Thank you for the opportunity to review this manuscript. It has already had 2 rounds of critical but constructive reviews. These, and the authors' occasionally somewhat defensive responses, expose two things: (1) the mathematical complexity of the chosen approach for the anticipated user community, that is not necessarily equipped with the required math skills and (2) the different views within that same community about how to move forward with improving TEX_{86} as a proxy for SST. About the former, may I congratulate the author team to have taken upon themselves the difficult task of bringing multiple research fields together in a multidisciplinary product. This is greatly applauded! The authors have succeeded to make the mathematical concept of the approach understandable for the target audience.

For the second point, the community as a whole has two problems to solve, as the authors rightfully state: reducing the residual error in the core-top calibration and understanding the GDGT-temperature relationship extrapolated beyond modern ranges. While the authors provide crucial improvements on the residual error, the manuscript itself is strongly directed towards the 'extrapolation problem', for which their method provides no new insights or improvements. I'll explain further below.

As proxy-applier I look purely at the applicability of Optimal to reconstruct past environmental change, as I am sure that is what the authors designed Optimal for. My general outset for the isoGDGT-based proxies would be for now to keep things as simple as possible. This is because we - and I fully agree with the authors here- lack the full mechanistic understanding of proxy functioning and confounding factors. Although I also agree with Dr. Tierneys review that there definitely are some strong hints on this front, these are as yet too poorly constrained and quantified to be properly accounted for.

Optimal demonstrated me that the residual error is in part due to mathematical oversimplification of the TEX86 index, but the other part must come from other reasons: non-thermal effects, Archaeal community changes, etc. At the same time, their study has also strengthened my conviction of the validity of the TEX_{86} index as good, simple, easy-to-use first-order approach to reconstruct past SSTs from isoGDGTs, although the authors do not give this aspect too much credit.

Most of the TEX86 users will, like me, be unable to fully understand the mathematical background (I expect the paper will lose the majority of the paleoceanographic readership at sentences like "*We choose a zero-mean Matern 3/2 kernel for the applications below. Note, however, that since we are working in ilr-transformed coordinates, this corresponds to a prior assumption of uniform compositions at all temperatures, i.e. all components are equally abundant.*" (lines 478-481). Crucially, after their mathematical iterations, we have a smaller residual error (great!) but we still do not understand mechanistically which non-thermal or archaeal community effects cause scatter in the isoGDGT composition. Crucially, it is the mechanistic understanding of proxy functioning that the community is after: we may not only improve GDGT-SST relationships but at the same time reconstruct past ocean environmental conditions beyond 'just' SST. The non-thermal contributions to isoGDGT distributions are not a nuisance, but an opportunity to learn more about the paleoceanographic/depositional conditions in ancient sediment archives. I understand the prime scope of the paper is to improve the SST reconstructions, but I feel the above aspect is ill-acknowledged in the paper. OPTIMAL kicks out data that has non-analogue isoGDGT distributions, irrespective of the underlying reason for the non-analogy. It is a pity that the model cannot distinguish between these criteria at all.

My main point of criticism is this: Why is there so much emphasis in the manuscript on the extrapolation of TEX86 beyond modern-day range, while, and I cite the authors here: "*The machine learning tools described above, which are ultimately based on the analysis of nearby calibration data*

