, 45, https://doi.org/10.1029/2008WR007201, 2009.

Gao, H., Zhu, C., and Xu, W.: Environmental change and cultural response around 4200 cal. yr BP in the Yishu River Basin, Shandong, J GEOGR SCI, 17, 285–292, https://doi.org/10.1007/s11442-007-0285-5, 2007.

Gibbons, A.: How the Akkadian Empire Was Hung Out to Dry, Science, 261, 985, https://doi.org/10.1126/science.261.5124.985, 1993.

Hazel, J. R. and Eugene Williams, E.: The role of alterations in membrane lipid composition in enabling physiological adaptation of organisms to their physical environment, Progress in Lipid Research, 29, 167–227, https://doi.org/10.1016/0163-7827(90)90002-3, 1990.

Hofmann, K., Lamprecht, A., Pauli, H., and Illmer, P.: 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, https://doi.org/10.1007/s00248-016-0803-z, 2016.

973 Hu, C., Henderson, G. M., Huang, J., Xie, S., Sun, Y., and Johnson, K. R.: Quantification of
974 Holocene Asian monsoon rainfall from spatially separated cave records, Earth and Planetary
975 Science Letters, 266, 221–232, https://doi.org/10.1016/j.epsl.2007.10.015, 2008.

976 Huguet, A., Coffinet, S., Roussel, A., Gayraud, F., Anquetil, C., Bergonzini, L., Bonanomi, G.,
977 Williamson, D., Majule, A., and Derenne, S.: Evaluation of 3-hydroxy fatty acids as a pH and
978 temperature proxy in soils from temperate and tropical altitudinal gradients, Organic
979 Geochemistry, 129, 1–13, https://doi.org/10.1016/j.orggeochem.2019.01.002, 2019.

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

984 Kaniewski, D., Campo, E. V., Guiot, J., Burel, S. L., Otto, T., and Baeteman, C.: Environmental
985 Roots of the Late Bronze Age Crisis, PLOS ONE, 8, e71004,
986 https://doi.org/10.1371/journal.pone.0071004, 2013.

987 Knapp, A. B. and Manning, S. W.: Crisis in Context: The End of the Late Bronze Age in the
988 Eastern Mediterranean, 120, 99–149, https://doi.org/10.3764/aja.120.1.0099, 2016.

989 Li, C.-H., Li, Y.-X., Zheng, Y.-F., Yu, S.-Y., Tang, L.-Y., Li, B.-B., and Cui, Q.-Y.: A high-
990 resolution pollen record from East China reveals large climate variability near the
991 Northgrippian-Meghalayan boundary (around 4200 years ago) exerted societal influence,
992 Palaeogeography, Palaeoclimatology, Palaeoecology, 512, 156–165,
993 https://doi.org/10.1016/j.palaeo.2018.07.031, 2018.

994 Liu, Z., Zhu, J., Rosenthal, Y., Zhang, X., Otto-Bliesner, B. L., Timmermann, A., Smith, R. S.,
995 Lohmann, G., Zheng, W., and Elison Timm, O.: The Holocene temperature conundrum, Proc
996 Natl Acad Sci U S A, 111, E3501–E3505, https://doi.org/10.1073/pnas.1407229111, 2014.

997 Ljungqvist, F. C.: A new reconstruction of temperature variability in the extra-tropical northern
998 hemisphere during the last two millennia, 92, 339–351, https://doi.org/10.1111/j.1468-
999 0459.2010.00399.x, 2010.

1000 Loomis, S. E., Russell, J. M., Ladd, B., Street-Perrott, F. A., and Sinninghe Damsté, J. S.:
1001 Calibration and application of the branched GDGT temperature proxy on East African lake
1002 sediments, Earth and Planetary Science Letters, 357–358, 277–288,
1003 https://doi.org/10.1016/j.epsl.2012.09.031, 2012.

1004 Malhi, Y., Silman, M., Salinas, N., Bush, M., Meir, P., and Saatchi, S.: Introduction: Elevation
1005 gradients in the tropics: laboratories for ecosystem ecology and global change research, 16,
1006 3171–3175, https://doi.org/10.1111/j.1365-2486.2010.02323.x, 2010.

1007 Mann, M. E., Zhang, Z., Hughes, M. K., Bradley, R. S., Miller, S. K., Rutherford, S., and Ni,
1008 F.: Proxy-based reconstructions of hemispheric and global surface temperature variations over
1009 the past two millennia, Proceedings of the National Academy of Sciences, 105, 13252–13257,
1010 https://doi.org/10.1073/pnas.0805721105, 2008.

1011 Margesin, R., Jud, M., Tscherko, D., and Schinner, F.: Microbial communities and activities in
1012 alpine and subalpine soils: Communities and activities in alpine and subalpine soils, 67, 208–
1013 218, https://doi.org/10.1111/j.1574-6941.2008.00620.x, 2009.

Mayewski, P. A., Rohling, E. E., Curt Stager, J., Karlén, W., Maasch, K. A., David Meeker, L., Meyerson, E. A., Gasse, F., van Kreveld, S., Holmgren, K., Lee-Thorp, J., Rosqvist, G., Rack, F., Staubwasser, M., Schneider, R. R., and Steig, E. J.: Holocene climate variability, Quaternary Research, 62, 243–255, https://doi.org/10.1016/j.yqres.2004.07.001, 2004.

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

Nottingham, A. T., Whitaker, J., Turner, B. L., Salinas, N., Zimmermann, M., Malhi, Y., and Meir, P.: Climate Warming and Soil Carbon in Tropical Forests: Insights from an Elevation Gradient in the Peruvian Andes, 65, 906–921, https://doi.org/10.1093/biosci/biv109, 2015.

Nottingham, A. T., Fierer, N., Turner, B. L., Whitaker, J., Ostle, N. J., McNamara, N. P., Bardgett, R. D., Leff, J. W., Salinas, N., Silman, M. R., Kruuk, L. E. B., and Meir, P.: Microbes follow Humboldt: temperature drives plant and soil microbial diversity patterns from the Amazon to the Andes, 99, 2455–2466, https://doi.org/10.1002/ecy.2482, 2018.

Peterse, F., Kim, J.-H., Schouten, S., Kristensen, D. K., Koç, N., and Sinninghe Damsté, J. S.: Constraints on the application of the MBT/CBT palaeothermometer at high latitude environments (Svalbard, Norway), Organic Geochemistry, 40, 692–699, https://doi.org/10.1016/j.orggeochem.2009.03.004, 2009.

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.: Revised calibration of the MBT–CBT paleotemperature proxy based on branched tetraether membrane lipids in surface soils, Geochimica et Cosmochimica Acta, 96, 215–229, https://doi.org/10.1016/j.gca.2012.08.011, 2012.

Peterse, F., Moy, C. M., and Eglinton, T. I.: A laboratory experiment on the behaviour of soil-derived core and intact polar GDGTs in aquatic environments, Biogeosciences, 12, 933–943, https://doi.org/10.5194/bg-12-933-2015, 2015.

R Core Team, R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, 2014.

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

Russell, N. J., Evans, R. I., ter Steeg, P. F., Hellemons, J., Verheul, A., and Abee, T.: Membranes as a target for stress adaptation, International Journal of Food Microbiology, 28, 255–261, https://doi.org/10.1016/0168-1605(95)00061-5, 1995.

