



OPTiMAL: a new machine learning approach for GDGT-based palaeothermometry

Tom Dunkley Jones¹, Yvette L. Eley¹, William Thomson², Sarah E. Greene¹, Ilya Mandel^{3,4,5}, Kirsty Edgar¹, and James A. Bendle¹

¹School of Geography, Earth and Environmental Sciences, University of Birmingham, Edgbaston, B15 2TT, UK

²School of Mathematics, University of Birmingham, Edgbaston, B15 2TT, UK

³School of Physics and Astronomy, Monash University, Clayton, Vic. 3800, Australia

⁴The ARC Centre of Excellence for Gravitational Wave Discovery – OzGrav, Hawthorn, Australia

⁵Birmingham Institute for Gravitational Wave Astronomy, School of Physics and Astronomy, University of Birmingham, B15 2TT, Birmingham, UK

Correspondence: Tom Dunkley Jones (t.dunkleyjones@bham.ac.uk)

Received: 21 May 2019 – Discussion started: 6 June 2019

Revised: 16 November 2020 – Accepted: 18 November 2020 – Published: 23 December 2020

Abstract. In the modern oceans, the relative abundances of glycerol dialkyl glycerol tetraether (GDGT) compounds produced by marine archaeal communities show a significant dependence on the local sea surface temperature at the site of deposition. When preserved in ancient marine sediments, the measured abundances of these fossil lipid biomarkers thus have the potential to provide a geological record of long-term variability in planetary surface temperatures. Several empirical calibrations have been made between observed GDGT relative abundances in late Holocene core-top sediments and modern upper ocean temperatures. These calibrations form the basis of the widely used TEX₈₆ palaeothermometer. There are, however, two outstanding problems with this approach: first the appropriate assignment of uncertainty to estimates of ancient sea surface temperatures based on the relationship of the ancient GDGT assemblage to the modern calibration dataset, and second, the problem of making temperature estimates beyond the range of the modern empirical calibrations (> 30 °C). Here we apply modern machine learning tools, including Gaussian process emulators and forward modelling, to develop a new mathematical approach we call OPTiMAL (Optimised Palaeothermometry from Tetraethers via MACHine Learning) to improve temperature estimation and the representation of uncertainty based on the relationship between ancient GDGT assemblage data and the structure of the modern calibration dataset. We reduce the root mean square uncertainty on temperature predic-

tions (validated using the modern dataset) from $\sim \pm 6$ °C using TEX₈₆-based estimators to ± 3.6 °C using Gaussian process estimators for temperatures below 30 °C. We also provide a new quantitative measure of the distance between an ancient GDGT assemblage and the nearest neighbour within the modern calibration dataset, as a test for significant non-analogue behaviour.

1 Introduction

Glycerol dibiphytanyl glycerol tetraethers (GDGTs) are membrane lipids consisting of isoprenoid carbon skeletons ether-bound to glycerol (Schouten et al., 2013). In marine systems they are primarily produced by ammonia oxidising marine Thaumarchaeota (Schouten et al., 2013). In modern marine core-top sediments, the relative abundance of GDGT compounds with more ring structures increases with the mean annual sea surface temperature (SST) of the overlying waters (Schouten et al., 2002). This trend is most likely driven by the need for increased cell membrane stability and rigidity at higher temperatures (Sinninghe Damsté et al., 2002). On this basis, the TEX₈₆ (tetraether index of tetraethers containing 86 carbon atoms) ratio was derived to provide an index to represent the extent of cyclisation (Eq. 1; where GDGT-x represents the fractional abundance of GDGTs determined by liquid chromatography mass spec-

trometry (LC-MS) peak area, and cren' is the peak area of the isomer of crenarchaeol, cren (Schouten et al., 2002; Kim et al., 2010) and was shown to be positively correlated with mean annual SSTs:

$$\text{TEX}_{86} = \frac{(\text{GDGT-2} + \text{GDGT-3} + \text{cren}')}{(\text{GDGT-1} + \text{GDGT-2} + \text{GDGT-3} + \text{cren}')} \quad (1)$$

Early applications of TEX_{86} to reconstruct ancient SSTs were promising, especially in providing temperature estimates in environments where standard carbonate-based proxies are hampered by poor preservation (Schouten et al., 2003; Herfort et al., 2006; Schouten et al., 2007; Huguet et al., 2006; Sluijs et al., 2006; Brinkhuis et al., 2006; Pearson et al., 2007; Sluijs et al., 2009). The TEX_{86} approach also extended beyond the range of the widely used alkenone-based $U_{37}^{k'}$ thermometer, in both temperature space, where $U_{37}^{k'}$ saturates at $\sim 28^\circ\text{C}$ (Brassell, 2014), and back into the early Cenozoic (Bijl et al., 2009; Hollis et al., 2009; Bijl et al., 2013; Inglis et al., 2015) and Mesozoic (Schouten et al., 2002; Jenkyns et al., 2012; O'Brien et al., 2017) where haptophyte-derived alkenones are typically absent from marine sediments (Brassell, 2014). Initially, TEX_{86} was converted to SSTs using the core-top calibration (Schouten et al., 2002) (Eq. 2):

$$\text{TEX}_{86} = 0.015 \times \text{SST} + 0.28. \quad (2)$$

However, as the number and range of applications of TEX_{86} palaeothermometry grew, concerns arose about proxy behaviour at both the high (Liu et al., 2009) and low (Kim et al., 2008) temperature ends of the modern calibration. In response to these observations, a new expanded modern core-top dataset (Kim et al., 2010) was used to generate two new indices – $\text{TEX}_{86}^{\text{L}}$ (Eq. 3), an exponential function that does not include the crenarchaeol regio-isomer and was recommended for use across the entire temperature range of the new core-top data (-3 to 30°C , particularly when SSTs are lower than 15°C), and $\text{TEX}_{86}^{\text{H}}$ (Eq. 4), also exponential, and recommended for use when SSTs exceeded 15°C (Kim et al., 2010). $\text{TEX}_{86}^{\text{L}}$ also excludes GDGT abundance data from the high-temperature regimes of the Red Sea, which are somewhat anomalous and likely related to salinity effects on community composition in this region (Trommer et al., 2009; Kim et al., 2010).

$$\text{TEX}_{86}^{\text{L}} = \log \left(\frac{[\text{GDGT-2}]}{[\text{GDGT-1}] + [\text{GDGT-2}] + [\text{GDGT-3}]} \right) \quad (3)$$

$$\text{TEX}_{86}^{\text{H}} = \log \left(\frac{[\text{GDGT-2}] + [\text{GDGT-3}] + [\text{cren}']}{[\text{GDGT-1}] + [\text{GDGT-2}] + [\text{GDGT-3}] + [\text{cren}']} \right) \quad (4)$$

Both $\text{TEX}_{86}^{\text{H}}$ and $\text{TEX}_{86}^{\text{L}}$ were widely used and tested across a range of temperatures and palaeoenvironments, including comparisons against other palaeotemperature proxy systems (Hollis et al., 2012; Lunt et al., 2012; Bijl et al., 2013; Dunkley Jones et al., 2013; Zhang et al., 2014; Seki et al., 2014;

Douglas et al., 2014; Linnert et al., 2014; Hertzberg et al., 2016). In certain environments $\text{TEX}_{86}^{\text{L}}$ was subject to significant variability in derived temperatures that were not apparent in $\text{TEX}_{86}^{\text{H}}$ (Taylor et al., 2013). This was mostly due to changing GDGT-2-to-GDGT-3 ratios, which strongly influence $\text{TEX}_{86}^{\text{L}}$, and may be related to local non-thermal environmental conditions at the site of GDGT production, and deep-water lipid production (Taylor et al., 2013). As a result, $\text{TEX}_{86}^{\text{L}}$ is no longer regarded as an appropriate tool for palaeotemperature reconstructions, except in limited polar conditions (Kim et al., 2010; Tierney, 2012).

Ongoing work to strengthen GDGT-based palaeothermometry is focused on three key issues. The first is a concern about undetected non-analogue palaeo-GDGT assemblages, for which the modern calibration dataset is inadequate to provide a robust temperature estimation. Although various screening protocols, with independent indices and thresholds, have been proposed to test for an excessive influence of terrestrial lipids (branched and isoprenoid tetraether, BIT index; Hopmans et al., 2004), within sediment methanogenesis (methane index, MI; Zhang et al., 2011) and non-thermal effects such as nutrient levels and archaeal community structure to impact the weighted average of cyclopentane moieties (ring index, RI; Zhang et al., 2016), these do not provide a fundamental measure of the proximity between GDGT abundance distributions in the modern calibration data set and ancient GDGT abundance distributions recorded in sediment samples. The fundamental question remains – are measured ancient assemblages of GDGT compounds anything like the modern assemblages, from which palaeotemperatures are being estimated? Understanding this question cannot easily be addressed with the use of indices – TEX_{86} itself, or BIT and MI – that collapse the dimensionality of GDGT abundance relationships onto a single axis of variation.

Second, from the earliest applications of the TEX_{86} proxy to deep-time warm climate states (Schouten et al., 2003) it was recognised that reconstructed temperatures beyond the range of the modern calibration ($> 30^\circ\text{C}$) were highly sensitive to model choice within the modern calibration range. Thus, Schouten et al. (2003) restricted their calibration data for deep-time temperature estimates to core-top data in the modern era with mean annual SSTs over 20°C . However, this problem of model choice, and its impact on temperature estimation beyond the modern calibration range, persists (Hollis et al., 2019), with current arguments focused on whether there is an exponential (e.g. Cramwinckel et al., 2018) or linear (Tierney and Tingley, 2015) dependency of TEX_{86} on SSTs, and the effect of these models on temperature estimates over 30°C .

Culture and mesocosm studies are sometimes cited in support of extrapolations beyond the modern calibration range when reconstructing ancient SSTs (Kim et al., 2010; Hollis et al., 2019). While there is a basic underlying trend for more rings within GDGT structures at higher temperatures (Zhang et al., 2016; Qin et al., 2015), the lack of a uniform

response to archaeal GDGT production in response to increasing growth temperatures (e.g. Elling et al., 2015; Qin et al., 2015) suggests that this does not easily translate into a simple linear model at the community scale (i.e. the core-top calibration dataset). In natural systems, it is likely that aggregated GDGT abundance variations in response to growth temperatures result from changing compositions of archaeal populations as well as the physiological response of individual strains to growth temperature (Elling et al., 2015; Polik et al., 2018). For deep-time applications it is even more difficult, where there is no independent constraint on the archaeal strains dominating production or their evolution through time (Elling et al., 2015). What is notable, however, is that the RI – calculated using all commonly measured GDGTs (Zhang et al., 2016) – has a more robust relationship with culture temperature between archaeal strains than TEX_{86} , indicating a potential loss of information within the TEX_{86} index (Elling et al., 2015).

Finally, under some conditions, the original TEX_{86} proxy had a relatively poor representation of the true uncertainty associated with palaeotemperature estimates, as it included no assessment of non-analogue behaviour relative to the modern core-top data. Instead, uncertainty was typically based on the residuals on the modern calibration, with no reference to the relationship between GDGT distributions of an ancient sample and the modern calibration data. An improved Bayesian uncertainty model “BAYSPAR” is now in widespread use for SST estimation, which models TEX_{86} to SST regression parameters, and associated uncertainty, as spatially varying functions (Tierney and Tingley, 2015). The Bayesian approach, as with all approaches based on the TEX_{86} index, however, still does not include an uncertainty that reflects how well modelled ancient GDGT assemblages are by the modern calibration – i.e. the degree to which they are non-analogue – as it still functions on one-dimensional TEX_{86} index values.

All empirical calibrations of GDGT-based proxies assume that mean annual SST is the master variable on GDGT assemblages both today and in the past. Mean annual SST, however, is strongly correlated with many other environmental variables (e.g. seasonality, pH, mixed layer depth and productivity). In the modern calibration dataset, mean annual SST shows the strongest correlation with TEX_{86} index (Schouten et al., 2002), but this does not preclude an important (but undetectable) influence of these other environmental variables. The use of empirical GDGT calibrations to infer ancient sea surface temperatures thus implicitly assumes that the relationships between mean annual SST and all other GDGT-influencing variables are invariant through time. This assumption is inescapable until, and unless, a more complete biological mechanistic model of GDGT production emerges.

Here, we return to the primary modern core-top GDGT assemblage data (Tierney and Tingley, 2015), and systematically explore the relationships between the modern GDGT distributions and surface ocean temperatures using powerful

mathematical tools. These tools can investigate correlations without prior assumptions on the best form of relationship or a priori selection of GDGT compounds to be used. This analysis is then extended through the exploration of the relationships between the modern core-top GDGT distributions and two compilations of ancient GDGT datasets, one from the Eocene (Inglis et al., 2015) and one from the Cretaceous (O’Brien et al., 2017). We explore simple metrics to answer the fundamental question – are modern core-top GDGT distributions good analogues for ancient distributions? We propose the first robust methodology to answer this question, and so screen for significantly non-analogue palaeo-assemblages. From this, we go on to derive a new machine learning approach “OPTiMAL” (Optimised Palaeothermometry from Tetraethers via MACHine Learning) for reconstructing SSTs from GDGT datasets, which outperforms previous GDGT palaeothermometers and includes robust error estimates that, for the first time, accounts for model uncertainty.

