Can we reliably reconstruct the mid-Pliocene Warm Period with sparse data and uncertain models?

Annan et al. *submitted to Climate of the Past* (https://doi.org/10.5194/egusphere-2023-1941)

In this contribution, Annan et al. attempt to reconstruct mid-Pliocene Warm Period (mPWP) surface temperatures by combining an ensemble of PMIP-based model simulations with prior compilations of mPWP sediment proxies (namely, alkenones and Mg/Ca). They report a likely mean global atmospheric surface warming of ~3.6 ± 1˚C (for their "preferred" solution) relative to the pre-industrial state, with estimates ranging anywhere from 1.0 to 4.7˚C average warming depending on the particulars of their chosen proxies (Table 2).

The analysis is interesting, insofar as it represents the first (that I'm aware) attempt at assimilating the mPWP. However, numbers like those mentioned above matter — they get passed on in the literature for generating model boundary conditions or climate-relevant constraints like ECS. Given such, I cannot in good faith recommend this manuscript for publication. The overarching concerns I have are that 1) the reconstruction decisions being made seem almost entirely *ad hoc* and otherwise unsupported in the present context and, to this end, 2) no validation efforts have apparently been attempted by the authors. From a paleoclimate- and proxy-interpretative standpoint, the study's scientific explorations feel rather cursory (e.g., the Discussion comprises only a paragraph or two of text), and I do not feel Annan et al. clearly answer their manuscript's own title, "Can we reliably reconstruct the mid-Pliocene Warm Period with sparse data and uncertain models?" I elaborate:

1. As a statistical methodology, offline data assimilation permits one to turn various "knobs" when generating paleoclimate reconstructions. While to first order this includes decisions on *which* models and proxies to include in the assimilation (aspects I am generally fine with in the authors' study, though please see Specific Point c), below), other important knobs include the degree of covariance localisation applied and the amount of uncertainty attributed to the individual proxy observations. Despite having considerably less proxy constraints as well as different models, Annan et al. are largely content to simply follow the empirical methodology presented in AHM22 without providing supporting evidence (such as careful validation testing; see Point 2, below) relevant to the present datasets / context. Further, Annan et al. include several pre-processing steps in their assimilation approach that do not appear supported in this study or in the prior paleoclimate literature outside of AHM22. I elaborate:

   i. The authors apply a localisation of 2500km as in AHM22. While I am highly sceptical of such a low value in general (based on prior validation efforts using similar data/methods, e.g., Tardiff et al., 2019, Tierney et al., 2020, Osman et al., 2021, , Erb et al., 2022 etc., each of whom use localisation ≥12,000 to 24,000 km for significantly larger proxy compilations), I am willing to accept its use in AHM22 where some, if limited, validation efforts were performed and where proxy data coverage was well over an order-of-magnitude larger. This is not the case in the present study, wherein only a *maximum* of 23 locations are assimilated (and only 14 in the "preferred" set-up), and thus the majority of points on Earth (given regional clustering of those proxies that do exist) never even "feel" an update in the assimilation process. Indeed, I suspect this small localisation radius is likely why the authors feel compelled to bias-adjust ("re-center", see Point *iii*. below) their multi-model ensemble mean to the proxies in the first place, given all grid cells outside the collective proxies' 2500km radii will simply remain at the proxy-adjusted multi-model ensemble mean state. See also Specific Point e), below.
   ii. Second, given the core importance of error quantification in data assimilation, the authors are rather *ad hoc* in their treatment of proxy uncertainties in assuming a constant 2˚C error (1σ) across their proxies. While I understand this is an approach the authors have attempted to justify in their past work (again, AHM22) and even later in the present manuscript (see Specific Point f), below), I would appreciate greater effort be taken here in exploring the

influence of proxy uncertainties across sites and proxy types. Prior authors (e.g., Tardif et al., 2019, Tierney et al., 2020, Erb et al., 2021) have gone through considerable effort to validate proxy uncertainties and test their sensitivity. Further, modern proxy system models have been designed specifically to attempt to formalise (in the Bayesian sense, albeit still using empirical relationships) the magnitude-dependent uncertainties assigned to alkenones and Mg/Ca either in temperature or proxy space, yet the authors curiously disregard use of such tools (e.g., page 5, lines 10-13).

iii. Third, the authors invoke various pre-assimilation steps involving EOF's truncations and bias-adjustments of their multi-model mean. It is noted, however, that this re-centering step is not used nor recommended by the vast majority of paleo-data assimilation practitioners, given that it can precondition the posterior result significantly (c.f., the authors' Fig. 3a and Fig. 3b vs. 3c), and thus potentially bias the answer *further* from the "true" state which is unknown. While I can imagine such re-centering may be permissible in certain instances in a modern context where the true state is more or less understood (e.g., operational weather forecasting), who's to say that proxies are perfectly reliable (non-biased) representations in the paleo world? Indeed, this paper's subsequent claims that many of their proxies are *not* reliable indicators of the mPWP (Pg. 7, L34-35) implies a logical inconsistency with their re-centering step more generally.

Thus, the authors' pre-assimilation steps should be either supported by mathematical derivation (not provided in AHM22), reference to the prior relevant literature (also not provided in AHM22*), or, at minimum, empirical validation testing (provided in some capacity in AHM22). Furthermore, the authors' suggestion that rank histograms (Fig. 2, Pg. 6, L15-25) provide such missing validation to support their re-centering step is flawed: discovering a more uniform rank histogram (indicating that roughly half the proxy-derived temperatures sit below the re-centered median PlioMIP temperature, half above) after re-centering the model priors around the proxies is nothing more than the expected outcome of the same re-centering process! And, there are other caveats to the rank histograms still: for example, whereas prior approaches (e.g., Tierney et al., 2020) have tested rank histograms using withheld validation data only – a much more challenging test – this is almost certainly not the case here given the authors' lack of validation testing. In fact, it's unclear how these rank histograms were even calculated. Overall, if the authors wish to use their re-centering step, then use this technique should be subject to stringent validation testing using withheld or independent proxy constraints, mathematical underpinning, and (or) reference to the prior relevant literature showcasing its utility*.