Scalercio, S., Bonacci, T., Mazzei, A., Pizzolotto, R., and Brandmayr, P.: Better up, worse down: bidirectional consequences of three decades of climate change on a relict population of Erebia cassioides, J Insect Conserv, 18, 643–650, https://doi.org/10.1007/s10841-014-9669-x, 2014.

1053 Schouten, S., Hopmans, E. C., and Sinninghe Damsté, J. S.: The organic geochemistry of
1054 glycerol dialkyl glycerol tetraether lipids: A review, Organic Geochemistry, 54, 19–61,
1055 https://doi.org/10.1016/j.orggeochem.2012.09.006, 2013.

1056 Siles, J. A. and Margesin, R.: Abundance and Diversity of Bacterial, Archaeal, and Fungal
1057 Communities Along an Altitudinal Gradient in Alpine Forest Soils: What Are the Driving
1058 Factors?, Microb Ecol, 72, 207–220, https://doi.org/10.1007/s00248-016-0748-2, 2016.

1059 Sinensky, M.: Homeoviscous Adaptation—A Homeostatic Process that Regulates the Viscosity
1060 of Membrane Lipids in Escherichia coli, PNAS, 71, 522–525,
1061 https://doi.org/10.1073/pnas.71.2.522, 1974.

1062 Singer, S. J. and Nicolson, G. L.: The Fluid Mosaic Model of the Structure of Cell Membranes,
1063 175, 720–731, https://doi.org/10.1126/science.175.4023.720, 1972.

1064 Sinninghe Damsté, J. S., Rijpstra, W. I. C., Hopmans, E. C., Weijers, J. W. H., Foesel, B. U.,
1065 Overmann, J. and Dedysh, S. N.: 13,16-Dimethyl Octacosanedioic Acid (iso-Diabolic Acid), a
1066 Common Membrane-Spanning Lipid of Acidobacteria Subdivisions 1 and 3, Appl. Environ.
1067 Microbiol., 77(12), 4147–4154, doi:10.1128/AEM.00466-11, 2011.

1068 Sinninghe Damsté, J. S., Rijpstra, W. I. C., Hopmans, E. C., Foesel, B. U., Wüst, P. K.,
1069 Overmann, J., Tank, M., Bryant, D. A., Dunfield, P. F., Houghton, K. and Stott, M. B.: Ether-
1070 and Ester-Bound iso-Diabolic Acid and Other Lipids in Members of Acidobacteria Subdivision
1071 4, Appl. Environ. Microbiol., 80(17), 5207–5218, doi:10.1128/AEM.01066-14, 2014.

1072 Sinninghe Damste, J. S. S., Rijpstra, W. I. C., Foesel, B. U., Huber, K. J., Overmann, J.,
1073 Nakagawa, S., Kim, J. J., Dunfield, P. F., Dedysh, S. N. and Villanueva, L.: An overview of the
1074 occurrence of ether- and ester-linked iso-diabolic acid membrane lipids in microbial cultures of
1075 the Acidobacteria: Implications for brGDGT paleoproxies for temperature and pH, Org.
1076 Geochem., 124, 63–76, doi:10.1016/j.orggeochem.2018.07.006, 2018.

1077 Staubwasser, M., Sirocko, F., Grootes, P. M., and Segl, M.: Climate change at the 4.2 ka BP
1078 termination of the Indus valley civilization and Holocene south Asian monsoon variability,
1079 Geophysical Research Letters, 30, 1425, https://doi.org/10.1029/2002GL016822, 2003.

1080 Steinhilber, F., Beer, J., and Fröhlich, C.: Total solar irradiance during the Holocene, 36,
1081 https://doi.org/10.1029/2009GL040142, 2009.

1082 Szponar, B., Kraśnik, L., Hryniewiecki, T., Gamian, A., and Larsson, L.: Distribution of 3-
1083 Hydroxy Fatty Acids in Tissues after Intraperitoneal Injection of Endotoxin, Clin Chem, 49,
1084 1149–1153, https://doi.org/10.1373/49.7.1149, 2003.

1085 Tierney, J. E. and Russell, J. M.: Distributions of branched GDGTs in a tropical lake system:
1086 Implications for lacustrine application of the MBT/CBT paleoproxy, Organic Geochemistry,
1087 40, 1032–1036, https://doi.org/10.1016/j.orggeochem.2009.04.014, 2009.

1088 Tierney, J. E. and Tingley, M. P.: A Bayesian, spatially-varying calibration model for the
1089 TEX86 proxy, Geochimica et Cosmochimica Acta, 127, 83–106,
1090 https://doi.org/10.1016/j.gca.2013.11.026, 2014.

1091 Todaro, L., Andreu-Hayles, L., D'Alessandro, C., Gutiérrez, E., Cherubini, P., and Saracino,
1092 A.: Response of Pinus leucodermis to climate and anthropogenic activity in the National Park

1093 of Pollino (Basilicata, Southern Italy), Biological Conservation, 137, 507–519,
1094 https://doi.org/10.1016/j.biocon.2007.03.010, 2007.

1095 Véquaud, P., Derenne, S., Anquetil, C., Collin, S., Poulenard, J., Sabatier, P., and Huguet, A.:
1096 Influence of environmental parameters on the distribution of bacterial lipids in soils from the
1097 French Alps: Implications for paleo-reconstructions, Organic Geochemistry, 153, 104194,
1098 https://doi.org/10.1016/j.orggeochem.2021.104194, 2021.

1099 Wakeham, S. G., Pease, T. K., and Benner, R.: Hydroxy fatty acids in marine dissolved organic
1100 matter as indicators of bacterial membrane material, Organic Geochemistry, 34, 857–868,
1101 https://doi.org/10.1016/S0146-6380(02)00189-4, 2003.

1102 Wang, C., Bendle, J., Yang, Y., Yang, H., Sun, H., Huang, J., and Xie, S.: Impacts of pH and
1103 temperature on soil bacterial 3-hydroxy fatty acids: Development of novel terrestrial proxies,
1104 Organic Geochemistry, 94, 21–31, https://doi.org/10.1016/j.orggeochem.2016.01.010, 2016.

1105 Wang, C., Bendle, J. A., Zhang, H., Yang, Y., Liu, D., Huang, J., Cui, J., and Xie, S.: Holocene
1106 temperature and hydrological changes reconstructed by bacterial 3-hydroxy fatty acids in a
1107 stalagmite from central China, Quaternary Science Reviews, 192, 97–105,
1108 https://doi.org/10.1016/j.quascirev.2018.05.030, 2018.

1109 Wang, J.-T., Cao, P., Hu, H.-W., Li, J., Han, L.-L., Zhang, L.-M., Zheng, Y.-M., and He, J.-Z.:
1110 Altitudinal Distribution Patterns of Soil Bacterial and Archaeal Communities Along Mt.
1111 Shegyla on the Tibetan Plateau, Microb Ecol, 69, 135–145, https://doi.org/10.1007/s00248-
1112 014-0465-7, 2015.

1113 Weber, Y., De Jonge, C., Rijpstra, W. I. C., Hopmans, E. C., Stadnitskaia, A., Schubert, C. J.,
1114 Lehmann, M. F., Sinninghe Damsté, J. S., and Niemann, H.: Identification and carbon isotope
1115 composition of a novel branched GDGT isomer in lake sediments: Evidence for lacustrine
1116 branched GDGT production, Geochimica et Cosmochimica Acta, 154, 118–129,
1117 https://doi.org/10.1016/j.gca.2015.01.032, 2015.