2 Models for GDGT-based temperature reconstruction

Our new analyses use the modern core-top data compilation, and satellite-derived estimates of SSTs, of Tierney and Tingley (2015) as well as compilations of Eocene (Inglis et al., 2015) and Cretaceous (O’Brien et al., 2017) GDGT assemblages. Within these fossil assemblages, only data points with full characterisation of individual GDGT relative abundances were used. We also note that, in the first instance, all available fossil assemblage data were included, although later comparisons between BAYSPAR and our new temperature predictor excludes fossil data that was regarded as unreliable based on standard pre-screening indices, as noted within the original compilations (Inglis et al., 2015; O’Brien et al., 2017). All data used in this study are tabulated in the Supplement.

In order to enable meaningful comparison between new and existing temperature predictors, we use the following consistent procedure for evaluating all predictors throughout this paper. We divide the modern core-top dataset of 854 data points into 85 validation data points (chosen randomly) and 769 calibration points (as we require fractional abundances for all 6 commonly measured GDGTs, we excluded those data points for which these values were not reported). We calibrate the predictor on the calibration points and then judge its performance on the validation points using the root mean square error:

$$\delta T = \sqrt{\frac{1}{N_v - 1} \sum_{k=1}^{N_v} \left(\hat{T}(x_k) - T(x_k) \right)^2}, \quad (5)$$

where the sum is taken over each of $N_v = 85$ validation points, T is the known measured temperature (which we refer to as the true temperature) and \hat{T} is the predicted temperature. For conciseness, we refer to δT as the predictor stan-

standard error. It is useful to compare the accuracy of the predictor to the standard deviation of all temperatures in the dataset σT , which corresponds to using the mean temperature as the predictor in Eq. (1); for the modern dataset, $\sigma T = 10.0^\circ\text{C}$. The coefficient of determination, R^2 , provides a measure of the fraction of the fluctuation in the temperature explained by the predictor. To facilitate performance comparisons between different methods of predicting temperature, we use the same subset of validation points for all analyses. To avoid sensitivity to the choice of validation points, we repeat the calibration–validation procedure for 10 random choices from the validation dataset.

2.1 Nearest neighbours

We begin with an agnostic approach to using some combination of the proportions of each of the six observables – GDGT-0, GDGT-1, GDGT-2, GDGT-3, crenarchaeol and cren', which we will jointly refer to as GDGTs – to predict sea surface temperatures. Whatever functional form the predictor might take, it can only provide accurate temperature predictions if nearby points in the six-dimensional observable space – i.e. the distribution of all of the six commonly reported GDGTs – can be translated to nearby points in temperature space. Conversely, if nearby points in the observable space correspond to vastly different temperatures, then no predictor, regardless of which combination of GDGTs are used, will be able to provide a useful temperature estimate. In other words, the structuring of GDGT distributions within multi-dimensional space must have some correspondence to the temperatures of formation (or rather the mean annual SSTs used for standard calibrations).

We therefore consider the prediction offered by the temperature at the nearest point in the GDGT parameter space. Of course, nearness depends on the choice of the distance metric. For example, it may be that sea surface temperatures are very sensitive to a particular GDGT, so even a small change in that GDGT corresponds to a significant distance, and rather insensitive to another, meaning that even with a large difference in the nominal value of that GDGT the distance is insignificant. In the first instance, we use a very simple Euclidian distance estimate $D_{x,y}$ where the distance along each GDGT is normalised by the total spread in that GDGT across the entire dataset. This normalisation ensures that a dimensionless distance estimate can be produced even when observables have very different dynamical ranges, or even different units. Thus, the normalised distance D between parameter data points x and y is

$$D_{x,y}^2 \equiv \sum_{i=0}^5 \frac{(\text{GDGT}_i(x) - \text{GDGT}_i(y))^2}{\text{var}(\text{GDGT}_i)}. \quad (6)$$

We show the distribution of nearest distances of points in the modern dataset, excluding the sample itself, in Fig. 1.

The nearest-sample temperature predictor is $\hat{T}_{\text{nearest}}(x) = T(y)$, where y is the nearest point to x over the calibration

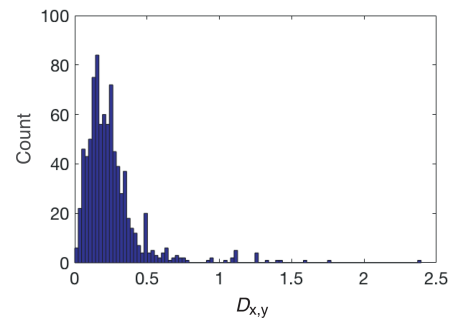


Figure 1. A histogram of the normalised distance to the nearest neighbour in GDGT space ($D_{x,y}$) for all samples in the modern calibration dataset of Tierney and Tingley (2015).

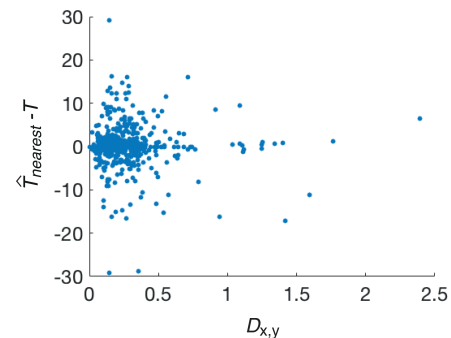


Figure 2. The error of the nearest-neighbour temperature ($D_{x,y}$) predictor, for modern core-top data, as a function of the distance to the nearest calibration sample.

dataset, i.e. one that minimises $D_{x,y}$. Figure 2 shows the scatter in the predicted temperature when using the temperature of the nearest data point to make the prediction. Overall, the failure of the nearest-neighbour predictor to provide accurate temperature estimates even when the normalised distance to the nearest point is small, $D_{x,y} \leq 0.5$, casts doubt on the possibility of designing an accurate predictor for temperature based on GDGT observations. This is most likely due to additional environmental controls on GDGT abundance distributions in natural systems, in particular the water depth (Zhang and Liu, 2018), nutrient availability (Hurley et al., 2016; Polik et al., 2018; Park et al., 2018), seasonality, growth rate (Elling et al., 2014; Hurley et al., 2016) and ecosystem composition (Polik et al., 2018), that obscure a predominant relationship to mean annual SSTs.

On the other hand, the standard error for the nearest-neighbour temperature predictor is $\delta T_{\text{nearest}} = 4.5^\circ\text{C}$. This is less than half of the standard deviation σT in the temperature values across the modern dataset. Thus, the temperatures corresponding to nearby points in GDGT observable space also cluster in temperature space. Consequently, there is hope that we can make some useful, if imperfect, temperature predictions. The value of $\delta T_{\text{nearest}}$ will also serve as a useful benchmark in this design: while we may hope to do

better by, say, suitably averaging over multiple nearby calibration points rather than adopting the temperature at one nearest point as a predictor, any method that performs worse than the nearest-neighbour predictor is clearly suboptimal.

2.2 TEX_{86} and Bayesian applications

The TEX_{86} index reduces the six-dimensional observable GDGT space to a single number. While this has the advantage of convenience for manipulation and the derivation of simple analytic formulae for predictors, as illustrated below, this approach has one critical disadvantage: it wastes significant information embedded in the hard-earned GDGT distribution data. Figure 3 illustrates both the advantage and disadvantage of TEX_{86} . On the one hand, there is a clear correlation between TEX_{86} and temperature (Fig. 3a), with a correlation coefficient of 0.81 corresponding to an overwhelming statistical significance of 10^{-198} . On the other hand, very similar TEX_{86} values can correspond to very different temperatures. We can apply the nearest-neighbour temperature prediction approach to the TEX_{86} value alone rather than the full GDGT parameter space; this predictor yields a large standard error of $\delta T_{\text{nearest } \text{TEX}_{86}} = 8.0^\circ\text{C}$ (bottom panel of Fig. 3). While smaller than σT , this is significantly larger than $\delta T_{\text{nearest}}$ (Fig. 2), consistent with the loss of information in TEX_{86} . We therefore do not expect other predictors based on TEX_{86} to perform as well as those based on the full available dataset.

Indeed, this is what we find when we consider predictors of the form $\hat{T}_{1/\text{TEX}} = a + b/\text{TEX}_{86}$ and $\hat{T}_{\text{TEXH}} = c + d \log_{\text{TEX}_{86}}$ (Liu et al., 2009; Kim et al., 2010), i.e. the established relationships between GDGT distributions and SST. We fit the free parameters a , b , c and d by minimising the sum of squares of the residuals over the calibration datasets (least squares regression). We find that $\delta T_{1/\text{TEX}} = 6.1^\circ\text{C}$ (note that this is slightly better than using the fixed values of a and b from Kim et al. (2010), which yield $\delta T_{1/\text{TEX}} = 6.2^\circ\text{C}$). We note that the corresponding R^2 value associated with these TEX_{86} -based predictors is 0.64, which is lower than the R^2 values in Kim et al. (2010). We attribute this to the fact that we are using a larger dataset based on Tierney and Tingley (2015), including data from the Red Sea (Kim et al., 2010).

Tierney and Tingley (2014) proposed a more sophisticated approach to obtaining the transfer function from TEX_{86} to temperature, continuing to use simple linear regression, but with the addition of Gaussian processes to model spatial variability in the temperature– TEX_{86} relationship and working with a forward model which is subsequently inverted to produce temperature predictions. This forward model “BAYSPAR” is capable of generating an infinite number of calibration curves relating TEX_{86} to sea surface temperatures (Tierney and Tingley, 2014). In order to derive a calibration for a specific dataset, the user edits a range of parameters which vary depending on whether the dataset in question is

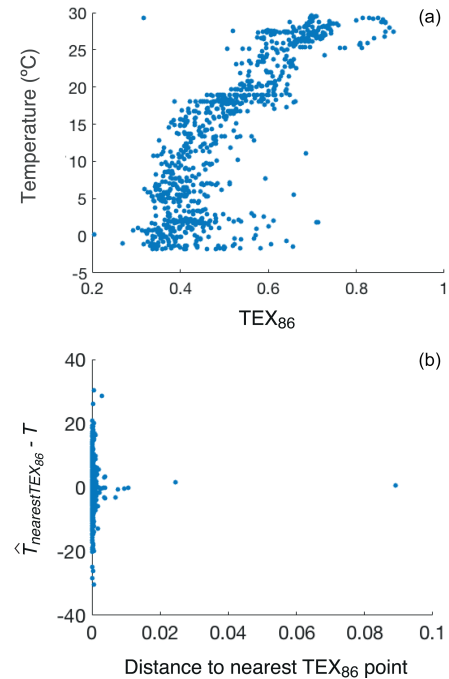


Figure 3. (a) The temperature of the modern dataset as a function of the TEX_{86} value, showing a clear linear correlation between the two, but also significant scatter. (b) The error of the predictor based on the nearest TEX_{86} calibration point.

from the relatively recent past or deep time (Tierney and Tingley, 2014). For deep time applications, the authors propose a modern analogue-type approach, in which they search the modern data for $20^\circ \times 20^\circ$ grid boxes containing “nearby” TEX_{86} measurements and subsequently apply linear regression models calibrated on the analogous samples for making predictions.

However, along with the simpler TEX_{86} -based models described above, this approach still suffers from the reduction of a six-dimensional dataset to a single number. Therefore, it is not surprising that even the simplest nearest-neighbour predictor (such as the one described above) that makes use of the full six-dimensional dataset outperforms single-dimensional forward modelling approaches. Additionally, uncertainty estimates do not account for the fact that TEX_{86} is, fundamentally, an empirical proxy, and so its validity outside the range of the modern calibration is not guaranteed. This is a fundamental issue for attempts to reconstruct surface temperatures during greenhouse climate states, when tropical and sub-tropical SSTs were likely hotter than those observed in the modern oceans.

2.3 Machine learning approaches – random forests

There are a number of options to improve on nearest-neighbour predictions using machine learning techniques such as artificial neural networks and random forests. These

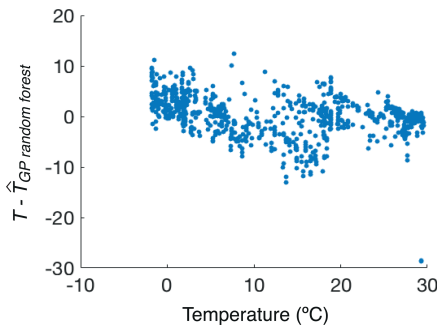


Figure 4. The error of a random forest predictor as a function of the true temperature.

flexible, non-parametric models would ideally be based on the underlying processes driving the GDGT response to temperature, but since these processes remain unconstrained at present, we choose to deploy models which can reasonably reflect predictive uncertainty and will be sufficiently adaptable in future (as new information regarding controls on GDGTs emerge). These machine learning approaches are all based on the idea of training a predictor by fitting a set of coefficients in a sufficiently complex multi-layer model in order to minimise residuals on the calibration dataset. As an example of the power of this approach, we train a random forest of decision trees with 100 learning cycles using a least-squares boosting to fit the regression ensemble. Figure 4 shows the prediction accuracy for this random forest implementation. This machine learning predictor yields $\delta T = 4.1^\circ\text{C}$, outperforming the naive nearest-neighbour predictor by effectively applying a suitable weighted average over multiple near neighbours. This corresponds to a very respectable $R^2 = 0.83$, meaning that 83 % of the variation in the observed temperature is successfully explained by our GDGT-based model.

