Response to Interactive comment on **"OPTiMAL: A new machine learning approach for GDGT-based palaeothermometry"** ***by*** **Yvette L. Eley et al.**

Dear Dr Reyes,

Thank you for your letter and for the reviews of our resubmission. We are pleased to note the "accept as is" judgement of Anonymous Reviewer 1, and their assessment of the manuscript as both excellent in its scientific significance and quality.

Here we address the points raised by new reviewer, Dr Tierney, which follow up on our response to her open discussion comments on our manuscript. Line numbers in our responses refer to the new "tracked changes" document.

Dr Tierney's comments below are highlighted in bold with our responses in plain text.

Kind regards,

Dr Tom Dunkley Jones, on behalf of all the co-authors.

General comments:

This is a re-review of Eley et al., who present the OPTIMAL model for converting GDGT distributions to SST. In the first round of review I brought up a number of issues which came down to 1) tone of the paper, 2) concerns about over-fitting and/or mathematical clarity and 3) lack of applications. In reading the revised version I think that the tone is now much more appropriate and respectful (with a few small exceptions noted in the specific comments below). I am still concerned about overfitting. The authors argue that their in-sample validation technique (leave out a portion of the coretop data for validation iteratively) is proof enough that the method is not overfitted. This is not the case, because they are still validating on the training dataset (the coretop data). Out-of-sample validation is the gold standard. This is one reason why I suggested last time that they add a number of applications of OPTIMAL to downcore datasets and compare the results to previous calibrations. They now provide an example in Figure 14 (application to ODP 806 and ODP 850), but this Figure needs comparison with previous calibrations (i.e. BAYSPAR) so that the readers can see the differences. It is not an ideal choice b/c the alkenones at ODP 806 are saturated for much of the record. The authors need to add more examples beyond this one. I suggest 1 high-resolution late Quaternary record, and 1 deeper time record (Eocene perhaps). There should be data out there that have all of the fractional GDGTs, or else they can email the authors and get this. Note that Figures 13 and 14 are not really helpful - it is very difficult for readers to take much away from these. Paleoceanographers need to see applications. This is standard for new calibration papers and not, as the authors claim, out of the scope of the study. Only through this out-of-sample testing can we assess whether OPTIMAL is performing well.

We have included three downcore records as requested – two from the late Pleistocene, with direct comparisons against independent alkenone-based $U^{k'}_{37}$ SSTs, and one from the Eocene. These are shown in new Figures 16 and 17 and discussed in Section 4 of the revised manuscript.

Regarding mathematical clarity, the paper is still lacking on this front. I noticed there are more details in the SOM - this should all be in the main text, including Fig. S1 which is incredibly helpful. I suggest reorganizing their Methods section so that everything is much more clear and straightforward: describe the math, define the terms.

The revised manuscript is already extensive, running to over 13,000 words, and it is our judgment that the methodological details of the Forward Model – which is included as an exploratory way of trying to provide improved calibrations, but is not a part of the final OPTiMAL GPR model – is better placed within the Supplementary Materials. These methods are clearly laid out in the Supplement for those who want to follow this detail and explore this model further, whereas inclusion in the text would likely reduce the accessibility of the core message and purpose of this manuscript from the general reader of *Climates of the Past*.

It is also now evident that inversion of the GPR model is Bayesian. This should be made explicit in the main text along with a description of the priors used for SST.

Dr Tierney is confused – the inversion is only relevant to the forward modelling not the GPR model. Only the GPR model is involved in generating the OPTiMAL SST estimates. This supports our contention to keep the forward model detailed methodology in the SOM to prevent confusion with the primary GPR model (OPTiMAL).

It also seems that the model can in fact be extrapolated (?) but that the uncertainties will be high unless an informative prior is made.

The forward model can be extrapolated, and uncertainties will be large, unless "an informativ prior is made". That is exactly our point – that SST estimation out of range of the modern calibration, is largely controlled by the choice of "an informative prior". We don't think there are, yet, informative priors to provide confidence outside of the modern calibration range, but these might become available through either further culture or mesocosm studies, or paleo proxy-proxy calibrations.

Could you use constraints from other proxies here? That might provide some hope for deep-time users.

Yes, that could be attempted, but is beyond the scope of this paper. It would require a detailed assessment of all the uncertainty within other proxy systems for these constraints to be robustly integrated into a calibration model for GDGT paleothermometry. It would also involve the loss of independence between proxy systems.

All in all, it will be much easier to understand the steps that the authors took if the math is laid out in detail, along with (as I asked for last time) a description of what platforms and codes were used (Matlab, Python packages).

The underlying maths is laid out in detail within the paper; the full details of the coding are provided in GitHub: [https://github.com/carbonatefan/OPTiMAL.](https://github.com/carbonatefan/OPTiMAL)

Also, my questions about collinearity (the fact that the GDGT predictors are not independent but in fact correlated with each other to a high degree) and possible regresssion dilution in the GPR model (or some other problem - maybe with the prior) were not explicitly addressed.

The machine learning methods we propose do not require inputs to be independent. Techniques such as Gaussian process regression learn from the data. If some inputs are uninformative (either because they are not correlated with the property we are trying to predict, or because they can be reconstructed from other inputs and therefore do not add value), they are recognised as such. They may not help, but they do not harm, either.

While we are, of course, sensitive to measurement noise in the independent variables, we do not perform linear regression and do not suffer from regression dilution in the usual sense of this term. Variability in both input and output variables is included in the uncertainty of the prediction.

GDGT abundance are definitely correlated, and that's expected — but not so correlated that you can effectively reduce the information to a single dimension, as TEX attempts to do. Here is the matrix of Spearman rank correlation coefficients (rows and columns are, in order, GDGT-0, GDGT-1, GDGT-2, GDGT-3, Cren, Cren'):

1.0000	-0.4165	-0.7898	-0.7986	-0.8937	-0.7266
-0.4165	1.0000	0.7807	0.4340	0.0364	0.6335
-0.7898	0.7807	1.0000	0.7443	0.4835	0.8844
-0.7986	0.4340	0.7443	1.0000	0.6690	0.6460
-0.8937	0.0364	0.4835	0.6690	1.0000	0.4399
-0.7266	0.6335	0.8844	0.6460	0.4399	1.0000

A lack of independence in the GDGT inputs is simply not an issue for Gaussian process regression (see C. E. Rasmussen & C. K. I. Williams, *Gaussian Processes for Machine Learning*, the MIT Press, 2006).

I welcome the inclusion of a comparison of Dnearest to BIT and MI, but deltaRI should be here too. In practice this is one of the most valuable metrics for identifying aberrant GDGT distributions.

Done.

Finally, it seems like the authors misunderstood my comment last time about providing some first-order constraints on their model. I was advocating for enforcing monotonicity, i.e. more rings = higher temperature. I was not advocating for a particular form for that (linear vs. non-linear). A monotonic constraint still seems appropriate to me. Couldn't this be built into OPTIMAL? Why would one not want to do so? The authors mention culture studies, most of which find more rings at higher T's. Even Bale et al. 2019 find more cren and cren' at higher T's, as expected. It seems like we have enough evidence in favor of monotonicity.

"*Why would one not want to do so?*" – because it would build-in unnecessary assumptions about the behaviour of the relationship into the calibration model. If this relationship is present in the data, our approach will find it; if it is not, then our approach won't be biased by a prior user-assumption about the form of the temperature dependency.

If the authors can revise their paper to make the math clear, discuss any limitations as appropriate, and demonstrate that OPTIMAL can perform reasonably well on out-of-sample data (downcore time series) that would make me (and I suppose all readers) much more comfortable in terms of using this new approach.

With the new examples that we have provided, we hope we have provided some of this assurance, and look forward to Dr Tierney making use of OPTiMAL in near future.

Specific comments:

Line 131: "Like Qin et al. (2015), we note the non-linear nature of the individual experiments in Wuchter et al. (2004; see Fig. 5)." Need to clarify here that you mean Wuchter et al's Figure 5 and not your own. However this statement is deceiving. What Figure 5 in that paper shows is no response of TEX to SST b/t 5-15 degrees, and then a linear response thereafter in the series I incubation, plus no response in series II. It is not a non-linear (i.e. exponential) relationship. Combined with Schouten '07, the mesocosm data are linear b/t 10C and 40C (Schouten et al., 2007, Figure 4). I agree that this does not preclude a non-linear relationship in the real world - but it's important to not misstate what the data show.

We added a direct reference to the figure. We have not changed our text – if experiments show no or varied response of TEX_{86} to temperature across the temperature range, as Tierney notes in this case, then we think it is valid to say that this is a "non-linear" response.

Line 150: "As such, these are collectively more representative of the community production contributing to samples in the global core-top TEX86 calibrations of Kim et al., (2010) and BAYSPAR (Tierney & Tingley, 2014), which predominantly sample continental margin environments, rather than deep ocean / pelagic environments." I don't agree with this. How could a single-strain culture be more representative that environmental samples, which likely reflect multiple strains? There is no way that it could be. We don't know what strains contribute to the coretop dataset but it is certainly more diverse than just N. maritimus. Please remove.

This is a misreading of the sentence. The sentence meaning is that culture strains recovered from epi-continental shelf environments are more likely to represent the strains contributing to production of GDGTs going into core top calibrations, which are from the same epi-continental margin environments. However, this is a minor point, and to avoid confusion we have deleted these lines.

Line 164: "To use the responses of single, selected archaeal strains in culture to validate

a particular model of community-level responses to growth temperature is problematic even in the modern system (Elling et al., 2015)." I agree, and this directly contradicts the statement on Line 150.

See comment above – it was a misreading of this section, but it is now removed.

Line 195" "Powerful mathematical tools." I asked you to please delete this last time, as it is hyperbolic and non-specific. The tools here are no more powerful than other mathematical approaches. Replace with "an analysis of distance metrics", "machine learning" "GPR" or something similar.

Yes – and we disagree and retain. These machine learning tools presented are considerably more powerful at modelling complex relationships within multi-dimensional datasets than, say, linear regression. They are powerful tools.

Line 252: "For example, it may be that sea surface temperatures are very sensitive to one observable." I think you mean the reverse here - the observables are the GDGTs, and they might be more or less sensitive to SST. This paragraph would be easier to understand if you just use the term "each GDGT" vs. "observable".

Changed for clarity.

Eq. 7: The description of this distance metric is not totally clear. Can you clarify what x and y are here? Also as pointed out by Yi Ge, there are only five degrees of freedom, so doesn't this need to be adjusted accordingly?

Line 259: "*Thus, the normalised distance D between parameter data points x and y is*:"

Points x and y are samples within the GDGT space; $D_{x,y}$ is the normalised Eucledian distance between these two points. The sum in Eq. 7 should be taken over 6 parameters (GDGT-0, GDGT-1, GDGT-2, GDGT-3, cren, cren') and we apologize for a typo in the equation where the sum should be taken from 0 to 5, not 0 to 6. This has been corrected.

Lines 289-300: To what extent is the gain in information from the individual GDGTs due simply to the use of more than one parameters? As I pointed out in my first review, adding more parameters will improve RMSE but doesn't necessarily mean an improvement in skill. This should be addressed here. You can't say that the NN predictor "outperforms" TEX unless you rule out the effect of using more parameters, which incidentally are also not independent from each other.

There are no free parameters in the nearest neighbour approach: the predictor is just the temperature of the closest training set point in input space. And, indeed, as we say, the gain lies precisely in having more information available in 6 (5 independent) coordinates of the input point than in their one-dimensional combination, TEX**.**

Line 345: This R^2 of 0.83 is identical to what BAYSPAR can do with all of the data (R^2 = 0.84, Figure 5 in TT14). So it seems like both the random forest and BAYSPAR model perform similarly, even though BAYSPAR uses TEX86. Worth noting.

BAYSPAR computes R^2 by using the same data for calibration and validation, which allows for over-fitting and over-predicts R^2**.**

Line 372: The authors haven't answered my question yet about the effects of collinearity (the correlation of the fractional GDGT abundances with each other). This seems like a good place to clarify this issue.

See points above, but have clarified in the main text (line 379-381): "*The accuracy of Gaussian process regression is not adversely affected by correlations between inputs (Rasmussen & Williams, 2006). Significantly correlated inputs that do not bring in new predictive power are appropriately down-weighted*."

Line 425: The DCs look like they are showing evidence of the "horseshoe effect" common to standard PCA, in that branches B and C are part of the same horseshoe. This would signal a strong linear response in the multivariate data that isn't well-separated (?) When this occurs in standard PCA, the PCs are not interpretable. Can the authors comment here on this as it applies to the diffusion map technique?

PCA is fundamentally a linear transformation — a rotation of the coordinates. The so-called horseshoe effect is an indication of nonlinear correlations, which the PCA cannot account for. Diffusion maps are nonlinear dimensionality reduction tools, and so are generally capable of handling nonlinear correlations.