1118 Weijers, J. W. H., Schouten, S., van den Donker, J. C., Hopmans, E. C., and Sinninghe Damsté,
1119 J. S.: Environmental controls on bacterial tetraether membrane lipid distribution in soils,
1120 Geochimica et Cosmochimica Acta, 71, 703–713, https://doi.org/10.1016/j.gca.2006.10.003,
1121 2007.

1122 Weiss, B.: The decline of Late Bronze Age civilization as a possible response to climatic
1123 change, Climatic Change, 4, 173–198, https://doi.org/10.1007/BF00140587, 1982.

1124 Whitaker, J., Ostle, N., Nottingham, A. T., Ccahuana, A., Salinas, N., Bardgett, R. D., Meir, P.,
1125 and McNamara, N. P.: Microbial community composition explains soil respiration responses to
1126 changing carbon inputs along an Andes-to-Amazon elevation gradient, 102, 1058–1071,
1127 https://doi.org/10.1111/1365-2745.12247, 2014.

1128 Wollenweber, H. W. and Rietschel, E. T.: Analysis of lipopolysaccharide (lipid A) fatty acids.,
1129 11, 195–211, 1990.

1130 Wollenweber, H.-W., Broady, K. W., Luderitz, O., and Rietschel, E. T.: The Chemical Structure
1131 of Lipid A, 124, 191–198, https://doi.org/10.1111/j.1432-1033.1982.tb05924.x, 1982.

1132 Yang, H., Lü, X., Ding, W., Lei, Y., Dang, X., and Xie, S.: The 6-methyl branched tetraethers
1133 significantly affect the performance of the methylation index (MBT′) in soils from an altitudinal
1134 transect at Mount Shennongjia, Organic Geochemistry, 82, 42–53,
1135 https://doi.org/10.1016/j.orggeochem.2015.02.003, 2015.

1136 Zelles, L.: Fatty acid patterns of phospholipids and lipopolysaccharides in the characterisation
1137 of microbial communities in soil: a review, Biol Fertil Soils, 29, 111–129,
1138 https://doi.org/10.1007/s003740050533, 1999.