2.4 Gaussian process regression

One downside of the random forest predictor is the difficulty of accurately estimating the uncertainty of the prediction (Mentch and Hooker, 2016), although this is possible with, for example, a bootstrapping approach (Coulston et al., 2016). Fortunately, Gaussian process (GP) regression provides a robust alternative. For full details on GP regression refer to Williams and Rasmussen (2006) and Rasmussen and Nickisch (2010). Loosely, the objective here is to search among a large space of smoothly varying functions of GDGT compositions for those functions which adequately describe temperature variability. This, essentially, is a way of combining information from all calibration data points, not just the nearest neighbours, assigning different weights to different calibration points depending on their utility in predicting the temperature at the input of interest. The trained Gaussian process learns the best choice of weights to fit the data. Typi-

cally, the GP will give greater weight to closer points, but, as we discuss below, it will learn the appropriate distance metric on the multi-dimensional GDGT input space.

The weighting coefficients learned by the GP emulator represent a covariance matrix on the GDGT parameter space. We can use this as a distance metric to provide meaningfully normalised distances between points, removing the arbitrariness from the nearest neighbour distance ($D_{x,y}$) definition used earlier, and this is the basis of the D_{nearest} metric described below. If the temperature is insensitive to a particular GDGT input coordinate (i.e. the value of that input has a minimal effect on the temperature) then points within GDGT space that have large differences in absolute input values in that coordinate are still near. We find that cren has very limited predictive power, and so points with large cren differences are close in terms of the normalised distance. Conversely, if the temperature is sensitive to small changes in a particular GDGT variant, then points with relatively nearby absolute input values in that coordinate are still distant. We find that most GDGT parameters other than cren are comparably useful in predicting temperature, with GDGT-0 and GDGT-3 marginally the most informative. We considered whether interdependency of percentage GDGT data could influence our calculations. Our analysis suggests that there are only five free parameters. Machine learning tools should be able to pick up this correlation and effectively ignore one of the parameters (or one parameter combination). For example, we do find that the GP emulator has a very broad kernel in at least one dimension, signalling this. In principle, we could have considered only five of six parameters. The smaller scale of some of the parameters is automatically accounted for by the trained kernel size in GP regression, or by normalising to the appropriate dynamical range in our initial investigation. In short, the accuracy of Gaussian process regression is not adversely affected by correlations between inputs (Rasmussen and Williams, 2006). Significantly correlated inputs that do not bring in new predictive power are appropriately down-weighted.

We use a Gaussian process model with a squared exponential kernel with automatic relevance determination (ARD) to allow for a separate length scale for each GDGT predictor. We fit the GP parameters with an optimiser based on quasi-Newtonian approximation to the Hessian. Prediction accuracy is shown in Fig. 5, and we find that $\delta T = 3.72^\circ\text{C}$, which is a substantial improvement over the existing indices, at least on the modern data. As mentioned, the GP framework provides a natural quantification of predictive uncertainty, which includes uncertainty about the learned function. This is in contrast to, for example, the TEX_{86} proxy, whereby the uncertainty associated with the selection of the particular functional form used for predictions is ignored. While Tierney and Tingley (2014) also use Gaussian processes to model uncertainty, they model spatial variability in the TEX_{86} –temperature relationship with a Gaussian process prior. While this is a valuable approach to understand re-

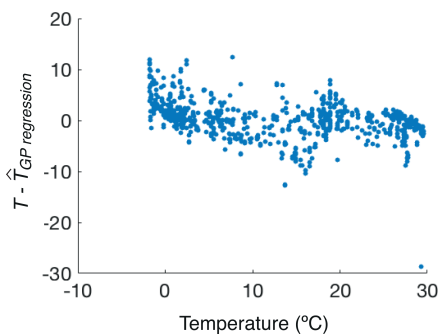


Figure 5. The error of the GPR (Gaussian process regression) predictor as a function of the true temperature.

gional effects in the TEX₈₆–temperature relationship, it does not deal with the “non-analogue” situations we are concerned with in this paper.

2.5 Data structure

The random forest (Sect. 2.3) and GPR approaches (Sect. 2.4) are agnostic about any underlying bio-physical model that might impart the observed temperature dependence on GDGT relative abundances produced by archaea. They are essentially optimised interpolation tools for mapping correlations between temperature and GDGT abundances within the range of the modern calibration dataset; they can make no sensible inference about the behaviour of this relationship outside of the range of this training data. To move from interpolation within, to extrapolation beyond, the modern calibration requires an understanding of, and model for, the temperature dependence of GDGT production. To explore these relationships and the extent to which the ancient and modern data reside in a coherent relationship within GDGT space, we employed two forms of dimensionality reduction to enable visualisation of the data in two or three dimensions. The fundamental point is that if temperature is the dominant control, all of the data should lie approximately on a one-dimensional curve in GDGT space, and the arclength along this curve should correspond to temperature; we will revisit this point below.

We first employed a version of principal component analysis (PCA) tailored to compositional data (Aitchison, 1982, 1983; Aitchison and Greenacre, 2002; Filzmoser et al., 2009a, b, 2012). Taking into account the compositional nature of the data is important because the sum-to-one constraint induces correlations between variables which are not accounted for by classical PCA. Furthermore, apparently nonlinear structure in Euclidean space often corresponds to linearity in the simplex (i.e. the restricted space in which all elements sum to one) (Egozcue et al., 2003). Figure 6 shows the modern, Eocene and Cretaceous data projected onto the first two principal components. Aside from the obvious outlying cluster of Cretaceous data, characterised by GDGT-

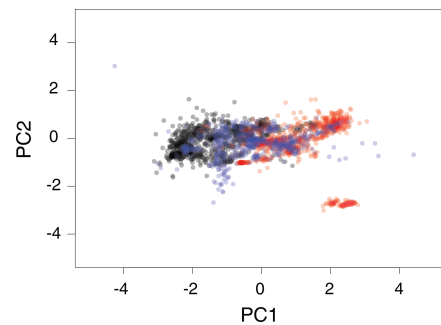


Figure 6. Modern and ancient data projected onto the first two compositional principal components. Black: modern; blue: Eocene (Inglis et al., 2015); red: Cretaceous (O’Brien et al., 2017).

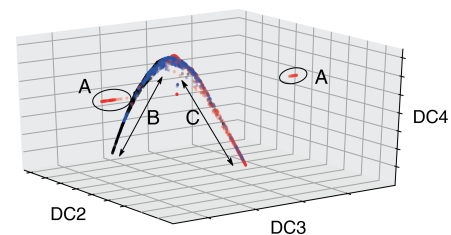


Figure 7. Diffusion map projection of the modern and ancient data. Black: modern; blue: Eocene (Inglis et al., 2015); red: Cretaceous (O’Brien et al., 2017). Separate clusters marked “A” are the outlying Cretaceous points with high GDGT-3 values. Branch “B” is dominated by modern data points; branch “C” by Cretaceous data.

3 fractions above 0.6, the bulk of the data occupy a two-dimensional point cloud with a small amount of curvature. The large majority of the Cretaceous data has more positive PC1 values relative to the modern data.

We also explored the data using diffusion maps (Coifman et al., 2005; Haghverdi et al., 2015), a nonlinear dimensionality reduction tool designed to extract the dominant modes of variability in the data. Such diffusion maps have been successfully used to infer latent variables that can explain patterns of gene expression. In the case of biological organisms, this latent variable is commonly developmental age (called pseudo-time) (Haghverdi et al., 2016). In our case, the assumption would be that this latent variable corresponds to temperature. Inspection of the eigenvalues of the diffusion map transition matrix suggests that four diffusion components are adequate to represent the data; we plot the second, third and fourth of these components in Fig. 7 for the modern and ancient data. The separate clusters marked “A” are the outlying Cretaceous points with high GDGT-3 values. The bulk of the modern data lies on the branch marked “B”, while the bulk of the Cretaceous data lies on the branch marked “C”. Notably, the majority of the modern points lying on branch C are from the Red Sea, which suggests that the Red Sea data are essential for understanding ancient climates (particularly Cretaceous climates).

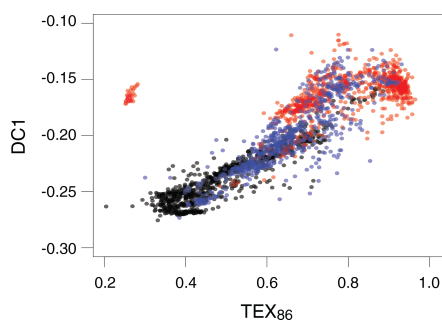


Figure 8. The first diffusion component as a function of TEX_{86} . Some outlying points have been excluded from the plot for the purposes of visualisation. Black: modern; blue: Eocene (Inglis et al., 2015); red: Cretaceous (O’Brien et al., 2017).

The relationship between the first diffusion component and TEX_{86} for all data is shown in Fig. 8. There is a clear correlation, despite the presence of some outlying Cretaceous points, some of which are not shown because they lie so far outside the majority data range within this projection. This suggests that TEX_{86} is, in one sense, a natural one-dimensional representation of the data. We also plot the first diffusion component for the modern data as a function of temperature (Fig. 9). We see a similar pattern emerging to that displayed by TEX_{86} – there is little sensitivity to temperature below 15°C , and between ~ 20 and 25°C . An interesting avenue for future research might be to explore the temperature–GDGT system from a dynamical systems perspective, i.e. use simple mechanistic mathematical models to explore the temperature dependence of steady-state GDGT distributions. It may be that such models suggest that only a few steady states exist, and that temperature is a bifurcation parameter, i.e. it controls the switch between the steady states. Note also the downward slope in the residual pattern in Fig. 4 between 0 and 15 – 17°C , and again at higher temperatures. This pattern is consistent with predictions that are biased towards the centre of each “cluster”, i.e. a system which is not very sensitive to temperature but can distinguish between high and low temperatures reasonably well. This observation also links to recent culture studies (Elling et al., 2015) and Pliocene–Pleistocene sapropel data (Polik et al., 2018), which support the existence of discrete populations with unique GDGT–temperature relationships and that temporal changes in population over time can drive changes in TEX_{86} .

2.6 Forward modelling

Based on the analysis of the combined modern and ancient data structure outlined above, there appears to be some consistency to underlying trends in the overall variance of GDGT relative abundances. These trends provide some hope that models of this variance, and its relationship to sea surface temperature, within the modern dataset could be devel-

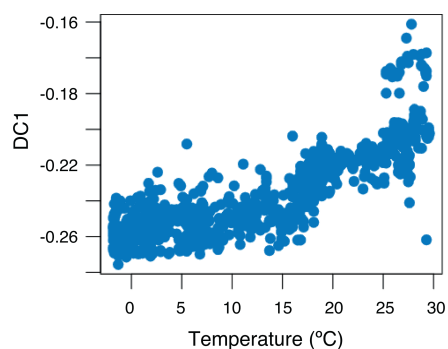


Figure 9. The first diffusion component as a function of temperature (modern data only).

oped to predict ancient SSTs. TEX_{86} and BAYSPAR are such models, but they are limited by, first, the reduction of six-dimensional GDGT space to a one-dimensional index, and second, by an ad hoc model choice – linear, exponential – that does not account for uncertainty in model fit to the modern calibration data, and the resultant uncertainty in the estimation of ancient SSTs relating to model choice. To overcome these issues, we develop a forward (Fwd) model based on a multi-output Gaussian process (Alvarez et al., 2012), which models GDGT compositions as functions of temperature, accounting for correlations between GDGT measurements. This model is then inverted to obtain temperatures which are compatible with a measured GDGT composition. In simple terms, we posit that a measured GDGT composition is generated by some unknown function of temperature and corrupted by noise, which may be due to measurement error or some unmodelled particularity of the environment in which the sample was generated. We proceed by defining a large (in this case infinite) set of functions of temperature to explore and compare them to the available data, throwing away those functions which do not adequately fit the data. This means, of course, that the behaviour of the functions we accept is allowed to vary more widely outside the range of the modern data than within it. With no mechanistic underpinning, choosing only one function (such as the inverse of TEX_{86}) based on how well it fits the modern data grossly underestimates our uncertainty about temperature where no modern analogue is available.