Line 477: "it does not account for the systematic uncertainty in the model when extrapolated beyond the calibration range" It does in that all possible regression lines are extrapolated, creating wide error bars and coalescing towards the prior if no other information is there. I would change to, "it still assumes a monotonic relationship between TEX and SST" which is a more accurate. However I don't think this is a bad assumption (see comment above).

Rephrased. Our point here is that BAYSPAR assumes that the linear model is intrinsically correct, and only uncertainties on the parameters within the model are accounted for, rather than any systematic uncertainty in the model itself.

Lines 481-486: So if I understand this correctly, the inversion of your model is Bayesian inference with priors placed on...what? This section needs some mathematical description to make this clear.

GPR need not be strictly Bayesian, though the calibration process can be viewed as Bayesian inference on the hyper-parameters of the model. For example, with the squared exponential kernel with automatic relevance determination that we use, these hyper-parameters are the kernel widths in each dimension. The priors referred to (e.g., in "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") are not the classical Bayesian inference priors on model parameters of interest to the user that the Dr Tierney may be thinking of**.**

Line 488: "The clear pattern in the residuals does not necessarily indicate model misspecification, since no explicit noise model is specified for temperatures". I noted this last time - this looks like regression dilution. But if the model was specified as TEX = f(SST) + error this shouldn't happen, unless...the prior is too tight? Please clarify what your priors are.

TEX is not used here at all. Regression dilution is simply not relevant as described above.

Line 521: This is the first mention of Dnearest, but it is not defined. Is Dnearest the same as D(x,y) described above? Please clarify.

No – as explained lines 523 - 525

Line 569: Bale et al. did observe an increase in cren and cren' in their culture experiments though, which suggests that there is some response of GDGTs to temperature, albeit not well-expressed in TEX86.

Fine. We state there is no correlation with TEX_{86}.

Line 592: Figures 13 and 14 don't really communicate well how the Optimal model performs. It would be more useful to show a couple of time series and compare Optimal vs. BAYSPAR.

Have added timeseries as requested.

Line 610: It's not necessary to italicize "parametric" here. I would rephrase this to more accurately describe the difference b/t BAYSPAR and Optimal: "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 still makes an assumption of a linear increase in SST at higher TEX values." As I said last time, BAYSPAR does account for model uncertainty, the issue is that the model we use is a linear one.

Parametric no longer italicized; rephrased in line with suggestion.

Line 615: deltaRI should be included here. It is arguably the most useful metric for ID'ing strange GDGT distributions.

Done.

Figure 16: I don't think this is an ideal application b/c UK37 is at its limit here at 806. Plus optimal should be compared to the previous calibrations (BAYSPAR, TEXH if you want, although the regression dilution in TEXH makes things hard to interpret).

Done

SOM: This should be in the main text, esp. Figure 1. Also the SOM appears to contain comments and incomplete references (refs).

See comments above about the forward model.

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

Tom Dunkley Jones[1], Yvette L. Eley[1], William Thomson[2], Sarah E. Greene[1], Ilya Mandel[3,4,5], Kirsty Edgar[1] and James A. Bendle[1]

Affiliations:

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

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

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

[4]The ARC Centre of Excellence for Gravitational Wave Discovery — OzGrav, Australia

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

* Responses to Tierney's comments and additions are shown in blue text.

Abstract

In the modern oceans, the relative abundances of Glycerol dialkyl glycerol tetraether (GDGTs) 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 TEX_{86} 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 data set; and second, the problem of making temperature estimates beyond the range of the modern empirical calibrations (>30 ºC). Here we apply modern machine-learning tools, including Gaussian Process Emulators and forward modelling, to develop a new mathematical approach we call OPTiMAL (**O**ptimised **P**alaeothermometry from **T**etraethers v**i**a **MA**chine **L**earning) 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 data set. We reduce the root mean square uncertainty on temperature predictions (validated using the modern data set) from ~± 6 ºC using TEX_{86} based estimators to ± 3.6 ºC using Gaussian Process estimators for temperatures below 30 ºC. 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. Finally, we advocate caution in the use of temperature estimates beyond the range of the modern empirical calibration dataset, given the lack of a robust predictive biological model or extensive and reproducible mesocosm experimental data in this elevated temperature range.

1. Introduction

Glycerol dibyphytanyl 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 TEX_{86} (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 GDGT-x determined by liquid chromatography mass spectrometery (LC-MS) peak area, and cren' is the peak area of the isomer of crenarchaeol) (Schouten et al., 2002; Liu et al. 2018) and was shown to be positively correlated with mean annual SSTs:

$$TEX_{86} = (\text{GDGT-2} + \text{GDGT-3} + \text{cren'}) / (\text{GDGT-1} + \text{GDGT-2} + \text{GDGT-3} + \text{cren'}) \quad \text{(Eq. 1)}$$

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; Slujis et al., 2009). The TEX_{86} approach also extended beyond the range of the widely used alkenone-based $U^{k'}_{37}$ thermometer, in both temperature space, where $U^{k'}_{37}$ saturates at ~28ºC (Brassell, 2014; Zhang et al., 2017), 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):

$$TEX_{86} = 0.015 * SST + 0.287 \quad \text{(Eq. 2)}$$

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 – TEX_{86}^{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 ºC, particularly when SSTs are lower than 15 ºC), and TEX_{86}^{H} (Eq. 4), also exponential, and recommended for use when SSTs exceeded 15 ºC (Kim et al., 2010). TEX_{86}^{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).

$$TEX_{86}^{L} = log\left(\frac{[GDGT2]}{[GDGT1]+[GDGT2]+[GDGT3]}\right) \qquad \text{Eq. 3}$$

$$TEX_{86}^{H} = log\left(\frac{[GDGT2]+[GDGT3\}+[Cren']}{[GDGT1]+[GDGT2]+[GDGT3]+[Cren']}\right) \text{ Eq. 4}$$

Despite the recommendations of Kim et al. (2010), both TEX_{86}^{H} and TEX_{86}^{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 2012 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). The rationale was that both TEX_{86}^{L} and TEX_{86}^{H} were calibrated across a full temperature range, with the exception of the inclusion or exclusion of Red Sea core-top data. The difference in model fit between the two proxy formulations to the calibration dataset was also minor (Kim et al. 2010). In certain environments, however, TEX_{86}^{L} was subject to significant variability in derived temperatures that were not apparent in TEX_{86}^{H} (Taylor et al., 2013). This was mostly due to changing GDGT2 to GDGT3 ratios, which strongly influence TEX_{86}^{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, TEX_{86}^{L} is no longer regarded as an appropriate tool for palaeotemperature reconstructions, except in limited Polar conditions (Kim et al., 2010; Tierney, 2012).

Three fundamental issues have troubled the TEX_{86} proxy. The first is a concern about undetected non-analogue palaeo-GDGT assemblages, for which the modern calibration data set 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, 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 recognized that reconstructed temperatures beyond the range of the modern calibration (>30 ºC), 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 with mean annual SSTs over 20 ºC. 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 & Tingley, 2015) dependency of TEX_{86} on SSTs, and the effect of these models on temperature estimates over 30 ºC.

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. 2015; 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). Wuchter et al. (2004) and Schouten et al. (2007) show a compiled linear calibration of TEX_{86} against incubation temperature (up to 40ºC in the case of Schouten et al., 2007) based on strains that were enriched from surface seawater collected from the North Sea and Indian Ocean respectively. Like Qin et al. (2015), we note the *non*-linear nature of the individual experiments in Wuchter et al. (see Fig. 5 in Wuchter et al. 204). Moreover, the relatively lower Cren' in these studies yield a very different intercept and slope compared to core-top calibrations (e.g. Kim et al. 2010) making direct comparisons problematic.

More recently, Elling et al. (2015) studied three different strains (*N. maritimus*, NAOA6, NAOA2) isolated from open ocean surface waters (South Atlantic) whilst Qin et al., (2015) studied a culture of *N. maritimus* and three *N. maritimus*-like strains isolated from Puget Sound. All strains are of marine, mesophilic, Thaumarchaeota within Marine Group 1 (equivalent to Crenarchaeota Group 1). Both of these papers clearly demonstrate distinctly different responses of membrane lipid composition to temperature in these strains, whilst Qin et al. (2015) additionally show that oxygen concentration is at least as important as temperature in controlling TEX_{86} values in culture. The impact of Thaumarchaeota community change on TEX_{86} in palaeoclimate studies is further suggested by the downcore study of Polik et al (2019). All of these culture studies, made on marine, mesophilic archaea demonstrate how community composition may have a significant impact on measured environmental TEX_{86} signatures. ~~In these cases (e.g., Zhang et al. 2015; Qin et al., 2015; Elling et al., 2015) cultured strains of Thaumarcheota were obtained from surface waters which overlie the epi-continental or continental shelf regions of the North Sea, Indian Ocean, South Atlantic and North Pacific - in addition to the pure culture strain *N. maritimus* in Qin et al. (2015) and Elling et al. (2015). As such, these are collectively more representative of the community production contributing to samples in the global core-top TEX_{86} calibrations of Kim et al., (2010) and BAYSPAR (Tierney & Tingley, 2014), which predominantly sample continental margin environments, rather than deep ocean / pelagic environments.~~

It is clear from the above discussion that there is evidence for more complex responses in GDGT-production to growth temperature in some instances, and across distinct strains of archaea (Elling et al., 2015). More fundamentally, 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). For instance, a multiproxy study of Mediterranean Pliocene-Pleistocene sapropels indicates that specific distributions of archaeal lipids might be reflective of temporal changes in thaumarchaeael communities rather than temperature alone (Polik et al., 2018). Indeed, the potential influence of community switching on GDGT composition can be seen in mesocosm studies, with different species preferentially thriving at different growth temperatures (e.g., Schouten et al., 2007). To use the responses of single, selected archaeal strains in culture to validate a particular model of community-level responses to growth temperature is problematic even in the modern system (Elling et al., 2015). 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 Ring Index (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, the original uses of the TEX_{86} proxy had a relatively poor representation of the true uncertainty associated with palaeotemperature estimates, as they included no assessment of non-analogue behavior 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 SSTs 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' (**O**ptimised **P**alaeothermometry from **T**etraethers v**i**a **MA**chine **L**earning) 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 supplementary information.

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 data set 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:

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

(Eq. 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 standard error. It is useful to compare the accuracy of the predictor to the standard deviation of all temperatures in the data set σT, which corresponds to using the mean temperature as the predictor in Equation 1; for the modern data set, σT= 10.0 ºC. 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 data set. 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

$$D_{x,y}^2 \equiv \sum_{i=0}^{5} \frac{(GDGT_i(x) - GDGT_i(y))^2}{var(GDGT_i)}$$

(Eq. 7)

We show the distribution of nearest distances of points in the modern data set, 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 data set, i.e., one that minimises $D_{x,y}$. Fig. 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 $\delta T_{\text{nearest}} = 4.5$ ºC. This is less than half of the standard deviation σT in the temperature values across the modern data set. 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. Fig. 3 illustrates both the advantage and disadvantage of TEX$_{86}$. On the one hand, there is a clear correlation between TEX$_{86}$ and temperature (top panel of Fig. 3), 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 $\delta T_{\text{nearestTEX86}} = 8.0$ ºC (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 data set.

Indeed, this is what we find when we consider predictors of the form $\hat{T}_{1/\mathrm{TEX}} = a + b/\mathrm{TEX}_{86}$ and $\hat{T}_{\mathrm{TEXH}} = c + d\ \log_{\mathrm{TEX86}}$ (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 data sets (least squares regression). We find that $\delta T_{1/\mathrm{TEX}}$ = 6.1 ºC (note that this is slightly better than using the fixed values of *a* and *b* from (Kim et al., 2010), which yield $\delta T_{1/\mathrm{TEX}}$ = 6.2 ºC). 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 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 20º x 20º 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 data set 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 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 data set. 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 δT = 4.1 ºC degrees, 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 on the prediction (Mentch and Hooker, 2016), although this is possible with, e.g., 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. Typically, 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 term 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, signaling 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 & 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-Newton approximation to the Hessian. Prediction accuracy is shown in Figure 5, and we find that δT = 3.72 ºC, 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 & 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 regional effects in the TEX_{86}-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 (Section 2.3) and GPR approaches (Section 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 optimized interpolation tools for mapping correlations between temperature and GDGT abundances within the range of the modern calibration data set; they can make no sensible inference about the behavior 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 (Aitcheson, 1982, 1983; Aitcheston and Greenacre, 2002; Filzmoser et al., 2009a; Filzmoser et al., 2009b; Filzmoser et al., 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-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 Figure 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 is essential for understanding ancient climates (particularly Cretaceous climates).

The relationship between the first diffusion component and TEX_{86} for all data is shown in Figure 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 (Figure 9). We see a similar pattern emerging to that displayed by TEX_{86} - there is little sensitivity to temperature below 15 ºC, and between ~20 and 25 ºC. 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 Figure 4 between 0 and 15-17 degrees celsius, 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 developed 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 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 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 Matern 3/2 kernel for the applications below. Note, however, that since we are working in ilr-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 Figure 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 (Figure 4) with two clear downward slopes, suggesting again that the data are clustered into temperatures above and below 16-17 ºC, 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 and lower limit on temperatures (see Supplementary Information). 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 ºC).

