UNIVERSITY OF BIRMINGHAM

Dr Yvette Eley
Geography, Earth and Environmental Sciences
t +44 (0) 121 414 6395
e y.eley@bham.ac.uk
14th Feb 2020

Dear Sirs

Re: cp-2019-60

Thank you for allowing us to present major revisions to our manuscript entitled "*OPTiMAL: A new machine learning approach for GDGT-based palaeothermometry"*, following on from an extensive first round of reviews.

We provide a revised submission, along with our detailed responses to the reviewers, and an annotated marked-up version of our manuscript showing the changes we have made. In addition to the specific comments, we have added new discussion of the most recent culture studies, which cast doubt on the linear relationship between TEX86 and growth temperatures above 30 degrees C. We also include a first comparison of the screening tools BIT, MI and our new nearest neighbor test (new Fig. 15) and we present new GDGT-derived OPTiMAL palaeotemperatures, for the late Miocene to Recent, from two sites in the Eastern (ODP Site 850) and Western (ODP Site 806) Equatorial Pacific (new Fig. 16). We thank all reviewers and commenters for their suggestions, which we feel have substantially improved our submission.

We thank you for collating the responses to our initial submission, and look forward to the reviews of our revised manuscript.

Yours faithfully

Yvette Eley

School of Geography, Earth and Environmental Sciences, University of Birmingham
Edgbaston, Birmingham, B15 2TT United Kingdom
Tel: +44 (0)121 414 3344 www.birmingham.ac.uk/schools/gees/index.aspx

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

Formatting notes:

Reviewers / Other comments are in bold italic

Our responses are in plain type

Proposed changes in manuscript or quotes from existing text are in italic

Response to Reviews:

Referee Yige Zhang: Received and published: 12 September 2019

I think TEX86 is wickedly charming. We all have heard about the misdeed of TEX86, but also realized that often times TEX86 is the only thing that could help us to learn about the ocean temperatures, especially in the greenhouse climates.

This proxy has probably already passed its pessimism stage, and currently in the realism phase. Eley et al is a timely contribution to help us better understand this important proxy. They applied the cutting-edge machine-learning tools to improve the SST estimates using GDGTs, with the concept of identifying nearest neighbors in the global core-top dataset. This is innovative, but I do have some concerns detailed below.

We thank Yige Zhang for his comments and agree that GDGT proxies are heading into a 'realism phase', and hope that our submission provides new analytical approaches for improving the robustness of GDGT-based paleothermometry.

I suggest moderate revision of the MS before it can be accepted by CP.

Although this was not explicitly explained in the MS, I assume the authors used the percentage of each individual GDGT, when [GDGT-0]+[GDGT-1]+[GDGT-2]+[GDGT3]+[Cren]+[Cren']=100%. If this is the case, these 6 variables are not independent from each other. Instead, they are often dominated by the variations of [GDGT-0] and [Cren], the two major GDGTs. To show this, I did a simple calculation of the T&T15 global core-top dataset, which yielded an average [GDGT-0]+[Cren] of 88% of all GDGTs. This means that the variability of [GDGT-1], [GDGT-2], [GDGT-3] or [Cren'] that you see might be largely explained by the changes of [GDGT-0] or [Cren]. This is one of the reasons that TEX86 only considers the minor GDGTs, and uses a ratio. Ratios are good, as demonstrated by numerous cases in geochemistry.

With that being said, I agree with the authors that by using a subset of GDGTs like the TEX86, we are losing some information. We also discussed this in the Rind Index paper of Zhang et al., 2016. We realized that Ring Index values are dominated by [GDGT-0] and [Cren], as illustrated in Fig. 2. So the real difference between OPTIMAL and TEX86 (and any TEX86 calibrations, BAYSPAR, Kim or Liu) is not 6 dimensions vs. 1.

If the authors would like to, they can try something like use 2 subgroups of GDGTs - major and minor ones, with the total of each equals 100%; or some other forms of 6 ratios, normalizing GDGTs to one in the major, and one in the minor category. There's still going to be some dependency between the 6 variables, but they are closer to the 6 dimensions than the original treatment.

If they decide not to pursue these alternatives, I'd like to see that at least they acknowledge the interdependency of GDGT% data.

We will amended the text in the final submission to make it explicit that we are using the percentage of each individual GDGT. Regarding the issue of dependency, our analysis suggests there are only five free parameters. Machine learning tools should be able to pick up on 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.

Another issue is the extrapolating from the modern calibration data set. Nobody likes extrapolations. But there are these mesocosm studies from NIOZ that demonstrated the response of GDGTs to ~40ºC temperatures. The archaea found in hot spring continues to increase their ring numbers until ~100ºC.

We refer back to the extensive responses to Tierney, concerning the nature of the constraints on the form of the temperature dependence of GDGT assemblages - that although there is evidence that ring number increases with temperature, the form of this relationship is characterised by empirical model fitting where we have modern environmental sampling (Kim et al. 2010; Tierney & Tingley et al. 2014; Zhang et al., 2016; OPTiMAL), but that extrapolation of these models beyond the calibration data is problematic, in the absence of other quantified constraints on the behaviour of this system.

I agree that we might not know the absolute temperature above ~30ºC very well, but I wouldn't call them "inappropriate use" that "impacts the confidence".

This is a matter of judgement. If the lack of understanding of the nature of GDGT temperature dependency above 30ºC results in paleo-SST estimates that we do "*not know ...very well*", that is grounds for a reduced confidence. In this context, as argued above, using such high temperature estimates to generate global mean surface temperature and climate sensitivity estimates (Zhu et al. 2019), in our judgment, is not appropriate without substantial caveats. We recognise however that the word 'inappropriate' may imply a subtle subtext of something being done in bad faith, and we have therefore amended this in the revised manuscript, replacing it with the words "provides absolute temperature reconstructions and uncertainty estimates that are relatively unconstrainted and therefore speculative".

In fact, the beauty of TEX86 is that it works in greenhouse climates and tropics when Uk'37 maxes out, carbonates are diagenetically C2 altered and seawater Mg/Ca is difficult to constrain.

As noted above: we do not wish to stop people using GDGT-based thermometry for the reconstruction of greenhouse climate states. We simply wish to urge caution in the application of this method when the fossil GDGT assemblages, from any time period, are strongly non-analogous to the modern calibration data on which this proxy rests. As we demonstrate this appears to be an increasing problem with increasing sample age, but it does not preclude the use of this proxy in greenhouse climate states where the fossil data are well-constrained by the modern calibration.

There is clearly an issue with the applicability of GDGT-based paleothermometry to high temperature conditions (>30ºC), but it is appropriate to end with a clear statement of our position: we would be truly delighted to secure a sound basis for the form of GDGT temperature dependency above 30ºC. As we, and all comments / reviewers note, applying GDGT paleothemometry at high temperatures would be extremely valuable for paleoclimate studies. Wishing for something, however, does not make it so. Instead, as a community we need to think and work creatively to address this problem, through culture or other manipulation studies, or through robust proxy-proxy inter-comparisons from ancient samples. Rather than the promotion of any single fix-all solution as *the answer,* we believe that advances will come by paying attention to the difficult questions. Amidst the complexities of nature, we should beware hubris.

Additional references (not found in our original submission) referred to in our response – these will be incorporated into our final revised submission:

Bale, N. et al. (2019) Applied and Environmental Microbiology 85(20) e01332-19
Cadillo-Qiuroz, H. et al. (2012) PLOS Biology, https://doi.org/10.1371/journal.pbio.1001265
Dunkley Jones, T. et al. (2013) Earth-Science Reviews, 125, 123-145
Elling, F. et al. (2017) Environmental Microbiology 19(7), 2681–2700
Hollis, C. et at. (2019) Geosci. Model Dev., 12, 3149–3206
Liu, X-L. et al. (2014) Marine Chemistry, 116, 1-8.
Lunt, D. J. et al. (2012) Clim. Past Discuss., 8, 1229- 787 1273.
Qin., W. et al. (2015) PNAS 112 (35) 10979-10984
Schouten, S. et al. (2007) Organic Geochemistry 38, 1537-1546
Wuchter, C. et al. (2004) Paleoceanography 19, PA4028, doi:10.1029/2004PA001041, 2004
Zhang, Y. G. et al. (2016) Paleoceanography, 31, 220-232
Zhu, J. et al. (2019) Science Advances 5 (9), eaax1874

.

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

Formatting notes:

Reviewers / Other comments are in bold italic

Our responses are in plain type

Proposed changes in manuscript or quotes from existing text are in italic

Response to Reviews:

Anonymous Referee #1 Received and published: 4 August 2019

The study by Eley et al. enriches the discussion on GDGT temperature calibrations by describing a novel machine learning approach. Previous calibrations including Bayspar relied on the reduction of GDGT abundance data into one dimension (TEX86) and used only a subset of the commonly quantified GDGTs. Compared to these prior approaches, the strength of Eley et al's approach lies in the development of a distance parameter of the full GDGT diversity. This approach has great potential for identifying anomalous GDGT distributions. However, I agree with some of the criticisms raised by J. Tierney and H. Yang.

Specifically, I think the authors should state more clearly the cons and pros of Optimal relative to Bayspar. In my view, the approaches taken in Bayspar and Optimal are complementary: From what I understand Bayspar is concerned with finding a regional TEX86-temperature calibration by subsampling the C1 modern dataset, Optimal seems to find the best fit to the whole modern dataset but ignores regional differences. A more balanced discussion should highlight these differences. Further, applying both approaches to the same paleorecord may yield novel information.

We thank the reviewer for their comments that our submission enriches the discussion surrounding GDGT paleothermometry. We note the reviewer's comments here regarding the comparison with BAYSPAR, and the regional variation, and respond to them in the context of the individual recommendations below.

Below I make recommendations for improving this study:

i) I missed a thorough exploration of the outliers, for example the large number of grey outlier data points in figure 13. Identifying patterns (e.g. similar depositional environments or similar environmental parameter space) in the distribution of outliers could advance our understanding of anomalous distributions.

We note again, and have made clearer in the revised manuscript through addition of a comment in the relavant figure caption, that, outside of the constraints of the training (modern calibration) dataset the GPR model temperature estimates revert to the mean value of the calibration dataset, with an uncertainty that reverts to the standard deviation of this dataset. In this sense the 'outliers'

in Figure 13 carry little meaning, except that these temperature estimates are more and more unconstrained, which is why they are greyed out and not considered. In these non-anologue samples BAYSPAR does continue to make temperature predictions.

ii) The agnostic approach of Optimal is both its greatest strength and weakness. Optimal implicitly ignores some of the things that we do know about GDGT distributions besides temperature-dependence. One important aspect that is not covered by Optimal is the existence of temperature-independent and regionally varying patterns of GDGT distributions and calibration slopes (see Fig. 5 of Pearson & Ingalls, 2013, Annu. Rev., and many papers concerning regional calibrations). Regionality is addressed to some degree by Bayspar and thus it would be worth exploring if the differences between Bayspar and Optimal could be attributed to regional effects. Further, addressing regionality as a source of noise could significantly reduce uncertainty and the issue of overfitting.

First, as noted above, we divide the data into training sets and validation sets, and report the errors on the validation sets that have not been used for training. That means that overfitting is not a concern: we do not fit to the data that we use to evaluate the errors in the prediction. The process is statistically robust. We repeat the process 10 times, choosing different validation subsets, to ensure that we are not biased by a fortuitous choice of the validation subset.

This is an interesting point. We do not agree with the reviewer that regionally varying patterns of GDGT distributions, non-temperature effects, and different calibration slopes are excluded from OPTiMAL. Rather, by not imposing a model form, and by considering the full 6-dimensional GDGT space, if there are strong regional signals within GDGT-temperature dependence, these will be well-modelled by the GPR approach. Further, in this instance, the agnostic approach has two further advantages: 1) geographical location is unlikely the determinant control on GDGT-temperature dependence, *per se*, rather a spatial signal will be the result of some spatially varying environmental parameter(s) (community structure, nutrient status, depth of production, oxygenation etc.). Modelling this spatial variation, using geographical location and extrapolating back through time, assumes a spatial uniformitarianism of the GDGT-temperature responses. In our model, we optimise the full GDGT assemblages themselves against temperature, such that if there are other strong competing influences on GDGT-temperature sensitivities, these will be more naturally modelled by OPTiMAL within the calibration process. 2) Building from point 1, with OPTiMAL there is no need to assume spatial-uniformitarianism through time. Rather, temperature predictions will be based on the similarity of samples in GDGT space - under the assumption that similar GDGT assemblages are derived from similar communities and environmental conditions - rather than by geographical location. An example of the above is the modelling of the Red Sea GDGT data. Whereas the Red Sea is typically an outlier in residual space for other TEX_{86} calibrations (Kim et al. 2010; Tierney & Tingley, 2014), OPTiMAL residuals show no significant outliers for this data - in other words, although unusual in the global ocean, the temperature sensitivity of Red Sea type GDGT assemblages are well-modelled by OPTiMAL. This is important for paleo-applications, where Red Sea-type GDGT assemblages may play an increasingly important role in past warm climate states (Inglis et al. 2015), and can be naturally identified by OPTiMAL.

iii) The final test of a proxy should be its application to a paleorecord. I recommend the authors test Optimal against established approaches (e.g., TEX86L/H, Bayspar, UK37, Mg/Ca) on both a deep-time (e.g., Eocene) and a recent (e.g., Pleistocene) temperature record. I believe that Optimal can become an important part of the GDGT toolbox if the above criticisms are addressed. I thank the authors for thinking about this problem in a novel way and pushing the field in a new and promising direction.

As noted in response to Tierney comments, we do show extensive comparisons between OPTiMAL and BAYSPAR in our paper, which includes two major compilations of both Eocene and Cretaceous GDGT data (Figs 13 and 14). We will provide applications of OPTiMAL to time-series from the Neogene, Paleogene and Cretaceous, and cross-comparisons to other proxy data, in submissions in the near future, but to properly discuss the behaviour of these proxy systems and the paleoclimate implications of such time-series is beyond the scope of this submission. We do, however, note that whilst key compilations of Eocene and Cretaceous GDGT data have strongly encouraged the release of full GDGT abundance data (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.

Additional comments below:

Line 20: Rather "site of deposition" than site of formation. C2 Line 43: from "dialkyl" to "dibiphytanyl"

We will amend this in the final revised submission

Line 45-46: Bacteria are not known to produce isoprenoidal GDGTs. You could also make this distinction by substituting "dialkyl" with "dibiphytanyl".

We will amend this in the final revised submission

Line 53: Based on new structural data by Liu et al. 2018, Organic Geochemistry, the "crenarchaeol regioisomer" is no longer considered to be a regioisomer of regular crenarchaeol. Consider renaming to "crenarchaeol isomer".

We will amend this in the final revised submission

Line 78-80: The original work on Red Sea GDGTs should be cited here (Trommer et al. 2009, Organic Geochemistry). This sentence may need to be rephrased. Trommer et al. suggested salinity as an indirect effect on TEX86 through its influence on community composition. Therefore, the actual difference between Red and Arabian Sea GDGT distributions may be the existence of an endemic population.

We will amend this in the final revised submission

Line 96: Unclear what "GDGT productivity environment" means: The habitat of Thaumarchaeota or the productivity of Thaumarchaeota? Line 110-111: I think "not helped" is a little harsh. BIT, MI and RI have valid use cases.

We will amend this in the final revised submission

Line 125-127: A bigger issue than the small number of cultures is the fact that all cultured planktonic species belong to the same genus, Nitrosopumilus, which is not necessarily representative for all Thaumarchaeota.

We agree that this is a significant issue, and goes directly to the heart of our assertion that at present, it is impossible to quantify the impact of community change on marine sediment GDGT distributions. We will amend our final revised submission to make this point more strongly.

Line 134-139: Elling et al. 2015 is not the only culture study on temperature response. Qin et al. 2015 (cited in line 127) also studied temperature response and found nonlinear temperature response. This study should be discussed here.

We will add some further discussion of Qin et al. (2015) in our final revised submission.

Line 162: Consider using more precise language than "wildly"

We do not feel that there is anything wrong with the word 'wildly' and we therefore have not amended this wording.

Line 280: Is this a linear correlation?

For the data shown in Fig. 3, this is a linear correlation.

Line 261: Missing citations for seasonality. Add Hurley et al. 2016, PNAS, for growth rate dependency. C3.

We will add this reference in our final revised submission.

Line 262: Trommer et al., 2009, Organic Geochemistry, was one of the first studies to suggest an effect of ecosystem composition.

We will add this reference in our final revised submission.

Consider Line 265: Not sure I understand why std errors are compared to standard deviations.

Where calibration data are available, the standard error refers to the standard deviation of the difference between the truth (calibration) and our predictions. It thus encompasses both the statistical scatter in the predictions and any potential systematic bias.

Line 340: Missing parenthesis after 2006.

We will make this correction in our final revised submission.

Line 507: What is the rationale behind using >0.5 as cutoff?

The exact choice of the cutoff is inevitably somewhat arbitrary. The general sense is that if there are no calibration points with inputs within roughly an (appropriately normalised) unit distance from the sample of interest in input space, we are inevitably extrapolating rather than interpolating, and should be extremely cautious. The specific choice of 0.5 rather than, say, 0.8 is further motivated by exploring when the GPR property of reverting to the mean in the prediction and, crucially, to the standard deviation of the calibration output in uncertainty estimates begins to significantly compromise the claimed uncertainty — i.e., when systematics are likely to dominate statistics.