1139
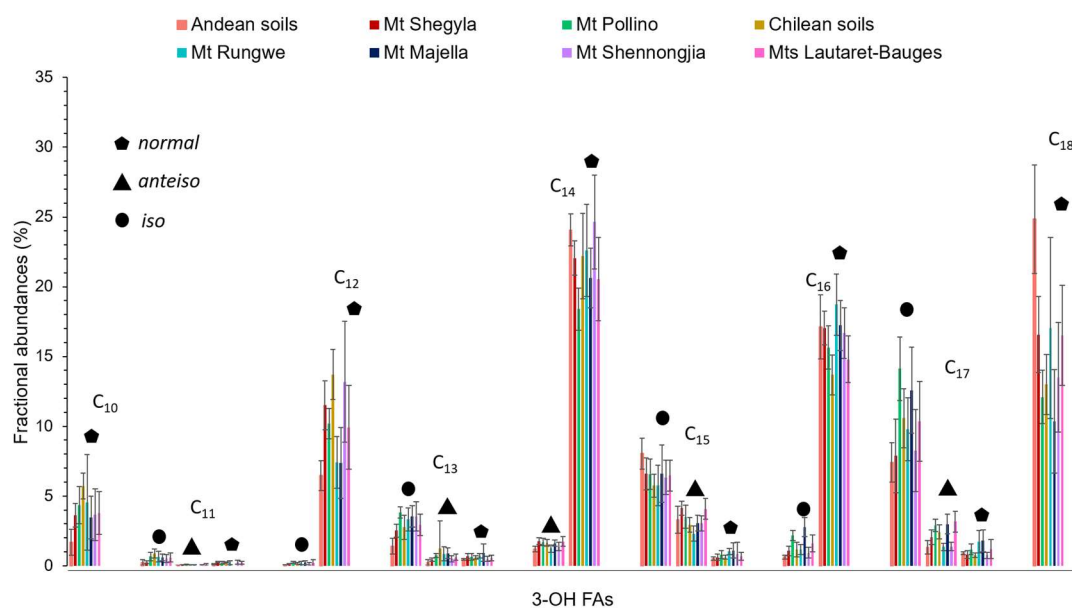
Figure 1. Average distribution of 3-OH FAs along the 8 altitudinal transects investigated in this study. Data from Mts. Majella and Rungwe were taken from Huguet et al. (2019). Data from Mt. Shennongjia were taken from Wang et al. (2016). Data from Mts. Lautaret-Galibier were taken from Véquaud et al. (2021).
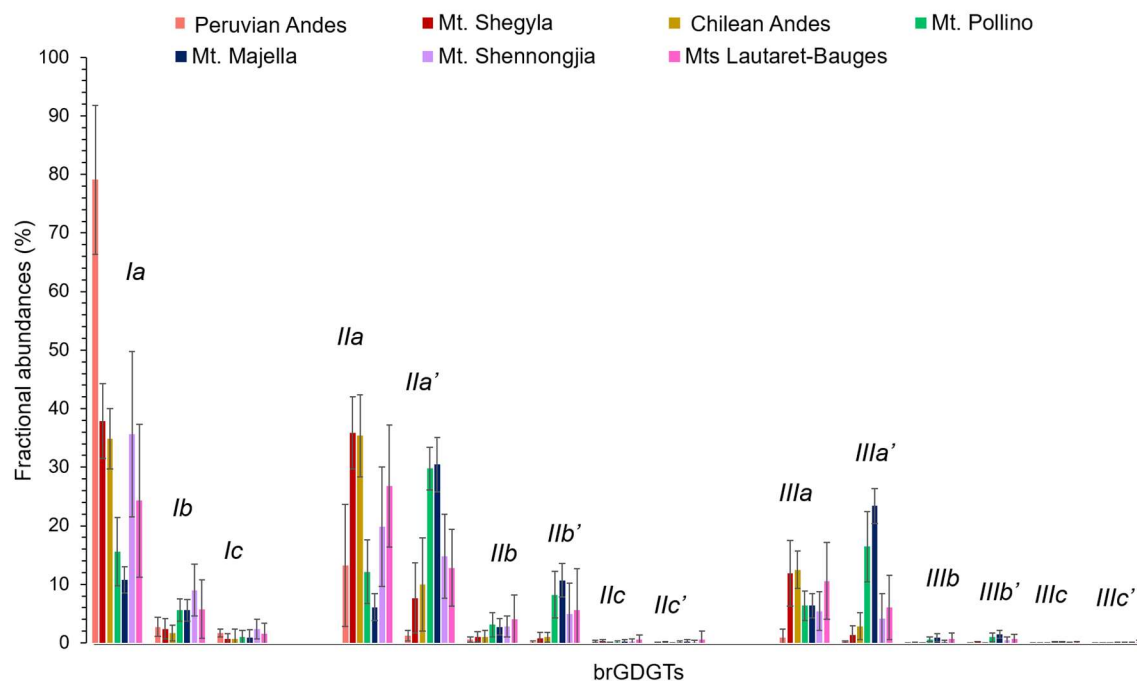
Figure 2. Average distribution of 5- and 6-methyl brGDGTs, along Mts. Shegyla, Pollino Majella, Lautaret-Bauges, Peruvian Andes and Chilean Andes. Data from Mt. Majella were taken from Huguet et al. (2019). Data from Mt. Shennongjia were taken from Yang et al. (2015). Data from Mts. Lautaret-Galibier were taken from Véquaud et al. (2021).
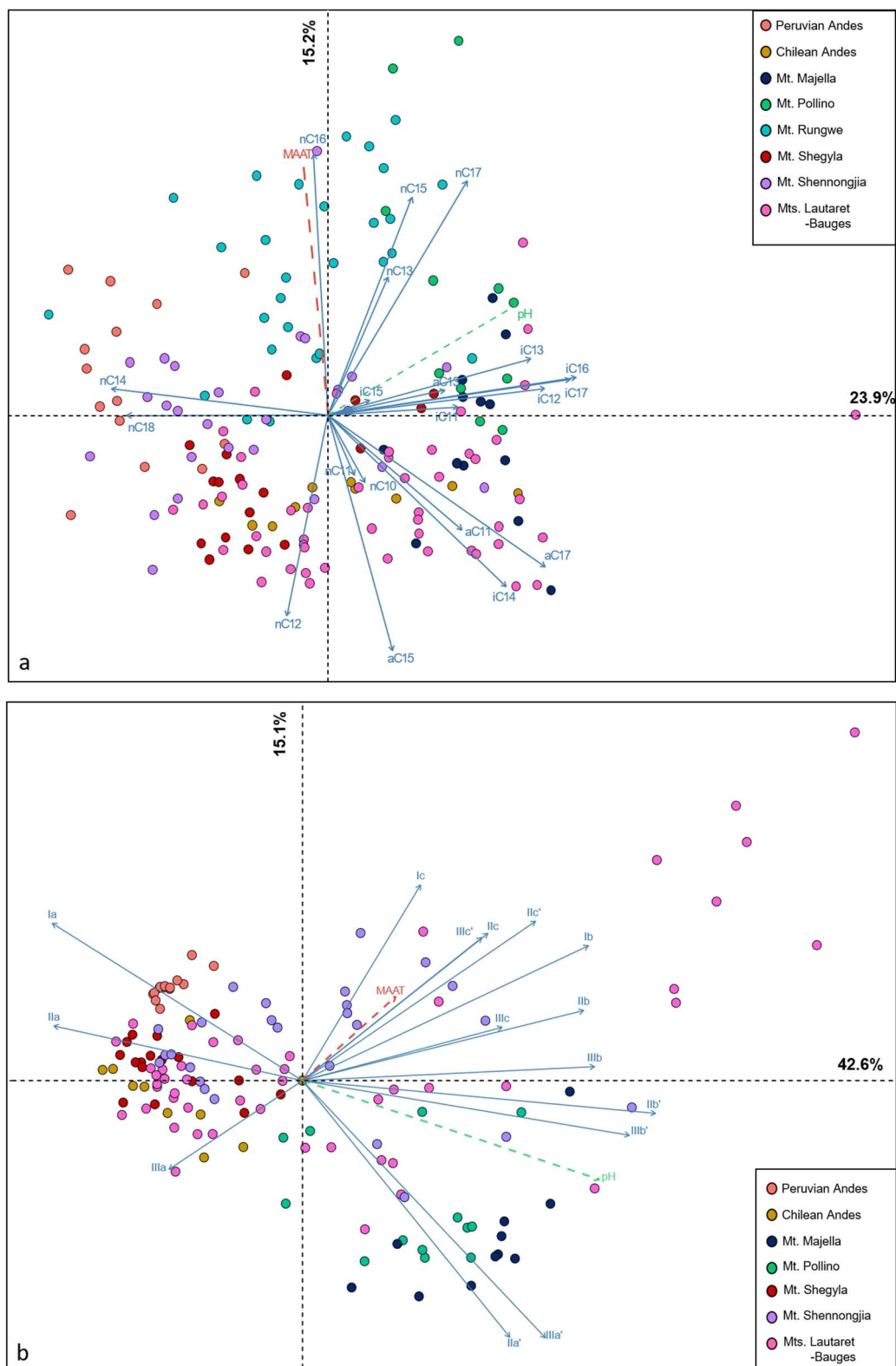
Figure 3. PCA biplot of (a) 3-OH FA fractional abundances in soil samples from the 8 altitudinal transects and (b) brGDGT fractional abundances in soil samples from 7 of the 8 altitudinal transects. BrGDGT data from Mt. Rungwe, for which 5- and 6-methyl isomers were not separated, were not included in the PCA.
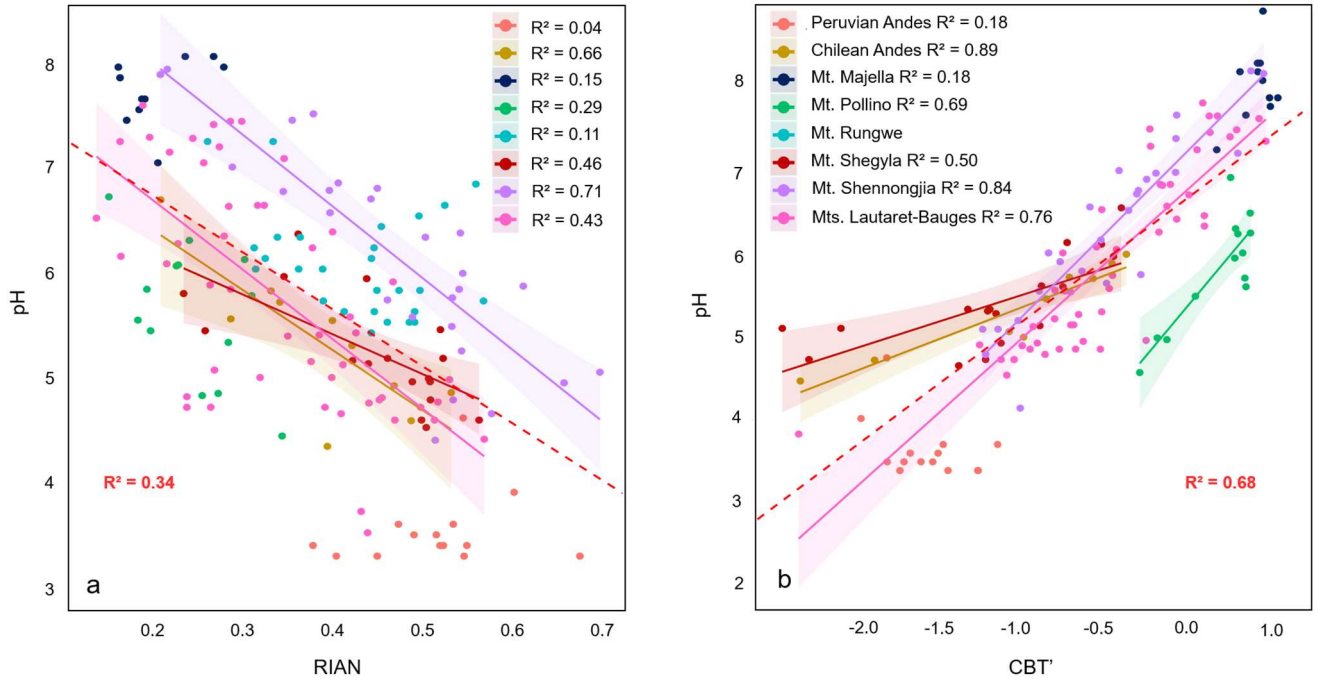
Figure 4. Linear regressions between (a) pH and RIAN and (b) pH and CBT' along the 8 altitudinal transects investigated. Dotted lines represent the 95% confidence interval for each regression and colored areas represent the 95% confidence interval for each regression. Data for Mts. Majella and Rungwe were taken from Huguet et al. (2019). Data from Mt. Shennongjia were taken from Yang et al. (2015) and Wang et al. (2016). Data from Mts. Lautaret-Galibier were taken from Véquaud et al. (2021). Only significant regressions (p < 0.05) are shown.