The forward modelling approach is similar to that of Haslett et al. (2006), who argue that it is preferable to model measured compositions as functions of climate, before probabilistically inverting the model to infer plausible climates given a composition. The cost of modelling the data in this more natural way is the loss of degrees of freedom – we are now attempting to fit a one-dimensional line through a multidimensional point cloud rather than fit a multidimensional surface to the GDGT data, which means that the predictive power of the model suffers, at least on the modern data. The existing BAYSPAR calibration also specifies the model in

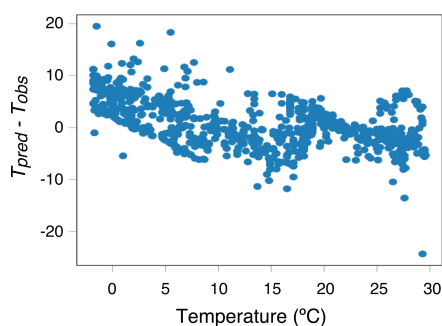


Figure 10. Temperature residuals for the forward model.

the forward direction; however, while BAYSPAR does model spatial variability, it assumes a monotonic relationship between TEX and SST, only accounting for uncertainties on the parameters within the model rather than any systematic uncertainty in the model itself. As with all GP models, the choice of kernel has a substantial impact on predictions (and their associated uncertainty) outside the range of the modern data, where predictions revert to the prior implied by the kernel. Given that we have no mechanistic model for the data-generating process, we recommend the use of kernels which do not impose strong prior assumptions on the form of the GDGT–temperature relationship (e.g. kernels with a linear component) and thus reasonably represent model uncertainty outside the range of the modern data. We choose a zero-mean Matérn 3/2 kernel for the applications below. Note, however, that since we are working in isometric log-ratio transformed coordinates, this corresponds to a prior assumption of uniform compositions at all temperatures, i.e. all components are equally abundant.

The residuals for the forward model are shown in Fig. 10. The clear pattern in the residuals does not necessarily indicate model misspecification, since no explicit noise model is specified for temperatures. Predictive distributions are to be interpreted in the Bayesian sense, in that they represent a “degree of belief” in temperatures given the model and the modern data. The residual pattern is similar to that of the random forest (Fig. 4) with two clear downward slopes, suggesting again that the data are clustered into temperatures above and below 16–17 °C, and that predictions tend towards temperatures at the centres of these clusters.

An advantage of the forward modelling approach is that the inversion can incorporate substantive prior information about temperatures for individual data points. In particular, other proxy systems can be used to elicit prior distributions over temperatures to constrain GDGT-based predictions, particularly when attempting to reconstruct ancient climates with no modern analogue in GDGT space. We emphasise that outside the range of the modern data, the utility of the models is almost solely due to the prior information included in the reconstruction. At present, the only priors being used in the forward model prescribe a reasonable upper limit

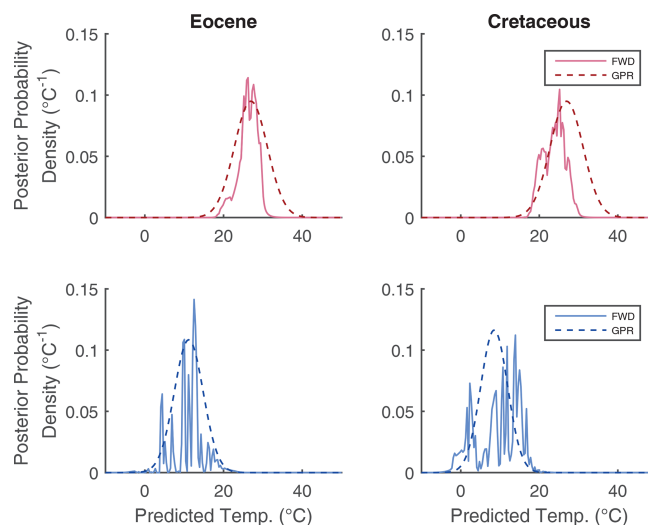


Figure 11. The posterior distributions over temperature from the forward model for selected examples of high and low temperature examples, from both the Eocene and Cretaceous. The Gaussian error envelope from the GPR model is shown for comparison.

and lower limit on temperatures (see Supplement). The only way to improve these reconstructions will be for future iterations to incorporate prior information from other proxies. It is worth noting that the predictive uncertainty, while reasonably well-described by the standard deviation in cases where ancient data lie quite close to the modern data in GDGT space, can be highly multimodal (Fig. 11). This is the case when estimates are significantly outside of the modern calibration dataset, such as low-latitude data in the Cretaceous, or where there is considerable scatter in the modern calibration data, for example in the low temperature range (< 5 °C).

3 Non-analogue behaviour and extrapolation

In principle, the predictors described above can be applied directly to ancient data, such as data from the Eocene or Cretaceous (Inglis et al., 2015; O’Brien et al., 2017). In practice, one should be careful with using models outside their domain of applicability. The machine learning tools described above, which are ultimately based on the analysis of nearby calibration data in GDGT space, are fundamentally designed for *interpolation*. To the extent that ancient data occupy a very different region in GDGT space, *extrapolation* is required, which the models do not adequately account for. The divergence between modern calibration data and ancient data is evident from Fig. 12, which shows histograms of minimum normalised distances between “high-quality” Eocene and Cretaceous data points (those that passed the screening tests applied by O’Brien et al., 2017 and Inglis et al., 2015) and the nearest point in the full modern dataset. We strongly recommend the use of the weighted distance metric (D_{nearest}) as a screening method to determine whether the modern core-

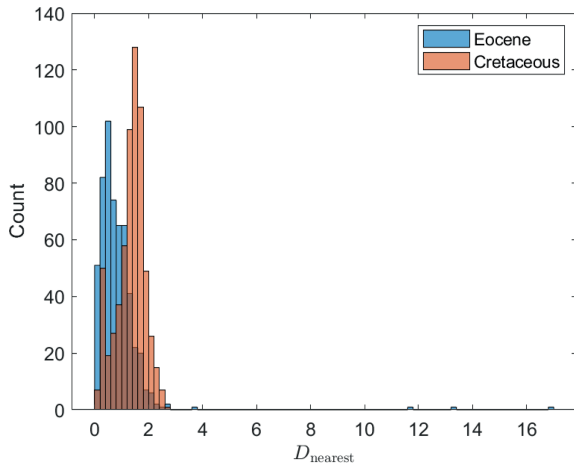


Figure 12. A histogram of normalised distances to the nearest sample in the modern dataset for Eocene and Cretaceous data, excluding samples that had been screened out in previous compilations using BIT, MI and RI following the approach of Inglis et al. (2015) and O’Brien et al. (2017).

top GDGT assemblage data is an appropriate basis for ancient SST estimation on a case-by-case basis. Note that this distance measure is weighted by the scale length of the relevant parameter as estimated by the Gaussian process emulator in order to quantify the relative position of ancient GDGT assemblages to the modern core-top data. By using the GP-estimated covariance as the distance metric, we account for the sensitivity of different GDGT components to temperature. Our inference is that samples with $D_{\text{nearest}} > 0.5$ are unlikely to be well constrained by any current calibration model. In these instances, in both our GPR and Fwd models, the constraints provided by the modern calibration dataset are such that estimates of temperature have large uncertainty bands that are dictated by model priors, i.e. are unconstrained by the calibration data (e.g. Figs. 13 and 14). This uncertainty is not apparent from estimates generated by BAYSPAR or $\text{TEX}_{86}^{\text{H}}$ models, although the underlying lack of constraints are the same. While 93 % of validation data points in the modern data have $D_{\text{nearest}} < 0.5$, this is the case for only 33 % of Eocene samples and 3 % of Cretaceous samples.

Where ancient GDGT distributions lie far from the modern calibration dataset ($D_{\text{nearest}} > 0.5$), we argue that there is no suitable set of modern analogue GDGT distributions from which to infer growth temperatures for this ancient GDGT distribution. Both the GPR and Fwd models revert to imposed priors once the distance from the modern calibration dataset increases. We also note that there are two broad, non-mutually exclusive categories of samples that lie far from the modern calibration dataset ($D_{\text{nearest}} > 0.5$): the first are samples that seem to lie “beyond” the temperature–GDGT calibration relationship, likely with (unconstrained) GDGT formation temperatures higher than the modern core-top calibrations; the second are samples with anomalous GDGT dis-

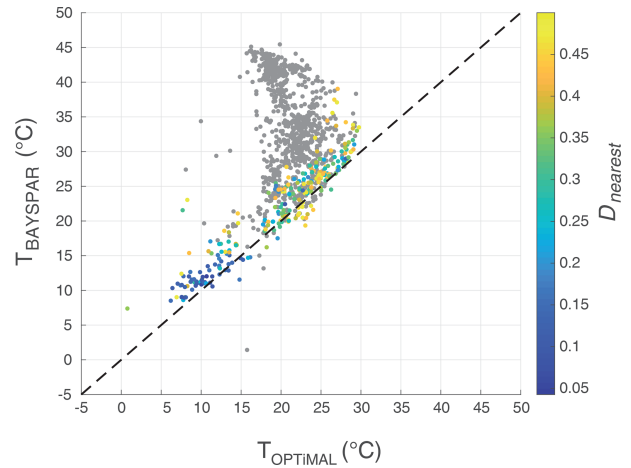


Figure 13. Comparison of temperature estimates for the BAYSPAR and the OPTiMAL GPR model. Greyed-out data fails the D_{nearest} test (> 0.5), and the colour scaling reflects D_{nearest} values for those data points that pass. Note that outside of the constraints of the modern calibration (training) dataset (D_{nearest} test > 0.5) the GPR model temperature estimates revert to the mean value of the calibration dataset, with an uncertainty that reverts to the standard deviation of the training data.

tributions lying on the margins of, or far away from, the main GDGT clustering in six-dimensional space (see outliers in Fig. 8).

Given the (current) limit on natural mean annual surface ocean temperatures of $\sim 30^\circ\text{C}$, extending the GDGT–temperature calibration might be possible through (1) integration of full GDGT abundance distributions produced in high-temperature culture, mesocosm or artificially warmed sea surface conditions into the models followed by (2) validation through robust inter-comparisons of any new GDGT palaeothermometer for high-temperature conditions with other temperature proxies from past warm climate states. As discussed in the introduction, the first approach is limited by the ability of culture or mesocosm experiments to accurately represent the true diversity and growth environments and dynamics of natural microbial populations. Such studies clearly indicate a more complex, community-scale control on changing GDGT relative abundances to growth temperatures (e.g. Elling et al., 2015). Community-scale temperature dependency can be modelled relatively well with analyses of natural production preserved in core-top sediments, especially with more sophisticated model fitting, including the GPR and Fwd model presented here. Above $\sim 30^\circ\text{C}$, however, the behaviour of even single strains of mesophilic archaea are not well-constrained by culture experiments, and the natural community-level responses above this temperature are, so far, unknown. While there is evidence for the temperature sensitivity of GDGT production by thermophilic and acidophilic archaea in older papers (de Rosa et al., 1980; Gliozzi et al., 1983), recent works, characterised by more

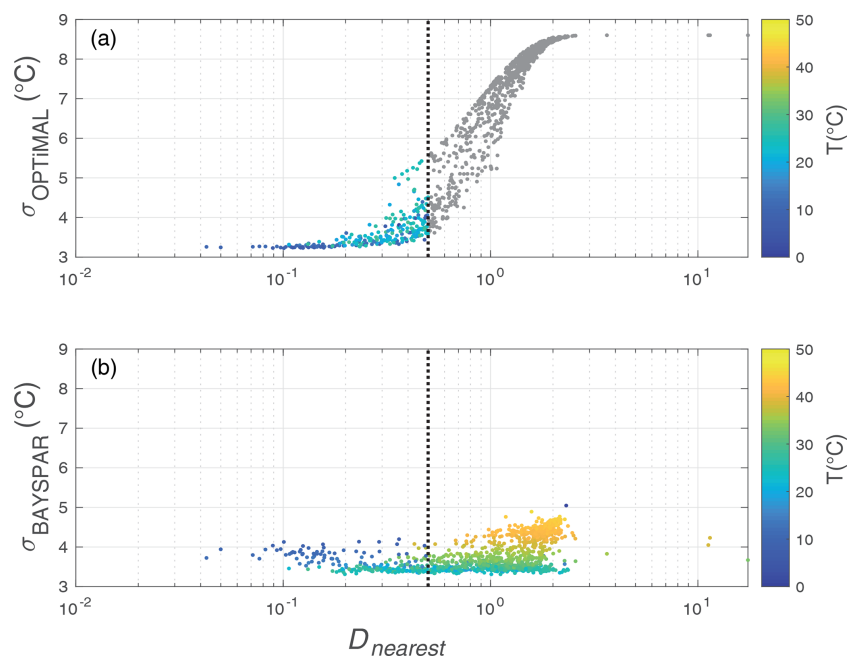


Figure 14. Inter-comparison of temperature estimates and standard errors (y axis) for compiled Eocene and Cretaceous data calculated using OPTiMAL (a) and BAYSPAR (b). Greyed-out data fails the D_{nearest} test (> 0.5), and the colour scaling reflects D_{nearest} values for those data points that pass. The black dashed line shows the D_{nearest} threshold (> 0.5).

precise phylogenetic and culturing techniques, show a more complex relationship between GDGT production and temperature. Elling et al. (2017) highlight that there is no correlation between TEX_{86} and growth temperature in a range of phylogenetically different thaumarchaeal cultures – including thermophilic species. Bale et al. (2019) recently cultured *Candidatus Nitrosotenuis uzonensis* from the moderately thermophilic order Nitrosopumilales (that contains many mesophilic marine strains). They found no correlation between TEX_{86} calibrations (either the Kim et al., 2010, core-top or Wuchter et al., 2004, and Schouten et al., 2008, mesocosm calibrations) with membrane lipid composition at different growth temperatures (37, 46 and 50 °C) and found that phylogeny generally seems to have a stronger influence on GDGT distribution than temperature. In view of these existing data, extrapolation of modern core-top calibration datasets into the unknown above 30 °C is uncertain, although the coherent patterns apparent across GDGT space, between modern, Eocene and Cretaceous data (Fig. 7), do provide some grounds for hope that the extension of GDGT palaeothermometry beyond 30 °C might be possible in future.