Additional references (not found in our original submission) referred to in our response – these will be incorporated into our final revised submission:

Bale, N. et al. (2019) Applied and Environmental Microbiology 85(20) e01332-19
Cadillo-Qiuroz, H. et al. (2012) PLOS Biology, https://doi.org/10.1371/journal.pbio.1001265
Dunkley Jones, T. et al. (2013) Earth-Science Reviews, 125, 123-145
Elling, F. et al. (2017) Environmental Microbiology 19(7), 2681–2700
Hollis, C. et at. (2019) Geosci. Model Dev., 12, 3149–3206
Liu, X-L. et al. (2014) Marine Chemistry, 116, 1-8.
Lunt, D. J. et al. (2012) Clim. Past Discuss., 8, 1229- 787 1273.
Qin., W. et al. (2015) PNAS 112 (35) 10979-10984
Schouten, S. et al. (2007) Organic Geochemistry 38, 1537-1546
Wuchter, C. et al. (2004) Paleoceanography 19, PA4028, doi:10.1029/2004PA001041, 2004
Zhang, Y. G. et al. (2016) Paleoceanography, 31, 220-232
Zhu, J. et al. (2019) Science Advances 5 (9), eaax1874

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

Formatting notes:

Reviewers / Other comments are in bold italic

Our responses are in plain type

Proposed changes in manuscript or quotes from existing text are in italic

<u>Response to Open Comments:</u> Open comments from Jessica Tierney

Much of the exploratory analysis in this paper (distance metrics and PCA) is interesting. I like the idea of developing a unified distance index to ID strange GDGT assemblages, although I'd like to see some applications to time series data there to visualize how this is working compared to the traditional screening indices of BIT, MI, deltaRI, 2/3.

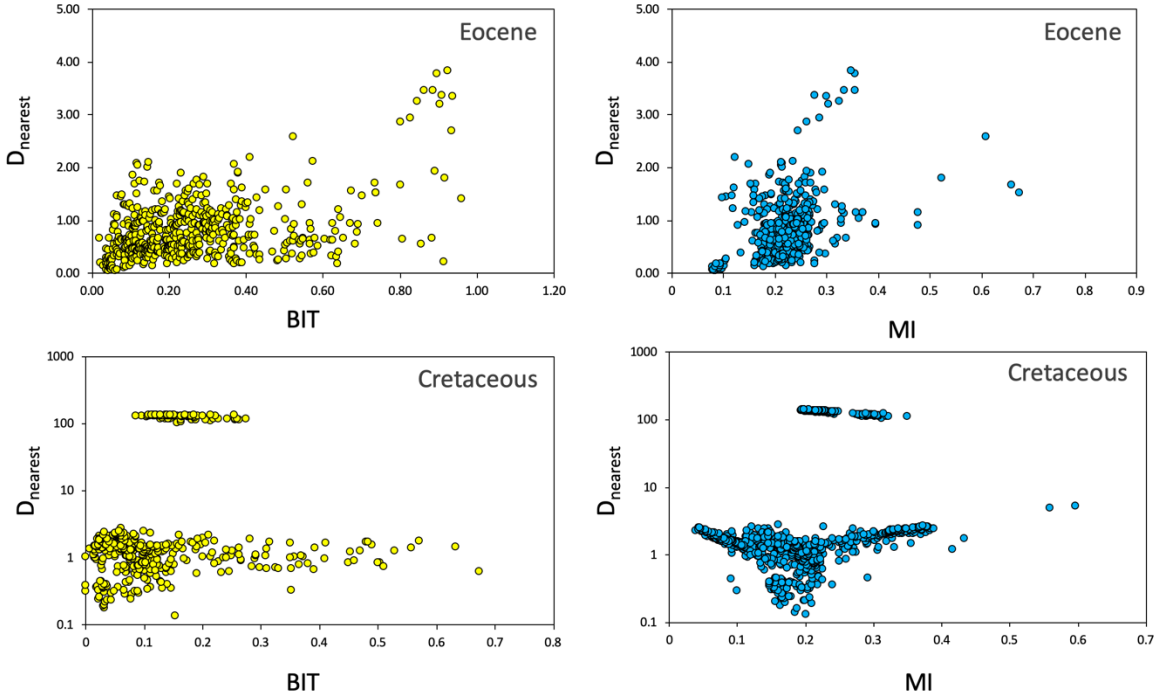
The distance metric that we present is focused on identifying non-analogue fossil GDGT assemblages, relative to those observed in the modern core-tops and scaled to the temperature-dependence observed in the modern GDGT - sea surface temperature (SST) relationship. Our aim is to determine how well-constrained a paleotemperature estimate is, based on its "nearness" to the modern calibration data. We believe it will operate differently from - but should be considered alongside - existing screening protocols in two key respects:

1. Non-marine contributions or diagenetic modifications to GDGT assemblages (e.g. methanogenesis) may not modify GDGT assemblages to an extent that they sit in a non-analogue space as identified by our screening metric, but might still have an impact on GDGT-based SST estimates. In this context, the continued assessment of these other indices is to be advised.

2. Conversely, some samples may pass these other screening methods, but may still have GDGT assemblages that are non-analogous to the modern calibration data set, for example in the high temperature range (>30ºC). As we argue in the manuscript, we suggest that these estimates are also treated with caution.

We have added a first assessment of the inter-relationship between these indices, with an additional figure (Fig. 15 in revised ms and also shown below) showing the relationship between $D_{nearest}$, BIT, and MI for the Eocene and Cretaceous compilations (where these data are available). 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, but with further study it may be possible to relax this requirement in future



I'm not convinced yet though that the OPTiMAL model provides good temperature prediction, for several reasons:

1) The authors take the view that being completely agnostic about the nature of GDGT response to temperature is a good thing. Why?

Elling 2015 show this, but the experiments go back to the late 70s/early 80s when thermoacidophiles were cultured at varying temperatures and the properties of the membranes were measured (c.f. de Rosa et al., 1980; Gliozzi et al., 1983). The Ring Index is the most experimentally-defensible way to quantify the relative amount of rings, but TEX86 is (non-linearly) one-to-one related to ring index in the modern coretop dataset (c.f. Zhang et al., 2015) which is why it works well as an index for the T response. Of course deviations occur between TEX and RI. . .which makes it up for debate whether one over the other is better in deep time.

We do know, after all, that higher temperatures should equal more rings. We know this from cultures, and from first principles about archaeal membrane structures.

We agree that there is a basic underlying trend for more rings within GDGT structures at higher temperatures (Zhang et al. 2015; Qin et al., 2015). What we dispute is that this translates into a simple linear model at the community scale (core top calibration dataset), or is yet reproduced with consistency between strains in laboratory cultures, including the temperature-dependence of GDGT ring numbers within the marine, mesophilic Thaumarchaeota in Marine Group 1 (broadly equivalent to the old Crenarchaeota Group 1) (Eilling et al., 2015; Qin et al., 2015; Wuchter et al., 2007). 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., 2004 (see Wuchter et al., 2004 Fig. 5). 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) meaning that the resulting calibrations for TEX_{86} cannot be applied to core-tops. This was recognised by Kim et al. (2010), who state "*but we may speculate that Marine Group I Crenarchaeota species in the enrichment cultures are not completely representative of those occurring in nature."*

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). The Elling and Qin 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 the above culture studies – made on marine, mesophilic archaea demonstrate how community composition may have 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 and other references above) 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) - and 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 (see Kim et al., 2010 Figure 2).

We agree that a robust mechanistic model is the most desirable scenario for any proxy-based reconstruction of environmental parameters, but in the absence of one for the temperature-dependence of GDGT assemblages, we contend that an agnostic approach is appropriate to discerning the underlying temperature-dependence of the actual, realized community-scale production of GDGTs in the modern oceans. We also note that if there is a strong underlying model within the calibration dataset this would be naturally recovered by our techniques, which would then make predictions at least as good as those that assume an *a priori* model choice (e.g. TEX_{86}, BAYSPAR). In the presence of this complexity, we believe OPTiMAL is more likely to recover the underlying temperature sensitivity of the system precisely *because* it is agnostic about model choice.

One does not necessarily need to reduce the problem down to the 1D space of TEX or Ring Index, but a model design that enforces this basic relationship is very defensible. Without this constraint, the model can not be extrapolated outside its calibration range at all - as is

the case with OPTiMAL. You can argue that that's the conservative choice, but I think most paleoceanographers want to use their proxies in greenhouse climates. d18O, Mg/Ca, and other thermometers are extrapolated away from modern conditions all the time, because we think we know something about how they should behave at higher temperatures. We know something about the GDGT response as well, although it's fuzzy, because of the limited number of culture studies. Still - it's seems prudent to use the information we have. The authors discuss "parametric" models (I think they mean the linear regression models that have been used thus far, because GPR is also technically parametric) as if they are bad thing - they are not, if the model form is based on scientific understanding.

As we state above, although we agree that there is a basic underlying trend of increasing ring number with increasing growth temperature, we do not agree that this is well enough known to be quantified into a "basic relationship" that can be "enforced" as a particular model form. Rather, there is uncertainty in the appropriate form of the relationship even within the modern calibration data (see Kim et al. 2010) which becomes substantial beyond the calibration range. The spatial structuring of residuals in global models of modern TEX_{86} temperature dependence (Tierney & Tingley, 2014) and clear structuring of residuals with temperature in our and other GDGT-temperature calibrations, are likely indications of transitions in the ecology, community make-up or habitat of modern GDGT producers that are not well constrained. We argue that this complexity in the GDGT temperature responses in the modern oceans should be grounds for caution when applying empirical models from the modern to ancient conditions, especially when working with the subset of ancient assemblage data for which there is no modern analogue.

With respect to other paleotemperature proxy systems, a key advantage of foraminifera-based proxies is the ability to precisely select individual species, and even size-classes, in order to maximize ecological stability of the proxy substrate through time. By reducing this biological variability, there is a greater chance that the underlying, and well-constrained, temperature-dependent thermodynamics of elemental or isotopic incorporation into biomineralized calcite will dominate analyses (Lea et al. 1999; Kim & O'Neil 1997). Although there are significant unknowns in deriving temperatures from these systems – most notably the seawater chemistry from which the shell calcified – these are "known unknowns" which can be more or less well circumscribed or modelled in any given application. In the case of the GDGT paleothermometer, uncertainty is located mostly in the realms of "unknown unknowns" - first in understanding what causes the scatter and variation in the temperature-dependency within the modern oceans, and then how this complex system might have evolved through time. It is this uncertainty that poses difficulty for an *a priori* choice of a particular model relationship for the modern system – which is also based on a summary variable which is non-unique with respect to the underlying data (GDGT assemblages) - and then extrapolating this beyond the range of the modern calibration and without an assessment of whether the underlying data retain some similarity of structure through time. It is in the light of this, that we wanted to provide a quantification of similarity between ancient and modern GDGT assemblages, to give some means of testing the uniformitarian assumption that underlies GDGT paleothermometry.

We also agree that the robust characterization of temperatures in greenhouse climate states is a research priority. And we acknowledge uncertainties with other proxy systems, but in most of these cases there is a fundamental understanding of the underlying temperature-dependence of the

chemistry of biomineralization and experimental data extending into temperature ranges beyond the modern surface ocean. We hope that this discussion, if it does anything, will spur new efforts to constrain the nature of GDGT production at high temperatures and/or within ancient greenhouse climates.

Furthermore, the argument that 6D space is better than 1D doesn't hold much weight. The authors' data exploration actually reveals that TEX does a great job in reflecting the GDGT response to temperature, as their DC1 is effectively similar. So it's not inherently bad to reduce GDGT response to something like TEX or Ring Index. In fact it can be good, because it reduces problems with collinearity (which I imagine is a problem for both RF and GPR?)

Although there is a structuring that appears to have some coherence between ancient and modern GDGT assemblages (Figures 7 and 8), we would not use this as a basis to argue that TEX_{86} does a 'great job' in reflecting the GDGT response to temperature. The relationships between TEX_{86} and temperature (SST) in the modern calibration dataset are explored in Figure 3, where the top panel shows the significant scatter in the relationship between SST and TEX_{86}, and the bottom panel highlights this, where samples with very similar TEX_{86} values can be produced from regions with substantially different SSTs. Although part of this scatter is inherent in the environmental sampling of complex systems, we argue that it is not improved by the reduction of all data to a single parameter. In other words, projecting the full complex data set onto a single axis can never improve predictive ability and is generally expected to degrade it unless other dimensions carry no information. In our case, they do, and we demonstrate this by showing that our temperature predictions have smaller standard errors (higher R^2, i.e., explain more of the data) than TEX_{86}-based predictors (see eq. 6, Figs 2 and 3 (bottom panel)).

2) I'm pretty concerned here about overfitting. The authors judge "model performance" by validation RMSE. This isn't the right metric - lower RMSE does not mean the model is better! An overfitted model will by nature give better RMSE. Instead they should use a metric like AIC, BIC, WAIC, etc that also penalizes over-parameterization. As it is, the models they fit have strong trends in the residuals, under-predicting T at low T's, over-predicting at high T's. This looks like a model problem to me.

We use a non-parametric, machine-learning model, so applying automatic Occam's razor penalties based on the parameter number as done in the variants of the Bayesian Information Criteria is not feasible. Instead, as described in the paper, we divide the data into training sets and validation sets (see Section 2), and report the errors on the validation sets that have not been used for training. That means that overfitting is not a concern: we do not fit to the data that we use to evaluate the errors in the prediction. The process is statistically robust. We repeat the process 10 times, choosing different validation subsets, to ensure that we are not biased by a fortuitous choice of the validation subset.

3) There are no example applications of OPTiMAL. The authors should apply it out of sample to paleoceanographic time series. I want to see how it does on Quaternary time series that

are within the calibration range - with comparisons to independent proxies like UK37 - and also some deep time series. The latter are outside the calibration so presumably the OPTiMAL predictions are just junk, but that needs to be transparent, so that folks don't start using it on Eocene data without thinking about it.

Figures 13 and 14 show the behaviour of OPTiMAL in deep time, as against BAYSPAR, for the most substantial compilations of Eocene and Cretaceous data available to date. As discussed at length in the paper, deep time applications of OPTiMAL are not 'junk' but rather require a careful assessment (aided by $D_{nearest}$) of whether the SST estimates generated have any validity based on the constraints provided by the modern calibration. When $D_{nearest}$ is <0.5 there is a clear trend to covariance between OPTiMAL and BAYSPAR in Figure 13 within the calibration range, although OPTiMAL generates SST estimates that are slightly cooler than BAYSPAR across the calibration range.

The Read Me for our model clearly advises against use for samples that do not meet the $D_{nearest}$ test (e.g. greyed out data in Fig. 13) and we would hope that users would read and take note of this guidance. If they do, we contend that SST predictions from OPTiMAL are at least as robust as any other GDGT-based methodology.

Overall I think the paper needs to be more cautious in selling this new approach as necessarily better than previous work. I.e., OPTiMAL is sold as circumventing the no-analogue problem, but it doesn't - indeed, this is a really insurmountable problem.

Our approach cannot and is not designed to circumvent the non-analogue problem. The very heart of our approach is to identify and quantify non-analogue conditions, and where they are identified (e.g. $D_{nearest}$ is >0.5) we explicitly caution about making inferences about ancient SSTs. Further, we explicitly state in the manuscript that OPTiMAL has no validity in such non-analogue conditions and should not be used - we can't be more "cautious" than that. Where we argue that OPTiMAL does have a potential advantage over current approaches is within the modern calibration range, in its ability to better match ancient samples to the most appropriate analogue GDGT assemblages in the modern calibration.

GPR is an interesting alternative, but It's not clear to me that it's better than the traditional TEX approach, and there are real limitations to the GPR model - the least of which is it cannot be extrapolated. This needs to be discussed in a more balanced way.

As explained in the manuscript, quantitative assessments of model performance show that the GPR model outperforms TEX_{86}-based approaches in the modern calibration data (see Figs. 13 and 14). It is not clear what Tierney is referring to by 'real limitations' in the GPR approach, with the only named one being the inability to extrapolate into non-analogue conditions. Again this "limitation" is at the heart of our paper, where we argue that we urgently need to improve constraints on the temperature-dependence of GDGT assemblages in these non-analogue conditions. At present the choice is whether to be cautious about system behaviour and temperature estimates outside of the modern calibration range, or to believe that one of a choice

of TEX_{86}-based models hold under these conditions, e.g. TEX_{86}^{H} (Cramwinckle et al. 2018) or BAYSPAR (Zhou et al. 2019).

For that matter, the authors argue that it's irresponsible to use TEX in greenhouse climates that are outside the calibration range. It certainly it not advisable to extrapolate any kind of calibration, but realistically: I don't think that paleoceanographers will stop using TEX86 in deep time. In many cases it gives similar estimates to independent proxies, which means there is temperature information in ancient GDGT assemblages. It seems to me that the way forward is a model that understands that there is recoverable information, but it is very uncertain. That is what we tried with the analogue method of BAYSPAR, which gives very large error bars for extrapolation, but there are probably other ways to go about this as well.

We agree that the wider community will need to make their own judgment on the validity of using models for GDGT-based SST estimation outside of the constraints of the modern calibration dataset. Our main hope is that this manuscript will prompt further rigorous evaluation of the behaviour of GDGT-based proxies, particularly under non-analogue conditions. Although there is agreement between BAYSPAR SSTs and independent proxy data in some locations, there are also locations where estimates diverge significantly (Hollis et al. 2019). The "recoverable information" which is encoded into BAYSPAR is a linear response of TEX_{86} to SSTs, our argument is that this is an *a priori* model choice and not based on a rigorous assessment of system behaviour in the modern.

Finally, I want to encourage the authors to adopt a more respectful and positive tone. There is a lot of hyperbolic language in here that implies that all previous calibration work, and use of TEX86, is "inappropriate" and has "eroded confidence" in the proxy. This isn't respectful to the Organic Geochemistry community, who has, in good faith, been trying very hard to understand whether there are other environmental controls on TEX and provide more laboratory-based evidence.