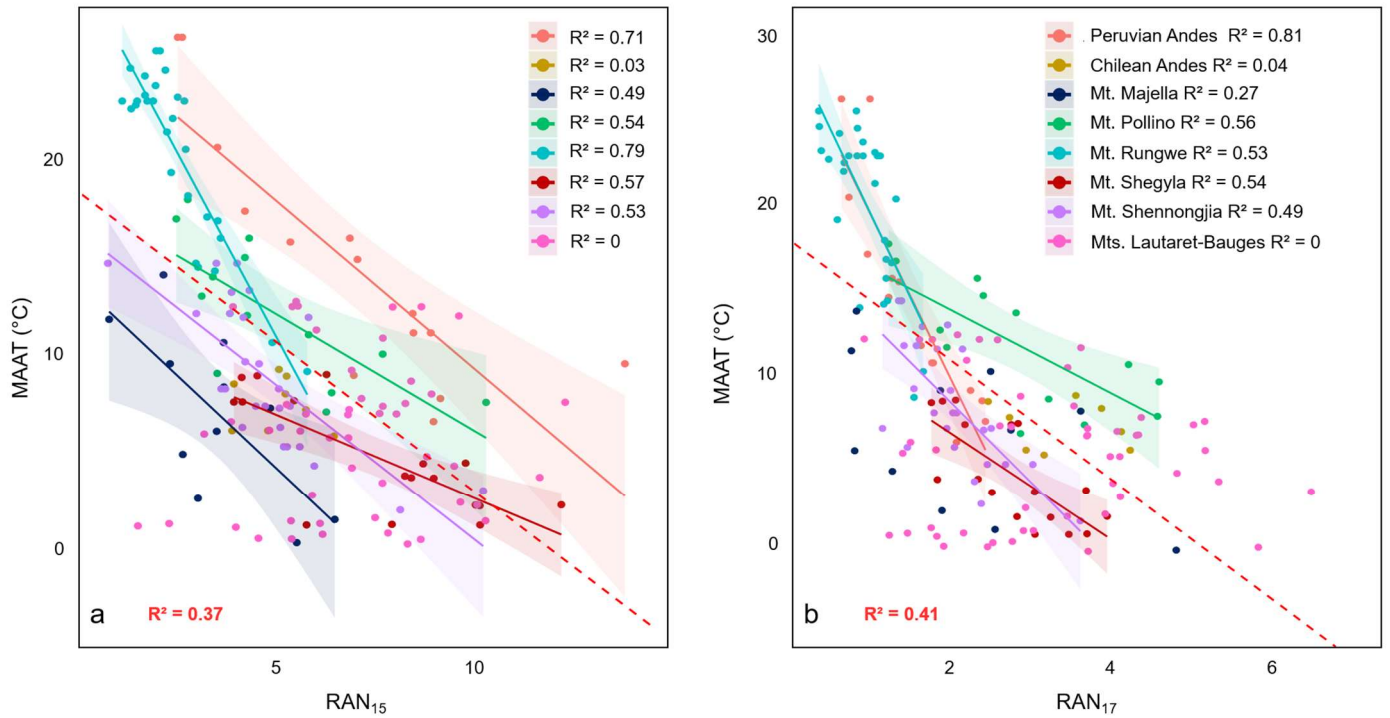
Figure 5. Linear regressions between (a) MAAT and RAN$_{15}$ and (b) MAAT and RAN$_{17}$ along the 8 altitudinal transects investigated. Dotted lines represent the 95% confidence interval for each regression and colored areas represent the 95% confidence interval for each regression. Data from Mts. Majella and Rungwe were taken from Huguet et al. (2019). Data from Mt. Shennongjia were taken from Wang et al. (2016). Data from Mts. Lautaret-Galibier were taken from Véquaud et al. (2021). Only significant regressions (p < 0.05) are shown.
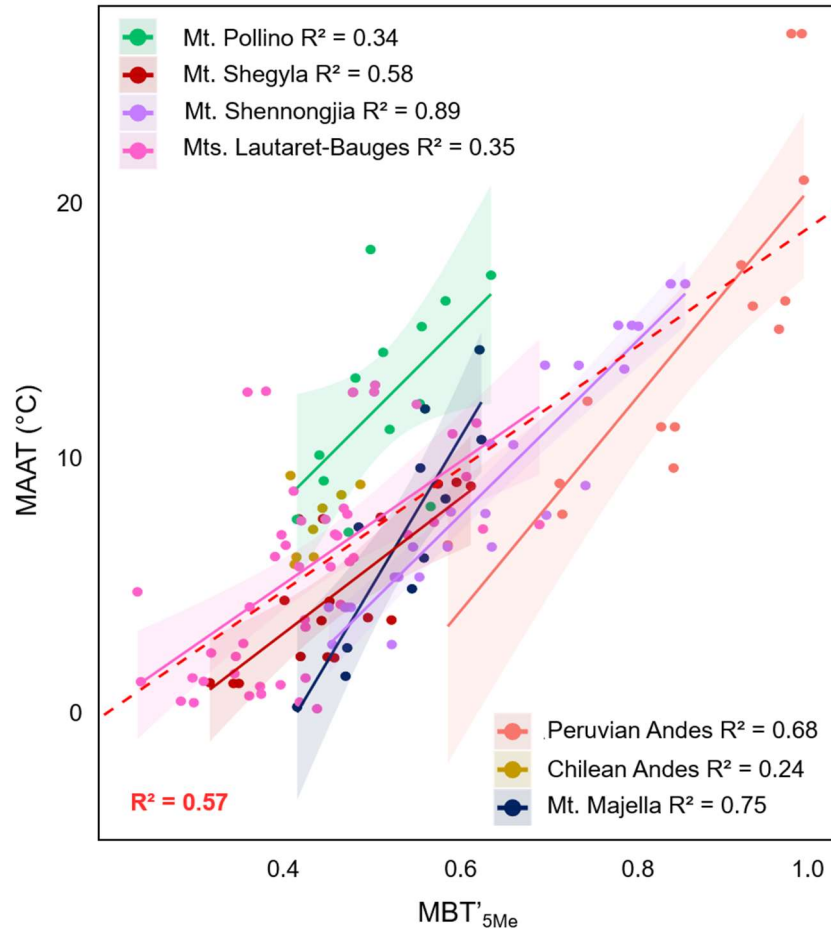
Figure 6. Linear regressions between (a) MAAT and MBT'$_{5Me}$ along 7 of the 8 altitudinal transects investigated. Data from Mt. Rungwe (Coffinet et al., 2014), for which 5- and 6-methyl brGDGTs were not separated, were not included in this graph. Dotted lines represent the 95% confidence interval for each regression and colored areas represent the 95% confidence interval for each regression. Data from Mt. Majella were taken from Huguet et al. (2019). Data from Mts. Lautaret-Galibier were taken from Véquaud et al. (2021). Data from Mt. Shennongjia were taken from Yang et al. (2015). The global soil calibration by De Jonge et al. (2014) was applied to all these transects. Only significant regressions (p < 0.05) are shown.
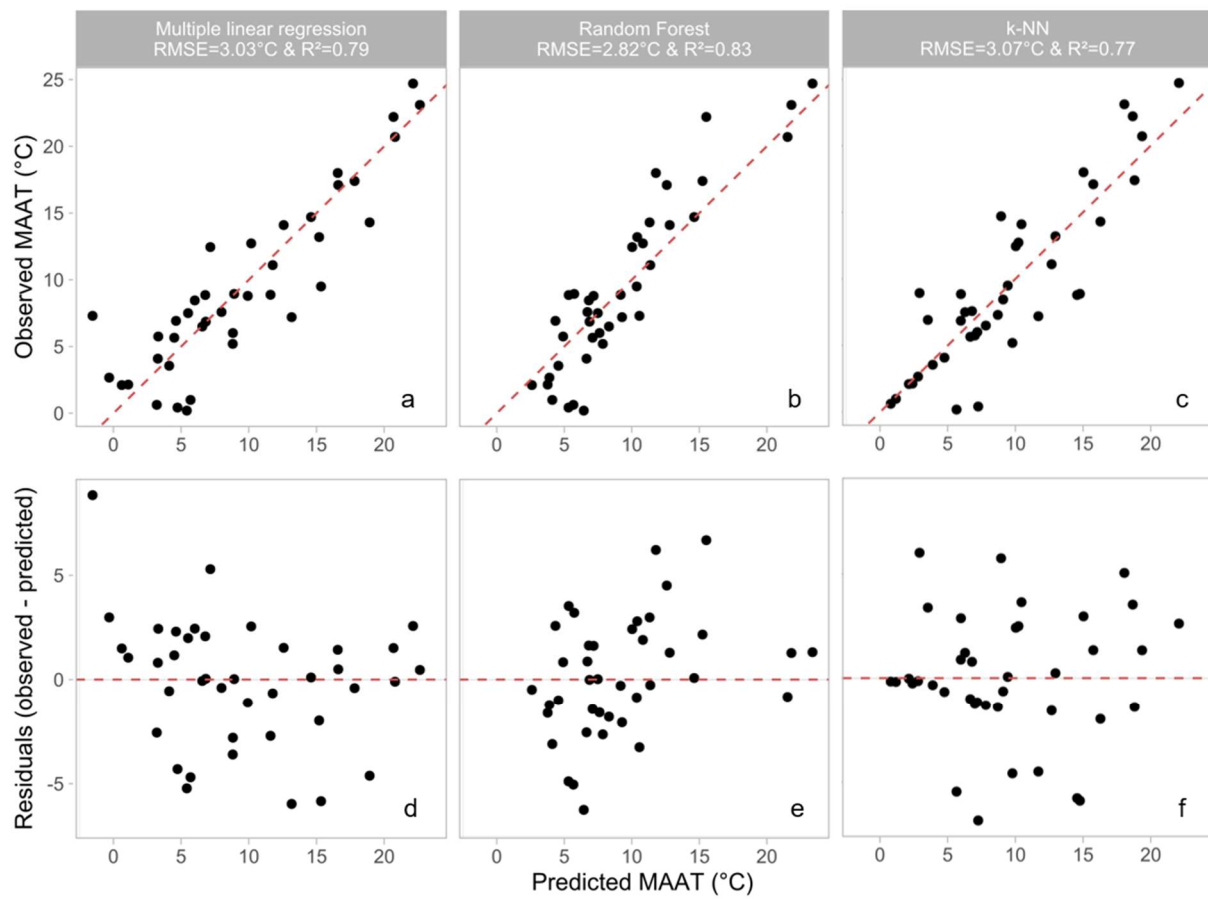