4 OPTiMAL and D_{nearest} : a more robust method for GDGT-based palaeothermometry

A more robust framework for GDGT-based palaeothermometry could be achieved with a flexible predictive model that uses the full range of six GDGT relative abundances and

has transparent and robust estimates of the prediction uncertainty. In this context, the Gaussian process regression model (GPR; Sect. 2.4) outperforms the forward model (Fwd; Sect. 2.6) within the modern calibration dataset and we recommend standard use of the GPR model, henceforth called OPTiMAL, over the Fwd model. Model code for the calculation of D_{nearest} values and OPTiMAL SST estimates (MATLAB script) and the Fwd model SST estimates (R script) are archived in the Zenodo repository, <https://doi.org/10.5281/zenodo.4293851> (Greene and Mandel, 2020).

Following Tierney and Tingley (2014) we use a reduced calibration dataset, with the exclusion of Arctic data with observed SSTs of less than 3 °C (“NoNorth/TT13” of Tierney and Tingley, 2014) but with the inclusion of additional core-top data from Seki et al. (2014). Full details of this calibration dataset are provided in the Supplement; to distinguish from the original OPTiMAL calibration data, which included the Arctic data < 3 °C, we refer to the original data as “Op1” and the new calibration dataset as “Op3”. An “Op2” is also available, which is the same as Op1 except that it excludes the Seki et al. (2014) data. In sensitivity tests to a range of applications across Quaternary and deep-time datasets, calibration Op1 and Op2 performed in almost identical fashion. The performances of Op1 and Op3 were very similar in most applications, except in applications to the palaeo-Arctic (see below), where the inclusion of modern Arctic calibration data (Op1) provided closer calibration constraints to the palaeo-data. Although the inclusion of modern Arctic data

may well be beneficial for the study of high-latitude palaeoclimate archives, we are initially cautious as in this instance the deep-time palaeo-data have previously been rejected because of a potential bias by non-marine inputs indicated by high BIT indices (Sluijs et al., 2020). In this case there could be some consistency between the modern and ancient GDGT production by marine archaea in the Arctic which may help in the understanding of GDGT-based palaeothermometry in this unusual environment (Sluijs et al., 2020), but we recommend further investigation of the modern Arctic core-top biomarker assemblages before their regular inclusion into the calibration dataset. The D_{nearest} methodology may prove useful in quantifying analogue and non-analogue behaviour through time in such conditions. For the purposes of this study, however, we take the conservative approach, and one that maintains a more consistent calibration basis with BAYSPAR, by using OPTiMAL calibration Op3 in the remainder of this discussion, and recommend its use in future applications of OPTiMAL.

To investigate the behaviour of the new OPTiMAL model, we compare temperature predictions including uncertainties for the Eocene and Cretaceous datasets, made by OPTiMAL and the BAYSPAR methodology of Tierney and Tingley (2014) (Figs. 13 and 14), using the default priors specified in the model code for the BAYSPAR estimation. The OPTiMAL model systematically estimates slightly cooler temperatures than BAYSPAR, with the biggest offsets below $\sim 15^\circ\text{C}$ (Fig. 13). Fossil GDGT assemblages that fail the D_{nearest} test are shown in grey, which clearly illustrate the regression to the mean in the OPTiMAL model, whereas BAYSPAR continues to make SST predictions up to and exceeding 40°C for these “non-analogue” samples due to the fact that BAYSPAR assumes that higher TEX_{86} values equate to higher temperatures as part of the functional form of the model, whereas the GPR model is agnostic on this. A comparison of error estimation between OPTiMAL and BAYSPAR is shown in Fig. 14. For most of the predictive range below the D_{nearest} cut-off of 0.5, OPTiMAL has smaller predicted uncertainties than BAYSPAR, especially in the lower temperature range. As D_{nearest} increases, i.e. as the fossil GDGT assemblage moves further from the constraints of the modern calibration dataset, the error on OPTiMAL increases, until it reaches the standard deviation of the modern calibration dataset (i.e. is completely unconstrained). In other words, OPTiMAL generates maximum-likelihood SSTs with robust confidence intervals, which appropriately reflect the relative position of an ancient sample used for SST estimation and the structure of the modern calibration dataset. Where there are strong constraints from near analogues in the modern data, uncertainties will be small; where there are weak constraints, uncertainty increases. In contrast, while uncertainty bounds do increase when BAYSPAR is used to extrapolate beyond the modern calibration, they are not as large as OPTiMAL because BAYSPAR assumes a linear increase in SST at higher TEX values.

We also provide an initial assessment of the inter-relationship between standard screening indices and D_{nearest} , for the Eocene and Cretaceous compilations, where the data are available to calculate these measures (Fig. 15). For ease of comparison between Eocene and Cretaceous datasets and visualisation of the majority of the data, extreme outliers ($D_{\text{nearest}} > 4.0$) are not shown. The metrics include the BIT index (Hopmans et al., 2004; Weijers et al., 2011), the methane index (MI; Zhang et al., 2011), the deviation between TEX_{86} and the ring index (ΔRI ; Zhang et al., 2016) and the %GDGT-0 (Blaga et al., 2009; Sinninghe Damsté et al., 2012). The standard screening levels for each of these metrics, as used in previous palaeo-compilations (O’Brien et al., 2017), are shown in the blue shaded areas in Fig. 15 (BIT > 0.5 ; MI > 0.5 ; $\Delta\text{RI} > 0.3$; %GDGT-0 $> 67\%$) – data points within these areas fail the standard screening. Also shown in Fig. 15 is the region where data pass our D_{nearest} screening requirement (grey shaded vertical region). In nearly all cases GDGT assemblages that fail these traditional screening tests also have D_{nearest} values that exceed 0.5 – i.e. “abnormal” GDGT assemblages are well-screened by D_{nearest} . The main exception to this is the BIT index in the Eocene dataset, where 15 samples have high BIT values (> 0.5) but have GDGT assemblages that are close to modern analogues in the calibration dataset ($D_{\text{nearest}} < 0.5$). Of these samples, 9 are from the Arctic Ocean between the Paleocene–Eocene thermal maximum (PETM) and Eocene thermal maximum 2 (ETM2), an interval noted for its relatively high BIT index values (Sluijs et al., 2020); 3 are from the Eocene–Oligocene transition of ODP Site 1218 (eastern Equatorial Pacific) (Liu et al., 2009); 2 are from the middle Eocene of Seymour Island (Douglas et al., 2014); and 1 is from the late Eocene of DSDP Site 511, which has been already noted as an individual sample with anomalously high BIT in this dataset (Liu et al., 2009; Inglis et al., 2015). Although high BIT at ODP Site 1218 has been inferred to represent “relatively high terrestrial input” (Inglis et al., 2015) this seems unusual for a fully pelagic site situated on oceanic crust > 3000 km away from the nearest continental landmass. Interpreting high BIT values as exclusively caused by terrestrial organic components appears problematic in this instance, especially as $D_{\text{nearest}} < 0.5$ give some assurance that these GDGT assemblages from ODP Site 1218 are well-modelled by the modern calibration dataset. GDGT assemblages from Seymour Island associated with high BIT values (> 0.4) appear to have an impact on the $\text{TEX}_{86}^{\text{H}}$ SST proxy (Inglis et al., 2015), but the 2 samples that fail BIT (> 0.5) but pass D_{nearest} (< 0.5) give OPTiMAL SSTs consistent ($5\text{--}6^\circ\text{C}$) with the SSTs from samples that pass all other screening and D_{nearest} ($\sim 4\text{--}7^\circ\text{C}$). In summary, the relationship between D_{nearest} and BIT suggests that BIT is not always closely coupled to GDGT assemblages that are strongly divergent from the modern calibration dataset.

With respect to the other screening indices there are clear indications that increased distance from the modern calibra-

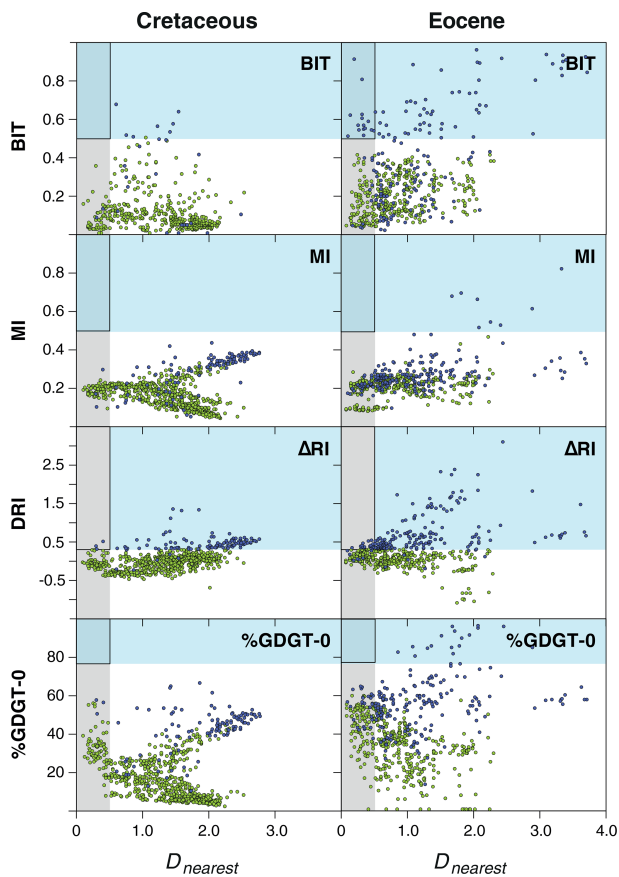


Figure 15. Comparison of D_{nearest} against standard screening indices: BIT and MI indices, ΔRI and %GDGT-0 for the Eocene (Inglis et al., 2015) and Cretaceous (O’Brien et al., 2017) datasets. Blue shaded regions show the standard failure thresholds for these indices (see text); grey shaded region highlights data that are below the D_{nearest} threshold of 0.5. Data points that fail any one of the standard screening indices are shown in blue, and data points that pass all standard screening indices are shown in green. The outlined black box is the region of data that fails traditional screening indices but passes $D_{\text{nearest}} (< 0.5)$.

tion (increased D_{nearest}) is associated with a trend towards the “thresholds of failure” in the screening indices. This pattern is most clear with the ΔRI in both the Cretaceous and the Eocene data, as increasing numbers of samples fail ΔRI as D_{nearest} increases. This supports ΔRI as a robust methodology for identifying samples that strongly diverge from the expected temperature dependence of GDGT assemblages as modelled by TEX_{86} in the modern calibration dataset. There are, however, samples that pass $D_{\text{nearest}} < 0.5$ but fail ΔRI in both the Eocene and Cretaceous datasets – these must have “near neighbours” in the modern calibration data but yet have a temperature sensitivity that is less well-modelled by TEX_{86} (divergence between RI and TEX_{86}). Conversely there are many Eocene and Cretaceous data points with $\Delta\text{RI} < 0.3$, but which fail $D_{\text{nearest}} (> 0.5)$. These data most likely represent

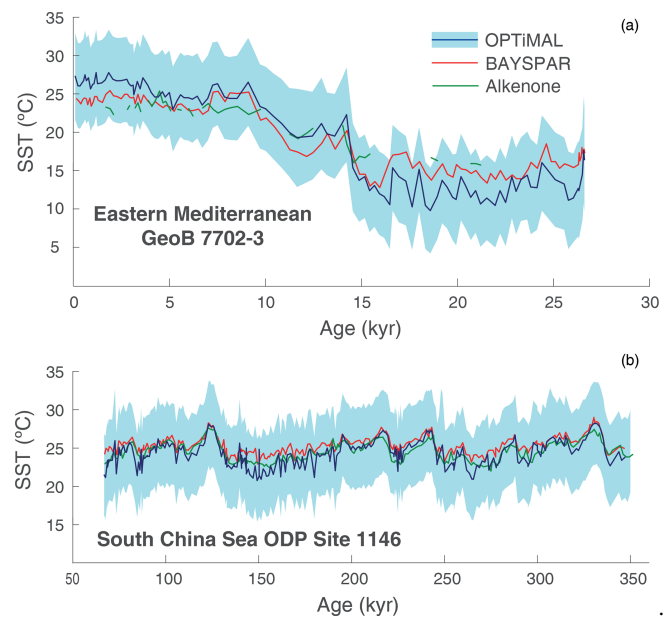


Figure 16. Late Pleistocene to Holocene GDGT-derived OPTiMAL palaeotemperatures compared to BAYSPAR and U_{37}^k SSTs. Shaded regions represent reported 5th and 95th percentile confidence intervals. (a) Eastern Mediterranean data from core GeoB 7702-3 (Castañeda et al., 2010); (b) South China Sea data from ODP Site 1146 (Thomas et al., 2014)

GDGT assemblages formed at high temperatures, beyond the range of the modern calibration data.