We do not mean to imply that previous calibration work has been inappropriate in a general sense, or that it was the quality of this calibration work that has eroded confidence in GDGT-based thermometry. We respect past calibration work as the fundamental basis for knowledge about the environmental controls on archaeal GDGT production, and it is the basis for a robust paleotemperature proxy system operating within the constraints provided by currently available data. Rather, we highlight that presuming a particular model form and extrapolating this beyond the available data constraints is problematic in the sense that it introduces an unknown error term, related to model choice. Whilst this may be less of a problem when resulting paleoclimate records are interpreted in terms of temporal climate dynamics and relative change, it is a more significant issue when absolute values are propagated through into quantifications of past mean global climate states (e.g. Zhu et al. 2019). We also note that a number of the authors of our submission are members of the Organic Geochemistry community and understand the great efforts and research advances made by many people in the application of biomarker proxies to palaeoclimate studies. Our hope is that the questions we raise improves and strengthens the science.

It's also not fair to TEX, which is no different from other temperature proxies in that there are complications associated with competing environmental factors and extrapolation to ancient time intervals. Take for example Mg/Ca, which is sensitive to T, S, pH, saturation state, laboratory cleaning methods, and changing Mg/Ca of seawater. It's hard to constrain paleo T estimates using it too! The bottom line is that using T proxies in deep time is full of challenges. The best we can do is to work together progressively to find a better solution.

We agree that all proxies have their own limitations. This paper discusses some of the limitations of TEX_{86}. We do not feel that this is being "unfair" to TEX_{86} and would hope that other proxy communities are also rigorous in the discussion and assessment of uncertainty. We agree that "*the best we can do is to work together progressively to find a better solution*" and we very much look forward to, and would be very supportive of, future work on the high-temperature behaviour of the GDGT system.

Below are some specific comments:

Line 45: Thaumarchaeota Line 46: Distinguish here that in this paper, you are speaking of isoprenoidal GDGTs (isoGDGTs). Bacteria do not produce isoGDGTs that I am aware of. Bacteria do produce branched GDGTs (Weijers et al., 2006).

We will amended this statement accordingly in the final submission.

Line 58: "were promising": this implies that use of TEX86 is no longer promising, which is not the case. TEX is widely used in both deep and shallow time applications and in many cases does a good job of describing past SSTs.

We see no issue with this use of language and we do not imply that TEX_{86} is no longer promising. As with any new proxy, however, the awareness of limitations and confounding factors always increase through time. Indeed, the fact that we have invested considerable time and effort to generate new tools to detect non-analogue behaviour and model SSTs with robust uncertainty estimates only makes sense if we valued GDGTs as tools for paleoclimate reconstruction.

Line 69: This is not the equation from Schouten 02. The correct equation is TEX86 = 0.015 *SST + 0.28.

We will correct this in the final submission

Line 73: These weren't criticisms, just observations that the calibration needed to be improved. rephrase.

We will rephrase this accordingly in the final submission.

Line 74: change "two new forms of the GDGT proxy" to "two new indices". the proxy itself has the same basis, these are just different indices to represent the cyclization. TEXL and H are a log10 transformation of TEX, not an exponential (however they assume that TEX is exponentially related to SST).

We will change the wording of this sentence to reflect this in the final submission.

Line 80: It's not clear that it's salinity per se - in any case I would cite Trommer et al., 2009 here, the original Red Sea TEX86 study.

We will add a citation of the Trommer et al. (2009) study to our revised final submission.

Line 100: "always troubled" - change the language to something less informal.

We will rephrase to "have troubled" in the final submission.

Line 110: I would not go so far to say that these indices are not helpful. In fact they are - the combo of BIT, MI, deltaRI can usually be used to identify suspect GDGT assemblages. It looks like you also rely of these later in the paper when you say you consider "screened" data. I agree that a unified distance index is also helpful, but there is a reason these indices were developed - they work and are easy to measure.

Our point here is misunderstood. We do not argue that these indices are unhelpful *per se*, but that they are not the most appropriate tools for identifying non-analogue GDGT assemblages. We will re-phrase this sentence in the final submission as follows, to make our point clearer:

"*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."*

Line 130: Actually the slope in these mesocosms was the same as the open ocean (0.015). It was the intercepts that were different.

This was wrongly phrased in the original submission, and we will update the wording as follows in our final revised submission:

"Early mesocosm data have been interpreted to indicate that a TEX86 to temperature relationship was maintained up to ~35-40ºC (Schouten et al. 2007; Wuchter et al. 2004), even though the primary data does not show a coherent response to incubation temperature (Figure 5; Wuchter et al., 2004)."

Line 155: To be fair to TEX, this is true for many paleoceanographic temperature proxies. We calibrate them based on our understanding of the modern system, and then hope that this understanding holds back in time.

Please see our responses to this argument above. Although the problem of non-analogue behaviour applies to all proxies, it is more acute when there is no clear biophysical model underlying proxy behaviour.

Line 160: sub-sampling? Not sure what you mean - rephrase. Our model uses formal Bayesian inference to estimate model uncertainty.

The deep-time settings for BAYSPAR search the modern core-top dataset for TEX86 values that are similar to the measured TEX86 value within a user-specified tolerance, and draws regression parameters from these modern locations (Hollis et al., 2019). This is what we mean by "sub-sampling" in line 160. We will amend this sentence in our final submission to more accurately reflect the functioning of BAYSPAR in deep time.

"*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, 2014)."*

Lines 161-164. Rephrase. It is not the Bayesian approach that is at fault - it's (arguably, because actually TEX is probably a very good index, as you have found in this paper) the use of TEX86 as the index, which does not in of itself detect non-analogue distributions. This said, we do attempt a rudimentary analogue approach for deep time applications.

We will rephrase ths as follows in the revised submission: "*The Bayesian approach, as with all approaches based on the* TEX_{86} *index, however, still does not model uncertainty based on the relationship of the underlying fossil GDGT abundances relative to modern data, and is insensitive to detecting substantially non-analogue behaviour in ancient GDGT distributions, as it still functions on one-dimensional* TEX_{86} *index values.*"

Line 162: "wildly insensitive" - use more formal language.

We did not use the phrase 'wildly insensitive' but 'wildly non-analogue'; this phrase accurately describes some of the Cretaceous data points with $D_{nearest}$ values of over 100. However, we are happy to change to the more neutral sounding: "substantial non-analogue" in the final revised submission.

Line 166: "master control" - use more formal language.

We will change the phrase 'master control' to 'master variable' in our final revised submission.

Line 178: "powerful mathematical tools" - hyperbole.

We do not see that there is a problem with the phrase 'powerful mathematical tools'. We will therefore not change this sentence.

Line 188: this is not the first time - BAYSPAR also accounts for model uncertainty.

We already discuss BAYSPAR in Section 2.2, lines 298-317. There, we point out that while BAYSPAR does model spatial variability, it does not account for the systematic uncertainty in the model when extrapolated beyond the range where it was calibrated.

Line 196: "interrogated" - not quite the right word. You mean all data were included in the analysis.

We will change this to "included" in the final revised submission.

Line 203: 854 data points - but there are 1095 in TT2015. Is this after doing some spatial averaging (for duplicates in the same location). Please describe any pre-processing that you did.

As we require fractional abundances for all 6 commonly measured GDGTs, we exclued those data points for which these values were not reported. We will add text to clarify this in our final revised submission.

Line 208: This is just root mean square error. No need to write it out - or redefine as something else.

We will modify this to read root mean square error in the revised final submission.

Line 215: "so-called"? That's what it's called. Again no need to define Rˆ2 as it's a common regression metric.

We will change this in the final revised submission.

Line 262: An alternative explanation is simply that the 6D GDGT space is not actually the right way to express the relationship b/t GDGT cyclization and temperature. The fact that TEX does show a great relationship to SST and the 6D space does not doesn't mean that the proxy is flawed, it means that the functional form of the model might be incorrect.

This is the first part of the exploration of temperature sensitivity across all components of the GDGT system by looking at the predictive capacity of based on nearest neighbours within the GDGT space. The GPR model develops from this and shows that a model based on all six components outperforms those based on a one dimensional index - it has smaller standard errors (higher R^2, i.e., explains more of the data) than TEX_{86}-based predictors (see section 4 and Figs. 13 and 14).

Line 265: RMSE is not a good metric of performance for any model. One can get a great RMSE and end up with the wrong form + overfit the data.

Please see above comments relating to overfitting, clearly setting out that it is not a problem with our response. Note further that Bayesian techniques on parameterized models are useful for model comparison, but do not provide an overall goodness of fit estimate. The calibration/validation approach that we use (with distinct data used for calibration and validation) does provide such an estimate. One of the co-authors on this submission (Mandel) has written several dozen papers on the methodology and applications of Bayesian statistics, and is very convinced of the merits of the Bayesian approach when there is a clear physical model with free parameters to be inferred, or a limited and enumerated set of alternative models to be compared. Alas, this is not the case here.

Line 271: Again you seem to be judging performance by RMSE and not considering model design and metrics of overfitting.

As we state above, we use a non-parametric, machine-learning model, so applying automatic Occam's razor penalties based on the parameter number as done in the variants of the Bayesian Information Criteria is not feasible. Instead, as described in the paper, we divide the data into training sets and validation sets, and report the errors on the validation sets that have not been used for training. That means that overfitting is not a concern: we do not fit to the data that we use to evaluate the errors in the prediction. The process is statistically robust. We repeat the process 10 times, choosing different validation subsets, to ensure that we are not biased by a fortuitous choice of the validation subset.

***Line 278: "wastes information" - I would actually argue it isolates T information, which is advantageous - unless you can prove otherwise. Do you, in fact, "get more information" out of using 6D space*?**

We are not clear what Tierney means here by the concept that TEX_{86} "*isolates temperature information*". All of the information within the TEX_{86} index is carried within the six-dimensional GDGT space, and any information about temperature-dependency will be naturally extracted by our GPR approach. The reverse is not true - the use of the TEX_{86} index can only represent a degradation of information as is shown in the improved temperature predictions of the GPR model - smaller standard errors (higher R^2, i.e., explain more of the data) than TEX_{86}-based predictors.

Line 291: "We fit the free parameters a, b, c, and d by minimising the sum of squares of the residuals over the calibration data sets" in other words, you did ordinary least squares regression. Just say that.

We are aiming for clarity, but are happy to add this qualifier "*(least squares regression)"* in our final revised submission.

Line 311: TEX86 is not a completely arbitrary index, and the proxy is not "fundamentally empirical". There is experimental basis for the proxy. Archaea produce more rings in their lipid membranes at high temperatures, and that the rings are there to change membrane fluidity, and this has been shown in laboratory settings going back 40 years. Elling 2015 show this, but the experiments go back to the late 70s/early 80s when thermoacidophiles were cultured at varying temperatures and the properties of the membranes were measured (c.f. de Rosa et al., 1980; Gliozzi et al., 1983). The Ring Index is the most experimentally-defensible way to quantify the relative amount of rings, but TEX86 is (non-linearly) one-to-one related to ring index in the modern coretop dataset (c.f. Zhang et al., 2015) which is why it works well as an index for the T response. Of course deviations occur between TEX and RI. . .which makes it up for debate whether one over the other is better in deep time.

We will remove the word "arbitrary" in our final submission. As noted above we accept that more rings are produced at higher temperatures, it is the form of this relationship, between species and in natural communities, that is uncertain.

Line 315: Validity outside the calibration range isn't guaranteed for any proxy (TEX is not unique here).

This statement is true, but (as discussed above) it is easier to defend extrapolation when there is a robust thermodynamic model that underlies a proxy (e.g. Mg/Ca, Lea et al. 1999; $\delta^{18}O$ Kim & O'Neil 1997). This is not the case for the temperature-dependence of GDGT assemblages. Proxy systems based on the analysis of fossil biominerals also have the advantage of the highly selective analysis of single species, and even morphological sub-groups within species (e.g. size categories), to control for species-specific or ecological variability in the chemistry of biomineralization. One of our major concerns for GDGT-based proxies is the potential impact of non-analogue conditions on the taxonomic assemblage of archaea and/or their ecology. It is very difficult to discern such behaviour through the study of ancient GDGT assemblages, but the $D_{nearest}$ metric is a step towards quantifying the differences between ancient and modern conditions.

Line 337: I have limited experience with Gaussian process regression but my understanding is non-linearity in predictor response might be important (?) I'm asking because, TEX has the nice property of making the relationship between GDGT assemblages and T linear (in the modern calibration dataset). But fractional abundances of each GDGT have non-linear relationships to T. Are the data transformed before the regression to account for this? What about collinearity?

Gaussian process regression does not rely on any assumptions of linearity. It can be thought of as using a weighted sum of nearby calibration (training) data to make predictions. The optimisation of the Gaussian process consists of finding the optimal choices for these weights. The beauty of the approach lies precisely in its robustness, which does not require any modification of the input data. If there are relationships between the input data points, they will be discovered by GP regression, and the weights will be adjusted appropriately.

We did not entirely follow the comments about the linearity between TEX_{86} and temperature. TEX_{86} is a ratio of the sums of two subsets of GDGTs, and so is not linear in the GDGT inputs. Furthermore, while the Schouten et al. (2002) model for connecting TEX_{86} and temperature is linear, non-linear models have been used to connect TEX_{86} and temperature, such as the models of Liu et al. (2009) and Kim et al. (2010) that we discuss around line 290.

Also in general this section and the random forests section need more info on how the data were treated, what algorithms were used in which programming language etc.

We used a Matlab implementation of an ensemble of decision trees with a fitensemble function, using the 'LBoost' training algorithm and 100 learning cycles. See Freund, https://arxiv.org/abs/0905.2138, for a discussion of the training algorithm. We do not feel that this information is required in the main text, but will add it to the SI in the final revised submission.

Line 371: Sure, I don't see how non-analogues are dealt with in the Gaussian process regression? You are training the model on the modern calibration dataset, so fundamentally the model can't know what to do with non-analogue data.

Tierney is correct: GP regression does not allow us to deal with non-analogue data. It does, however, use $D_{nearest}$ to robustly determine what data are similar to the calibration data (and apply our predictive model to it), and what data fall into the non-analogue category (as described in Section 3). We clearly recommend against using our technique, or any other technique, in the absence of physical models that could be meaningfully extrapolated beyond the calibration regime for non-analogue data (see line 507). To make the model as user friendly as possible, the Matlab code even makes it clear for the user which predictions should not be trusted, by plotting those data points in grey.

Line 378: "they can make no sensible inference about the behavior of this relationship outside of the range of this training data" - exactly. Just like all previous models (Kim's regression, BAYSPAR).

We agree with this statement. We make it clear throughout the manuscript that we cannot make any sensible inference about the form of the relationship between GDGT distribution and SST beyond the range of the training data. We further agree with Tierney that this is true, given the current lack of a robust mechanistic model, of all other previous models including Kim's regression and BAYSPAR. Indeed, this supports our assertion that it is problematic to use of these models, including OPTiMAL, to extrapolate beyond the modern calibration range.

Line 388-410: Do any of these outlying clusters have unusual BIT, MI, deltaRI, 2/3 values (?)

As noted above, associated with plots of $D_{nearest}$ agaisnt MI and BIT, the most anomalous Cretaceous data points are not associated with anomalous MI or BIT values.

Line 415: "This suggests that TEX86 is, in one sense, a natural one-dimensional representation of the data." So after all this - TEX is pretty great! This isn't surprising, for the reason I alluded to above, which is that TEX is a good index for relative cyclization that also happens to minimize non-T influences on the GDGT data.

We agree with Tierney, for a 1D representation, TEX does a reasonable job in representing the data. We maintain, however, that it is a representation that involves the loss of information about the temperature dependence of GDGT assemblages. It is unclear to us on what empirical basis TEX "miminises non-temperature influences on the GDGT data" - this would seem to rely on an *a priori* assumption that TEX is the correct way to model temperature-dependence, and any divergence from this is "noise". The fact that the GPR model outperforms TEX is evidence to the contrary.

Line 422: "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." Maybe, but this could also indicate a problem with the model.

Machine-learning approaches will generally tend to favour regression toward the mean for extreme data. This is expected behaviour and therefore in no sense indicative of a problem with the model. We will provide additional text clarifying this for the reader in our final revised submission. This behaviour is very similar to patterns of residuals on TEX_{86}-based SST models (e.g. the Kim et al., 2008 and 2010). Finally, we also emphasize that we show actual residuals rather than the standardised residuals used in the Kim et al. studies, which scale with the absolute temperature value and generate an apparent 'reduction' of error towards the high temperatures that can be easily misinterpreted.

Line 436: "They are limited by, first, the reduction of six-dimensional GDGT space to a one-dimensional index" The fact that DC1 is very similar to TEX86 suggests that this is not actually a limitation.

As noted above, OPTiMAL outperforms TEX-based approaches as clearly demonstrated by temperature predictions have smaller standard errors (higher R^2, i.e., explain more of the data) than TEX_{86}-based predictors.

Line 440: again: are the fractional abundances transformed to account for non- linearities? Also how is the model inverted?

Full details of the log-ratio transformations, the diffusion maps for nonlinear dimensionality reduction and visualisation, and details regarding the build, algorithms and dimensionality of the forward model, are all contained in the SI.

Line 445: "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 seems like a process that could be prone to overfitting. How do you assess overfitting?

Please see earlier responses with regard to the fact that overfitting is not a concern with these models.

Line 452: We argue this in our BAYSPAR paper as well (TT2014). Haslett also uses Bayesian methods.

We cite the Haslett study in our manuscript, and thank Tierney for pointing this out.

Line 457: "The existing BAYSPAR calibration also specifies the model in the forward direction, but ignores model uncertainty." Rephrase. what you mean to say it that we assume a certain model form (linear, spatially-varying regression). But of course we do estimate model uncertainty in that we estimate the parameters.

We have rephrased this to clarify what we mean in relation to model uncertainty.

Line 464: So is this GP model formally Bayesian?

It is possible to interpret the parameters specifying the GP kernel as model parameters and therefore to formulate GP regression optimisation as a Bayesian inference problem on these parameters. However, this is not a fruitful approach because these parameters are not physically interesting. Moreover, the intrinsic stochasticity of the Gaussian process already provides robust estimates of the prediction uncertainty, without the need to marginalise over the posterior distributions on the kernel parameters, so maximum-posterior optimisation is typically used.