3. Non-analogue behavior 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/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 data set. We strongly recommend the use of the weighted distance metric ($D_{nearest}$) as a screening method to determine whether the modern core 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 $D_{nearest}$ >0.5, *regardless of the calibration model or approach applied*, are unlikely to generate temperature estimates that are much better than informed guesswork. In these instances, in both our GPR and Fwd models, the constraints provided by the modern calibration data set are so weak that estimates of temperature have large uncertainty bands that are dictated by model priors; i.e. are unconstrained by the calibration data (e.g., Figure 13 and Figure 14). This uncertainty is not apparent from estimates generated by BAYSPAR or TEX_{86}^{H} models, although the underlying and fundamental lack of constraints are the same. While 93% of validation data points in the modern data have $D_{nearest}$ <0.5, this is the case for only 33% of Eocene samples and 3% for Cretaceous samples.

Where ancient GDGT distributions lie far from the modern calibration data set ($D_{nearest} > 0.5$), 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 propose that this is more rigorous and justified model behavior than extrapolation of TEX_{86} or BAYSPAR predictors to non-analogue samples far from the modern calibration data. As a result, the predictive models can only be applied to a subset of the Eocene and Cretaceous data. We also note that there are two broad, non-mutually-exclusive categories of samples that lie far from the modern calibration dataset ($D_{nearest} > 0.5$), 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 distributions lying on the margins of, or far away from the main GDGT clustering in 6-dimensional space (see outliers in Fig. 8).

Given the (current) limit on natural mean annual surface ocean temperatures of ~30 ºC, 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 temperatures 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 ~30ºC, however, the behavior 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, completely 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 work, characterised by more 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., core-top or Wuchter et al. 2004 and Schouten et al., 2008 mesocosm calibrations) with membrane lipid composition at different growth temperatures (37°C, 46°C, 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, we see no robust justification at present for the extrapolation of modern core-top calibration data sets into the unknown above 30 ºC, although the coherent patterns apparent across GDGT space, between modern, Eocene and Cretaceous data (Figure 7), do provide some grounds for hope that the extension of GDGT palaeothermometry beyond 30ºC might be possible in future.

4. OPTiMAL and $D_{nearest}$: A more robust method for GDGT-based paleothermometry

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; Section 2.4) outperforms the Forward model (Fwd; Section 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 GITHUB repository, https://github.com/carbonatefan/OPTiMAL.

Following Tierney and Tingley (2014) we use a reduced calibration data set, with the exclusion of Arctic data with observed SSTs less than 3ºC ("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 Supplementary Information; to distinguish from the original OPTiMAL calibration data, which included the Arctic data <3ºC, 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 performance of Op1 and Op3 were very similar in most applications, except in applications to the paleo-Arctic (see below), where the inclusion of modern Arctic calibration data (Op1) provided closer calibration constraints to the paleo-data. Although superficially this may be regarded as beneficial, in these instances the paleo-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, either the modern Arctic calibration data is impacted by similar non-thermal processes, generating unusual GDGT abundance patterns, which are not appropriate to use for SST calibration, or, 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 paleothermometry in this unusual environment (Sluijs et al. 2020). The $D_{nearest}$ methodology may prove useful in quantifying analogue and non-analogue behavior 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) (Figures 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 ~15 ºC (Figure 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 ºC 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 Figure 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 data set. 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 (Figure 15). For ease of comparison between Eocene and Cretaceous datasets and visualization of the majority of the data, extreme outliers ($D_{nearest}$ > 4.0) are not shown. The metrics include the BIT index

(Hopmans et al., 2004; Weijers et al., 2006), 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 paleo-compilations (O'Brien et al. 2017), are shown in the blue shaded areas on Figure 15 (BIT > 0.5; MI > 0.5; ΔRI > 0.3; %GDGT-0 > 67%) – data points within these areas fail the standard screening. Also shown on Figure 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 $D_{nearest}$. The main exception to this is the BIT index in the Eocene data set, where 15 samples have high BIT values (>0.5) but have GDGT assemblages that are close to modern analogues in the calibration dataset ($D_{nearest}$ <0.5). Of these samples, 9 are from the Arctic Ocean between the PETM and 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 anomalous 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 $D_{nearest}$ <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 TEX_{86}^{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 (5-6ºC) with the SSTs from samples that pass all other screening and $D_{nearest}$ (~4-7ºC). 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 calibration (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 $D_{nearest}$ <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 $\Delta RI < 0.3$, but which fail $D_{nearest}$ (>0.5). These data most likely represent GDGT assemblages formed at high temperatures, beyond the range of the modern calibration data.

~~These plots show little relationship between the BIT and MI screening indices and $D_{nearest}$ values. Whilst in the Eocene, samples with the highest $D_{nearest}$ values (>3) also show very elevated BIT values (>0.8), in the Cretaceous the exceptionally anomalous assemblages ($D_{nearest}$ values >100) are not anomalous in either BIT or MI. Conversely, in the Eocene there are many samples with relatively high BIT (>0.3) that are below the $D_{nearest}$ threshold of 0.5. The behaviour of these systems needs to be examined in detail in future studies, but a conservative approach would be to apply all three screening indices (BIT, MI and $D_{nearest}$) to have the most confidence in resulting temperature estimates.~~

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 behavior of OPTiMAL within established SST time series, we provide three examples two from the late Pleistocene to Holocene (Figure 16) and one from the Eocene (Figures 17 and 18). ~~, where full GDGT assemblage data were made available, and there are comparison alkenone-based $U^{k'}_{37}$ data from the same sampling location – ODP 806 and ODP 850, respectively in the West and Eastern Equatorial Pacific (Figure 16; Zhang et al. 2014)~~. For the Pleistocene to Holocene examples OPTiMAL SSTs are shown against estimates from BAYSPAR and the alkenone-based $U^{k'}_{37}$ temperature proxy. The first of these timeseries 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 $U^{k'}_{37}$ SST proxy and both BAYSPAR and OPTiMAL. In the Eastern Mediterranean OPTiMAL SSTs are slightly cooler in the glacial 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 $U^{k'}_{37}$ SST proxy through most of the record.

In both these examples, we show the 5^{th} and 95^{th} percentiles for OPTiMAL and those reported by the BAYSPAR methodology.

The final example is from the latest Paleocene to early Eocene of IODP Expedition 302 Hole 4A on Lomonosov Ridge (Sluijs et al. 2006; Sluijs et al. 2009; Sluijs et al. 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 (Figure 17). We also show both $D_{nearest}$ values and OPTiMAL SST estimates for two models – one with modern Arctic data with SST < 3ºC included in the calibration ($OPTiMAL_{Arctic}$; equivalent to calibration dataset Op1 first present by Eley et al. 2019) and one with this data excluded ($OPTiMAL_{noArctic}$; equivalent 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 paleo-Arctic, to the extent that substantially more samples pass the $D_{nearest}$ < 0.5 constraint, especially in pre-ETM2 interval from ~372 to 376 mcd. 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 paleo-Arctic. More work needs to be done to constrain the reliability of temperature-dependence and archaeal GDGT production in these modern high latitude systems before we can have confidence in their inclusion in calibration datasets for paleo-SST estimation. It is on the basis that we recommend users of OPTiMAL use the the "noArctic" (Op3) calibration for the time being. 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 TEX_{86}^{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 ~5ºC cooler than TEX_{86}^{H} and BAYSPAR (e.g. PETM SSTs <20ºC for OPTiMAL and > 25ºC for TEX_{86}^{H} and BAYSPAR), and show more cooling post-PETM, with SST estimates of ~10ºC ($OPTiMAL_{noArctic}$) as opposed to ~20ºC for TEX_{86}^{H} and BAYSPAR.

~~are not at the limit of alkenone saturation in ODP 806, OPTiMAL and $U^{k'}_{37}$ agree well in the Plio-Pleistocene. In ODP 850, there is a strong agreement in the reproduction of a long-term late Miocene to Recent cooling trend in both $U^{k'}_{37}$ and OPTiMAL of ~5ºC. There is, however a consistent ~2º offset between cooler OPTiMAL and warmer $U^{k'}_{37}$ temperatures at this location. This offset is very similar in magnitude and direction as that between TEX_{86}^{H} and $U^{k'}_{37}$ used in the original study. and is more likely due to an inherent feature (seasonality or depth) of archaeal versus eukaryotic production at this site (Zhang et al. 2014).~~

5. Conclusions

Although the fundamental issue of non-analogue behaviour is a key problem for GDGT-temperature estimation, it has an undue impact on the community's general confidence in this method. In part, this is because these issues have not been clearly stated and circumscribed - rather they have been allowed to erode confidence in the GDGT-based methodology through the use of GDGT-based palaeothermometry far outside the modern constraints on the behavior of this system. 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 paleoclimate 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. With no effective means of separating the "good from the bad" can lead to either false confidence and inappropriate inferences in non-analogue conditions, or a false pessimism when ancient samples are actually well constrained by modern core-top assemblages.

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 differ so significantly from modern as to call into question SSTs derived from these assemblages. For data that does show sufficient similarity to modern, 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: 1) that these uncertainty estimates are intrinsically linked to the strength of the relationship between the fossil GDGT distributions and the modern calibration data set, 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 palaeothermeters with independent uncertainties and biases, to improve calibration of ancient GDGT-derived SST reconstructions.

Acknowledgements:

TDJ, JAB, IM, KME and YE acknowledge NERC grant NE/P013112/1. SEG was supported by NERC Independent Research Fellowship NE/L011050/1 and NERC large grant NE/P01903X/1. WT acknowledges the Wellcome Trust (grant code: 1516ISSFFEL9, www.wellcome.ac.uk/) for funding a parameterisation workshop at the University of Birmingham (UK). WT, TDJ and IM would like to thank the BBSRC UK Multi-Scale Biology Network Grant No. BB/M025888/1. IM is a recipient of the Australian Research Council Future Fellowship, FT190100574. We would like to thank the extensive and constructive reviews of Jessica Tierney, Yige Zhang, Huan Yang and an anonymous reviewer, as well as the great patience of Editor Alberto Reyes. We would also like to give especial thanks to Elizabeth Thomas and Isla Castaneda for digging up primary GDGT data when access to labs was far from straight forward.

Figure Captions:

Figure 1. A histogram of the normalised distance to the nearest neighbour in GDGT space ($D_{x,yt}$) 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.

Figure 3. Top: The temperature of the modern data set as a function of the TEX_{86} value, showing a clear linear correlation between the two, but also significant scatter. Bottom: the error of the predictor based on the nearest TEX_{86} calibration point.

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

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

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.

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).

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

Figure 10. Temperature residuals for the forward model.

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

Figure 12. A histogram of normalised distances to the nearest sample in the modern data set 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; O'Brien et al., 2017).

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 datapoints 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.

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

Figure 15. Comparison of $D_{nearest}$ against standard screening indices, BIT and MI index, ΔRI and %GDGT-O for the Eocene (Inglis et al., 2015) and Cretaceous (O'Brien et al., 2017) datasets. Blue shaded regions show the standard cut-off points for these indices (see text); grey shaded region highlights data that are below the $D_{nearest}$ threshold of 0.5. The outlined black box is the region of data that fails traditional screening indices but passes $D_{nearest}$ (<0.5).

Figure 16. Late Pleistocene to Holocene GDGT-derived OPTiMAL palaeotemperatures compared to BAYSPAR and $U^{k'}_{37}$ SSTs. Shaded regions represent reported 5th and 95th percentile confidence intervals. Top panel - Eastern Mediterranaean data from core GeoB 7702-3 (Castaneda et al. 2010); bottom panel – South China Sea data from ODP Site 1146 (Thomas et al. 2014).

Figure 17. Comparison of GDGT screening indices, TEX_{86}^{H}, BAYSPAR and OPTiMAL SSTs from the Eocene Arctic Site IODP Expedition 302 Hole 4A. Data and figures modified from the most recent reassessment by Sluijs et al. (2020).

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., Greenacre, M.: Biplots of compositional data. J. R. Stat. Soc. Ser. C Appl. Stat. 51, 375–392, 2002.

Álvarez, M.A., Rosasco, L., 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., Sinnighe Damste, J. S.: Membrane lipid composition of the moderately thermophilic ammonia-oxidizing Archaeon "*Candidatus* Nitrosotenius uzonensis" at different growth temperatures, Applied and Environmental Microbiolpogy, DOI: 10.1128/AEM.01332-19, 2019.

Bijl, P. K., S. Schouten, A. Sluijs, G.-J. Reichart, J. C. Zachos, and H. Brinkhuis.: 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., Slujis, A., Escutia, C., Brinkhuis, H. and Expedition 318 Scientists.: Eocene cooling linked to early flow across the Tasmanian Gateway. Proc. Natl. Acad. Sci. U.S.A., 110, 9645–9650, 2013.