To investigate these behaviours requires the publication of the full range GDGT abundance data. Whilst key compilations of Eocene and Cretaceous GDGT data have strongly encouraged the release of such datasets (Lunt et al., 2012; Dunkley Jones et al., 2013; Inglis et al., 2015; O’Brien et al., 2017), most Neogene studies only publish TEX_{86} values. Without full GDGT assemblage data neither OPTiMAL nor other detailed assessments of GDGT behaviour and type can be made, and we would strongly encourage authors, reviewers and editors to ensure the publication of full GDGT assemblages in future.

Finally, to test the behaviour of OPTiMAL within established SST time series, we provide three examples: two from the late Pleistocene to Holocene (Fig. 16) and one from the Eocene (Fig. 17). For the Pleistocene to Holocene examples OPTiMAL SSTs are shown against estimates from BAYSPAR and the alkenone-based U_{37}^k temperature proxy. The first of these time series is from GeoB 7702-3 in the eastern Mediterranean and spans the last 26 kyr, including data spanning Termination I (Castañeda et al., 2010). The second is from ODP Site 1146 in the South China Sea and spans the last 350 kyr (Thomas et al., 2014). In both records the long-term dynamics are consistent between the independent U_{37}^k SST proxy and both BAYSPAR and OPTiMAL. In the eastern Mediterranean OPTiMAL SSTs are slightly cooler in the

Late Paleocene to early Eocene Arctic Ocean IODP Expedition 302 Hole 4A

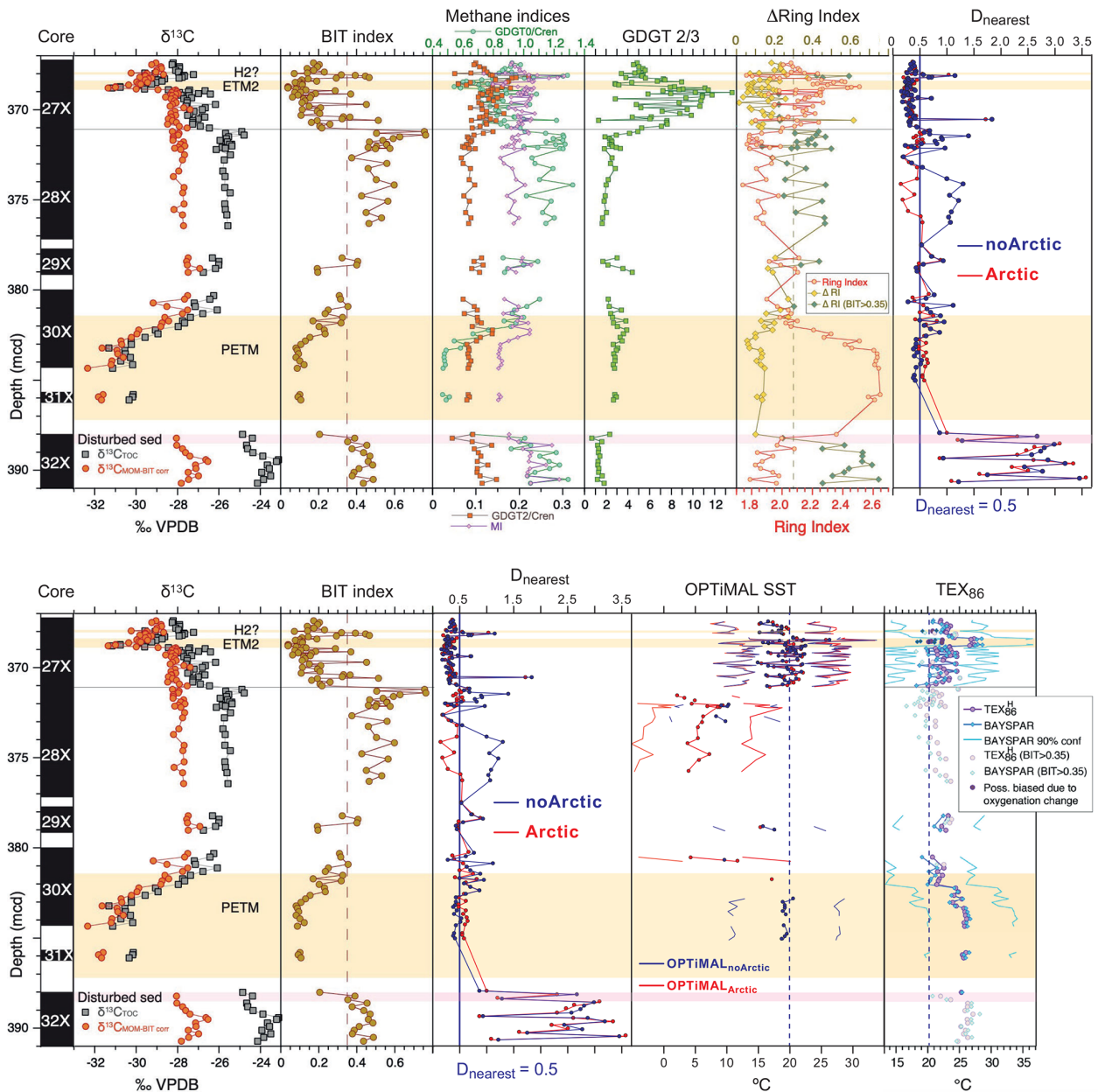


Figure 17. Comparison of GDGT screening indices, $\text{TEX}_{86}^{\text{H}}$, BAYSPAR and OPTiMAL SSTs from the Arctic Site IODP Expedition 302 Hole 4A. Depths are metres composite depth (mcd). Data and figures modified from the most recent reassessment by Sluijs et al. (2020).

glacial era and warmer in the Holocene than the other proxies. In the South China Sea, OPTiMAL is again cooler than BAYSPAR during glacial intervals but at this location is in closer agreement than BAYSPAR with the U_{37}^k SST proxy through most of the record.

The final example is from the latest Paleocene to early Eocene of IODP Expedition 302 Hole 4A on Lomonosov Ridge (Sluijs et al., 2006, 2009, 2020). This site is useful as it has been the focus of detailed reassessment and reanalysis,

using most of the available screening methodologies to detect aberrant GDGT assemblages (Sluijs et al., 2020). Here we use this recently published data to compare the new D_{nearest} screening metric against multiple other screening protocols (Fig. 17). We also show both D_{nearest} values and OPTiMAL SST estimates for two models – one with modern Arctic data with $\text{SST} < 3^{\circ}\text{C}$ included in the calibration (OPTiMAL_{Arctic}; equivalent to calibration dataset Op1 presented in this paper) and one with this data excluded (OPTiMAL_{noArctic}; equiva-

lent to the new calibration dataset Op3). It is clear from the pattern of D_{nearest} for these two options that the inclusion of modern Arctic data provides more calibration data that are closer to the Eocene palaeo-Arctic, to the extent that substantially more samples pass the $D_{\text{nearest}} < 0.5$ constraint, especially in pre-ETM2 interval from ~ 372 to 376 mcd (metres composite depth). This interval contains, however, samples with the highest BIT values of the succession (> 0.4), and elevated ΔRI (> 0.3). With these other “warning signs” concerning the reliability of GDGT assemblages for SST estimation in this interval, the relatively low D_{nearest} values are most likely to represent some similarity in the non-thermal controls on GDGT assemblages between the modern and palaeo-Arctic. More work needs to be done to constrain the reliability of temperature dependence and archaeal GDGT production in these modern high-latitude systems so that we can have confidence in their inclusion in calibration datasets for palaeo-SST estimation. It is on this basis that we recommend users of OPTiMAL use the the “noArctic” (Op3) calibration as the default. The OPTiMAL methodology does, however, offer a simple means to integrate new robust calibration data, and a method to explore the distance relationships between modern and ancient GDGT production.

Considering the “noArctic” D_{nearest} and OPTiMAL SSTs for Exp. 302 Hole 4A, it is clear that of all the screening methods, D_{nearest} shows the strongest similarity to ΔRI – with high (“failure”) values in the pre-PETM and then again between ~ 371 and 376 mcd, and even picking up the same short-lived “failure” intervals, or spikes, between 368 and 371 mcd. SST estimates based on OPTiMAL show broadly similar trends to $\text{TEX}_{86}^{\text{H}}$ and BAYSPAR, with a warm PETM, cooling post-PETM and then warming again into ETM2. It should be noted, however, that peak temperatures for OPTiMAL are $\sim 5^\circ\text{C}$ cooler than $\text{TEX}_{86}^{\text{H}}$ and BAYSPAR (e.g. PETM SSTs $< 20^\circ\text{C}$ for OPTiMAL and $> 25^\circ\text{C}$ for $\text{TEX}_{86}^{\text{H}}$ and BAYSPAR) and show more cooling post-PETM, with SST estimates of $\sim 10^\circ\text{C}$ (OPTiMAL_{noArctic}) as opposed to $\sim 20^\circ\text{C}$ for $\text{TEX}_{86}^{\text{H}}$ and BAYSPAR.

5 Conclusions

The use of GDGT abundances to estimate temperatures in clearly non-analogue conditions is, at present, problematic on the basis of the available calibration constraints or a good understanding of underlying biophysical models. We hope that this study prompts further investigations that will improve these constraints for the use of GDGTs in deep-time palaeoclimate studies, where they clearly have substantial potential as temperature proxies. Temperature estimates based on fossil GDGT assemblages that are within range of, or similar to, modern GDGT calibration data, do, however, rest on a strong, underlying temperature dependence observed in the empirical data.

In this study, we apply modern machine learning tools, including Gaussian process emulators and forward modelling, to improve temperature estimation and the representation of uncertainty in GDGT-based SST reconstructions. Using our new nearest-neighbour test, we demonstrate that $> 60\%$ of Eocene and $> 90\%$ of Cretaceous fossil GDGT distribution patterns are poorly constrained by the modern core-top calibration data. For data that do show sufficient similarity to modern data, we present OPTiMAL, a new multi-dimensional Gaussian process regression tool which uses all six GDGTs (GDGT-0, -1, -2, -3, cren and cren') to generate an SST estimate with associated uncertainty. The key advantages of the OPTiMAL approach are as follows: (1) these uncertainty estimates are intrinsically linked to the strength of the relationship between the fossil GDGT distributions and the modern calibration dataset, and (2) by considering all GDGT compounds in a multi-dimensional regression model it avoids the dimensionality reduction and loss of information that takes place when calibrating single parameters (TEX_{86}) to temperature. The methods presented above make very few assumptions about the data. We argue that such methods are appropriate with the current absence of any reasonable mechanistic model for the data-generating process, in that they reflect model uncertainty in a natural way. Finally, we note the potential for multi-proxy machine learning approaches, synthesising data from other palaeothermometers with independent uncertainties and biases, to improve calibration of ancient GDGT-derived SST reconstructions.

Data availability. All data used in this study are publicly available at <https://doi.org/10.5281/zenodo.4293851> (Greene and Mandel, 2020). Modern calibration data are taken from Tierney and Tingley (2015) with additional data from Seki et al. (2014). Eocene data are from Inglis et al. (2015), and Cretaceous data from O'Brien et al. (2017). Eocene Arctic data are from Sluijs et al. (2020). Quaternary data are from Thomas et al. (2014) and Castañeda et al. (2010).

Supplement. The supplement related to this article is available online at: <https://doi.org/10.5194/cp-16-2599-2020-supplement>.

Author contributions. TDJ and IM conceived this study; TDJ, IM, WT and YLE formulated the methodological approach. YLE and JAB provided expertise on organic biomarker synthesis, culture studies and environmental controls on GDGT production. SEG, WT, IM and YLE extensively analysed the performance of the models. SEG generated user code and “how to” support. All authors contributed to the writing of the manuscript.