Line 468: These residuals look even worse than the other models. What is going on? Are you sure this isn't a problem linked to your model structure?

The forward model is based on an attempt to predict all GDGTs based on the temperature measurements for training data, and then invert these predictions. There is no unique solution to the forward problem and in that sense, as discussed in detail in this section (see especially lines 468 — 474), the performance of the forward model is not less than the GPR model within the range of calibration data. As above, we also note that we show actual residuals, rather than standardised residuals as shown in other studies (e.g. Kim et al. 2010), because this is a truer representation of error across the data range. Comparisons of residuals between studies should thus take care to note which residuals are plotted.

Line 507: I guess this cut-off is based on Figure 1 (?) Can you show some kind of graphic, based on the modern calibration dataset, that demonstrates whether this cut- off correctly identifies GDGT distributions that deviate from expected (?) Also please compare how ID'ing on distance performs vis a vis using BIT, MI, deltaRI, etc. Have a look at the type of examples in Zhang et al., 2015 (the Ring Index paper).

Given that this parameter is based on a scaled-distance to a nearest neighbour within the modern calibration data set, it is difficult to interpret what is meant here by "deviates from expected" within the modern system - it would be effectively asking whether the modern calibration dataset substantially deviates from itself. This may identify rare and extreme outliers in the modern calibration dataset, but would not be informative in deciding a cut-off in the application of the OPTiMAL model to ancient GDGT assemblages. Instead, we refer to Figure 14, which demonstrates OPTiMAL model behaviour - in terms of temperature predictions and the error on these predictions - for compiled Cretaceous and Eocene GDGT assemblages, against the distance metric. This is informative, as it clearly shows an inflexion point in the error structure, with sharply increasing errors, at values above ~0.5 in $D_{nearest}$. Although the OPTiMAL model is explicit in its loss of predictive power as $D_{nearest}$ increases - through the rapid increase in uncertainty on temperature predictions to the standard deviation of the calibration data (and a regression in predictors to the mean of the calibration data i.e. towards no predictive power), we recommend the use of a cut-off around this value for the practical application of GDGT paleothermometry to generate paleoclimate time-series. In this context, the $D_{nearest}$ cut-off allows the rapid discrimination between well and poorly constrained temperature estimates.

Line 511: "This uncertainty is not apparent from estimates generated by BAYSPAR or TEX_H models, although the underlying and fundamental lack of constraints are the same.." The underlying model forms are actually not the same. Previous models make some assumptions about functional form. And it isn't bad to make some assumptions about response if you think you know what to expect: e.g. there should be more rings at higher T.

We do not state that the model forms are the same, rather, that the "lack of constraints" above 30ºC are the same for all models. We do understand Tierney's point, that models based on an empirical model fit within the modern calibration range can be assumed to reflect an underlying biophysical process, and that this knowledge can be extrapolated beyond the constraints of the modern calibration (30ºC). This does however encounter two fundamental problems: 1) that multiple model fits can be equally efficient *within* the modern calibration range, but that can then diverge significantly beyond this range. This results in predictions outside of the modern calibration range being highly dependent on model choice, whilst not providing quantifiable criteria on which to make this choice; and, 2) it assumes that GDGT-temperature dependency is a smooth function that can be extrapolated. We argue that this is an assumption that needs to be rigorously tested before extrapolation is valid, given the substantial evidence for multi-state conditions, likely associated with switches in archaeal communities, within the modern calibration data. We may *expect* more rings at higher temperatures, but this does not yet translate into a functional form of GDGT-temperature dependence above 30ºC with robust predictive power.

Figures 4 and 5: There are definitely some trends in the residuals! Please quantify these and discuss.

Please see the previous responses relating to overfitting and the fact that we are plotting the actual residuals rather than standardised residuals.

Figure 9 (and other similar figures): Dots are too big and it's hard to see the differences b/t data points.

This will be corrected in the revised manuscipt.

Line 540: "Above ~30ºC, however, the behavior of even single strains of archaea are not well- constrained by culture experiments, and the natural community-level responses above this temperature are, so far, completely unknown." Not true if you consider the substantial literature on thermo- and acidophile archaea. This is true only for the mesophilic pelagic guys. Specify that you are talking about mesophilic strains (but who knows, it could be that thermophile strains ARE relevant for greenhouse climates).

There is evidence for the temperature-sensitivity of GDGT production by thermophilic and acidophilic archaea in older papers (*c.f. de Rosa et al., 1980; Gliozzi et al., 1983*). However, more recent work, characterised by more precise phylogenetic and culturing techniques show a more complex relationship between GDGT production and temperature. Elling et al., (2017 Env. Microbio – see Table 1) highlight that there is no correlation between TEX_{86} and growth temperature in a range of phylogenetically different thaumarchaeal cultures - including thermophilic species. In fact the thermo- and acidophile *N. yellowstonii* grown at 72ºC had the fewest rings and thus yielded the lowest TEX_{86} temperature (calc. TEX_{86} temp of 24ºC). 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 (both 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, in comparing a wide range of thaumarchaeotal core lip compositions, that phylogeny generally seems to have a stronger influence on GDGT distribution than temperature.

Furthermore, thermo- and acidophile archaea do not grow in typical oceanic conditions and we dispute that data from these unique systems is appropriate to validate extrapolation of the behaviour of mesophillic pelagic archaea to temperatures beyond 30ºC. While there is no way to prove that extremophiles are irrelevant for greenhouse oceans, there is also no reason to suppose that they were. Thermophilic archaea are identified in conditions such as geothermal hot springs, and are often polyextremophiles in that they thrive in both hot and acidic environments (e.g., Cadillo-Quiroz et al., 2012). The studies cited by Tierney - de Rosa et al. (1980); Gliozzi et al. (1983) - are also of thermoacidophiles, species that typically require strongly acidic conditions (~pH <4). Based on the above, we do not consider that there is any strong evidence for a linear TEX_{86} temperature sensitivity above 30ºC in the modern or ancient oceans.

Line 544: "Until such data exist, we see no robust justification for any particular extrapolation of modern core-top calibration data sets into the unknown above 30C, although the coherent patterns apparent across GDGT space, between modern, Eocene and Cretaceous data (Figures 7), does provide some grounds for hope that the extension of GDGT palaeothermometry beyond 30C might be possible in future." This is true only if you assume - as you do in this paper - that there is no knowledge about the functional form of GDGT response to temperature.

As stated several times above, we have not yet seen, or been presented with any convincing evidence that would support a particular functional form of the community-level GDGT response to temperature. We would argue that this is not an assumption - the absence of convincing evidence holds as a fact until such evidence can be provided. Extrapolating a particular form of the GDGT-temperature relationship beyond the calibration range without evidence, however, *is* an assumption.

And you are not going to convince people to stop using TEX in greenhouse climates - it's still going to happen.

We do not wish to stop people using GDGT-based thermometry for the reconstruction of greenhouse climate states. We simply wish to urge caution in the application of this method when the fossil GDGT assemblages, from any time period, are strongly non-analogous to the modern calibration data on which this proxy rests. As we demonstrate this appears to be an increasing problem with increasing sample age, but it does not preclude the use of this proxy in greenhouse climate states where the fossil data are well-constrained by the modern calibration. Again, this was a primary goal of this paper, to separate well- from poorly-constrained SST estimates; it is a concern to us if the community do not wish to make the same distinction but that is yet to be decided.

So a more optimistic way forward would be to consider what we do know about GDGT response and see if we can model that in an acceptable way that would allow for some extrapolation. This is effectively what we did with our analogue mode of BAYSPAR - but this is just one way to approach the problem.

As above, we agree that qualitatively, higher temperatures generally drive more rings in archaeal membrane lipids, but we do not consider there to be any well-grounded biophysical model, or unique empirical model fit that would yet provide a robust basis for extrapolation of system behaviour to temperatures.

Line 560: This section needs applications to actual time series data - from both the Quaternary and deep time.

As noted above, Figures 13 and 14 show the behaviour of OPTiMAL in deep time, as against BAYSPAR, for the most substantial compilations of Eocene and Cretaceous data available to date. There is a clear trend to covariance in Figure 13 (when $D_{nearest}$ is <0.5) within the calibration range, although OPTiMAL generates SST estimates that are consistently slightly cooler than BAYSPAR across the calibration range. We will provide applications of OPTiMAL to time-series from the Neogene, Paleogene and Cretaceous, and cross-comparisons to other proxy data, in submissions in the near future, but to properly discuss the behaviour of these proxy systems and the paleoclimate implications of such time-series is well beyond the scope of this submission.

Line 562: "The OPTiMAL model systematically estimates slightly cooler temperatures than BAYSPAR, with the biggest offsets below ~15 oC (Figure 13)" That's because of your residual trends (Fig. 5).

Although at very low temperatures there is a model bias towards warmer predicted temperatures, this is similar for all the models when applied across the full range of modern calibration data (TEX_{86}^{H}, TEX_{86}, BAYSPAR) and cannot explain the offset in the two models. We suggest it is more likely driven by our inclusion of all data present in our modern calibration set, including data north of 70° N which is excluded from the BAYSPAR model (Tierney and Tingley, 2015).

Line 565: "whereas BAYSPAR continues to make SST predictions up to and exceeding 40ºC for these "non-analogue" samples" tell the readers why. It's because BAYSPAR assumes that higher TEX = warmer as part of the functional form of the model, whereas GPR is agnostic on this.

We will amend this sentence to include the requested information.

Line 575: "In contrast, BAYSPAR, because it is fundamentally based on a parametric linear model and therefore does not account for model uncertainty, assigns similar uncertainty intervals as to the rest of the data, despite there being no way of reasonably testing whether the linear model is an appropriate description of the data far from the modern dataset." 1) being parametric is not inherently bad and 2) not true. When BAYSPAR is asked to extrapolate, the error bars get bigger - very big in fact, as long as the prior is also big. Panel (b) of Figure 14 should specify the priors used for BAYSPAR estimation.

We used the default priors for the BAYSPAR estimation, as would most people who have either run the model from the original website or downloaded the model from github. We will specify this in the final revised text to make it clear for the reader.

Lines 582-596: Tone this down. Non-analogue behavior is problem for all proxy systems used in deep time. c.f. for example the work that David Evans has done to understand the myriad uncertainties in the Mg/Ca system (Mg/Ca of seawater, pH, dissolution, salinity, etc). It is difficult for all us to use proxies in deep time - it is not a problem unique to TEX.

Please see our responses above to this point. But in the spirit of conciliation we will removethe words "*entire*" and "*inappropriate*" from line 585, and change them to "*difficult to justify*" to *"problematic*" in line 587

It is also not considerate to those of us who have worked on this proxy for a long time to dismiss all previous work as "inappropriate use of GDGT paleothermometry" or claim that it has "eroded confidence". I actually remain optimistic about TEX, and I think many others are too. It has done a lot for us in terms of revealing temperatures in greenhouse climates. I do agree that calibration for deep time needs more work and that no-analogue assemblages are a problem. I also think that leveraging other proxy information with TEX is a nice way forward, which you mention here and there. Perhaps adopt a more positive conclusion along these lines?

We already include text at lines 530 – 541 discussing the potential for leveraging other proxy information to improve GDGT-based paleothermometry. We will add a sentence reflecting these possibilities to the conclusion in our final revised submission.

Additional references (not found in our original submission) referred to in our response – these will be incorporated into our final revised submission:

Bale, N. et al. (2019) Applied and Environmental Microbiology 85(20) e01332-19
Cadillo-Qiuroz, H. et al. (2012) PLOS Biology, https://doi.org/10.1371/journal.pbio.1001265
Dunkley Jones, T. et al. (2013) Earth-Science Reviews, 125, 123-145
Elling, F. et al. (2017) Environmental Microbiology 19(7), 2681–2700
Hollis, C. et at. (2019) Geosci. Model Dev., 12, 3149–3206
Liu, X-L. et al. (2014) Marine Chemistry, 116, 1-8.
Lunt, D. J. et al. (2012) Clim. Past Discuss., 8, 1229- 787 1273.
Qin., W. et al. (2015) PNAS 112 (35) 10979-10984
Schouten, S. et al. (2007) Organic Geochemistry 38, 1537-1546
Wuchter, C. et al. (2004) Paleoceanography 19, PA4028, doi:10.1029/2004PA001041, 2004
Zhang, Y. G. et al. (2016) Paleoceanography, 31, 220-232
Zhu, J. et al. (2019) Science Advances 5 (9), eaax1874

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

Formatting notes:

Reviewers / Other comments are in bold italic

Our responses are in plain type

Proposed changes in manuscript or quotes from existing text are in italic

Response to Open Comments of HUAN YANG

This manuscript presents a new approach, 'OPTiMAL' to improve the accuracy of palaeo sea surface temperature reconstruction, in particularly focusing on the GDGT data from greenhouse worlds. The effort of improving the sea surface temperature reconstruction is welcome. As an organic geochemist without enough knowledge of machine learning, it is difficult for me to understand some parts of the manuscript as most parts of the discussion show how the model was established and what the performance of this model was. I can understand the rationale of the model used in this manuscript: GDGT data should have 6-dimensional information and machine learning approach used in this study can capture the 6-dimensional information, which appears to be better than one dimensional information shown by TEX86. However, I strongly suggest the authors should be cautious when discussing the TEX86 and its derivatives because these proxies are still applicable in most cases and in most work of palaeoSST reconstruction in the greenhouse world, these proxies are still applicable.

We thank Huan Yang for their comments regarding the need to improve SST reconstructions. We have gone through the manuscript again, to ensure clarity of phraseology. We note the suggestion that we should exercise caution when discussing TEX_{86} and other similar indices, and the argument that even in greenhouse worlds, these proxies are still applicable. In this context we refer back to our response to Tierney above:

"We do not wish to stop people using GDGT-based thermometry for the reconstruction of greenhouse climate states. We simply wish to urge caution in the application of this method when the fossil GDGT assemblages, from any time period, are strongly non-analogous to the modern calibration data on which this proxy rests. As we demonstrate this appears to be an increasing problem with increasing sample age, but it does not preclude the use of this proxy in greenhouse climate states where the fossil data are well-constrained by the modern calibration. Again, this was a primary goal of this paper, to separate well- from poorly-constrained SST estimates; it is a concern to us if the community do not wish to make the same distinction but that is yet to be decided."

The authors should also carefully check the manuscript as there are a number of typos. In particular, the reference list should be checked carefully. Some minor points are appended below.

Line 52 'LC-MS' Full name should be shown here.

We will amend this in the final revised submission

Line 60-61 References should be arranged in the time order.

We will amend this in the final revised submission

Line 63 Zhaung should be Zhang.

We will amend this in the final revised submission

Line 73 'in response to these criticisms' Better to use other words.

We are not sure why the reviewer feels that these words are inappropriate, and therefore we have not amended them

Line 77 'Kim et al. 2010' should be 'Kim et al., 2010'.

We will amend this in the final revised submission

Line 90 'Hollis et al. 2012; Dunkley Jones et al. 2013; Lunt 2012' should be 'Hollis et al., 2012; Dunkley Jones et al., 2013; Lunt 2012' . The references should be listed in the time order.

We will amend this in the final revised submission

Line 95 Add comma after 'et al.' Line 96 Delete the comma after 'production'.

We will amend this in the final revised submission

Line 97 'ð˙IS´ Gð˘ ˙IRÿð ˘ ˙IS´ N86L is no longer regarded ´ as an appropriate tool for palaeotemperature reconstructions'. This proxy is still applicable in most marine regimes and could be used to infer SSTs.

This comment relates to TEX_{86}^{L}. A recent seminal study on proxy methodologies for paleoclimate reconstruction in deep time (Hollis et al., 2019) highlights that TEX_{86}^{L} does not accurately reflect the degree of cyclization of GDGTs, and therfore lacks a biological rationale. Equally, TEX_{86}^{L}shows enhanced sensitivity to biases driven by contributions from other subsurface archaea (Hollis et al., 2019). We therefore do not agree with this comment that TEX_{86}^{L} is "applicable in most marine regimes" and will not amend the sentence in our revised manuscript.

Line 103 'BIT' should be shown in full name

We will amend this in the final revised submission

Line 104 Add comma after 'et al.'.

We will amend this in the final revised submission

Line 131 'crenarchaeol regioisomer'

We will amend this in the final revised submission

Line 192-193 I think data from Tierney and Tingley (2015) do not cover all the published core-top GDGT data. A number of other data published recently should also be included.

In the first instance we stick to the Tierney and Tingley (2015) data to be comparable to the BAYSPAR calibration dataset, but we will explore further integrations of new core top data.

Line 228 Delete 'the crenarchaeol regio-isomer' and you can use cren' here.

We will amend this in the final revised submission

Additional references (not found in our original submission) referred to in our response – these will be incorporated into our final revised submission:

Bale, N. et al. (2019) Applied and Environmental Microbiology 85(20) e01332-19
Cadillo-Qiuroz, H. et al. (2012) PLOS Biology, https://doi.org/10.1371/journal.pbio.1001265
Dunkley Jones, T. et al. (2013) Earth-Science Reviews, 125, 123-145
Elling, F. et al. (2017) Environmental Microbiology 19(7), 2681–2700
Hollis, C. et at. (2019) Geosci. Model Dev., 12, 3149–3206
Liu, X-L. et al. (2014) Marine Chemistry, 116, 1-8.
Lunt, D. J. et al. (2012) Clim. Past Discuss., 8, 1229- 787 1273.
Qin., W. et al. (2015) PNAS 112 (35) 10979-10984
Schouten, S. et al. (2007) Organic Geochemistry 38, 1537-1546
Wuchter, C. et al. (2004) Paleoceanography 19, PA4028, doi:10.1029/2004PA001041, 2004
Zhang, Y. G. et al. (2016) Paleoceanography, 31, 220-232
Zhu, J. et al. (2019) Science Advances 5 (9), eaax1874

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

Yvette L. Eley,[1,] William Thomson[2], Sarah E. Greene[1], Ilya Mandel[3,4], Kirsty Edgar[1], James A. Bendle[1], and Tom Dunkley Jones[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]Institute of Gravitational Wave Astronomy, School of Physics and Astronomy, University of Birmingham, Edgbaston, B15 2TT, UK

[4]Monash Centre for Astrophysics, School of Physics and Astronomy, Monash University, Clayton, Victoria 3800, Australia

* Responses to Tierney's comments are in shown red; Response to Zhang's review are shown in blue blue; Response to Yang's comments appear in Green; and the response to Anon reviewer are in purple. Other changes made by the authors are highlighted in yellow.