Brassell, S. C.: Climatic influences on the Paleogene evolution of alkenones, Paleoceanography, 29, 255-272, doi: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.

Cadillo-Quiroz, H., Didelot, X., Held, N. L., Herrera, A., Darling, A., Reno, M. L., Krause, D. J., Whitaker, R. J.: Patterns of gene flow define species of thermophilic Archaea, PLOS, Cadillo-Qiuroz, H. et al. (2012) PLOS Biology, https://doi.org/10.1371/journal.pbio.1001265, 2012.

Castañeda, I. S., E. Schefuß, J. Pätzold, J. S. Sinninghe Damsté, S. Weldeab, and S. Schouten.: Millennial-scale sea surface temperature changes in the eastern Mediterranean (Nile River Delta region) over the last 27,000 years, Paleoceanography, 25, PA1208, doi:10.1029/2009PA001740, 2010.

Coifman, R.R., Lafon, S., Lee, A.B., Maggioni, M., Nadler, B., Warner, F., Zucker, S.W.: Geometric diffusions as a tool for harmonic analysis and structure definition of data: diffusion maps. Proc. Natl. Acad. Sci. U. S. A. 102, 7426–7431, 2005.

Coulston, J.W., Blinn, C.E., Thomas, V.A.: Approximating prediction uncertainty for random forest regression models. Photogrammetric Engineering & Remote Sensing, Volume 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., Bu'Lock, J. D.: Effects of temperatures on ether lipid composition of *Caldariella acidophilia*, 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., Pagani, M.: Pronounced zonal heterogeneity in Eocene southern high-latitude sea surface temperatures. Proceedings of the National Academy of Sciences, 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-Science Reviews, 125, 123-145, 2013.

Egozcue, J.J., Pawlowsky-Glahn, V., Mateu-Figueras, G., Barceló-Vidal, C.: Isometric Logratio Transformations for Compositional Data Analysis. Math. Geol. 35, 279–300, 2003.

Eley, Y. L., Thompson, W., Greene, S. E., Mandel, I., Edgar, K., Bendle, J. A., and Dunkley Jones, T.: OPTiMAL: A new machine learning approach for GDGT-based palaeothermometry, Clim. Past Discuss., https://doi.org/10.5194/cp-2019-60, in review, 2019.

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 20 distributions in the marine environment, Geochimica et Cosmochimica Acta, 141, 579-597, 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 TEX_{86} of marine planktonic thaumarchaeal isolates, Geochimica et Cosmochimica Acta, 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., Hinrichs, K-U.: Chemotaxonomic characterisation of the thaumarchaeal lipidome, Environmental Microbiology 19, 2681–2700 , 2017.

Filzmoser, P., Hron, K., Reimann, C.: Principal component analysis for compositional data with outliers. Environmetrics 20, 621–632, 2009a.

Filzmoser, P., Hron, K., Reimann, C., Garrett, R.: Robust factor analysis for compositional data. Comput. Geosci. 35, 1854–1861, 2009b.

Filzmoser, P., Hron, K., Reimann, C.: Interpretation of multivariate outliers for compositional data. Comput. Geosci. 39, 77–85, 2012.

Gliozzi, A., Paoli, G., De Rosa, M., Gambacorta, A.: Effect of isoprenoid cyclization on the transition temperature of lipids in thermophilic archaebacteria, Biochimica et Biophysica Acta (BBA) Biomembranes, 735, 234 – 242, 1983.

Haghverdi, L., Buettner, F., 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., Theis, F.J.: Diffusion pseudotime robustly reconstructs lineage branching. Nat. Methods 13, 845–848, 2016.

Haslett, J., Whiley, M., Bhattacharya, S., Salter-Townshend, M., Wilson, S.P., Allen, J.R.M., Huntley, B., Mitchell, F.J.G.: Bayesian palaeoclimate reconstruction. J Royal Statistical Soc A 169, 395–438;, 2006.

Herfort, L., Schouten, S., Boon, J. P., and Sinninghe Damsté, J. S.: Application of the TEX_{86} temperature proxy to the southern North Sea, Organic Geochemistry, 37, 1715-1726, 2006.

Hertzberg, J. E., Schmidt, M. W., Bianchi, T. S., Smith, R. K., Shields, M. R., & Marcantonio, F.: Comparison of eastern tropical Pacific TEX_{86} and Globigerinoides ruber Mg/Ca derived sea surface temperatures: Insights from the Holocene and Last Glacial Maximum. Earth and Planetary Science Letters, 434, 320–332, 2016.

Hollis, C. J., Taylor, K. W. R., Handley, L., Pancost, R. D., Huber, M., Creech, J. B., Hines, B. R., Crouch, E. M., Morgans, 25 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 and Planetary Science Letters, 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., Meckler, N., Naafs, B. D. A., Pälike, H., Pancost, R. D., Pearson, P., Royer, D. L., Salzmann, U., Schubert, B., Seebeck, H., Sluijs, A., Speijer, R., Stassen, P., Tierney, J., Tripati, 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. Discuss., https://doi.org/10.5194/gmd-2018-309, in review, 2019.

Hollis, C. J., Handley, L., Crouch, E. M., Morgans, H. E., Baker, J. A., Creech, J., Collins, K. S., Gibbs, S. J., Huber, M., Schouten, S.: Tropical sea temperatures in the high-latitude South Pacific during the Eocene. Geology, 37, 99–102, 2009.

Hopmans, E. C., Weijers, J. W. H., Schefuss, E., Herfort, L., Sinninghe Damsté, J. S., Schouten, S.: A novel proxy for terrestrial organic matter in sediments based on branched and isoprenoid tetraether lipids, Earth and Planetary Science Letters, 224, 107-116, 2004.

Huguet C, Kim J-H, Sinninghe Damste´ J.S., Schouten S: Reconstruction of sea surface temperature variations in the Arabian Sea over the last 23 kyr using organic proxies (TEX$_{86}$ and UK 0 37 . Paleoceanography 21(3): PA3003, 2006.

Hurley, S. J., Elling, F. J., Könneke, M., Buchwald, C., Wankel, S. D., Santoro, A. E., Lipp. J.S., Hinrichs, K., Pearson, A.: Influence of ammonia oxidation rate on thaumarchaeal lipid composition and the TEX$_{86}$ temperature proxy. Proceedings of the National Academy of Sciences, 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., Damsté J.S.S.: Warm Middle Jurassic-Early Cretaceous high-latitude sea-surface temperatures from the Southern Ocean. Clim Past 8 (1):215–226, 2012.

Kim, J.-H., Schouten, S., Hopmans, E. C., Donner, B., and Sinninghe Damsté, J. S.: Global sediment core-top calibration of the TEX$_{86}$ paleothermometer in the ocean, Geochimica et Cosmochimica Acta, 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, Geochimica et Cosmochimica Acta, 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., Russell, E. E. : Evidence for global cooling in the Late Cretaceous. Nature Communications, 5, 1–7, 2014.

Liu, X-L., Zhu, C., Wakeham, S.G., Hinrichs, K-U.: In situ production of branched glycerol dialkyl glycerol tetraethers in anoxic marine water columns, Marine Chemistry, 166, 1 – 8, 2014.

Lunt, D. J., Dunkley Jones, T., Heinemann, M., Huber, M., LeGrande, A., Winguth, A., Loptson, C., Marotzke, J., Tindall, J., 15 Valdes, P., Winguth, C.: A model-data comparison for a multi-model ensemble of early Eocene atmosphere-ocean simulations: EoMIP, Clim. Past Discuss., 8, 1229-1273, 2012.

Mentch, L., Hooker, G.: Quantifying Uncertainty in Random Forests via Confidence Intervals and Hypothesis Tests. Journal of Machine Learning Research, 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, 20 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., Wrobel, N. E.: Cretaceous sea-surface temperature evolution: Constraints from TEX_{86} and planktonic foraminiferal oxygen isotopes, Earth-Science Reviews, 172, 224-247, 2017.

Park, E., Hefter, J., Fischer, G., Mollenhauer, G.: TEX_{86} in sinking particles in three eastern Atlantic upwelling regimes. Organic Geochemistry, 124, 151–163, 2018.

Pearson, P. N., van Dongen, B. E., Nicholas, C. J., Pancost, R. D., Schouten, S., Singano, J. M., Wade, B. S.: Stable warm tropical climate through the Eocene Epoch. Geology, 35, 211-214, 2007.

Polik, C. A., Elling, F. J., Pearson, A.: Impacts of Paleoecology on the TEX_{86} Sea Surface Temperature Proxy in the Pliocene-Pleistocene Mediterranean Sea. Paleoceanography and Paleoclimatology, 33, 1472–1489, 2018.

Qin, W., Amin, S. A., Martens-Habbena, W., Walker, C. B., Urakawa, H., Devol, A. H., Ingalls, A.E., Moffett, J.W., Ambrust, E.V., Stahl, D. A.: Marine ammonia-oxidizing archaeal isolates display obligate mixotrophy and wide ecotypic variation. Proceedings of the National Academy of Sciences of the United States of America, 111, 12504–12509, 2014.

Qin, W., Carlson, L. T., Armbrust, E. V., Stahl, D. A., Devol, A. H., Moffett, J. W., Ingalls, A. E.: Confounding effects of oxygen and temperature on the TEX 86 signature of marine Thaumarchaeota. Proceedings of the National Academy of Sciences, 112, 10979–10984, 2015.

Rasmussen, C.E., Nickisch, H.: Gaussian Processes for Machine Learning (GPML) Toolbox. J. Mach. Learn. Res. 11, 3011–3015, 2010.

Rasmussen, C. E. & Williams, C. K. I.: Gaussian Processes for Machine Learning, the MIT Press, 2006, ISBN 026218253X.

Sangiorgi, F., van Soelen Els, E., Spofforth David, J. A., Pälike, H., Stickley Catherine, E., St. John, K., Koç, N., Schouten, S., Sinninghe Damsté Jaap, S., Brinkhuis, H.: Cyclicity in the middle Eocene central Arctic Ocean sediment record: Orbital forcing and environmental response, Paleoceanography, 23, 10.1029/2007PA001487, 2008.

Schouten, E., Hopmans, E.C., Forster, A., Van Breugel, Y., Kuypers, M.M.M., 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., Forster, A., Panoto, F. E., and Sinninghe Damsté, J. S.: Towards calibration of the TEX_{86} palaeothermometer for 20 tropical sea surface temperatures in ancient greenhouse worlds, Organic Geochemistry, 38, 1537-1546, 2007.

Schouten, S., Hopmans, E. C., Sinninghe Damsté, J. S.: The organic geochemistry of glycerol dialkyl glycerol tetraether lipids: A review, Organic Geochemistry, 54, 19-61, 2013.

Schouten, S., Hopmans, E. C., Schefuß, E., Sinninghe Damsté, J. S.: Distributional variations in marine crenarchaeotal membrane lipids: a new tool for reconstructing ancient sea water temperatures? Earth and Planetary Science Letters, 204, 15 265-274, 2002.

Seki, O., Bendle, J. A., Harada, N., Kobayashi, M., Sawada, K., Moossen, H., Sakamoto, T.: Assessment and calibration of TEX_{86} paleothermometry in the Sea of Okhotsk and sub-polar North Pacific region: Implications for paleoceanography. Progress in Oceanography, 126, 254–266, 2014.

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: Subtropical 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., Brinkhuis, H.: Warm and wet conditions in the Arctic region during Eocene Thermal Maximum 2, Nature Geosci, 2, 777-780, 2009.

Taylor, K. W. R., Willumsen, P. S., Hollis, C. J., Pancost, R. D.: South Pacific evidence for the long-term climate impact of the Cretaceous/Paleogene boundary event, Earth-Science Reviews, 179, 287-302, 2018.

Taylor, K. W., Huber, M., Hollis, C. J., Hernandez-Sanchez, M. T., Pancost, R. D.: Re-evaluating modern and Palaeogene GDGT distributions: Implications for SST reconstructions, Global and Planetary Change, 108, 158-174, 2013.

Thomas, E. K., S. C. Clemens, W. L. Prell, T. D. Herbert, Y. Huang, Z. Liu, J. S. S. Damsté, Y. Sun, and X. Wen,: Temperature and leaf wax δ^2H records demonstrate seasonal and regional controls on Asian monsoon proxies, Geology, 42(12), 1075–1078, doi:10.1130/G36289.1, 2014.

Tierney, J. E.: GDGT Thermometry: Lipid Tools for Reconstructing Paleotemperatures. Retrieved from https://www.geo.arizona.edu/~jesst/resources/TierneyPSP_GDGTs.pdf, 2012

Tierney, J. E., and Tingley, M. P.: A Bayesian, spatially-varying calibration model for the TEX_{86} proxy. Geochimica et Cosmochimica Acta, 127, 83-106, 2014.

Tierney, J. E., and Tingley, M. P.: A TEX_{86} surface sediment database and extended Bayesian calibration, Scientific data, 2, 150029, 2015.