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

Acknowledgements. Tom Dunkley Jones, James A. Bendle, Ilya Mandel, Kirsty Edgar and Yvette L. Eley acknowledge NERC grant NE/P013112/1. Sarah E. Greene was supported by NERC Independent Research Fellowship NE/L011050/1 and NERC large grant NE/P01903X/1. William Thomson acknowledges the Wellcome Trust (grant 1516ISSFFEL9) for funding a parameterisation workshop at the University of Birmingham (UK). William Thomson, Tom Dunkley Jones and Ilya Mandel would like to thank the BB-SRC UK Multi-Scale Biology Network grant no. BB/M025888/1. Ilya Mandel is a recipient of the Australian Research Council Future Fellowship, FT190100574. We are grateful for the extensive and constructive reviews of Yige Zhang, Jessica Tierney, Peter Bijl and an anonymous reviewer, the comments of Huan Yang, and the great patience of Editor Alberto Reyes. We would also like to give special thanks to Elizabeth Thomas and Isla Castañeda for providing primary GDGT data during a global pandemic when access to laboratories was far from straightforward.

Financial support. This research has been supported by the Natural Environment Research Council (grant nos. NE/P013112/1, NE/P01903X/1, and NE/L011050/1), the Biotechnology and Biological Sciences Research Council (grant no. BB/M025888/1), the Wellcome (grant no. 1516ISSFFEL9), and the Australian Research Council Future Fellowship (grant no. FT190100574).

Review statement. This paper was edited by Alberto Reyes and reviewed by Yige Zhang, Jessica Tierney, Peter Bijl, and one anonymous referee.

References

- Aitchison, J.: The Statistical Analysis of Compositional Data, *J. R. Stat. Soc. Series B Stat. Methodol.*, 44, 139–160, 1982.
- Aitchison, J.: Principal component analysis of compositional data, *Biometrika*, 70, 57–65, 1983.
- Aitchison, J. and Greenacre, M.: Biplots of compositional data, *J. R. Stat. Soc. Ser. C Appl. Stat.*, 51, 375–392, 2002.
- Álvarez, M. A., Rosasco, L., and Lawrence, N. D.: Kernels for Vector-Valued Functions: A Review, *Foundations and Trends® in Machine Learning*, 4, 195–266, 2012.
- Bale, N. J., Palatinszky, M., Rijpstra, I. C., Herbold, C. W., Wagner, M., and Sinnighe Damste, J. S.: Membrane lipid composition of the moderately thermophilic ammonia-oxidizing Archaeon “*Candidatus Nitrosotenus uzonensis*” at different growth temperatures, *Appl. Environ. Microb.*, 85, e01332-19, <https://doi.org/10.1128/AEM.01332-19>, 2019.
- Bijl, P. K., Schouten, S., Sluijs, A., Reichart, G.-J., Zachos, J. C., and Brinkhuis, H.: Early Palaeogene temperature evolution of the southwest Pacific Ocean, *Nature*, 461, 776–779, 2009.
- Bijl, P. K., Bendle, J. A. P., Bohaty, S. M., Pross, J., Schouten, S., Tauxe, L., Stickley, C., McKay, R. M., Röhl, U., Olney, M., Sluijs, A., Escutia, C., Brinkhuis, H., and Expedition 318 Scientists: Eocene cooling linked to early flow across the Tasmanian Gateway, *P. Natl. Acad. Sci. USA*, 110, 9645–9650, 2013.
- Blaga, C. I., Reichart, G. J., Heiri, O., and Sinninghe Damsté, J. S.: Tetraether membrane lipid distributions in water-column particulate matter and sediments: a study of 47 European lakes along a north–south transect, *J. Paleolimnol.*, 41, 523–540, <https://doi.org/10.1007/s10933-008-9242-2>, 2009.
- Brassell, S. C.: Climatic influences on the Paleogene evolution of alkenones, *Paleoceanography*, 29, 255–272, <https://doi.org/10.1002/2013PA002576>, 2014.
- Brinkhuis, H., Schouten, S., Collinson, M. E., Sluijs, A., Damsté, J. S. S., Dickens, G. R., Huber, M., Cronin, T. M., Onodera, J., Takahashi, K., Bujak, J. P., Stein, R., van der Burgh, J., Eldrett, J. S., Harding, I. C., Lotter, A. F., Sangiorgi, F., Cittert, H. V. K.-V., de Leeuw, J. W., Matthiessen, J., Backman, J., Moran, K., and the Expedition Scientists: Episodic fresh surface waters in the Eocene Arctic Ocean, *Nature*, 441, 606–609, 2006.
- Castañeda, I. S., Schefuß, E., Pätzold, J., Sinninghe Damsté, J. S., Weldeab, S., and Schouten, S.: Millennial-scale sea surface temperature changes in the eastern Mediterranean (Nile River Delta region) over the last 27,000 years, *Paleoceanography*, 25, PA1208, <https://doi.org/10.1029/2009PA001740>, 2010.
- Coifman, R. R., Lafon, S., Lee, A. B., Maggioni, M., Nadler, B., Warner, F., and Zucker, S. W.: Geometric diffusions as a tool for harmonic analysis and structure definition of data: diffusion maps, *P. Natl. Acad. Sci. USA*, 102, 7426–7431, 2005.
- Coulston, J. W., Blinn, C. E., and Thomas, V. A.: Approximating prediction uncertainty for random forest regression models, *Photogramm. Eng. Rem. S.*, 82, 189–197, <https://doi.org/10.14358/PERS.82.3.189>, 2016.
- Cramwinckel, M. J., Huber, M., Kocken, I. J., Agnini, C., Bijl, P. K., Bohaty, S. M., Frieling, J., Goldner, A., Hilgen, F. J., Kip, E. L., Peterse, F., van der Ploeg, R., Röhl, U., Schouten, S., and Sluijs, A.: Synchronous tropical and polar temperature evolution in the Eocene, *Nature*, 559, 382–386, 2018.
- De Rosa, M., Esposito, E., Gambacorta, A., Nicolaus, B., and Bu’Lock, J. D.: Effects of temperatures on ether lipid composition of *Caldariella acidophila*, *Phytochemistry*, 19, 827–831, 1980.
- Douglas, P. M. J., Affek, H. P., Ivany, L. C., Houben, A. J. P., Sijp, W. P., Sluijs, A., Schouten, S., and Pagani, M.: Pronounced zonal heterogeneity in Eocene southern high-latitude sea surface temperatures, *P. Natl. Acad. Sci. USA*, 111, 6582–6587, 2014.
- Dunkley Jones, T., Lunt, D. J., Schmidt, D. N., Ridgwell, A., Sluijs, A., Valdes, P. J., and Maslin, M.: Climate model and proxy data constraints on ocean warming across the Paleocene–Eocene Thermal Maximum, *Earth-Sci. Rev.*, 125, 123–145, 2013.
- Egozcue, J. J., Pawłowsky-Glahn, V., Mateu-Figueras, G., and Barceló-Vidal, C.: Isometric Logratio Transformations for Compositional Data Analysis, *Math. Geol.*, 35, 279–300, 2003.
- Elling, F. J., Könneke, M., Lipp, J. S., Becker, K. W., Gagen, E. J., and Hinrichs, K.-U.: Effects of growth phase on the membrane lipid composition of the thaumarchaeon *Nitrosopumilus maritimus* and their implications for archaeal lipid distributions in the marine environment, *Geochim. Cosmochim. Ac.*, 141, 5790–5797, 2014.
- Elling, F. J., Könneke, M., Mußmann, M., Greve, A., and Hinrichs, K.-U.: Influence of temperature, pH, and salinity on membrane lipid composition and TEX₈₆ of marine planktonic thaumarchaeal isolates, *Geochim. Cosmochim. Ac.*, 171, 238–255, 2015.
- Elling, F. J., Konnecke, M., Nicol, G. W., Stieglmeier, M., Bayer, B., Spieck, E., de la Torre, J. R., Becker, K. W., Thomm, M. Prosser,

- J. I., Herndl, G., Schleper, C., and Hinrichs, K.-U.: Chemotaxonomic characterisation of the thaumarchaeal lipidome, *Environ. Microbiol.*, 19, 2681–2700, 2017.
- Filzmoser, P., Hron, K., and Reimann, C.: Principal component analysis for compositional data with outliers, *Environmetrics*, 20, 621–632, 2009a.
- Filzmoser, P., Hron, K., Reimann, C., and Garrett, R.: Robust factor analysis for compositional data, *Comput. Geosci.*, 35, 1854–1861, 2009b.
- Filzmoser, P., Hron, K., and Reimann, C.: Interpretation of multivariate outliers for compositional data, *Comput. Geosci.*, 39, 77–85, 2012.
- Gliozzi, A., Paoli, G., De Rosa, M., and Gambacorta, A.: Effect of isoprenoid cyclization on the transition temperature of lipids in thermophilic archaeobacteria, *BBA-Biomembranes*, 735, 234–242, 1983.
- Greene, S. and Mandel, I.: carbonatefan/OPTiMAL: First release. (Version v1.0.0), Zenodo, <https://doi.org/10.5281/zenodo.4293851>, 2020.
- Haghverdi, L., Buettner, F., and Theis, F. J.: Diffusion maps for high-dimensional single-cell analysis of differentiation data, *Bioinformatics*, 31, 2989–2998, 2015.
- Haghverdi, L., Büttner, M., Wolf, F. A., Buettner, F., and Theis, F. J.: Diffusion pseudotime robustly reconstructs lineage branching, *Nat. Methods*, 13, 845–848, 2016.
- Haslett, J., Whitley, M., Bhattacharya, S., Salter-Townshend, M., Wilson, S. P., Allen, J. R. M., Huntley, B., and Mitchell, F. J. G.: Bayesian palaeoclimate reconstruction, *J. Roy. Stat. Soc. A*, 169, 395–438, 2006.
- Herfort, L., Schouten, S., Boon, J. P., and Sinninghe Damsté, J. S.: Application of the TEX₈₆ temperature proxy to the southern North Sea, *Org. Geochem.*, 37, 1715–1726, 2006.
- Hertzberg, J. E., Schmidt, M. W., Bianchi, T. S., Smith, R. K., Shields, M. R., and Marcantonio, F.: Comparison of eastern tropical Pacific TEX₈₆ and Globigerinoides ruber Mg/Ca derived sea surface temperatures: Insights from the Holocene and Last Glacial Maximum, *Earth Planet. Sc. Lett.*, 434, 320–332, 2016.
- Hollis, C. J., Handley, L., Crouch, E. M., Morgans, H. E., Baker, J. A., Creech, J., Collins, K. S., Gibbs, S. J., Huber, M., and Schouten, S.: Tropical sea temperatures in the high-latitude South Pacific during the Eocene, *Geology*, 37, 99–102, 2009.
- Hollis, C. J., Taylor, K. W. R., Handley, L., Pancost, R. D., Huber, M., Creech, J. B., Hines, B. R., Crouch, E. M., Morgans, H. E. G., Crampton, J. S., Gibbs, S., Pearson, P. N., and Zachos, J. C.: Early Paleogene temperature history of the Southwest Pacific Ocean: Reconciling proxies and models, *Earth Planet. Sc. Lett.*, 349–350, 53–66, 2012.
- Hollis, C. J., Dunkley Jones, T., Anagnostou, E., Bijl, P. K., Cramwinckel, M. J., Cui, Y., Dickens, G. R., Edgar, K. M., Eley, Y., Evans, D., Foster, G. L., Frieling, J., Inglis, G. N., Kennedy, E. M., Kozdon, R., Lauretano, V., Lear, C. H., Littler, K., Lourens, L., Meckler, A. N., Naafs, B. D. A., Pälike, H., Pancost, R. D., Pearson, P. N., Röhl, U., Royer, D. L., Salzmann, U., Schubert, B. A., Seebeck, H., Sluijs, A., Speijer, R. P., Stassen, P., Tierney, J., Tripathi, A., Wade, B., Westerhold, T., Witkowski, C., Zachos, J. C., Zhang, Y. G., Huber, M., and Lunt, D. J.: The DeepMIP contribution to PMIP4: methodologies for selection, compilation and analysis of latest Paleocene and early Eocene climate proxy data, incorporating version 0.1 of the DeepMIP database, *Geosci. Model Dev.*, 12, 3149–3206, <https://doi.org/10.5194/gmd-12-3149-2019>, 2019.
- Hopmans, E. C., Weijers, J. W. H., Schefuss, E., Herfort, L., Sinninghe Damsté, J. S., and Schouten, S.: A novel proxy for terrestrial organic matter in sediments based on branched and isoprenoid tetraether lipids, *Earth Planet. Sc. Lett.*, 224, 107–116, 2004.
- Huguet, C., Kim, J.-H., Sinninghe Damsté, J. S., and Schouten, S.: Reconstruction of sea surface temperature variations in the Arabian Sea over the last 23 kyr using organic proxies (TEX₈₆ and UK₃₇), *Paleoceanography*, 21, PA3003, <https://doi.org/10.1029/2005PA001215>, 2006.
- Hurley, S. J., Elling, F. J., Könneke, M., Buchwald, C., Wankel, S. D., Santoro, A. E., Lipp, J. S., Hinrichs, K., and Pearson, A.: Influence of ammonia oxidation rate on thaumarchaeal lipid composition and the TEX₈₆ temperature proxy, *P. Natl. Acad. Sci. USA*, 113, 7762–7767, 2016.
- Inglis, G. N., Farnsworth, A., Lunt, D., Foster, G. L., Hollis, C. J., Pagani, M., Jardine, P. E., Pearson, P. N., Markwick, P., Galsworthy, A. M. J., Raynham, L., Taylor, K. W. R., and Pancost, R. D.: Descent toward the Icehouse: Eocene sea surface cooling inferred from GDGT distributions, *Paleoceanography*, 30, 1000–1020, 2015.
- Jenkyns, H. C., Schouten-Huibers, L., Schouten, S., and Sinninghe Damsté, J. S.: Warm Middle Jurassic–Early Cretaceous high-latitude sea-surface temperatures from the Southern Ocean, *Clim. Past*, 8, 215–226, <https://doi.org/10.5194/cp-8-215-2012>, 2012.
- Kim, J.-H., Schouten, S., Hopmans, E. C., Donner, B., and Sinninghe Damsté, J. S.: Global sediment core-top calibration of the TEX₈₆ paleothermometer in the ocean, *Geochim. Cosmochim. Ac.*, 72, 1154–1173, 2008.
- Kim, J.-H., van der Meer, J., Schouten, S., Helmke, P., Willmott, V., Sangiorgi, F., Koç, N., Hopmans, E. C., and Sinninghe Damsté, J. S.: New indices and calibrations derived from the distribution of crenarchaeal isoprenoid tetraether lipids: Implications for past sea surface temperature reconstructions, *Geochim. Cosmochim. Ac.*, 74, 4639–4654, 2010.
- Linnert, C., Robinson, S. A., Lees, J. A., Bown, P. R., Perez-Rodriguez, I., Petrizzo, M. R., Falzoni, F., Littler, K., Antonio Arz, J., and Russell, E. E.: Evidence for global cooling in the Late Cretaceous, *Nat. Commun.*, 5, 1–7, 2014.
- Liu, Z., Pagani, M., Zinniker, D., DeConto, R., Huber, M., Brinkhuis, H., Shah, S. R., Leckie, R. M., and Pearson, A.: Global cooling during the Eocene-Oligocene climate transition, *Science*, 323, 1187–1190, <https://doi.org/10.1126/science.1166368>, 2009.
- Lunt, D. J., Dunkley Jones, T., Heinemann, M., Huber, M., LeGrande, A., Winguth, A., Loptson, C., Marotzke, J., Roberts, C. D., Tindall, J., Valdes, P., and Winguth, C.: A model-data comparison for a multi-model ensemble of early Eocene atmosphere-ocean simulations: EoMIP, *Clim. Past*, 8, 1717–1736, <https://doi.org/10.5194/cp-8-1717-2012>, 2012.
- Mentch, L. and Hooker, G.: Quantifying Uncertainty in Random Forests via Confidence Intervals and Hypothesis Tests, *J. Mach. Learn. Res.*, 17, 1–41, 2016.
- O’Brien, C. L., Robinson, S. A., Pancost, R. D., Sinninghe Damsté, J. S., Schouten, S., Lunt, D. J., Alsenz, H., Bornemann, A., Bottini, C., Brassell, S. C., Farnsworth, A., Forster, A., Huber, B. T.,