in GDGT space, are fundamentally designed for interpolation." Indeed, the authors approach does not really solve the outstanding question of the continuation of the regression beyond modern-day SST range, other than "be careful with that" and "consider the 'true' error bars". Of course, these points are valid, but result from their review of the isoGDGT literature rather than from their mathematical approach. Other papers, like Hollis et al., 2019 already reviewed in detail the uncertainties in TEX86 calibrations at high SSTs. Crucially, the evidence of a robust temperature control on TEX86 index values above modern SST range is overwhelming, albeit with a warm bias compared to other proxies (e.g., Bijl et al., 2010; Frieling et al., 2017; Tierney et al., 2017; Hollis et al., 2019; Crouch et al., 2020). Extrapolating a linear, Bayesian or exponential regression outside of the modern core-top data, with or without the Red Sea data, are all wrong models, and the uncertainty in the resulting SST reconstructions for the Paleogene should actually include the combined errors of all these models. The discrepancy between mesocosm, hydrothermal vent and fossil GDGT communities from warm climates are also amply reported (e.g., Schouten et al., 2013; Cramwinckel et al., 2018; Hollis et al., 2019), and do not really give conclusive evidence for which kind of extrapolation model fits best. If the approach used in OPTIMAL is specifically designed for interpolation, and the approach is not really providing new constraints on how to extrapolate the TEX86-SST relationship beyond modern-day range, why then focus on that aspect so much in this paper? Why would you not focus on presenting a great new approach to reconstruct SST from isoGDGTs for data that fall within the modern-day SST range, and explicitly discourage users in non-analogue, warm studies to use OPTIMAL? I think that is how this work will be used in the community. To give an exact example, a new dataset of Eocene isoGDGT data with TEX86 index values over 0.73 will not use Optimal because the data will be deemed unreliable by the model. Any reviewer of Eocene GDGT data who means to disqualify Eocene SST reconstructions from GDGTs based on this manuscript will receive the response that the OPTIMAL approach is unqualified for that rejection because it is specifically designed for interpolation. So what is the contribution of this paper to the 'high-SST community' that justifies so much attention about that in this paper?

I find the problem statement of lines 107-108 illustrative in this sense. Because, by default, any SST reconstructions using TEX86 that exceed modern SSTs will have strong distance from modern core top data. But that is just another way of saying there are uncertainties in the extension of the TEX86-SST relationship beyond the modern-day range (e.g., the exponential/linear discussion). Indeed, in many publications, these uncertainties are stated (see, e.g., Hollis et al., 2019), and currently we cannot resolve this issue, but neither can Optimal, other than quantifying the mathematical distance between isoGDGT compositions in hothouse climates from those in the modern core-top dataset. Sure, there were Baobab trees on Antarctica during the Eocene, the world looked fundamentally non-analogue! And of course, the absolute SSTs come with considerable errors and uncertainties. I totally agree with the authors that the errors that are used on these SSTs do not give proper credit to the uncertainties in absolute SSTs. However, also above the modern TEX86 index value range, the proxy functions as expected when it comes to trends (e.g., Bijl et al., 2013; Crouch et al., 2020), which means that temperature must have a strong control over isoGDGT distributions also in those hothouse climate states. That this does not always come out of the modern mesocosm studies under warm SSTs does not discredit that geologic observation.

In short, my suggestion would be to focus on the strength of the Optimal approach, notably its application in analogue environmental settings, and steer away from the extrapolation problem that advanced mathematics is simply unable to solve.

That said, I find the term 'non-analogue' confusing and vague. It seems to refer to isoGDGT assemblages that have no near analogue in the core top data (irrespective of the underlying reason). In general, most paleoclimate records are non-analogue, if you include geographic information. For instance, the ACEX dataset may have isoGDGT assemblages that have an analogue in the modern

core top data, but that analogue will come from the warm-temperate ocean and not from the Arctic. If you include paleolatitude in the assessment of analogy, most paleo-GDGT assemblages are non-analogue. This is not specifically an issue for OPTIMAL, but the term non-analogue needs further specification in this sense.

Cheers,

Peter Bijl

Specific comments:
Lines 85-89: The rationale for using TEXL outside of its modern limit of 15 degrees has also been because it was unknown what factor limited its use: SST itself or a factor that correlates to SST? Bijl et al., 2013 plotted both TEXH and TEXL because of the high paleolatitude position of the site. The use of TEXL to temps below 15 degrees modern SSTs could also have been a latitudinal restriction (i.e., use TEXL South/North of 55 degrees latitude, because SSTs of 15 degrees correspond to that latitude. See explanation in Bijl et al., 2013. It was however strange that while proxy fit at high SSTs was similar, TEXH and TEXL were consistently offset when applied in the Eocene sample set.