Figure 7. Results of the three different models tested to reconstruct the MAAT from 3-OH FA distribution: observed MAAT (°C) vs Predicted MAAT (°C) for (a) the multiple linear regression model, (b) the random forest model and (c) the k-NN method. MAAT residuals plotted against the predicted MAAT for (d) the multiple linear regression model, (e) the random forest model and (f) the k-NN method.
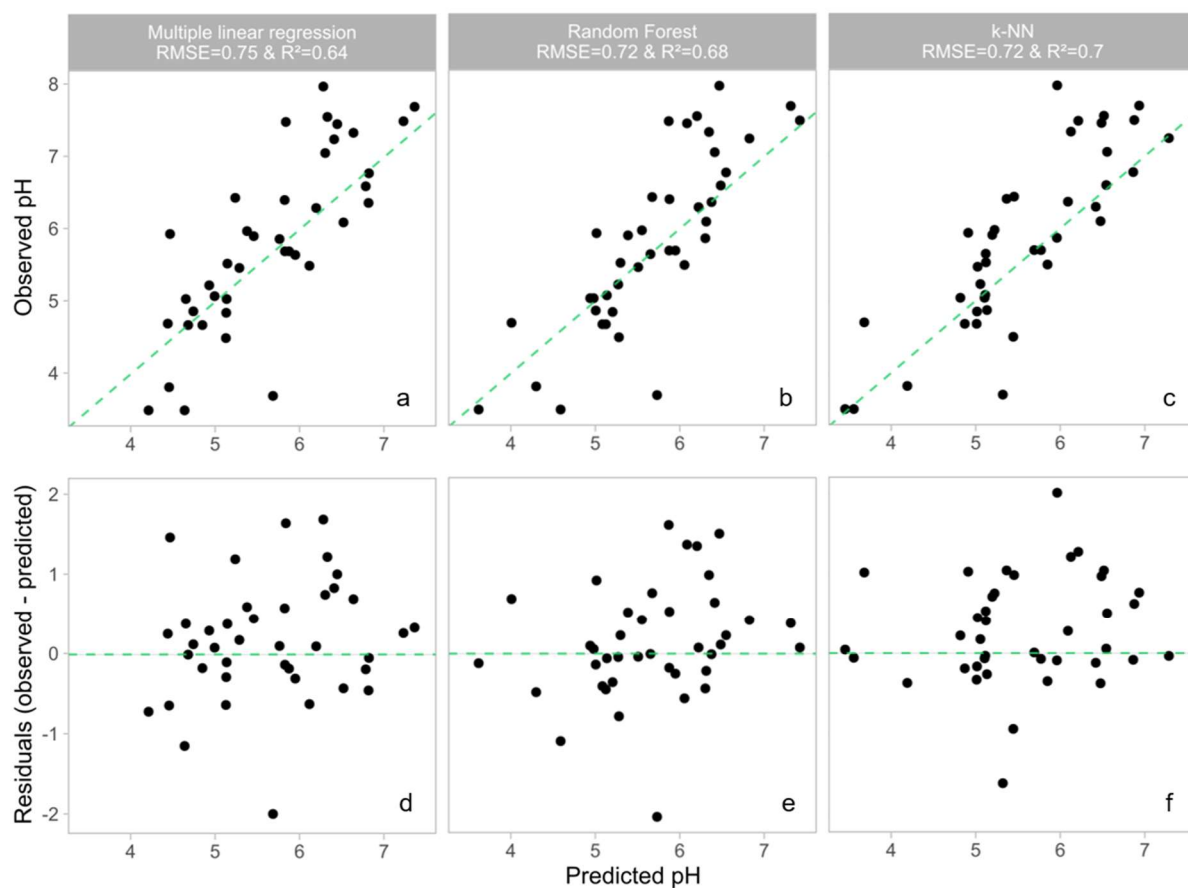
Figure 8. Results of the three different models tested to reconstruct the pH from 3-OH FA distribution: observed pH vs predicted pH for (a) the multiple linear regression model, (b) the random forest model, (c) the k-NN method. pH residuals plotted against the predicted pH for (d) the multiple linear regression model, (e) the random forest model and (f) the k-NN method.
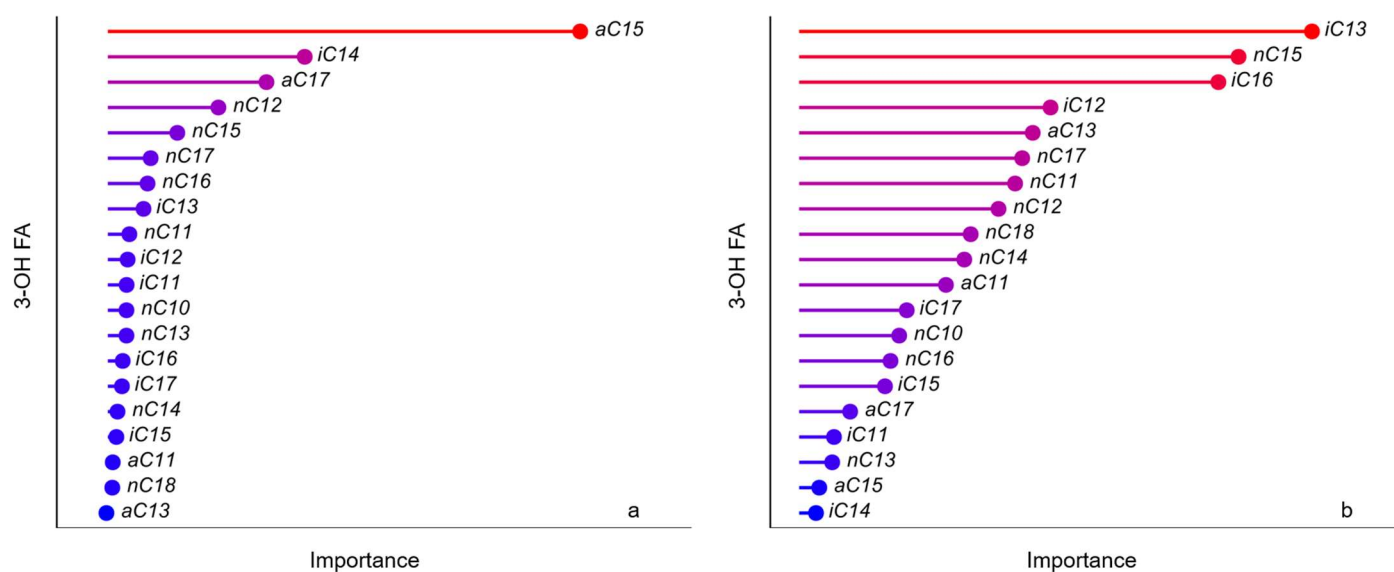
Figure 9. Importance (arbitrary unit) of the 3-OH FAs used to estimate (a) MAAT and (b) pH in the random forest models proposed in this study according to the permutation importance method (Breiman, 2001).
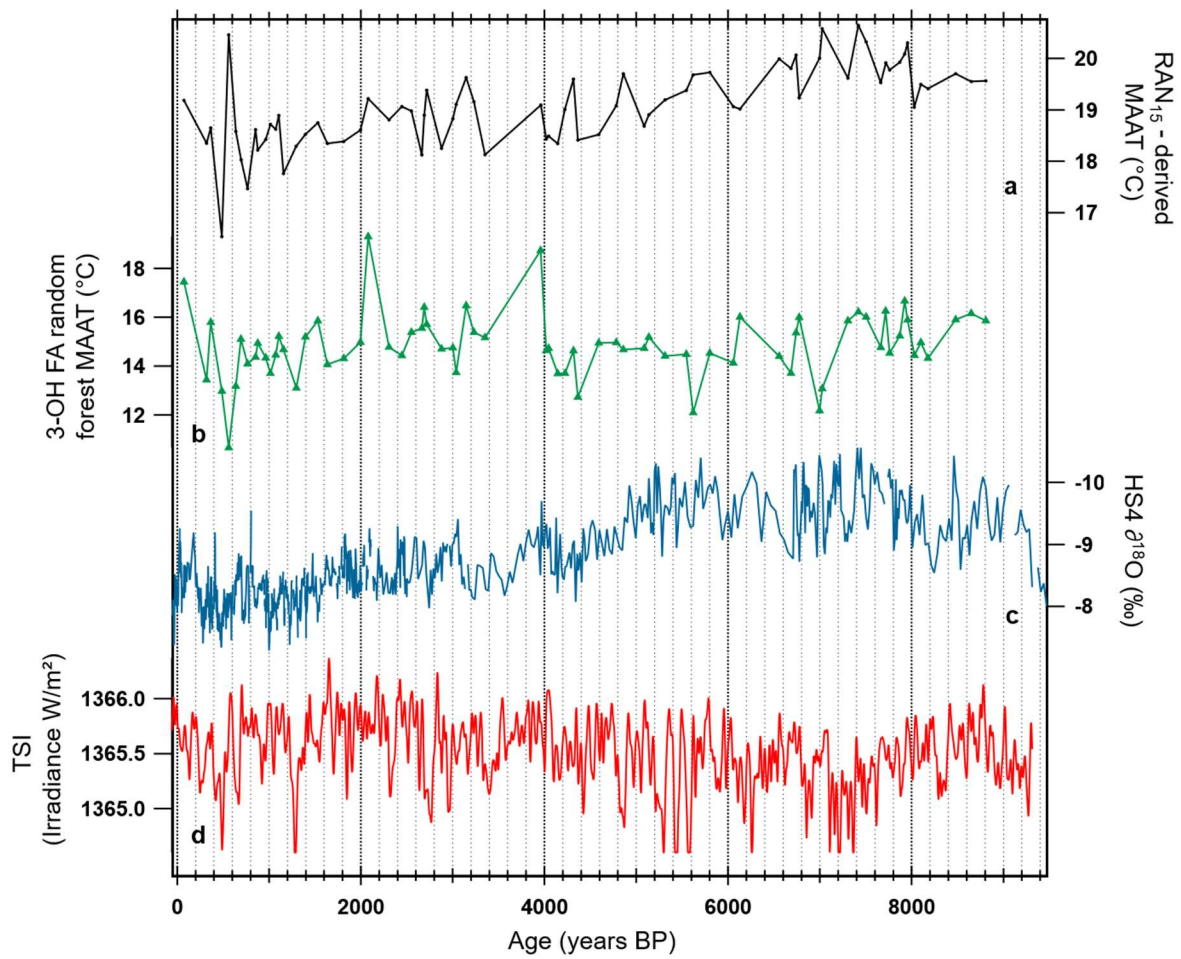
Figure 10. Comparison of the 3-OH FA model-MAAT record with other time-series and proxy records for the HS4 speleothem (Wang et al., 2018). (a) RAN$_{15}$-MAAT record reconstructed using a local Chinese calibration (Wang et al., 2016; Wang et al., 2018). (b) 3-OH FA random forest model-MAAT. (d) The CaCO$_3$ oxygen isotope record (Hu et al., 2008b). (e) Total solar irradiance (TSI; W/m²) during the Holocene (past 9300 years) based on a composite described in Steinhilber et al. (2009).