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**ia** **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:

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

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 (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).

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

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

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. (2004; see Fig. 5). 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**ia** **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:

$$\delta T = \sqrt{\frac{1}{N_v - 1} \sum_{k=1}^{N_v} (\hat{T}(x_k) - T(x_k))^2} \qquad \text{(Eq. 5)}$$

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 one observable, so even a small change in that observable corresponds to a significant distance, and rather insensitive to another, meaning that even with a large difference in the nominal value of that observable the distance is insignificant. In the first instance, we use a very simple Euclidian distance estimate $D_{x,y}$ where the distance along each observable is normalised by the total spread in that observable 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

$$D_{x,y}^2 \equiv \sum_{i=0}^{6} \frac{GDGT_i(x) - (GDGT_i(y))^2}{var(GDGT_i)} \quad \text{(Eq. 7)}$$

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 $\hat{T}_{1/TEX} = a + b/TEX_{86}$ and $\hat{T}_{TEXH} = c + d \log_{TEX86}$ (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 $\delta T_{1/TEX} = 6.1$ ºC (note that this is slightly better than using the fixed values of *a* and *b* from (Kim et al., 2010), which yield $\delta T_{1/TEX} = 6.2$ 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. 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.

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

404 employed two forms of dimensionality reduction to enable visualisation of the data in two or three
405 dimensions. The fundamental point is that if temperature is the dominant control, all of the data should lie
406 approximately on a one-dimensional curve in GDGT space, and the arclength along this curve should
407 correspond to temperature; we will revisit this point below.
408
409 We first employed a version of principal component analysis (PCA) tailored to compositional data
410 (Aitcheson, 1982, 1983; Aitcheston and Greenacre, 2002; Filzmoser et al., 2009a; Filzmoser et al., 2009b;
411 Filzmoser et al., 2012). Taking into account the compositional nature of the data is important because the
412 sum-to-one constraint induces correlations between variables which are not accounted for by classical PCA.
413 Furthermore, apparently nonlinear structure in Euclidean space often corresponds to linearity in the simplex
414 (i.e. the restricted space in which all elements sum to one) (Egozcue et al., 2003). Figure 6 shows the
415 modern, Eocene and Cretaceous data projected onto the first two principal components. Aside from the
416 obvious outlying cluster of Cretaceous data, characterised by GDGT-3 fractions above 0.6, the bulk of the
417 data occupy a two-dimensional point cloud with a small amount of curvature. The large majority of the
418 Cretaceous data has more positive PC1 values relative to the modern data.
419
420 We also explored the data using diffusion maps (Coifman et al., 2005; Haghverdi et al., 2015), a nonlinear
421 dimensionality reduction tool designed to extract the dominant modes of variability in the data. Such
422 diffusion maps have been successfully used to infer latent variables that can explain patterns of gene
423 expression. In the case of biological organisms, this latent variable is commonly developmental age (called
424 pseudo-time) (Haghverdi et al., 2016). In our case, the assumption would be that this latent variable
425 corresponds to temperature. Inspection of the eigenvalues of the diffusion map transition matrix suggests
426 that four diffusion components are adequate to represent the data; we plot the second, third and fourth of
427 these components in Figure 7 for the modern and ancient data. The separate clusters marked `A' are the
428 outlying Cretaceous points with high GDGT-3 values. The bulk of the modern data lies on the branch
429 marked `B', while the bulk of the Cretaceous data lies on the branch marked `C'. Notably, the majority of
430 the modern points lying on branch C are from the Red Sea, which suggests that the Red Sea data is essential
431 for understanding ancient climates (particularly Cretaceous climates).
432
433 The relationship between the first diffusion component and TEX_{86} for all data is shown in Figure 8. There
434 is a clear correlation, despite the presence of some outlying Cretaceous points, some of which are not shown
435 because they lie so far outside the majority data range within this projection. This suggests that TEX_{86} is,
436 in one sense, a natural one-dimensional representation of the data. We also plot the first diffusion
437 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 does not account for the systematic uncertainty in the model when extrapolated beyond the calibration range. 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 degrees celsius, 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 nearest neighbor 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 (Figures 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.

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). We suggest that this is driven by our inclusion of all data present in our modern calibration dataset, including data north of 70ºN, which is excluded from the BAYSPAR model (Tierney and Tingley, 2015). 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 TEX86 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 errors 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, BAYSPAR, because it is fundamentally based on a *parametric* linear model and therefore does not account for model uncertainty, assigns similar uncertainty intervals as to the rest of the data, despite there being no way of reasonably testing whether the linear model is an appropriate description of the data far from the modern dataset.

We further provide a first assessment of the inter-relationship between standard screening indices and $D_{nearest}$, with an additional figure (Figure 15) showing the relationship between $D_{nearest}$, BIT, and MI for the Eocene and Cretaceous compilations (where these data are available). 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, we provide two example time series from the Neogene to modern, 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). Where $U^{k'}_{37}$ temperatures 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.

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 (see Section 2.3). This machine learning based predictor yields δT 4.1 ºC degrees, by applying a suitable weighted average over multiple near neighbours. R2 = 0.83, meaning that 83% of the variation in the observed temperature is successfully explained by our GDGT-based model.

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 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 $D_{nearest}$ threshold (>0.5), BIT and MI index for the Eocene (Inglis et al., 2015) and Cretaceous (O'Brien et al., 2017) datasets.

Figure 16. GDGT-derived OPTiMAL palaeotemperatures, for the late Miocene to Recent, from two sites in the Eastern (ODP Site 850) and Western (ODP Site 806) Equatorial Pacific; uncertainty envelopes are one standard deviation each side of the maximum likelihood temperature estimator. Also shown for comparison are $U^{k'}_{37}$ data from the same sites and modern mean annual SSTs.

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.

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.

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.

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.

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.

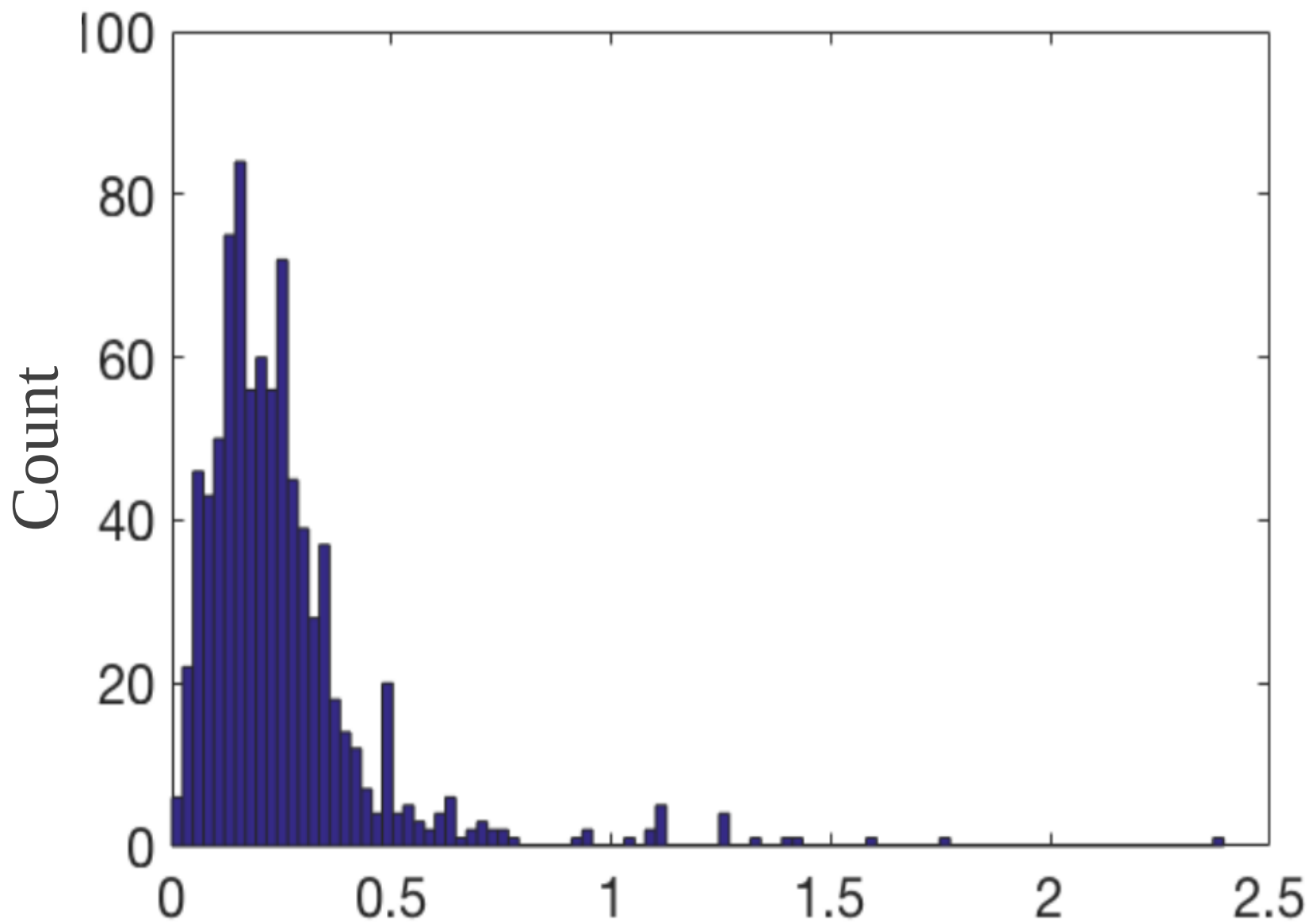
Zhang, Y. G., Pagani, M., Wang, Z.: Ring Index: A new strategy to evaluate the integrity of TEX_{86} paleothermometry, Paleoceanography, 31, 220-232, 2016.

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



$$D_{x,y}$$

Figure 2

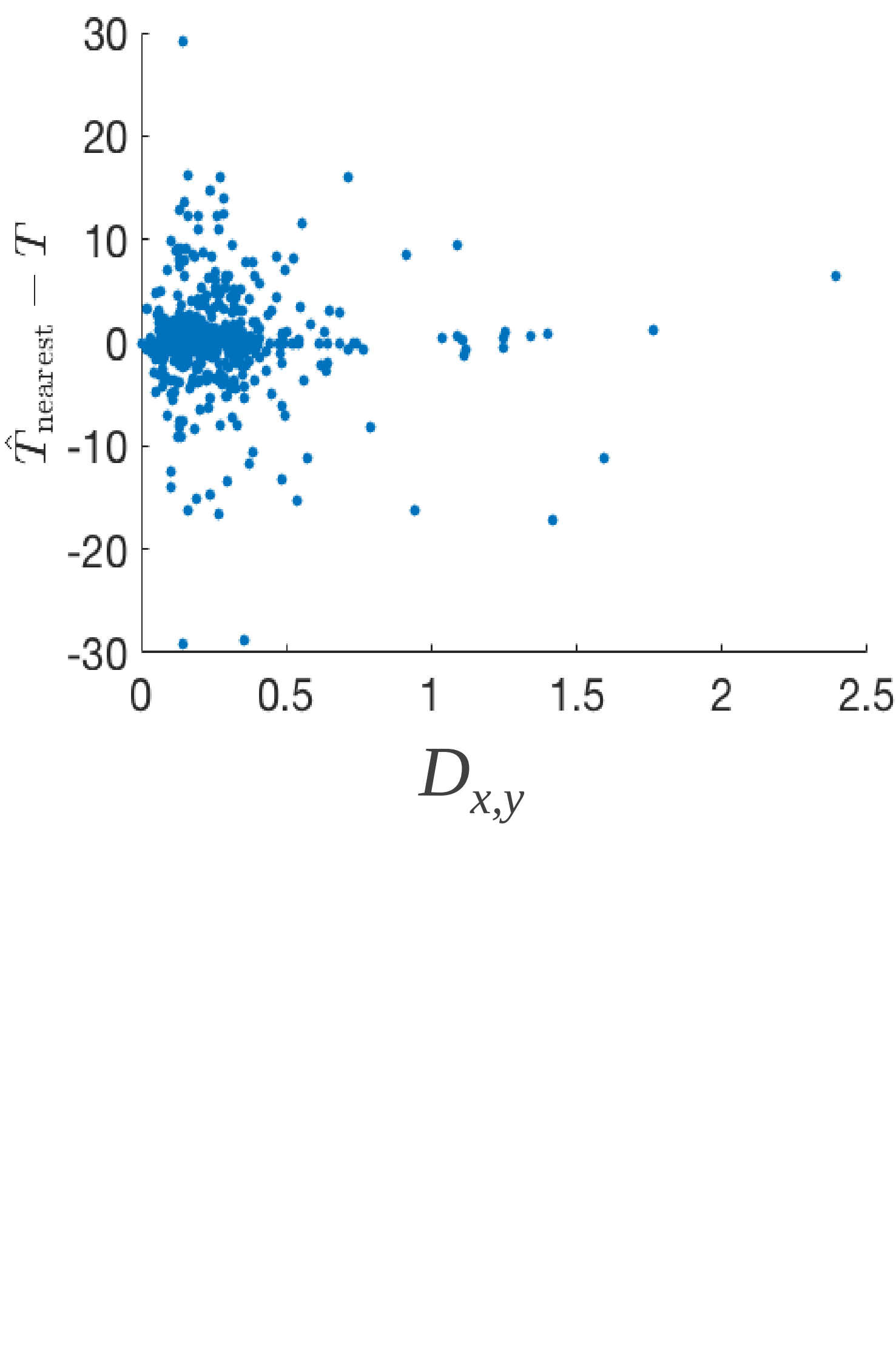


Figure 3

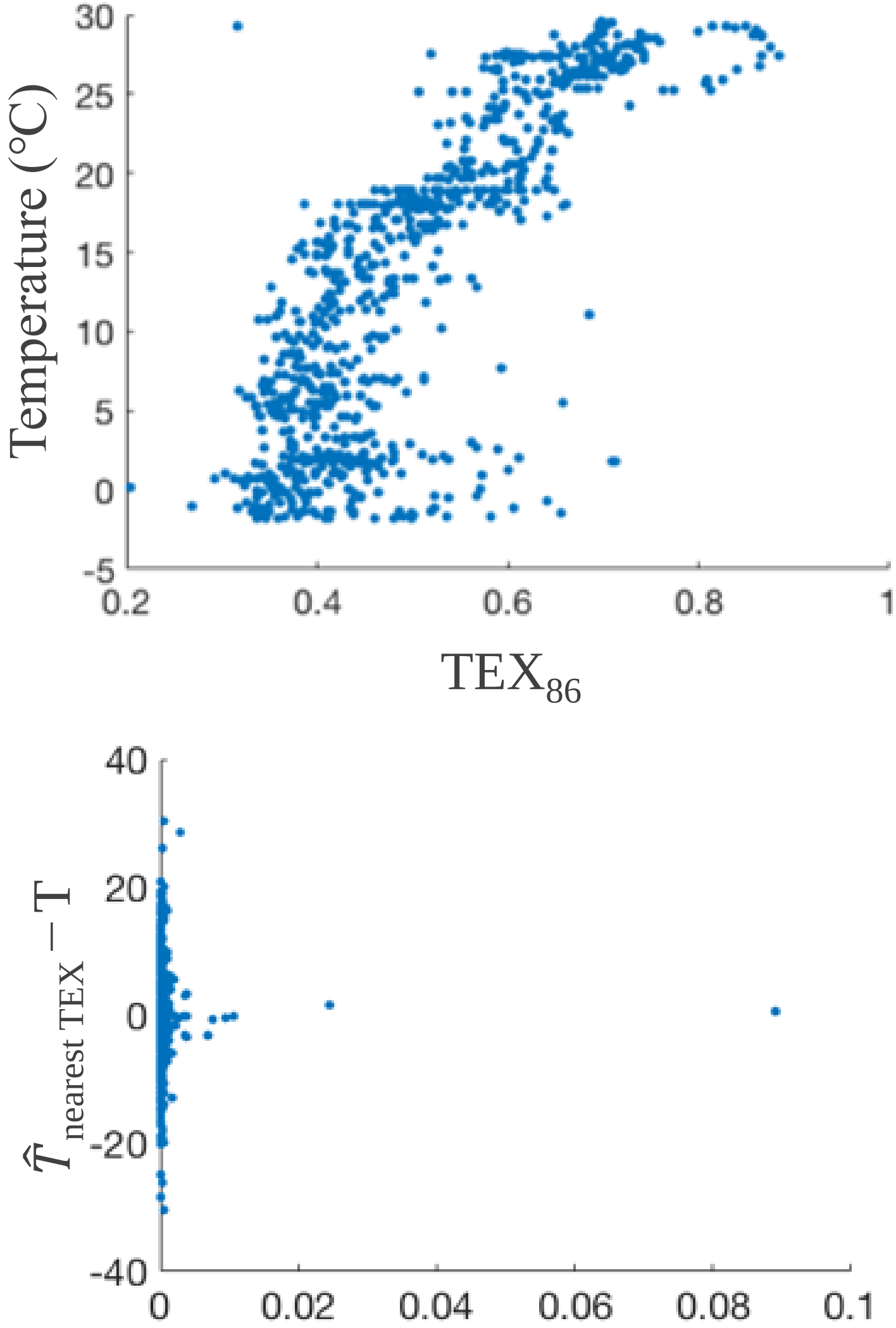

Temperature (°C)
30
25
20
15
10
5
0
-5
0.2
0.4
0.6
0.8
1
TEX86
$\hat{T}_{\text{nearest TEX}} - T$
40
20
0
-20
-40
0
0.02
0.04
0.06
0.08
0.1
Distance to nearest TEX86 point