Williams, C.K.I., and Rasmussen, C.E.: Gaussian processes for machine learning. MIT Press Cambridge, MA, 2006.

Wuchter, C., Schouten, S., Coolen, M. J. L., and Sinninghe Damsté, J. S.: Temperature-dependent variation in the distribution 30 of tetraether membrane lipids of marine Crenarchaeota: Implications for TEX_{86} paleothermometry, Paleoceanography and Paleoclimatology. doi:10.1029/2004PA001041, 2004.

Zhang, Y. G., and Liu, X.: Export Depth of the TEX_{86} Signal. Paleoceanography and Paleoclimatology. doi.org/10.1029/2018PA003337, 2018.

Zhang, Y. G., Pagani, M., Wang, Z.: Ring Index: A new strategy to evaluate the integrity of TEX_{86} paleothermometry, Paleoceanography, 31, 220-232, 2016.

Zhang, Y. G., Zhang, C. L., Liu, X.-L., Li, L., Hinrichs, K.-U., Noakes, J. E.: Methane Index: a tetraether archaeal lipid 15 biomarker indicator for detecting the instability of marine gas hydrates, Earth and Planetary Science Letters, 307, 525- 534, 2011.

Zhang, Y.G., Pagani, M., Liu, Z.: A 12-million-year temperature history of the tropical Pacific Ocean: Science, 343, 84-86, 2014.

Zhu, J., Poulsen, C. J., Tierney, J.: Simulation of Eocene extreme warmth and high climate sensitivity through cloud feedbacks, Science Advances, 5(9), eaax1874, 2019.

Figure 1

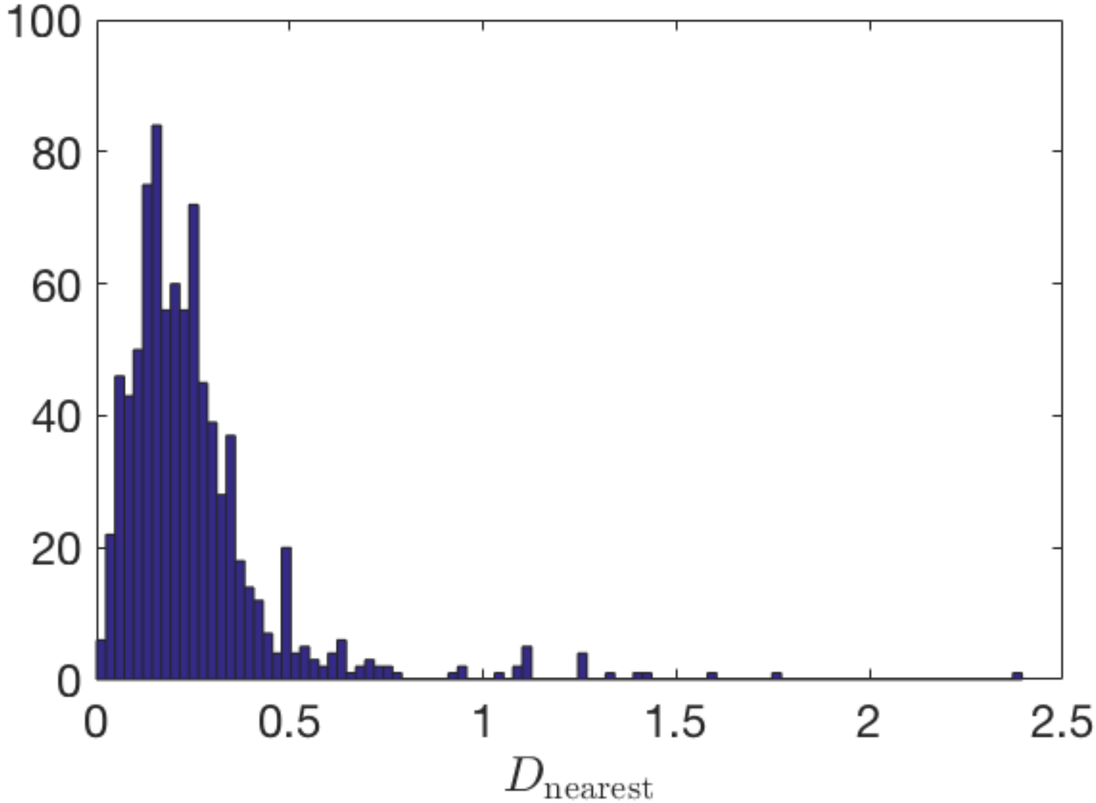

100
80
60
40
20
0
0
0.5
1
1.5
2
2.5
D_{nearest}

Figure 2

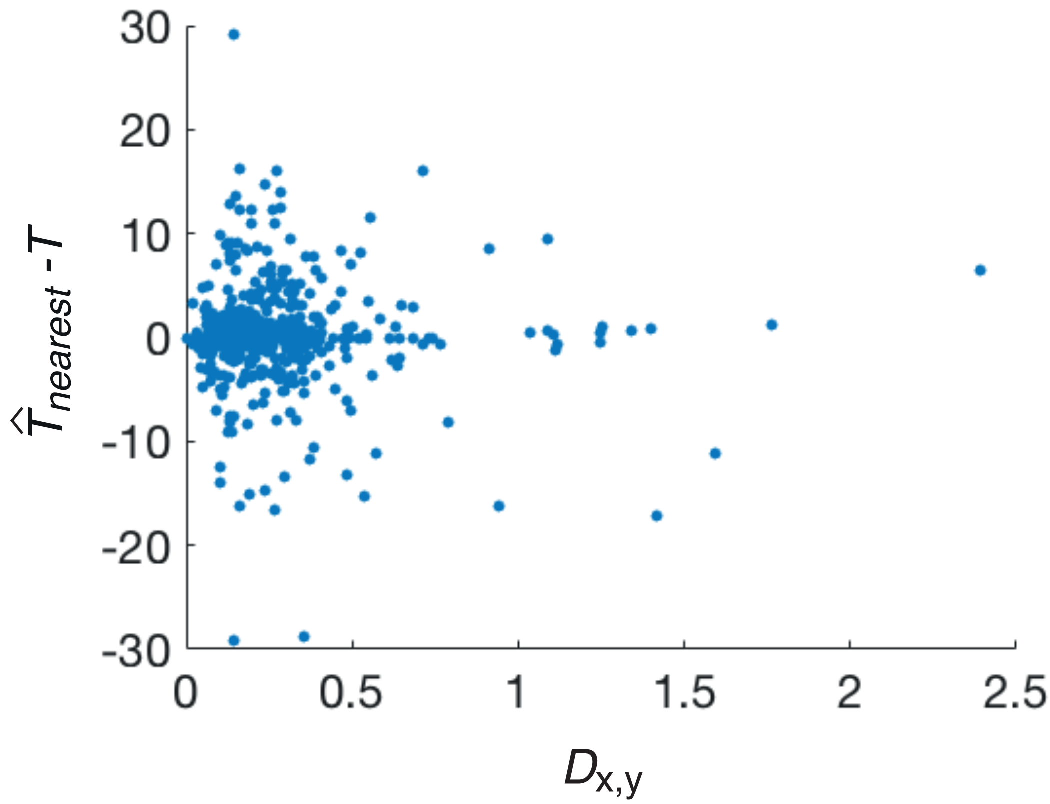
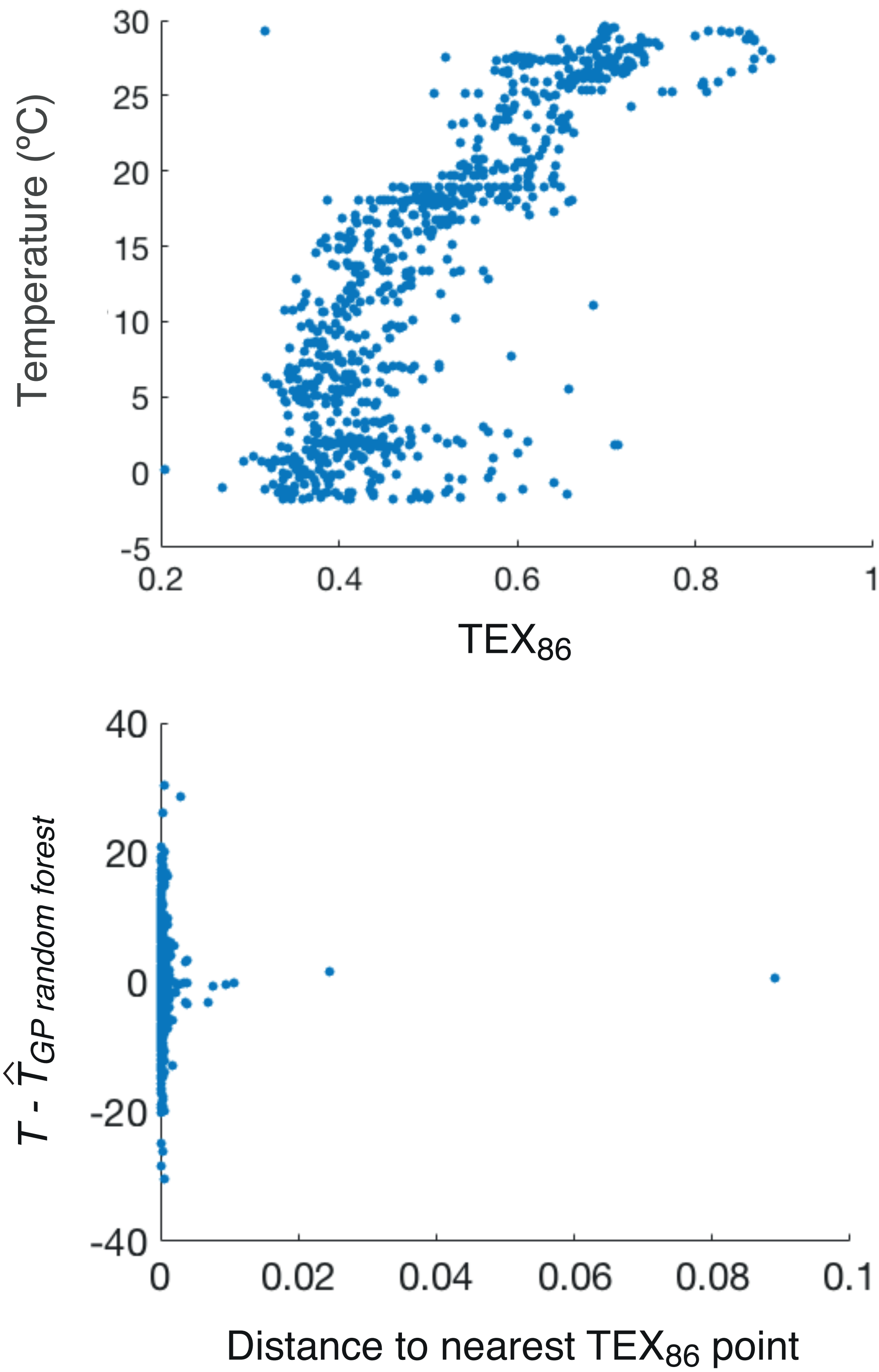
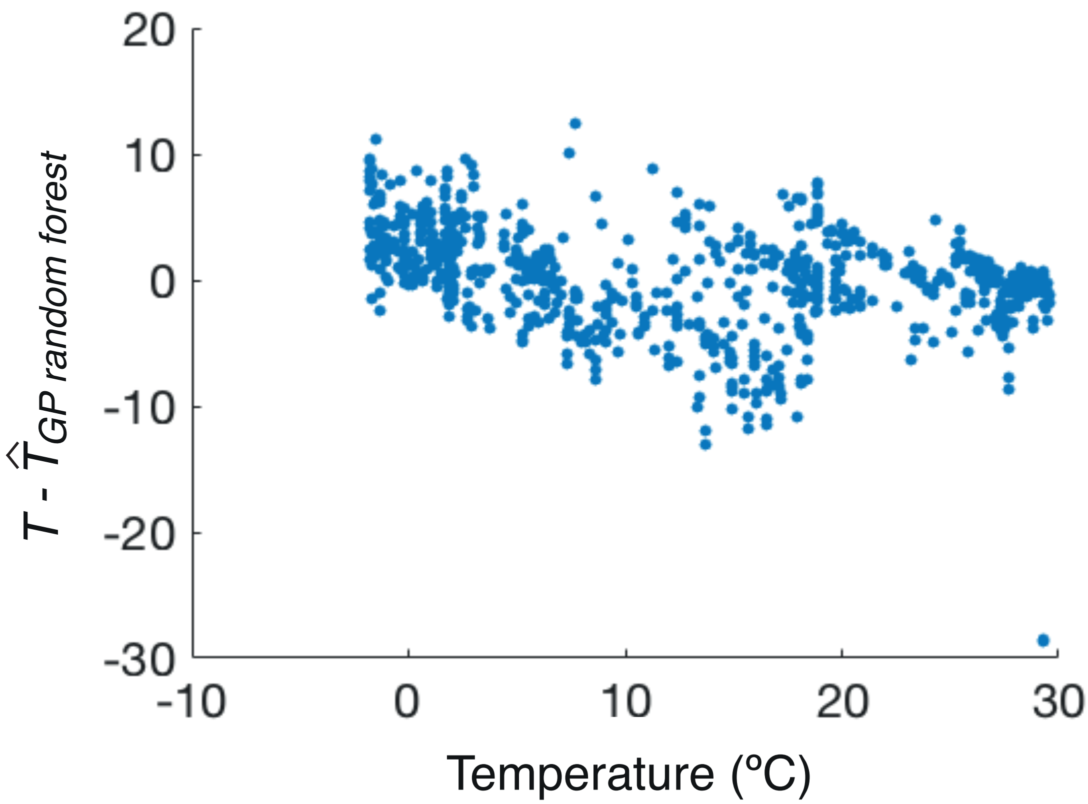