Lines 106-107: Here the authors should emphasize the abundant evidence of a temperature control on isoGDGT assemblages that have non-analogue compositions (see above). That is, if this section is retained after revision.

Lines 122-133: Focus this culture paragraph on what these show for the isoGDGT-temperature relationships in modern analogue temperatures, as that is what OPTIMAL is all about. You can still stress that there is complexity in GDGT production to growth temperature (lines 146-147) which explains the remaining scatter in the modern core top dataset rather than provides information on the extrapolation of its regression.

163-172: Does this point also hold true for SST reconstructions within the modern T range, or only for the extrapolation out of that? Clarify and specify.

Lines 702_704: So, if I understand correctly, Dnearest can be small for samples and modern sites that have structurally different SSTs? What is then the real significance of Dnearest as a criterium?

Lines 724: It is not necessarily non-analogue behaviour, but confounding environmental or biological factors, which as yet cannot be quantitatively corrected for.

Lines 726: I strongly disagree with the conclusions that 'issues have not been clearly stated and circumscribed'. There is no proxy in paleoclimate research that comes with a longer description of methods, analysis data, data scrutinization and critical evaluation than the biomarker-based proxies. On the front of error the field is evolving, along with that in other proxies. The community has been extremely careful in interpreting and presenting their data, against modern data, against other data in the same sediments, against other proxies for SST. And no, that has not eroded confidence, it has set a bar for critical assessment of assumptions in other proxies. Rephrase to make this point clear.

References

Bijl, P. K., Bendle, A. P. J., Bohaty, S. M., Pross, J., Schouten, S., Tauxe, L., et al. (2013). Eocene cooling linked to early flow across the tasmanian gateway. *Proceedings of the National Academy of Sciences of the United States of America, 110*(24), 9645-9650.

Bijl, P. K., Houben, A. J. P., Schouten, S., Bohaty, S. M., Sluijs, A., Reichart, G. -., et al. (2010). Transient middle eocene atmospheric carbon dioxide and temperature variations. *Science, 330*, 819-821.

Cramwinckel, M. J., Huber, M., Kocken, I. J., Agnini, C., Bijl, P. K., Bohaty, S. M., et al. (2018). Synchronous tropical and deep ocean temperature evolution in the eocene. *Nature, 559*, 382-386.

Crouch, E. M., Shepherd, C. L., Morgans, H. E. G., Naafs, B. D. A., Dallanave, E., Phillips, A., et al. (2020). Climatic and environmental changes across the early eocene climatic optimum at mid-waipara river, canterbury basin, new zealand. *Earth-Science Reviews, 200* doi:10.1016/j.earscirev.2019.102961

Frieling, J., Gebhardt, H., Huber, M., Adekeye, O. A., Akande, S. O., Reichart, G. -., et al. (2017). Extreme warmth and heat-stressed plankton in the tropics during the paleocene-eocene thermal maximum. *Science Advances, 3*(3) doi:10.1126/sciadv.1600891

Hollis, C. J., Dunkley Jones, T., Anagnostou, E., Bijl, P. K., Cramwinckel, M. J., Cui, Y., et al. (2019). The DeepMIP contribution to PMIP4: Methodologies for selection, compilation and analysis of latest paleocene and early eocene climate proxy data, incorporating version 0.1 of the DeepMIP database. *Geoscientific Model Development, 12*(7), 3149-3206. doi:10.5194/gmd-12-3149-2019

Schouten, S., Hopmans, E. C., & Sinninghe Damsté, J. S. (2013). The organic geochemistry of glycerol dialkyl glycerol tetraether lipids: A review. *Organic Geochemistry, 54*, 19-61. doi:10.1016/j.orggeochem.2012.09.006

Tierney, J. E., Sinninghe Damsté, J. S., Pancost, R. D., Sluijs, A., & Zachos, J. C. (2017). Eocene temperature gradients. *Nature Geoscience, 10*(8), 538-539. doi:10.1038/ngeo2997