Figure 4

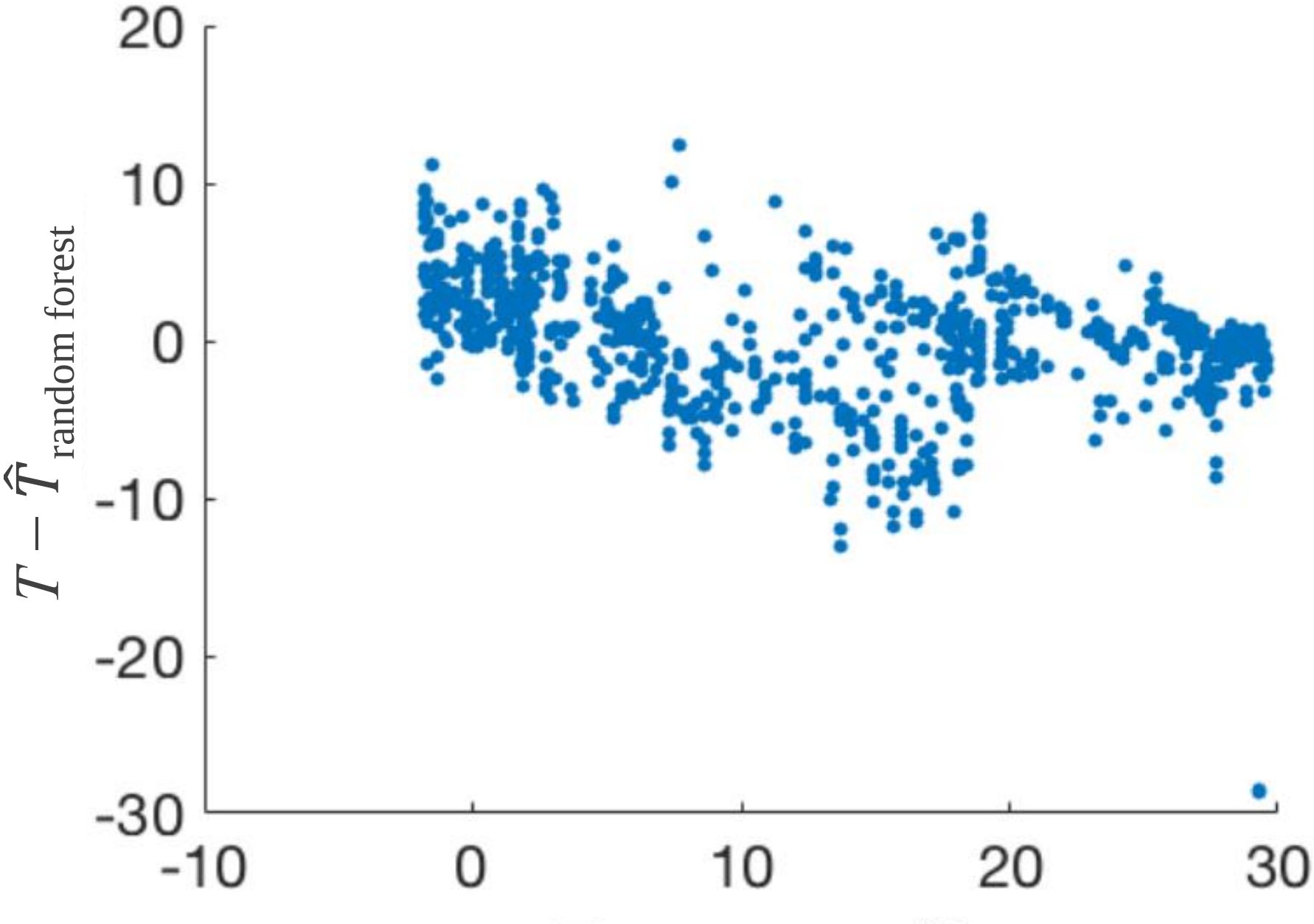

20
10
0
-10
-20
-30
-10
0
10
20
30
$T - \hat{T}_{\text{random forest}}$
Temperature (°C)