Figure 3



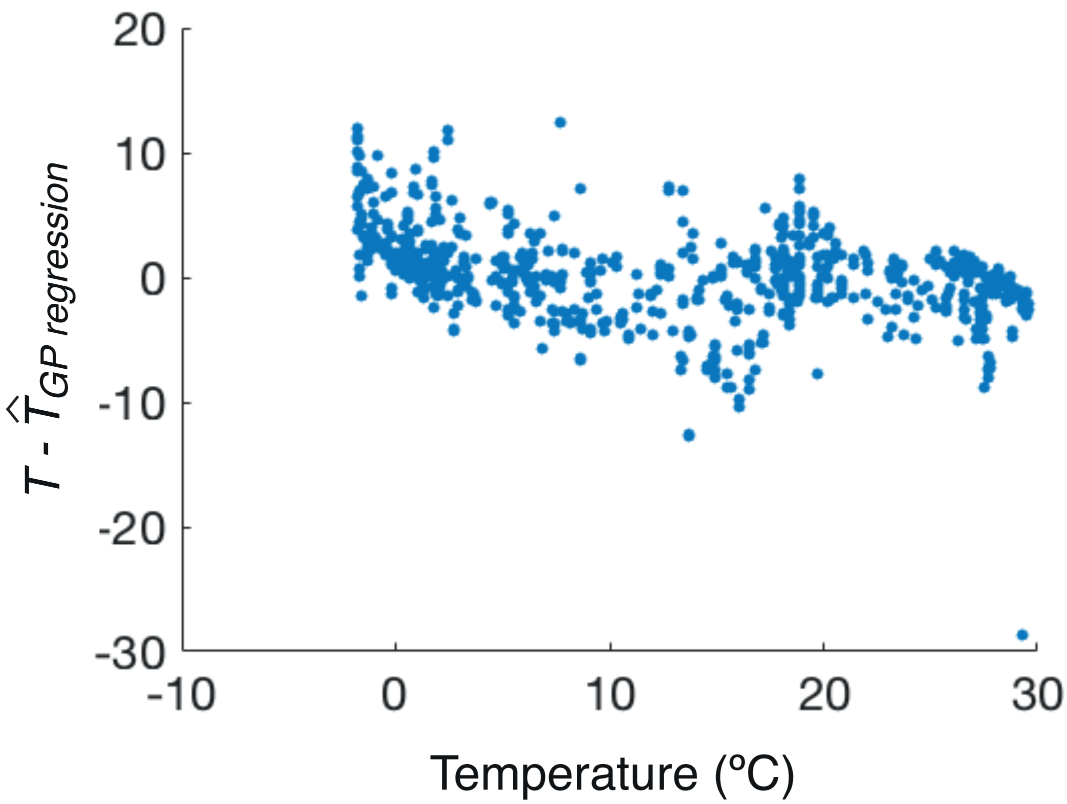
Temperature (°C)
30
25
20
15
10
5
0
-5
0.2
0.4
0.6
0.8
1
TEX86
$T - \hat{T}_{GP\ random\ forest}$
40
20
0
-20
-40
0
0.02
0.04
0.06
0.08
0.1
Distance to nearest TEX86 point

Figure 4



20
10
0
-10
-20
-30
-10
0
10
20
30
Temperature (ºC)
$T - \hat{T}_{GP\ random\ forest}$

Figure 5



20
10
0
-10
-20
-30
-10
0
10
20
30
T - $\widehat{T}_{GP\ regression}$
Temperature (ºC)