ID	Location	Altitude (m)	MAAT(°C)	pH	RAN$_{15}$	RAN$_{17}$	RIAN	MBT'$_{5Me}$	CBT'
1	Peruvian Andes	194	26.4	3.7	2.45	0.96	0.47	0.96	-1.09
2	Peruvian Andes	210	26.4	4	2.56	0.61	0.60	0.97	-1.92
3	Peruvian Andes	1063	20.7	4.7	3.46	0.70	0.54	0.98	-1.76
4	Peruvian Andes	1500	17.4	3.5	4.15	0.93	0.51	0.91	-1.55
5	Peruvian Andes	1750	15.8	3.6	5.30	1.32	0.51	0.92	-1.62
6	Peruvian Andes	1850	16	3.5	6.81	1.23	0.54	0.96	-1.76
7	Peruvian Andes	2020	14.9	3.4	7.00	1.19	0.54	0.95	-1.68
8	Peruvian Andes	2520	12.1	3.7	8.40	1.59	0.53	0.74	-1.42
9	Peruvian Andes	2720	11.1	3.6	8.42	1.73	0.48	0.83	-1.45
10	Peruvian Andes	3020	9.5	3.4	13.78	2.21	0.44	0.83	-1.21
11	Peruvian Andes	3200	8.9	3.5	6.91	2.35	0.37	0.71	-1.48
12	Peruvian Andes	3025	11.1	3.5	8.86	1.74	0.52	0.82	-1.66
13	Peruvian Andes	3400	7.7	3.4	9.10	2.39	0.40	0.71	-1.39
14	Peruvian Andes	3644	6.5	3.4	8.93	2.03	0.67	0.58	-1.21
15	Mt. Shegyla, Tibet	3106	8.9	5.53	6.22	2.02	0.51	0.59	-0.83
16	Mt. Shegyla, Tibet	3117	8.9	6.43	4.47	1.86	0.36	0.57	-0.35
17	Mt. Shegyla, Tibet	3132	8.8	6.01	4.07	1.72	0.43	0.61	-0.47
18	Mt. Shegyla, Tibet	3344	7.6	6.03	5.40	2.80	0.34	0.51	-0.67
19	Mt. Shegyla, Tibet	3355	7.5	5.87	4.09	2.71	0.23	0.44	-0.39
20	Mt. Shegyla, Tibet	3356	7.5	5.52	3.87	2.14	0.25	0.42	-0.70
21	Mt. Shegyla, Tibet	4030	3.7	5.21	8.21	3.64	0.43	0.49	-1.10
22	Mt. Shegyla, Tibet	4046	3.6	4.68	8.37	3.00	0.49	0.52	-1.17
23	Mt. Shegyla, Tibet	4050	3.6	4.61	8.94	2.47	0.50	0.44	-1.33
24	Mt. Shegyla, Tibet	3912	4.3	5.04	9.74	2.30	0.48	0.40	-2.39
25	Mt. Shegyla, Tibet	3918	4.3	4.68	8.67	1.80	0.56	0.45	-2.23
26	Mt. Shegyla, Tibet	4298	2.1	5.04	10.00	2.78	0.50	0.45	-2.04
27	Mt. Shegyla, Tibet	4295	2.2	4.87	12.17	3.90	0.50	0.42	-1.07
28	Mt. Shegyla, Tibet	4304	2.1	5.26	10.10	3.20	0.46	0.46	-1.14
29	Mt. Shegyla, Tibet	4479	1.1	5.26	10.11	3.42	0.52	0.35	-1.27
30	Mt. Shegyla, Tibet	4479	1.1	5.07	5.71	3.00	0.50	0.35	-0.84
31	Mt. Shegyla, Tibet	4474	1.1	5.24	7.88	3.65	0.42	0.32	-1.15
32	Mt. Pollino, Italy	0	18	6.78	2.71	1.19	0.15	0.50	0.31
33	Mt. Pollino, Italy	200	17	6.19	2.41	1.28	0.30	0.63	0.34
34	Mt. Pollino, Italy	400	16	6.13	4.26	2.29	0.22	0.58	0.35
35	Mt. Pollino, Italy	600	15	6.14	4.15	2.36	0.22	0.55	0.43
36	Mt. Pollino, Italy	800	14	4.53	3.34	2.77	0.34	0.51	-0.24
37	Mt. Pollino, Italy	1000	13	5.41	3.06	1.83	0.28	0.48	0.10
38	Mt. Pollino, Italy	1200	12	6.37	4.21	1.91	0.24	0.55	0.43
39	Mt. Pollino, Italy	1400	11	5.62	5.77	4.16	0.18	0.52	0.40
40	Mt. Pollino, Italy	1600	10	4.93	7.64	4.54	0.27	0.44	-0.13
41	Mt. Pollino, Italy	1800	9	4.91	3.45	3.17	0.25	0.45	-0.07
42	Mt. Pollino, Italy	2000	8	5.52	6.35	4.52	0.19	0.56	0.40
43	Mt. Pollino, Italy	2100	7.5	5.91	10.26	3.62	0.19	0.42	0.38
44	Mt. Pollino, Italy	2200	7	5.85	6.21	2.82	0.31	0.47	0.34
45	Chilean Andes	690	9.2	5.38	5.01	3.51	0.42	0.41	-0.80
46	Chilean Andes	870	8.9	5.62	5.21	2.43	0.39	0.49	-0.52
47	Chilean Andes	891	7.9	4.94	5.18	2.69	0.53	0.44	-0.94
48	Chilean Andes	915	NA	6.75	4.67	4.25	0.21	NA	NA
49	Chilean Andes	980	8.5	5.63	3.87	3.83	0.28	0.46	-0.66
50	Chilean Andes	985	5.8	4.67	6.41	3.12	0.48	0.41	-1.83
51	Chilean Andes	1125	6.0	5.00	3.83	4.18	0.46	0.42	-1.02
52	Chilean Andes	1151	6.0	5.89	4.74	2.89	0.33	0.43	-0.32
53	Chilean Andes	1196	7.1	5.79	5.70	4.07	0.34	0.43	-0.40
54	Chilean Andes	1385	NA	4.43	4.85	1.91	0.39	0.41	-2.28

