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 Reviews:</u>

<u>Anonymous Referee #1 Received and published: 4 August 2019</u>

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