Figure 6

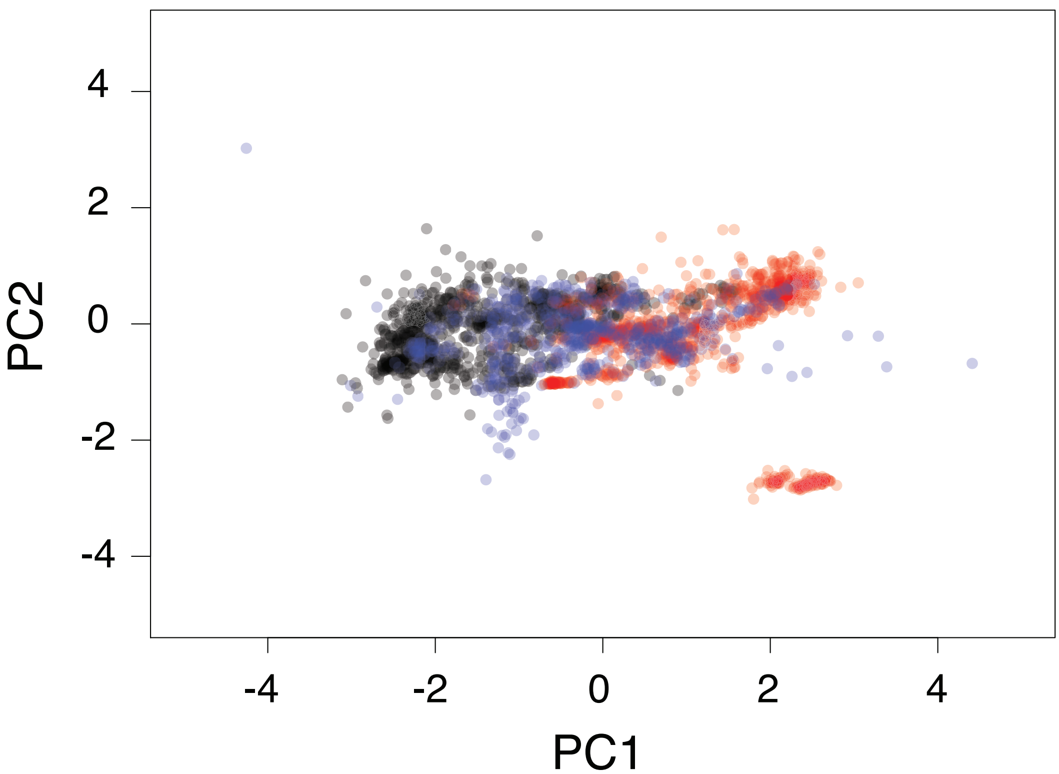


Figure 7

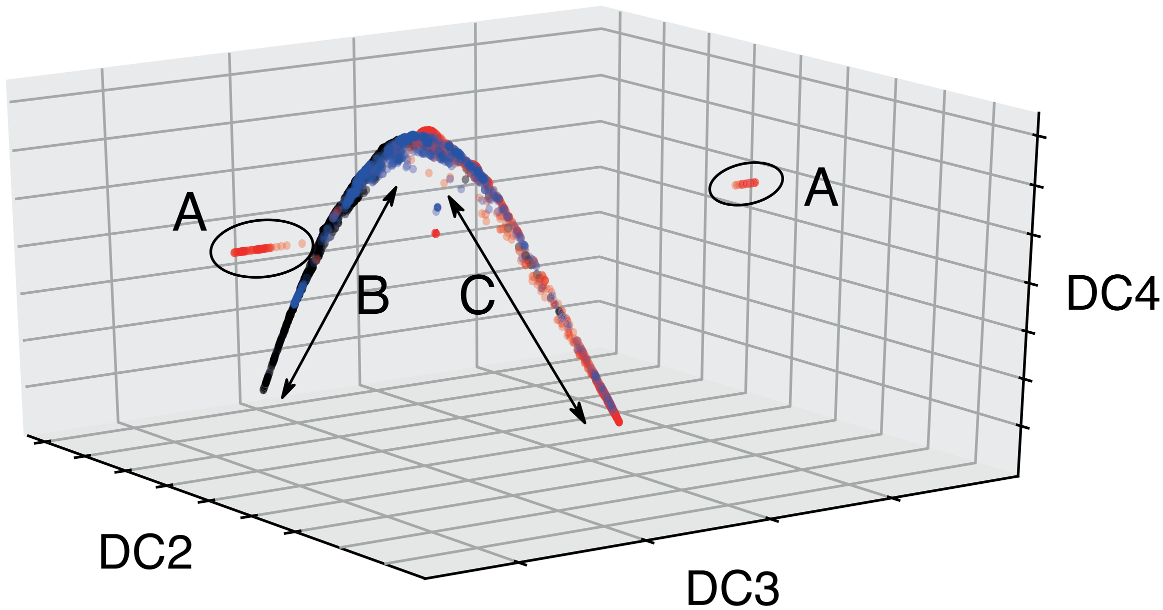


Figure 8

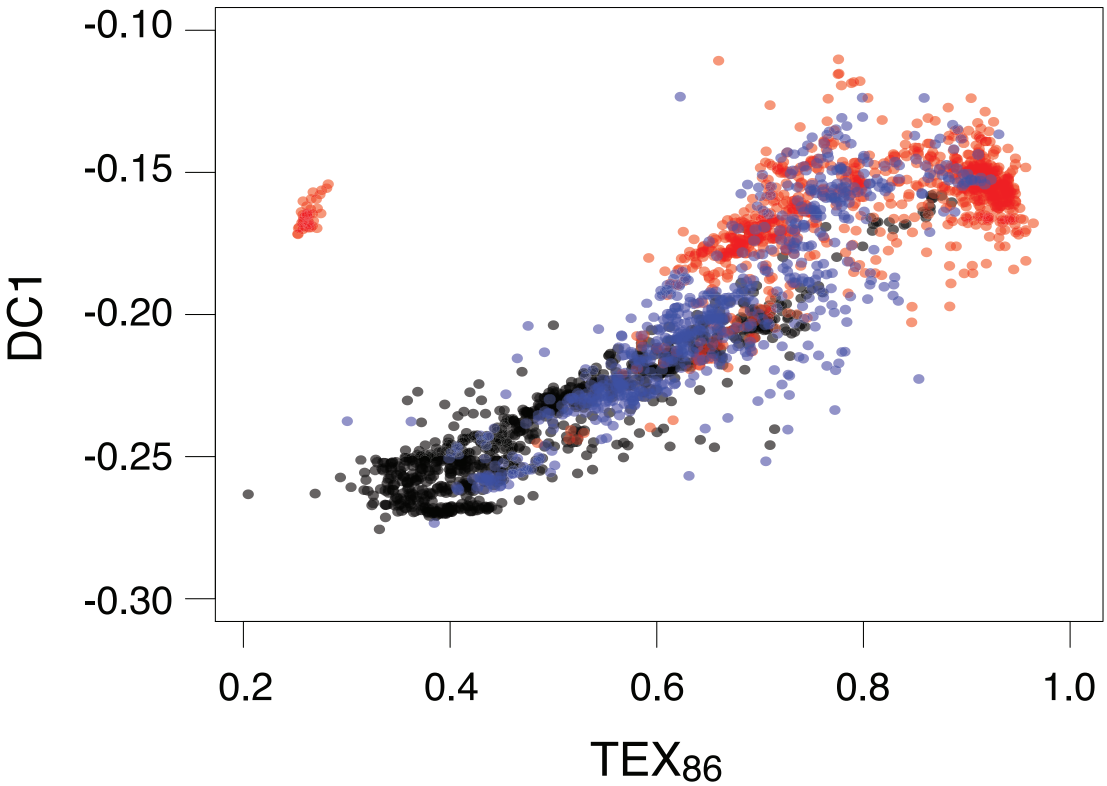


Figure 9

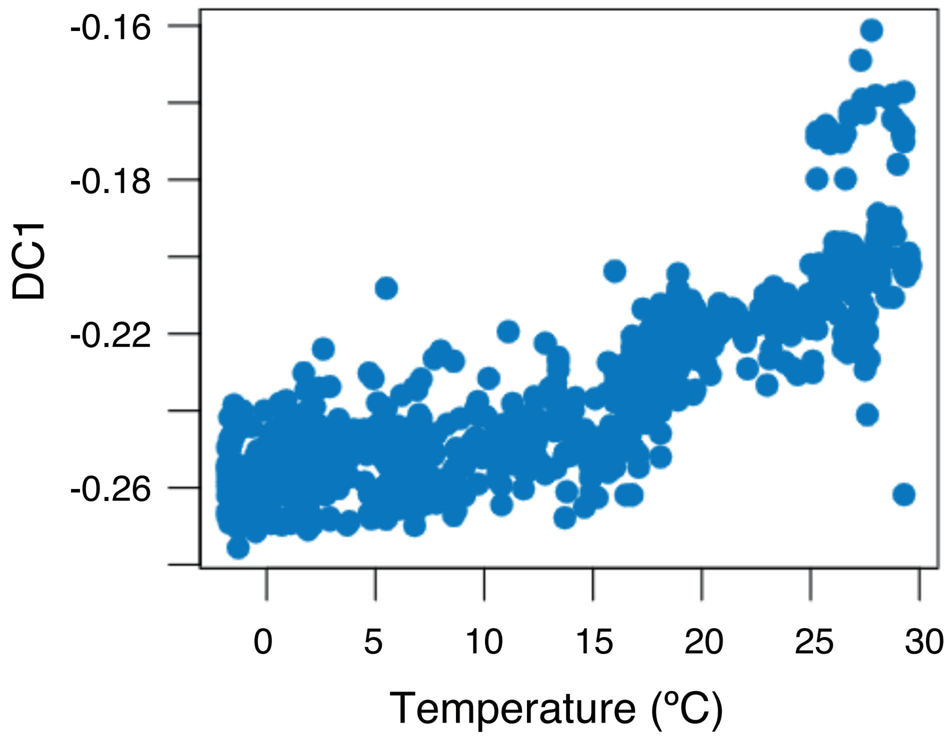


Figure 10

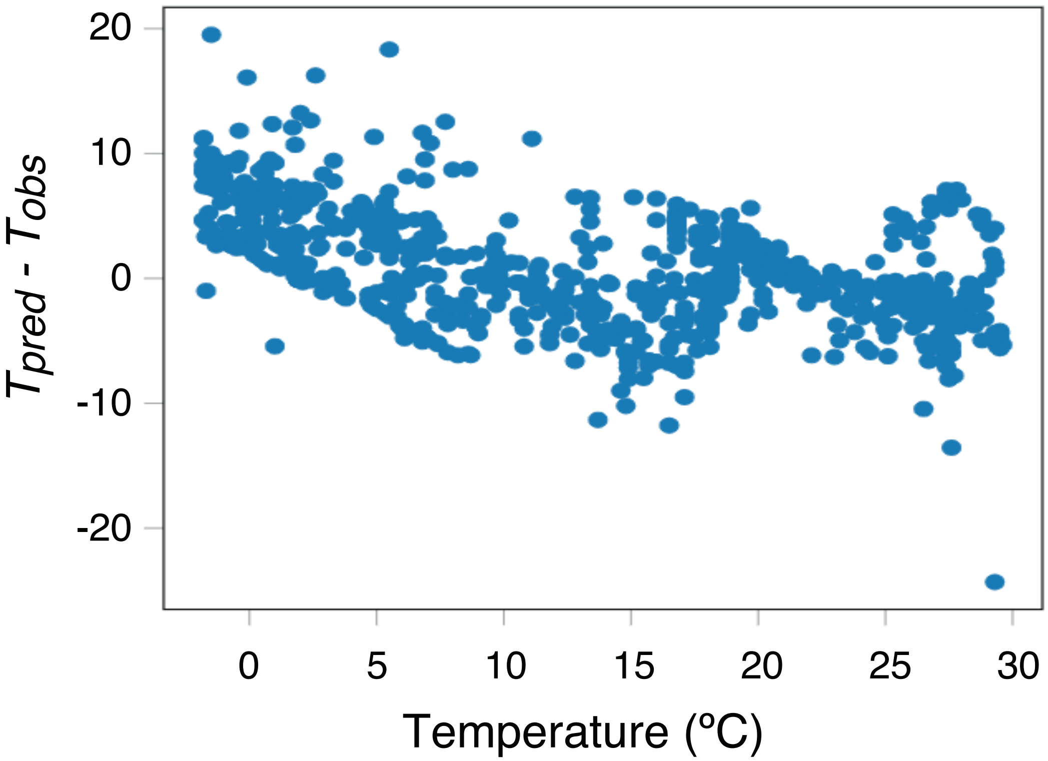


Figure 11

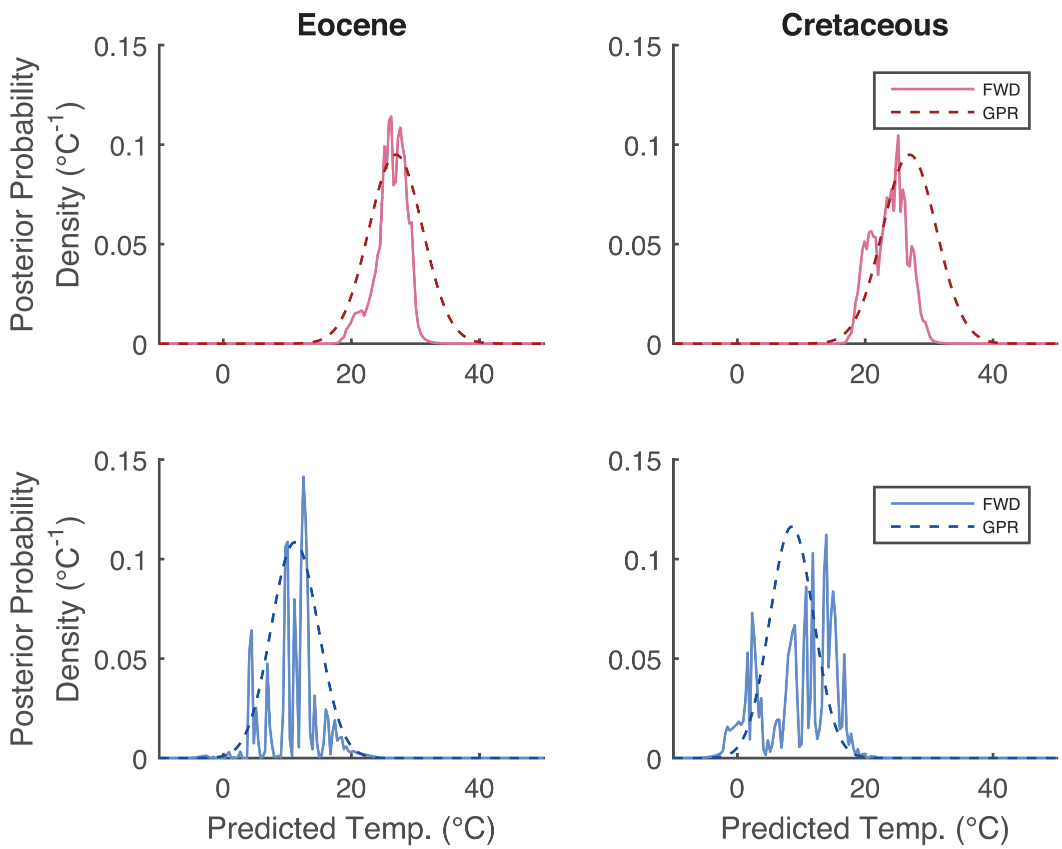

Eocene
Cretaceous
Posterior Probability Density (°C^{-1})
0.15
0.1
0.05
0
0
20
40
Predicted Temp. (°C)
FWD
GPR

Figure 11

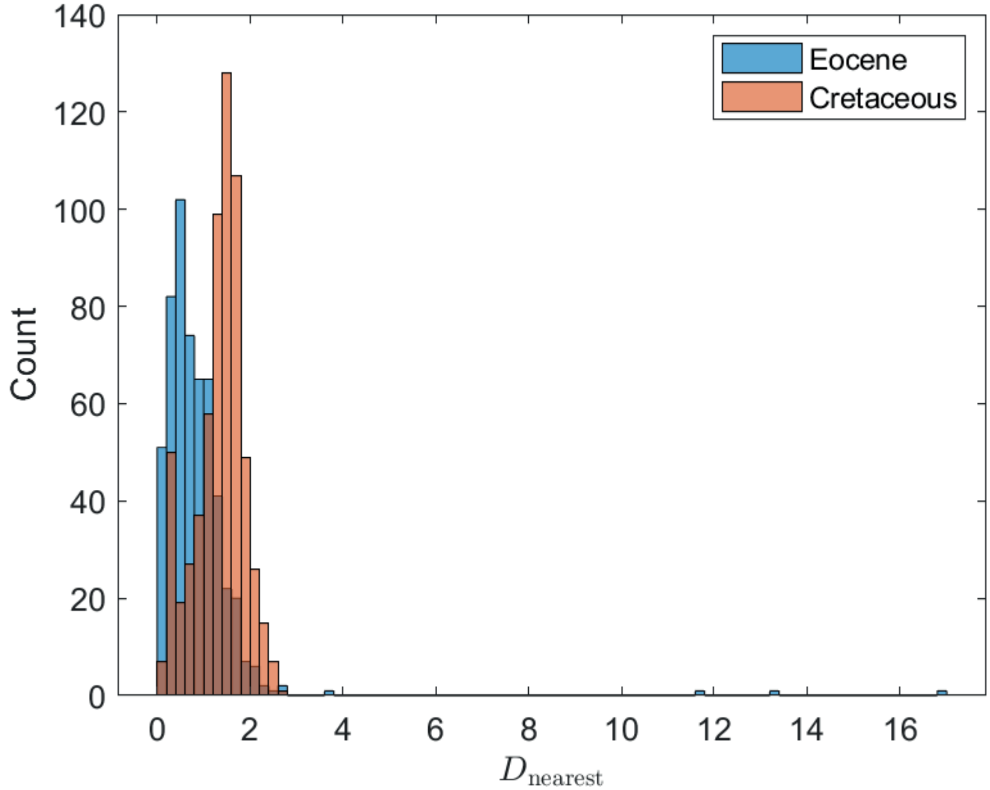
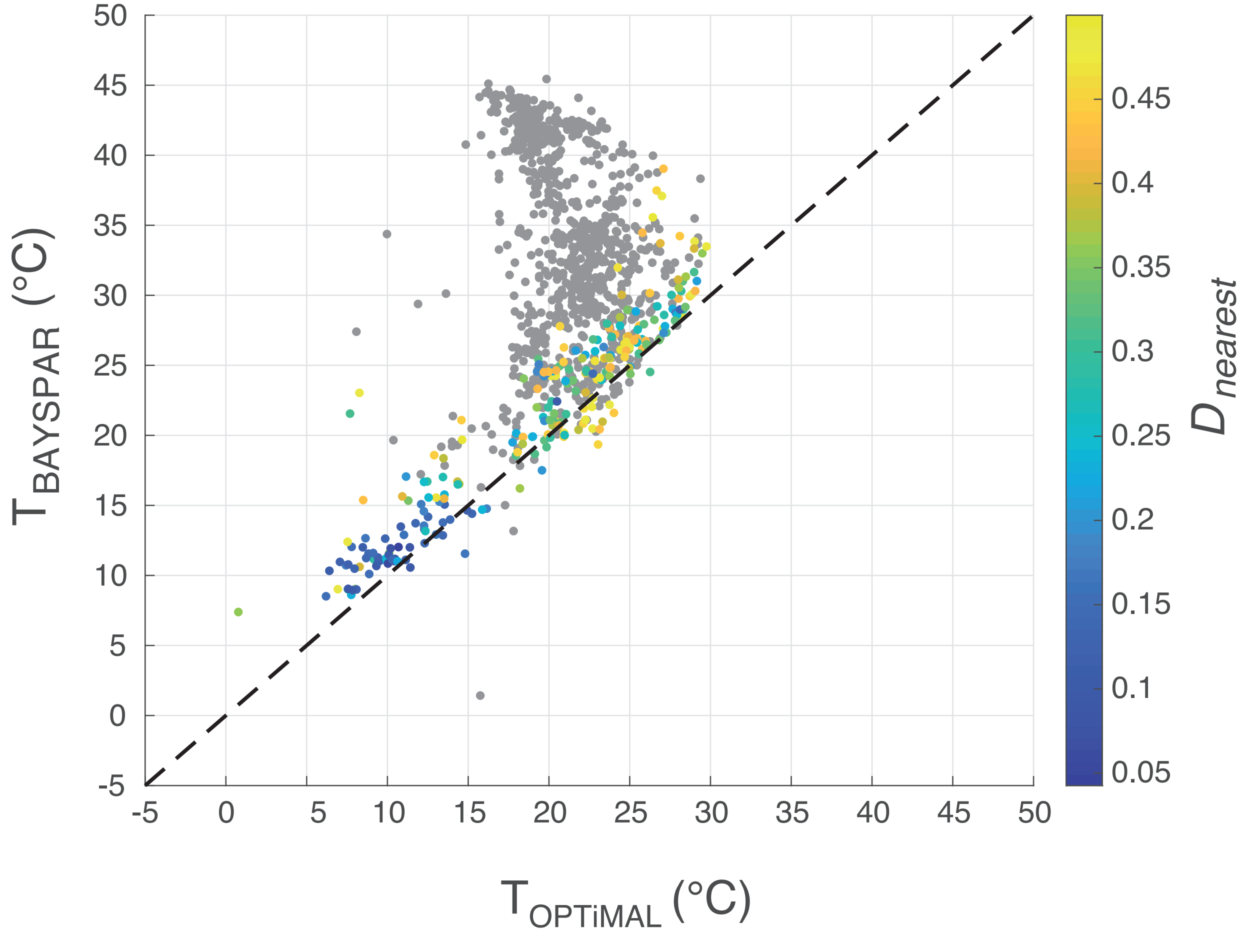