Table 1. List of the soil samples collected along Mts. Shegyla, Pollino, Peruvian Andes and Chilean Andes, with corresponding altitude (m), MAAT (°C), pH and 3-OH FA/brGDGT-derived indices.

	Model	n (training)	n (test)	R²	RMSE	Variance in residuals	Mean absolute error	Lower estimation limit	Upper estimation limit
MAAT (°C)	RAN$_{15}$	-	168	0.37	5.5	29.8	4.0	-3.1	17.2
	RAN$_{17}$	-	168	0.41	5.3	27.9	3.9	-4.3	17.0
	k-NN	128	40	0.77	3.1	9.4	2.3	0.5	25.0
	Multiple linear regression	128	40	0.79	3.0	9.2	2.3	-1.2	25.8
	Random forest	128	40	0.83	2.8	8.0	2.2	0.8	24.9
pH	RIAN	-	168	0.34	1.0	1.0	0.8	4.1	7.9
	k-NN	128	40	0.70	0.7	0.5	0.5	3.4	8.7
	Multiple linear regression	128	40	0.64	0.8	0.6	0.6	4.0	8.3
	Random forest	128	40	0.68	0.7	0.5	0.5	3.5	7.8

Table 2. Characteristics of the different models proposed in this study to estimate MAAT and pH: R², RMSE, variance of the residuals, mean absolute error (MAE) and the upper and lower limits of estimation. The "training" samples were used to develop the different machine learning models, which were then tested on a "test" sample set.