***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.

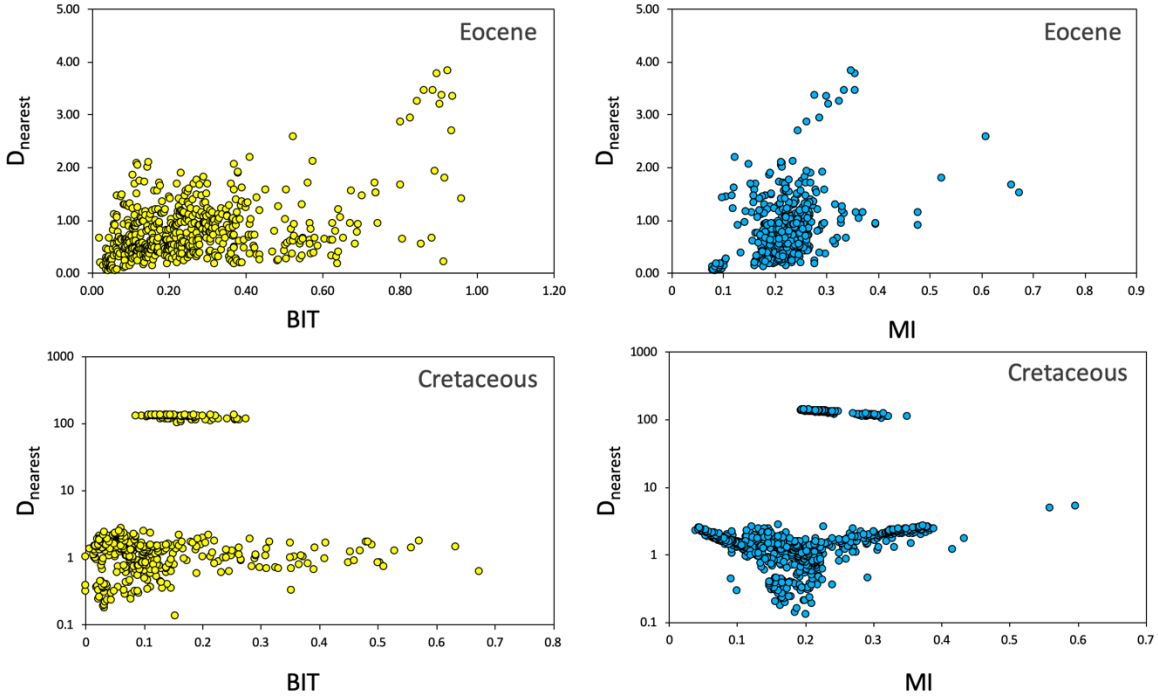
The distance metric that we present is focused on identifying non-analogue fossil GDGT assemblages, relative to those observed in the modern core-tops and scaled to the temperature-dependence observed in the modern GDGT - sea surface temperature (SST) relationship. Our aim is to determine how well-constrained 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