- Inglis, G. N., Jenkyns, H. C., Linnert, C., Littler, K., Markwick, P., McAnena, A., Mutterlose, J., Naafs, B. D. A., Püttmann, W., Sluijs, A., van Helmond, N. A. G. M., Vellekoop, J., Wagner, T., and Wrobel, N. E.: Cretaceous sea-surface temperature evolution: Constraints from TEX₈₆ and planktonic foraminiferal oxygen isotopes, *Earth-Sci. Rev.*, 172, 224–247, 2017.
- Park, E., Hefter, J., Fischer, G., and Mollenhauer, G.: TEX₈₆ in sinking particles in three eastern Atlantic upwelling regimes, *Org. Geochem.*, 124, 151–163, 2018.
- Pearson, P. N., van Dongen, B. E., Nicholas, C. J., Pancost, R. D., Schouten, S., Singano, J. M., and Wade, B. S.: Stable warm tropical climate through the Eocene Epoch, *Geology*, 35, 211–214, 2007.
- Polik, C. A., Elling, F. J., and Pearson, A.: Impacts of Paleocology on the TEX₈₆ Sea Surface Temperature Proxy in the Pliocene-Pleistocene Mediterranean Sea, *Paleoceanogr. Paleocl.*, 33, 1472–1489, 2018.
- Qin, W., Carlson, L. T., Armbrust, E. V., Stahl, D. A., Devol, A. H., Moffett, J. W., and Ingalls, A. E.: Confounding effects of oxygen and temperature on the TEX₈₆ signature of marine Thaumarchaeota, *P. Natl. Acad. Sci. USA*, 112, 10979–10984, 2015.
- Rasmussen, C. E. and Nickisch, H.: Gaussian Processes for Machine Learning (GPML) Toolbox, *J. Mach. Learn. Res.*, 11, 3011–3015, 2010.
- Rasmussen, C. E. and Williams, C. K. I.: Gaussian Processes for Machine Learning, MIT Press, Cambridge, MA, USA, ISBN 026218253X, 2006.
- Schouten, E., Hopmans, E. C., Forster, A., Van Breugel, Y., Kuypers, M. M. M., and Sinninghe Damsté, J. S.: Extremely high seasurface temperatures at low latitudes during the middle Cretaceous as revealed by archaeal membrane lipids, *Geology*, 31, 1069–1072, 2003.
- Schouten, S., Hopmans, E. C., Schefuß, E., and Sinninghe Damsté, J. S.: Distributional variations in marine crenarchaeotal membrane lipids: a new tool for reconstructing ancient sea water temperatures?, *Earth Planet. Sc. Lett.*, 204, 265–274, 2002.
- Schouten, S., Forster, A., Panoto, F. E., and Sinninghe Damsté, J. S.: Towards calibration of the TEX₈₆ palaeothermometer for 20 tropical sea surface temperatures in ancient greenhouse worlds, *Org. Geochem.*, 38, 1537–1546, 2007.
- Schouten, S., Hopmans, E. C., Baas, M., Boumann, H., Standfest, S., Könneke, M., Stahl, D. A., and Sinninghe Damsté, J. S.: Intact membrane lipids of “*Candidatus Nitrosopumilus maritimus*”, a cultivated representative of the cosmopolitan mesophilic group I crenarchaeota, *Appl. Environ. Microb.*, 74, 2433–2440, <https://doi.org/10.1128/AEM.01709-07>, 2008.
- Schouten, S., Hopmans, E. C., and Sinninghe Damsté, J. S.: The organic geochemistry of glycerol dialkyl glycerol tetraether lipids: A review, *Org. Geochem.*, 54, 19–61, 2013.
- Seki, O., Bendle, J. A., Harada, N., Kobayashi, M., Sawada, K., Moossen, H., and Sakamoto, T.: Assessment and calibration of TEX₈₆ paleothermometry in the Sea of Okhotsk and sub-polar North Pacific region: Implications for paleoceanography, *Progr. Oceanogr.*, 126, 254–266, 2014.
- Sinninghe Damsté, J. S., Hopmans, E. C., Schouten, S., Van Duin, A. C. T., and Geenevasen, J. A. J.: Crenarchaeol: the characteristic core glycerol dibiphytanyl glycerol tetraether membrane lipid of cosmopolitan pelagic crenarchaeota, *J. Lipid Res.*, 43, 1641–1651, <https://doi.org/10.1194/jlr.M200148-JLR200>, 2002.
- Sinninghe Damsté, J. S., Schouten, S., and Verschuren, D.: Distribution of tetraether lipids in the 25-ka sedimentary record of Lake Challa: extracting reliable TEX₈₆ and MBT/CBT palaeotemperatures from an equatorial African lake, *Quaternary Sci. Rev.*, 50, 43–54, <https://doi.org/10.1016/j.quascirev.2012.07.001>, 2012.
- Sluijs, A., Schouten, S., Pagani, M., Woltering, M., Brinkhuis, H., Sinninghe Damsté, J. S., Dickens, G. R., Huber, M., Reichart, G., Stein, R., Matthiessen, J., Lourens, L. J., Pedentchouk, N., Backman, J., Moran, K., and the Expedition 320 Scientists: Sub-tropical arctic ocean temperatures during the Palaeocene/Eocene thermal maximum, *Nature*, 441, 610–613, 2006.
- Sluijs, A., Schouten, S., Donders, T. H., Schoon, P. L., Rohl, U., Reichart, G.-J., Sangiorgi, F., Kim, J.-H., Sinninghe Damsté, J. S., and Brinkhuis, H.: Warm and wet conditions in the Arctic region during Eocene Thermal Maximum 2, *Nat. Geosci.*, 2, 777–780, 2009.
- Sluijs, A., Frieling, J., Inglis, G. N., Nierop, K. G. J., Peterse, F., Sangiorgi, F., and Schouten, S.: Late Paleocene–early Eocene Arctic Ocean sea surface temperatures: reassessing biomarker paleothermometry at Lomonosov Ridge, *Clim. Past*, 16, 2381–2400, <https://doi.org/10.5194/cp-16-2381-2020>, 2020.
- Taylor, K. W., Huber, M., Hollis, C. J., Hernandez-Sanchez, M. T., and Pancost, R. D.: Re-evaluating modern and Palaeogene GDGT distributions: Implications for SST reconstructions, *Global Planet. Change*, 108, 158–174, 2013.
- Thomas, E. K., Clemens, S. C., Prell, W. L., Herbert, T. D., Huang, Y., Liu, Z., Damsté, J. S. S., Sun, Y., and Wen, X.: Temperature and leaf wax $\delta^2\text{H}$ records demonstrate seasonal and regional controls on Asian monsoon proxies, *Geology*, 42, 1075–1078, <https://doi.org/10.1130/G36289.1>, 2014.
- Tierney, J. E.: GDGT Thermometry: Lipid Tools for Reconstructing Paleotemperatures, available at: https://www.geo.arizona.edu/~jesst/resources/TierneyPSP_GDGTs.pdf (last access: 14 December 2020), 2012.
- Tierney, J. E. and Tingley, M. P.: A Bayesian, spatially-varying calibration model for the TEX₈₆ proxy, *Geochim. Cosmochim. Ac.*, 127, 83–106, 2014.
- Tierney, J. E. and Tingley, M. P.: A TEX₈₆ surface sediment database and extended Bayesian calibration, *Sci. Data*, 2, 150029, <https://doi.org/10.1038/sdata.2015.29>, 2015.
- Trommer, G., Siccha, M., van der Meer, M. T., Schouten, S., Sinninghe Damsté, J. S., Schulz, H., Hemleben, C., and Kucera, M.: Distribution of Crenarchaeota tetraether membrane lipids in surface sediments from the Red Sea, *Org. Geochem.*, 40, 724–731, <https://doi.org/10.1016/j.orggeochem.2009.03.001>, 2009.
- Weijers, J. W. H., Lim, K. L. H., Aquilina, A., Sinninghe Damsté, J. S., and Pancost, R. D.: Biogeochemical controls on glycerol dialkyl glycerol tetraether lipid distributions in sediments characterized by diffusive methane flux, *Geochem. Geophys. Geos.*, 12, Q10010, <https://doi.org/10.1029/2011GC003724>, 2011.
- Williams, C. K. I. and Rasmussen, C. E.: Gaussian processes for machine learning, MIT Press Cambridge, MA, USA, 2006.
- Wuchter, C., Schouten, S., Coolen, M. J. L., and Sinninghe Damsté, J. S.: Temperature-dependent variation in the distribution of tetraether membrane lipids of marine Crenarchaeota: Implications for TEX₈₆ paleothermometry, *Paleoceanography*, 19, PA4028, <https://doi.org/10.1029/2004PA001041>, 2004.

- Zhang, Y. G. and Liu, X.: Export Depth of the TEX₈₆ Signal, *Paleoceanogr. Paleocl.*, 33, 666–671, <https://doi.org/10.1029/2018PA003337>, 2018.
- Zhang, Y. G., Zhang, C. L., Liu, X.-L., Li, L., Hinrichs, K.-U., and Noakes, J. E.: Methane Index: a tetraether archaeal lipid 15 biomarker indicator for detecting the instability of marine gas hydrates, *Earth Planet. Sc. Lett.*, 307, 525–534, 2011.
- Zhang, Y. G., Pagani, M., and Liu, Z.: A 12-million-year temperature history of the tropical Pacific Ocean, *Science*, 343, 84–86, 2014.
- Zhang, Y. G., Pagani, M., and Wang, Z.: Ring Index: A new strategy to evaluate the integrity of TEX₈₆ paleothermometry, *Paleoceanography*, 31, 220–232, 2016.