Figure 5

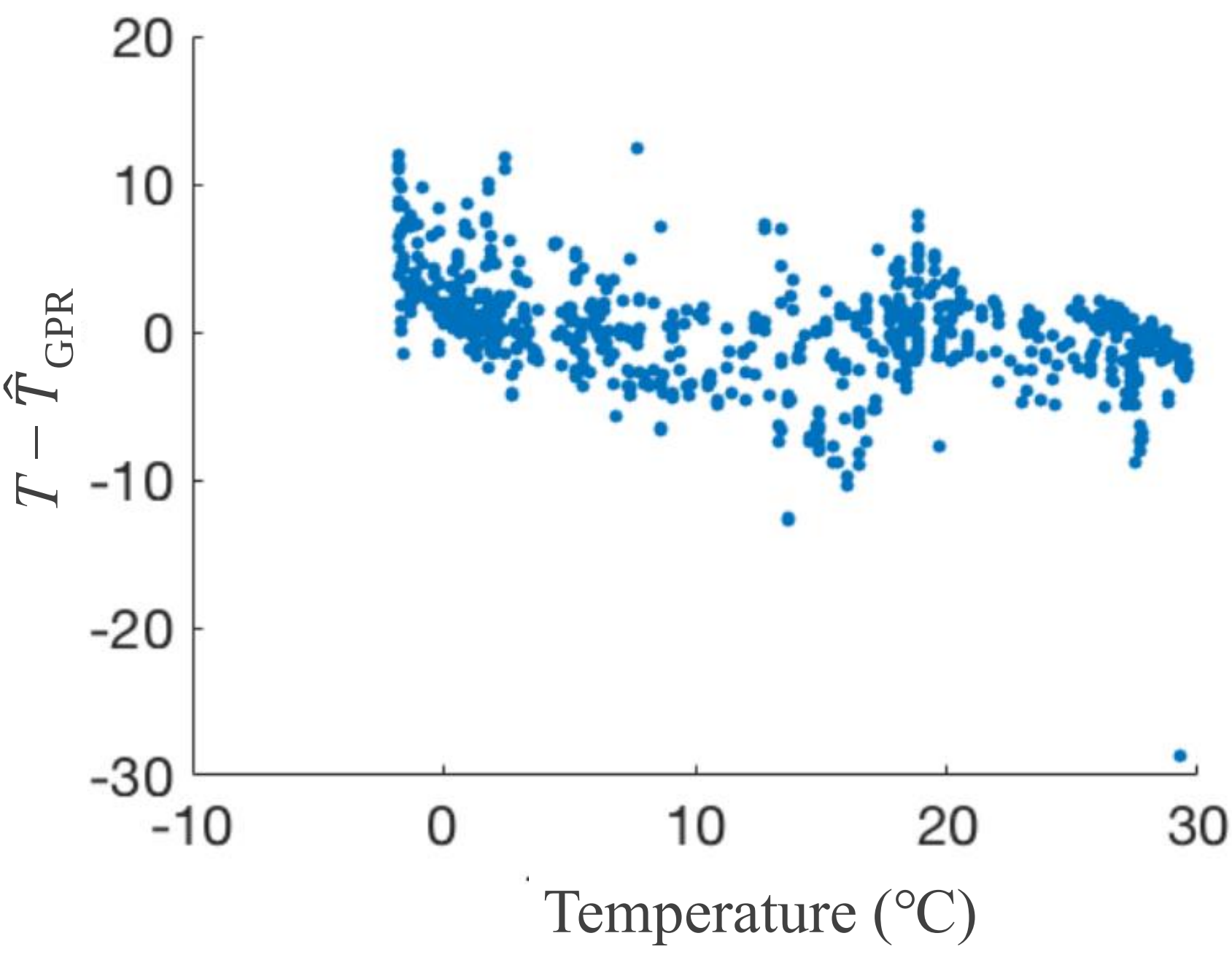


Figure 6

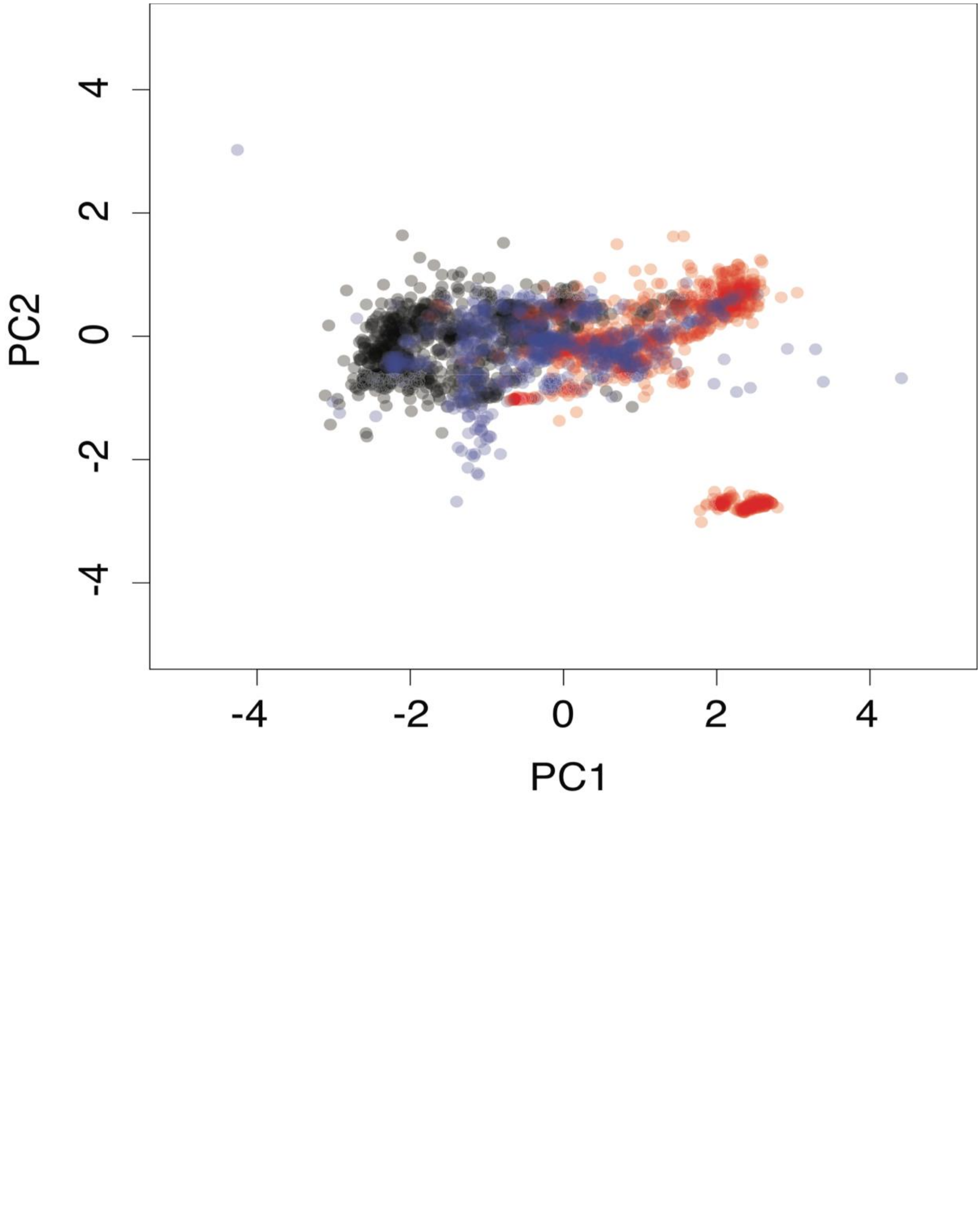


Figure 7

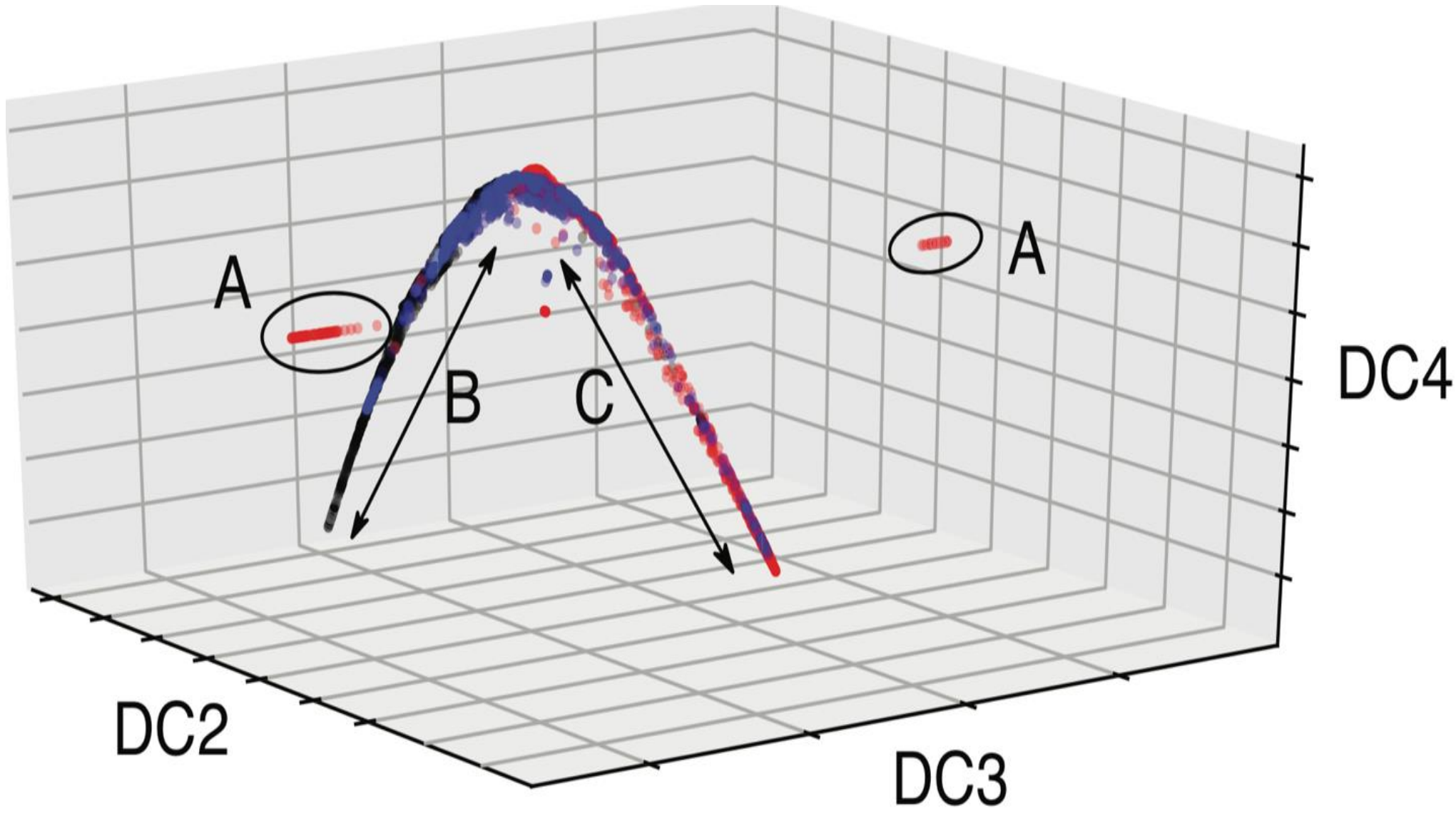


Figure 8

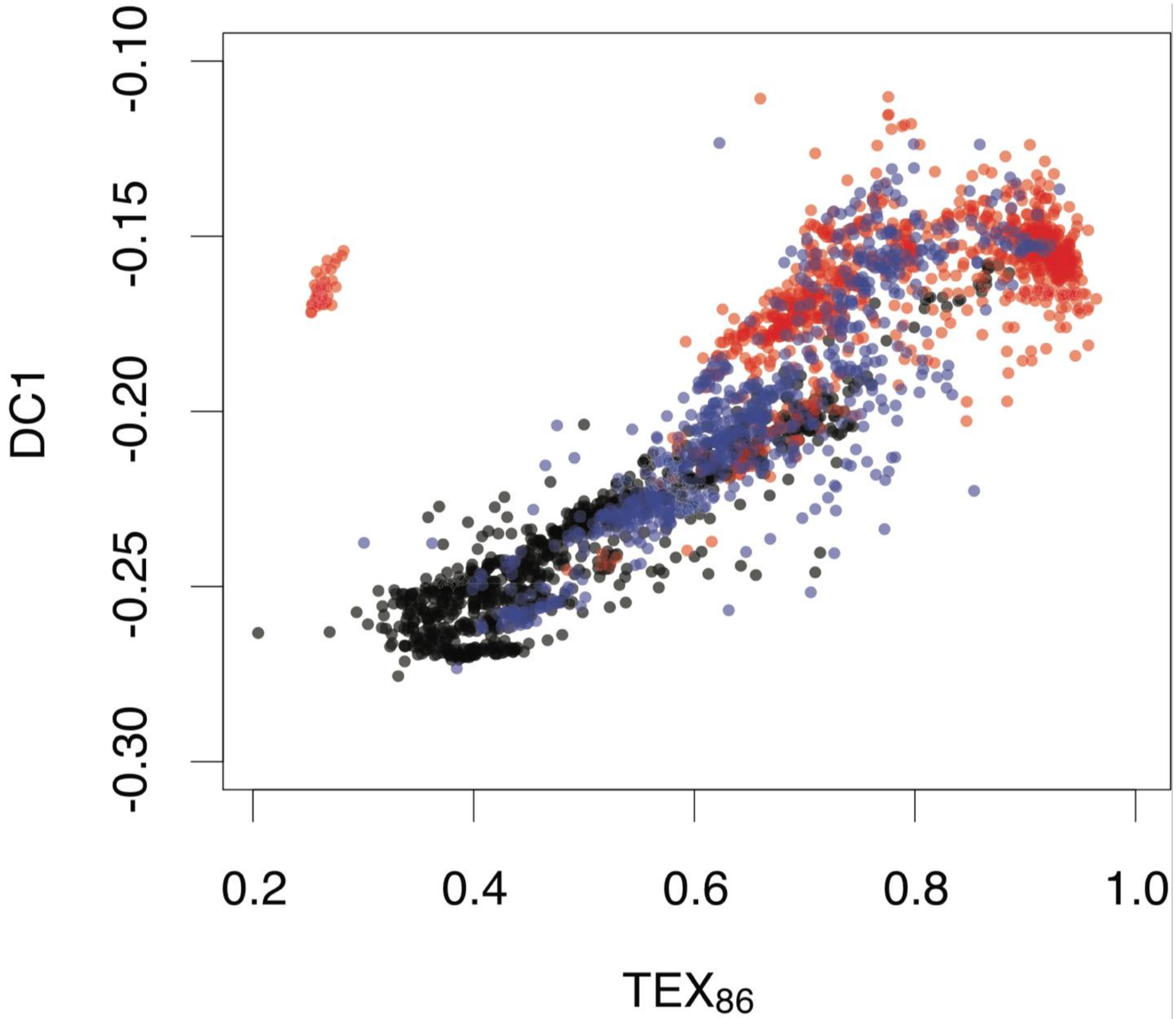


Figure 9

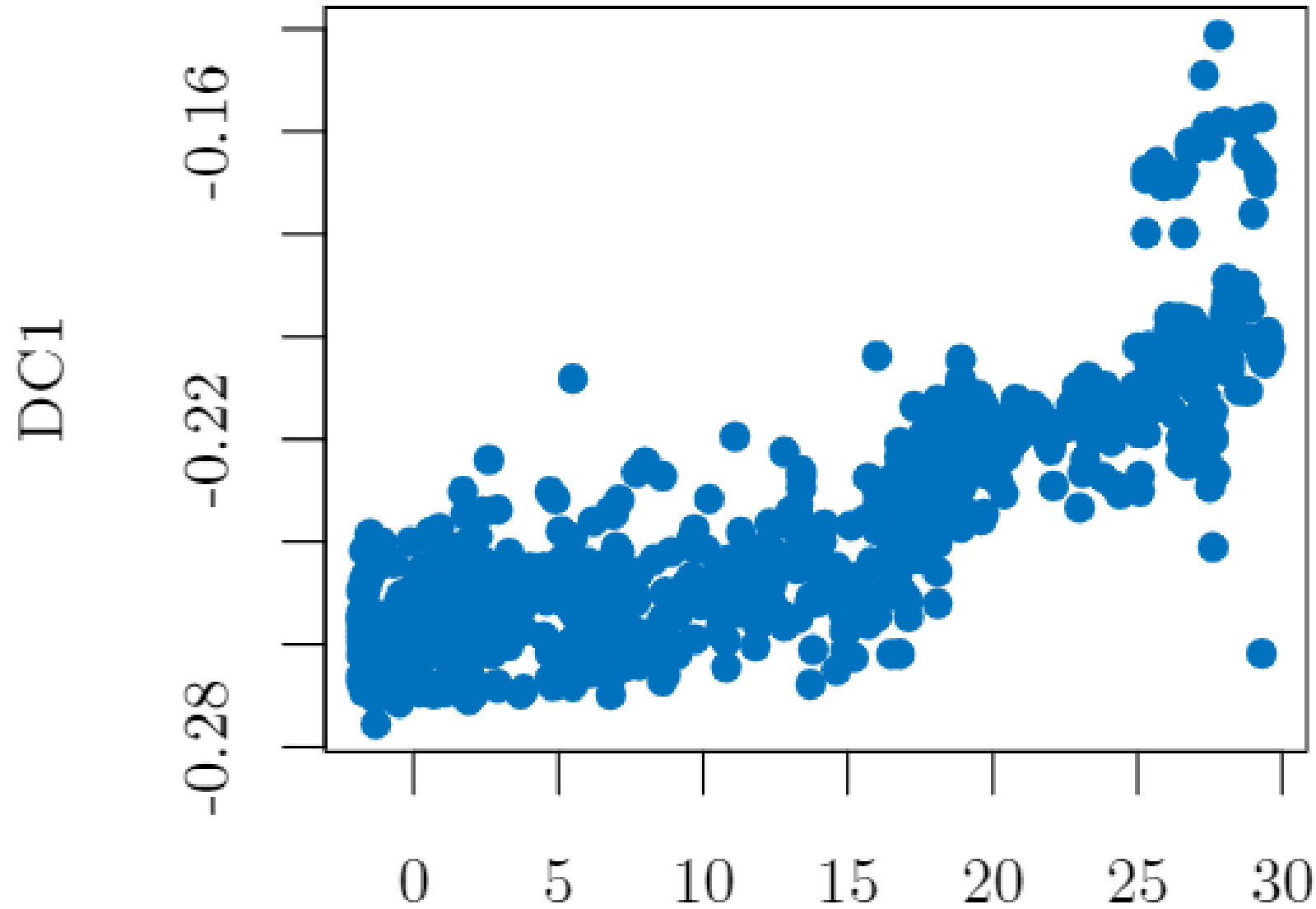


Figure 10

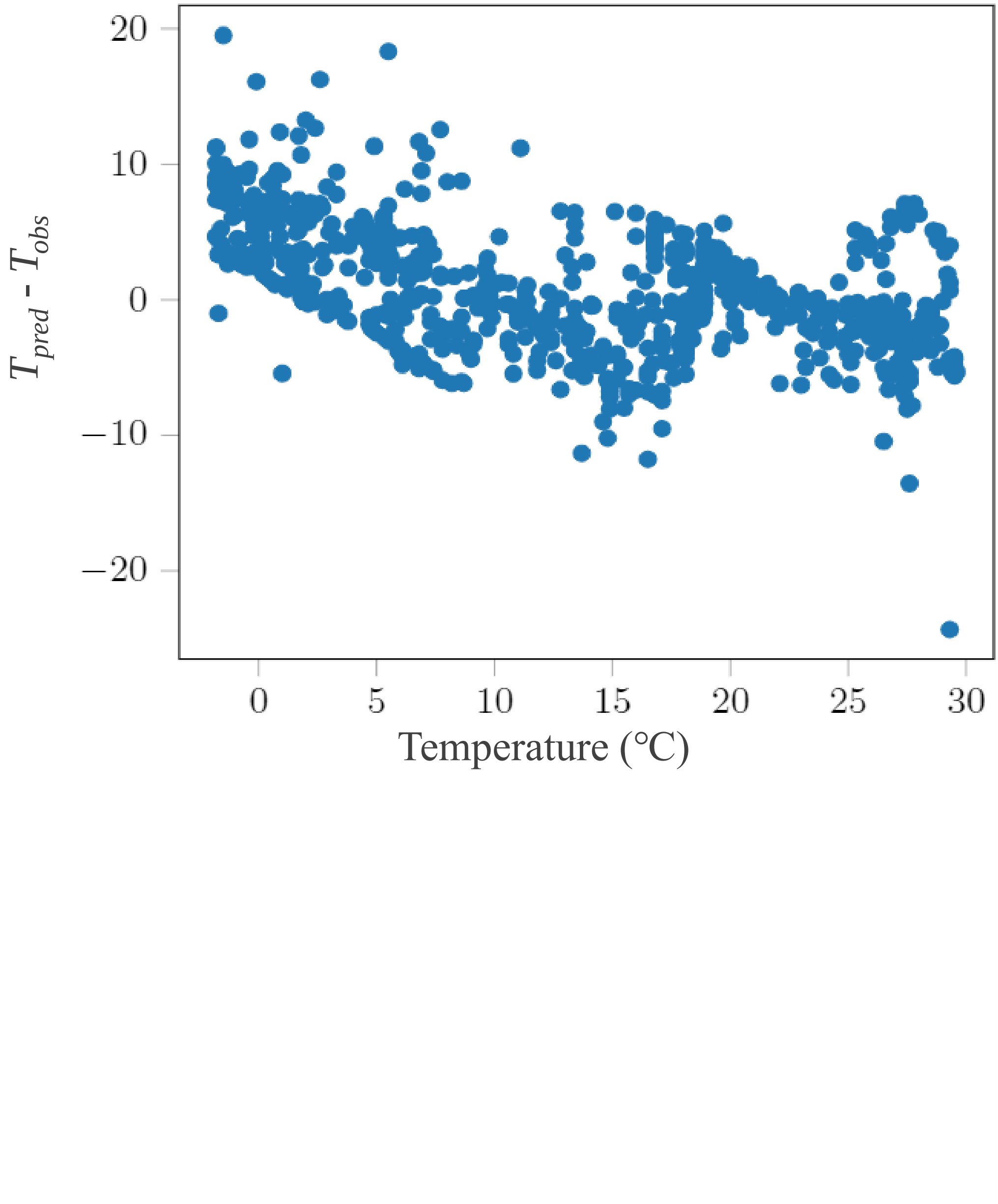


Figure 11

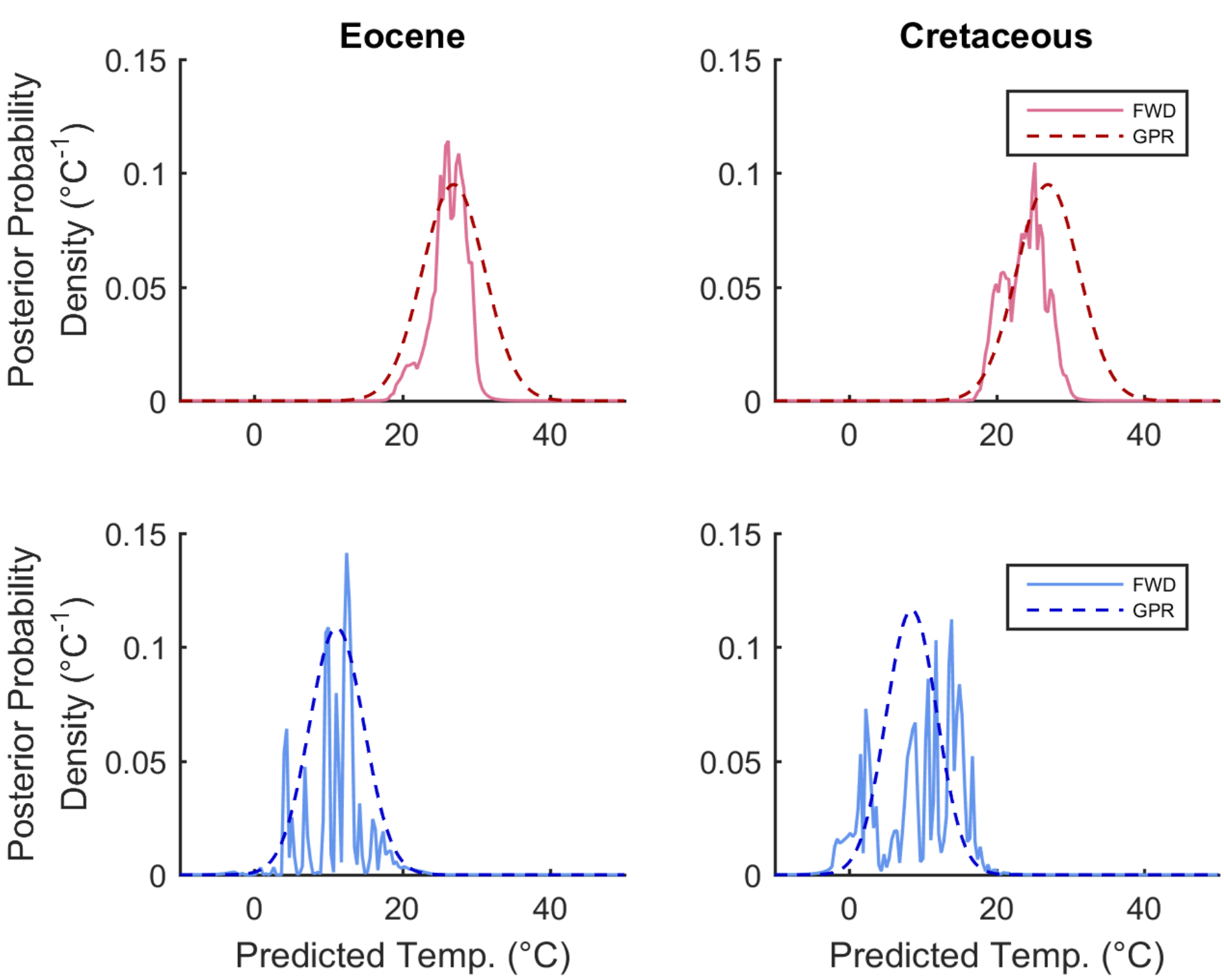

Eocene
Cretaceous
Posterior Probability Density (°C^{-1})
0.15
0.1
0.05
0
0
20
40
FWD
GPR
Predicted Temp. (°C)

Figure 12

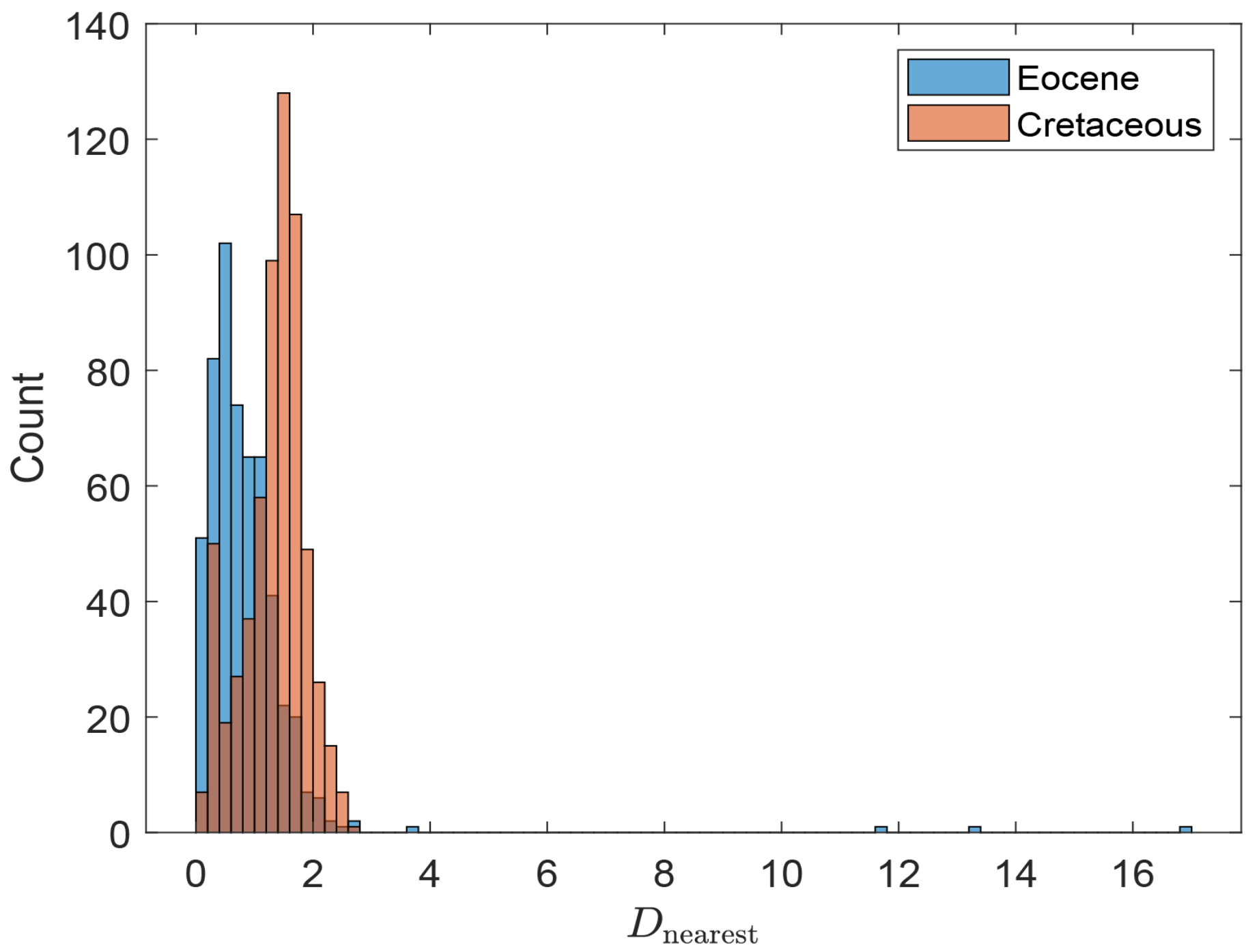

Eocene
Cretaceous
Count
140
120
100
80
60
40
20
0
0
2
4
6
8
10
12
14
16
D_{nearest}