Eocene
Cretaceous
Count
D_{nearest}
0
2
4
6
8
10
12
14
16
0
20
40
60
80
100
120
140

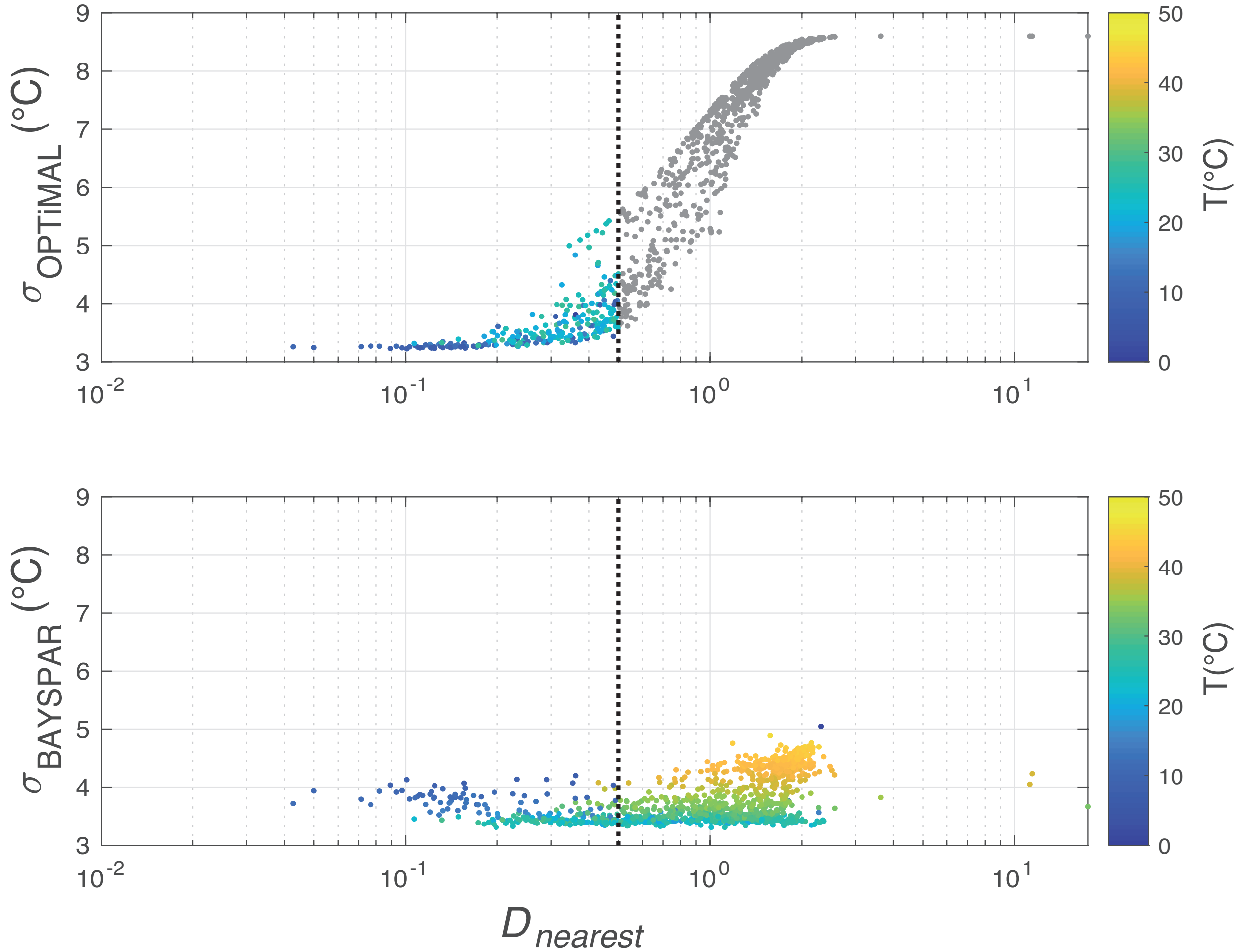
Figure 13



1128

1129

Figure 14



σ OPTiMAL (°C)
σ BAYSPAR (°C)
T(°C)
T(°C)
D nearest
9
8
7
6
5
4
3
10^-2
10^-1
10^0
10^1
50
40
30
20
10
0

Figure 15

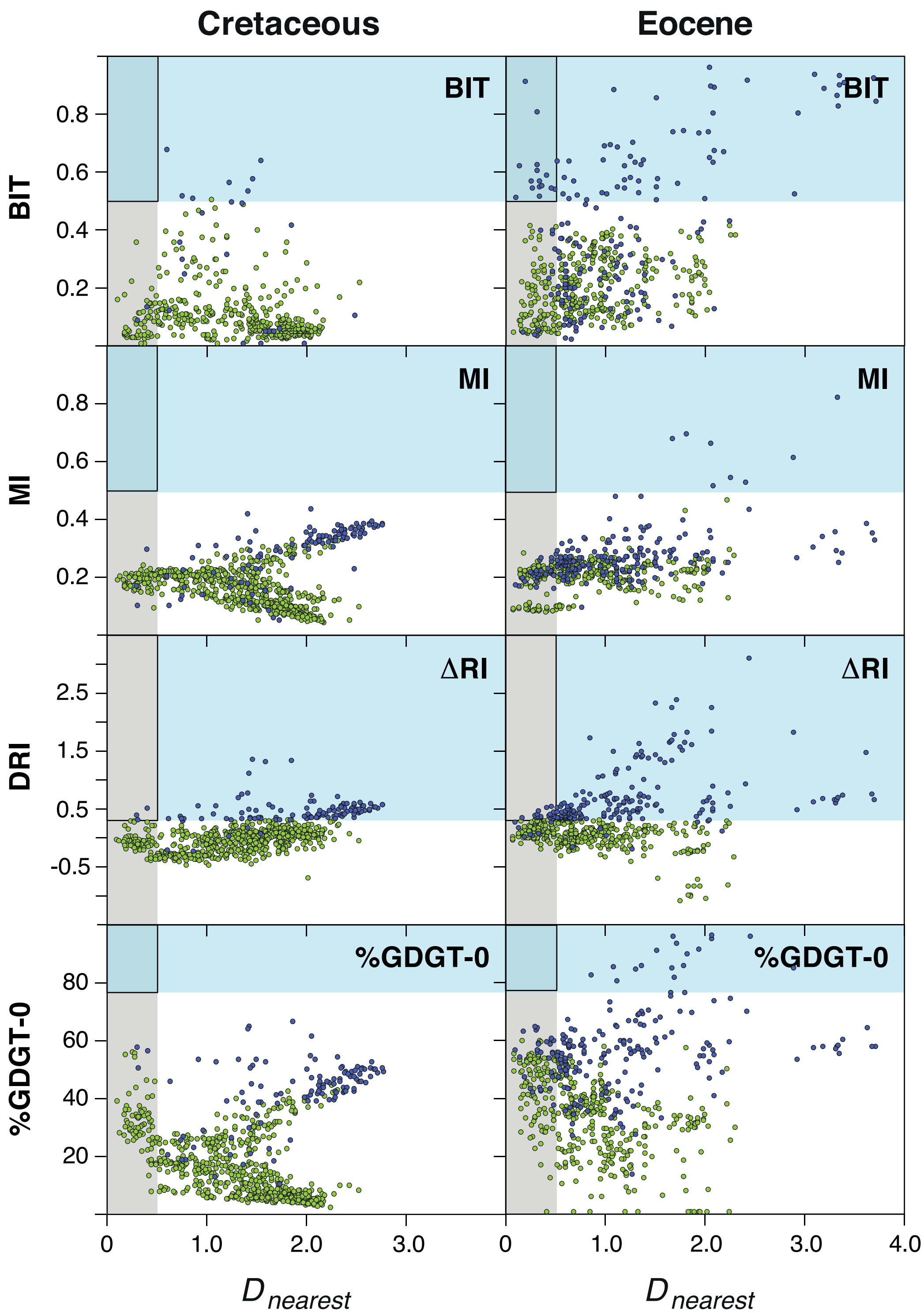
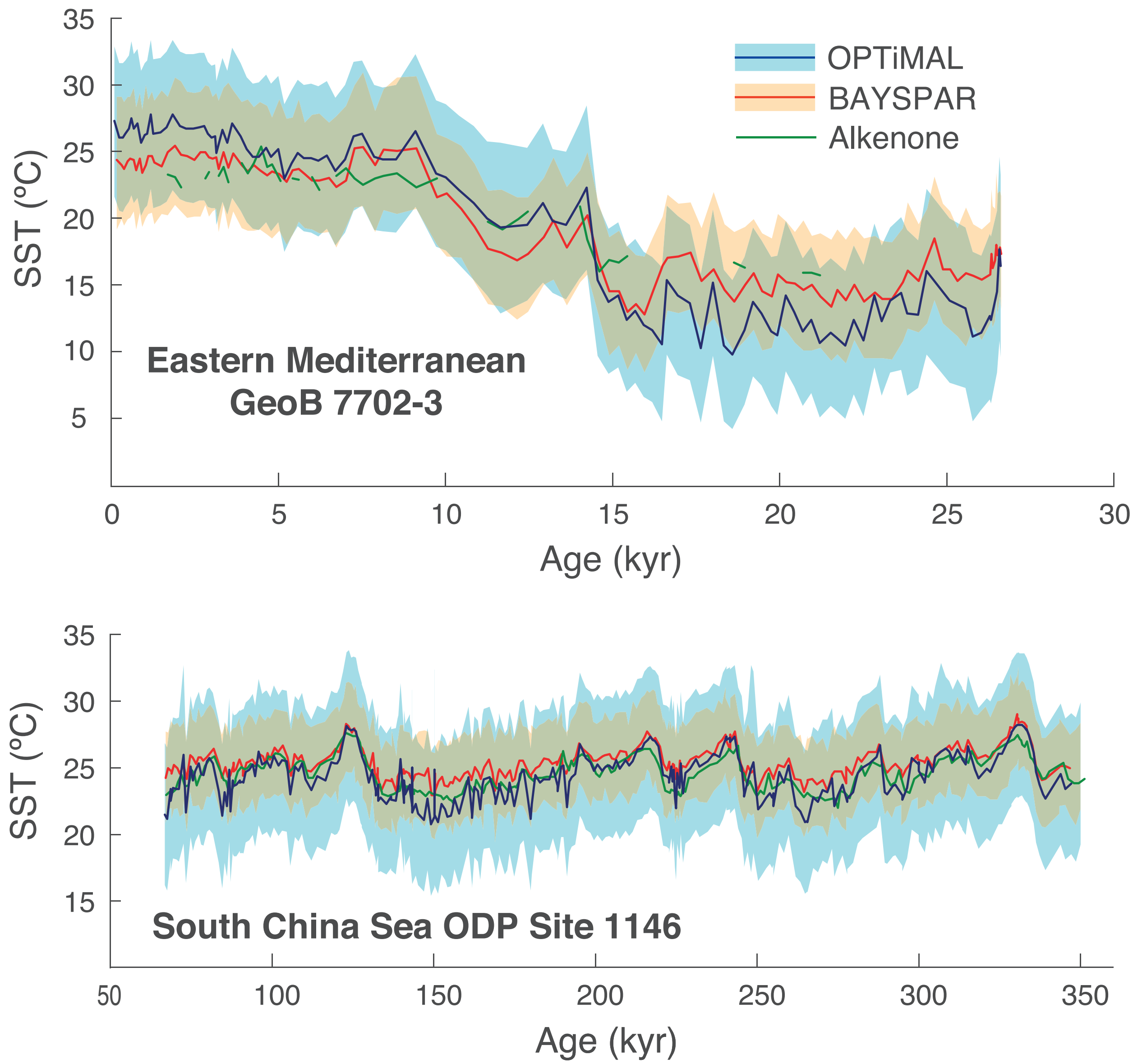

Cretaceous
Eocene
BIT
MI
ΔRI
%GDGT-0
BIT
MI
DRI
%GDGT-0
0.8
0.6
0.4
0.2
2.5
1.5
0.5
-0.5
80
60
40
20
0
1.0
2.0
3.0
4.0
D nearest
D nearest

Figure 16



OPTiMAL
BAYSPAR
Alkenone
Eastern Mediterranean
GeoB 7702-3
SST (ºC)
Age (kyr)
35
30
25
20
15
10
5
0
5
10
15
20
25
30
South China Sea ODP Site 1146
SST (ºC)
Age (kyr)
35
30
25
20
15
50
100
150
200
250
300
350

Figure 17

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