Clim. Past Discuss.,
https://doi.org/10.5194/cp-2019-60-SC1, 2019

Climate of the Past Discussions

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

Jessica Tierney

jesst@email.arizona.edu

Received and published: 6 July 2019

In this work, Eley et al explore both modern and ancient isoGDGT distributions and propose a new method to calibrate GDGTs to temperatures called OPTiMAL, based on Gaussian process regression. Since I'm pretty interested in TEX86 and stats, I decided to give this a read and a review. I'm glad to see a growing interest in improving GDGT proxy calibration and the use of different statistical approaches to do so. 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.

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? 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. 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.

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?)

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.

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.

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. 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.

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.

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. 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.

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).

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.

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

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

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).

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.

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

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.

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

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.

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

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.

Line 162: “wildly insensitive” - use more formal language.

Line 166: “master control” - use more formal language.

Line 178: “powerful mathematical tools” - hyperbole.

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

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

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.

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

Line 215: “so-called”? That’s what it’s called. Again no need to define R^2 as it’s a common regression metric.

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.

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.

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

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?

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.

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.

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

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?

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.

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.

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).

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

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.

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.

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

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?

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

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.

Line 464: So is this GP model formally Bayesian?

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?

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).

Line 511: "This uncertainty is not apparent from estimates generated by BAYSPAR or ðİŚĞðİŘÿðİŚŃ' 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.

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

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

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).

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

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

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).

Line 565: "whereas BAYSPAR continues to make SST predictions up to and exceeding 40 oC 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.

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.

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.

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?

Interactive comment on Clim. Past Discuss., https://doi.org/10.5194/cp-2019-60, 2019.