Figure 13

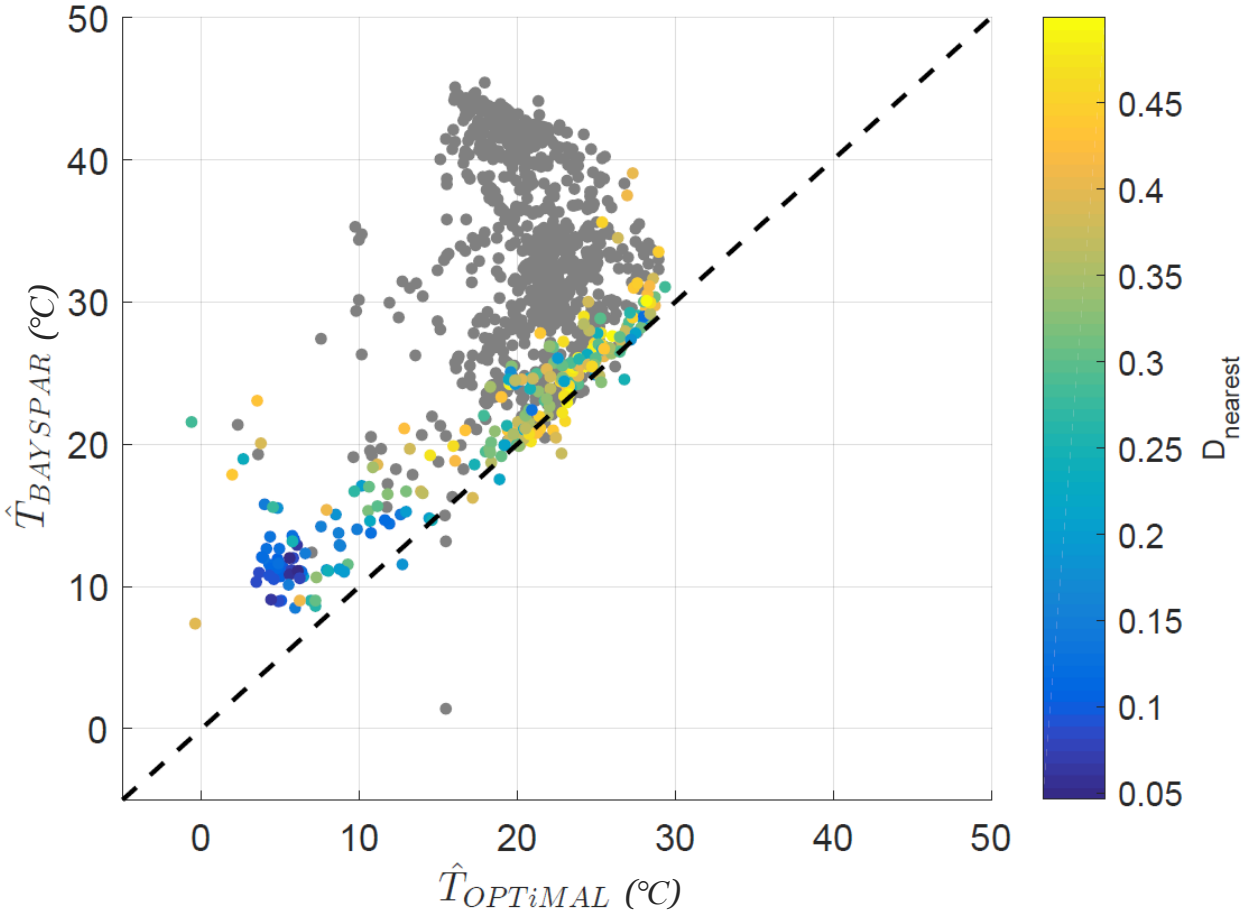

50
40
30
20
10
0
0
10
20
30
40
50
$\hat{T}_{BAYSPAR}$ (°C)
$\hat{T}_{OPTiMAL}$ (°C)
0.45
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
$D_{nearest}$

Figure 14

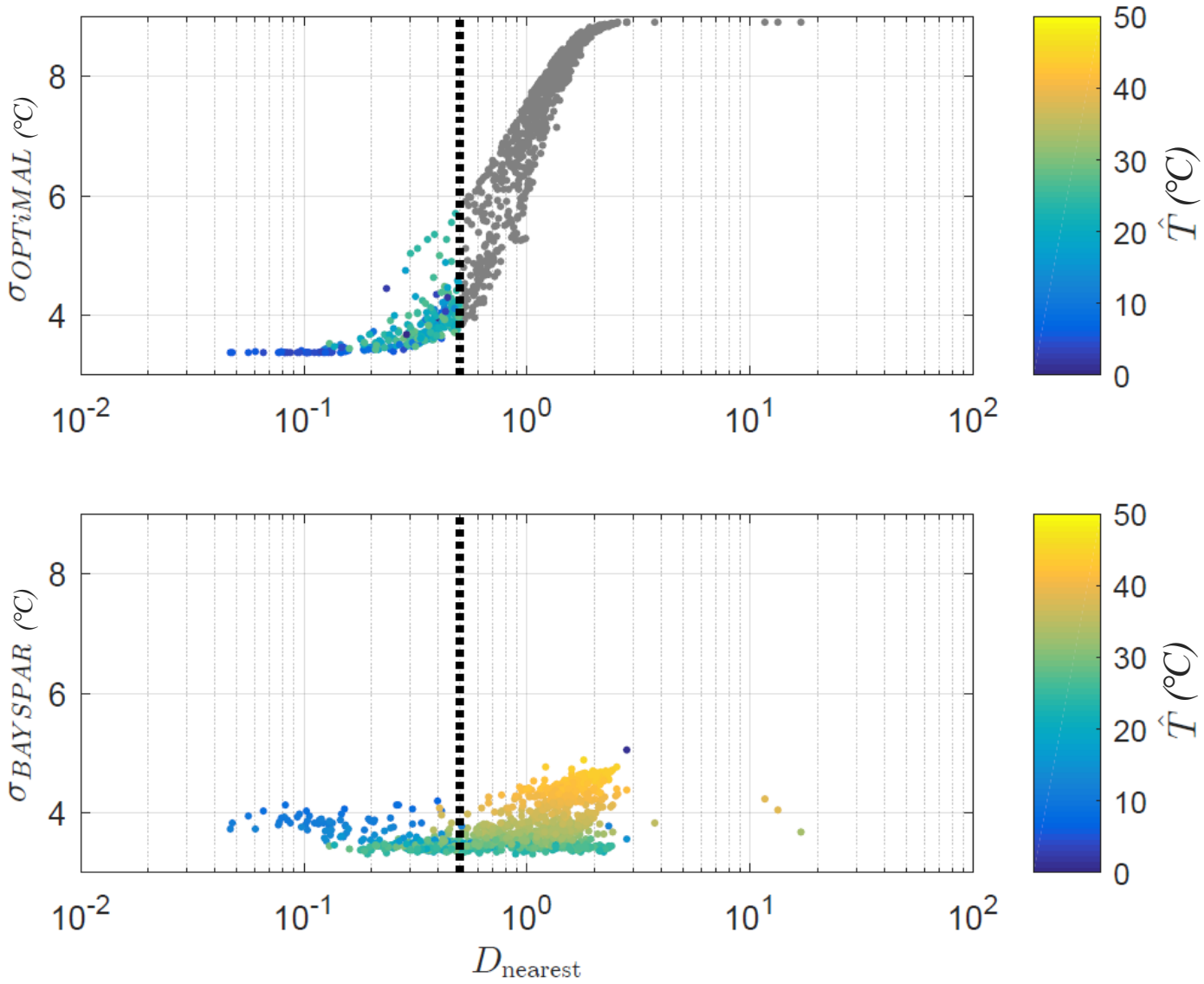
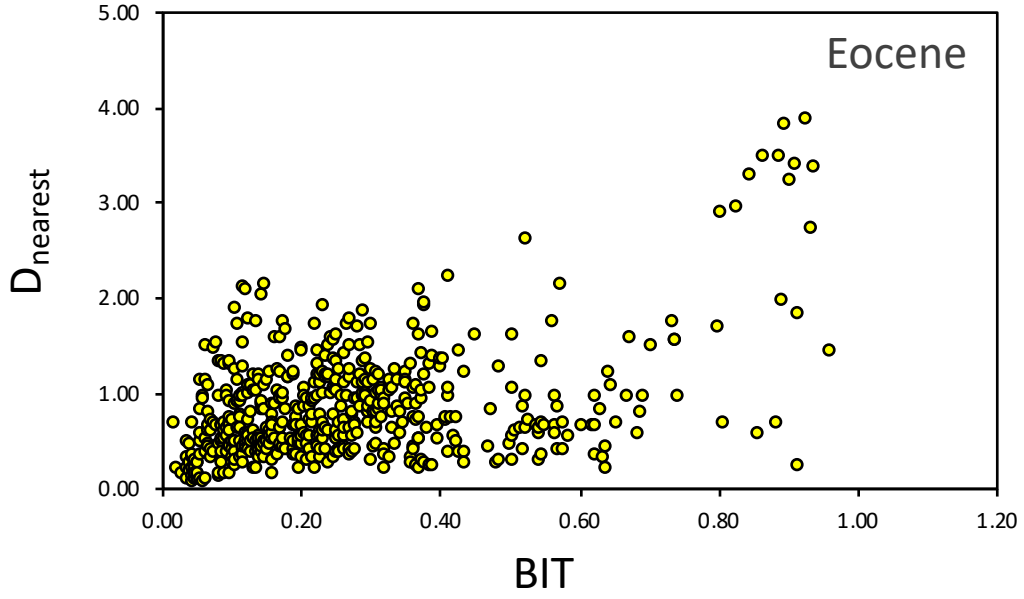
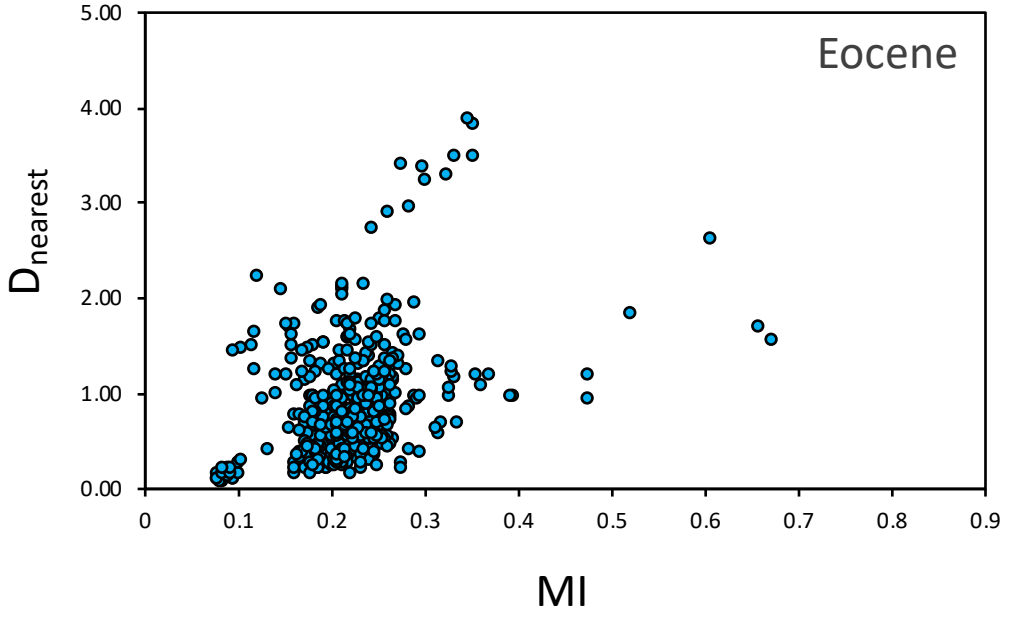
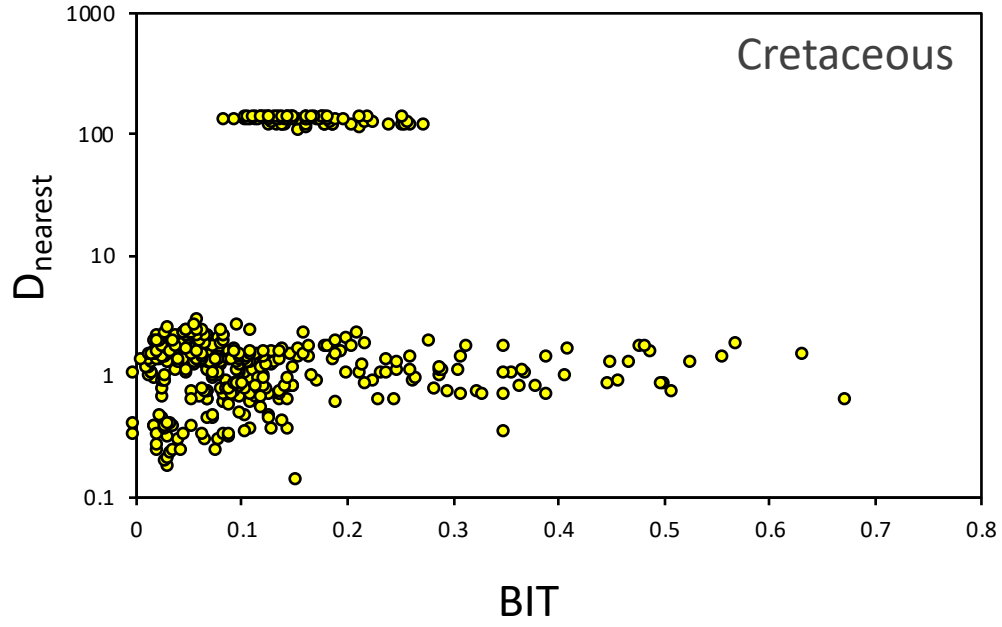
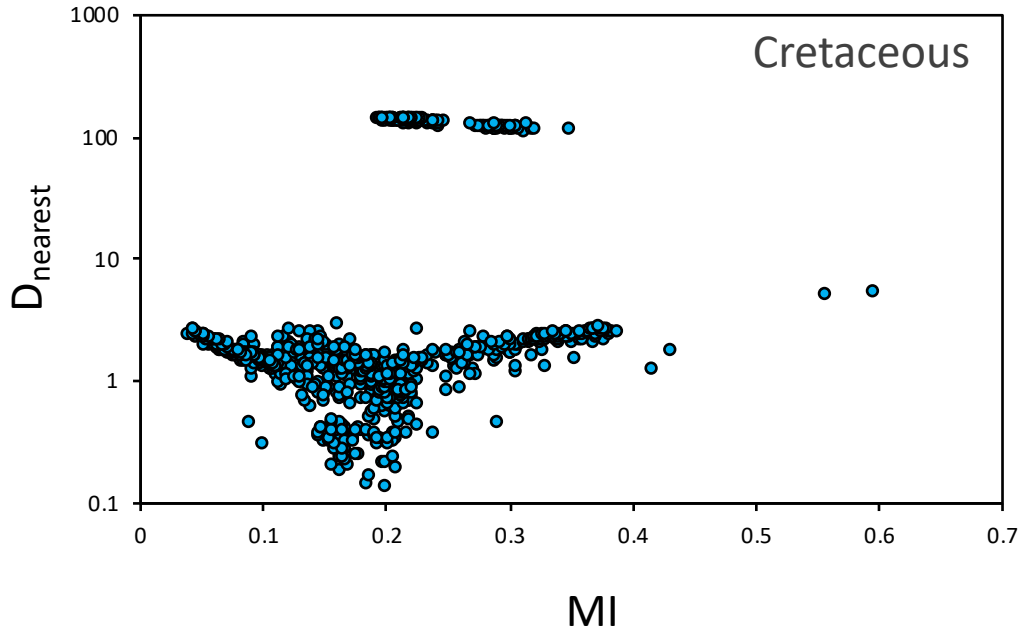

$\sigma_{OPTiMAL}$ (°C)
$\sigma_{BAYSPAR}$ (°C)
$\hat{T}$ (°C)
D_{nearest}
8
6
4
10^{-2}
10^{-1}
10^{0}
10^{1}
10^{2}
50
40
30
20
10
0

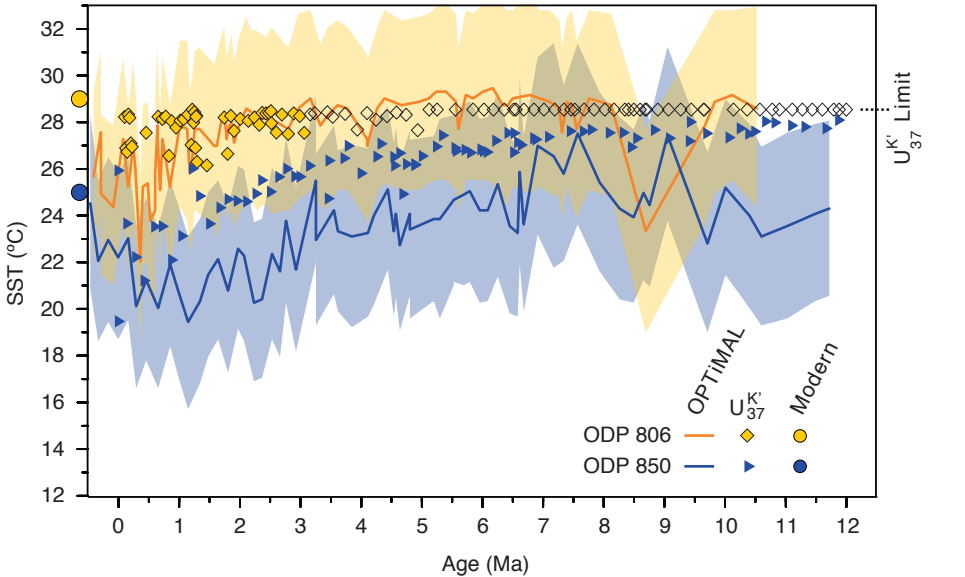
Figure 15


Eocene
$D_{nearest}$
BIT
0.00
1.00
2.00
3.00
4.00
5.00
0.00
0.20
0.40
0.60
0.80
1.00
1.20


Eocene
$D_{nearest}$
MI
0.00
1.00
2.00
3.00
4.00
5.00
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9


Cretaceous
$D_{nearest}$
BIT
0.1
1
10
100
1000
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8


Cretaceous
$D_{nearest}$
MI
0.1
1
10
100
1000
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7


SST (°C)
32
30
28
26
24
22
20
18
16
14
12
0
1
2
3
4
5
6
7
8
9
10
11
12
Age (Ma)
$U^{K'}_{37}$ Limit
OPTiMAL
$U^{K'}_{37}$
Modern
ODP 806
ODP 850