*In introducing their re-centering step in AHM22, the authors cite only their previous study, Annan and Hargreaves (2013), which does not entail a data assimilation-based reconstruction approach and thus does not appear to be a relevant citation in the present context.

2. The authors do not validate their assimilated results using withheld or independent data. I often tell colleagues: "Anyone can create a data assimilation reconstruction (given the various toolsets and data compilations now openly available to do so) but not every reconstruction can be reliably validated." To me, this is what Annan et al. have provided: an interesting data analysis, to be sure, but without validating their reconstructions it's impossible to gauge which (if any) of the authors' various assimilated results should be believed, or whether any of their results actually improve upon the models or proxies in isolation? There are numerous approaches to validating a paleoclimate data assimilated result – the least stringent I'd hazard being leave-one-out proxy validation, the more stringent options being random sample without replacement, regional proxy withholding, or validation using independent (e.g., terrestrial) proxies. When rationalizing each of the various assimilation decisions and "knobs" noted above, the authors should be showcasing targeted validation efforts that support their claims.

Specific Points (Authors' text in quoted italics)

a) Pg. 5 L27-29 "*The Mg/Ca might be considered less reliable as their relationship to SST may depend on site-specific factors including the species analysed, the calibration used, or the seasonality or depth*

habitat of the foraminifera." And, subsequently, Pg. 8 L16-18 "*…we (argue) that the Mg/Ca may be less reliable. We do not explore possible reasons for this here, but they may include site-specific factors including the species analysed, the calibration used, or the seasonality or depth habitat of the foraminifera (e.g. McClymont and Ho et al., 2023).*"

Yet, this is exactly what I feel this manuscript *should* be exploring. Several tools, imperfect or not, *do* exist to facilitate these aims (e.g., Tierney et al., 2019, Gray and Evans, 2019; see also review by Rosenthal et al., 2022). Indeed, understanding the sensitivity of proxy systems during past warm intervals is, to me, a key research area where paleoclimate data assimilation stands to permit real intellectual gains. Past efforts (Tardif et al., 2019, Tierney et al., 2020, Osman et al., 2021) have chosen to assimilate in proxy units precisely for the reason that doing so permits possibility of exploring sensitivity of the assimilation results to, e.g., underlying seasonal bias, depth-habitat influences, species sensitivities, pH, carbon dissolution effects, and sea surface salinity (SSS; among others). Assuming Mg/Ca reflects annual temperatures does not permit such explorations, unfortunately. Taking one example, recent results have suggested a much stronger change in SSS during the past several Ma than previously recognised, which could have influenced Mg/Ca-derived SST during the mPWP; this was not considered in the PlioVar data used here (Rosenthal et al., 2022). Finally, as a minor note, the authors' decision to omit BAYMAG-derived SST-estimates seems unfounded (Pg. 5, L10-13) especially given their decision to average alkenone-derived SST based on two separate estimates.

Speaking of …

b) Pg 5, L4-5. "*The UK37 SST values we take here are the simple average of the calibration of Müller et al. (1998), and the BAYSPLINE calculation, as presented in McClymont et al. (2020).*"

Averaging Muller and Tierney alkenone-derived SST values will be problematic for the tropics, where Tierney and Tingley (2018) illustrated a clear non-linearity of UK37 to SST as UK37 values approach their limit at 1. As this was not accounted for in the Muller et al. (1998) relations, Tierney should not be averaged with Muller-derived SST values in the tropics. (And, given the separation that will be required there, it would be more preferrable that alkenone-derived SST estimates be separated everywhere else.)

c) Pg 2, L31-34 "*While this has the unfortunate effect of reducing the number of usable data points, we note that the points that are masked in this way are coastal in location, which are potentially problematic for data-model comparisons due to the local nature of upwelling dynamics that is not always adequately captured by models.*"
And, subsequently, Pg 4 L22-24 "*This PRISM4 compilation contains 37 data points, reducing to 34 distinct grid points on the regular 5 × 5 degree grid that we use for our SST analysis, of which 23 locations remain after masking to the ocean grid of our ensemble.*"

Not all coastal regions are necessarily areas of downwelling, and in an interval as data sparse as the mPWP it would be useful to know where data are being lost and what effect they would have had on the assimilation should they have been kept. It is concerning that the authors are losing nearly 40% of their potential constraints due to their coarse re-gridding procedure alone. I'd ask that the authors explore this assumption further by either not masking their 5x5˚ grids that intersect land or, alternatively, assimilating these near-coastal points against the nearest available SST value.

d) Pg 6, L24-25 "*We then adopt this recentred ensemble as a prior for a standard Ensemble Kalman Filter (EnKF) assimilation step using the proxy data, similar to that of AHM22 and Tierney et al. (2020).*"

Tierney et al. (2020) did not apply this approach.

e) Pg 7, L18-23 "*If we do not perform the recentering and instead just perform the single step standard EnKF update to the original PlioMIP ensemble, the update is slightly greater than the EnKF step in the*

two-step method, and generally more positive, due to the data being warmer than the ensemble mean. However the increment is still small, as the data are sparse and uncertain. Large areas of the globe are almost unaffected by the assimilation, with a temperature change of less than 0.1˚C. This is an inevitable consequence of having limited sparse data, and points to the influence of the model prior on the final result. Thus, with so few and uncertain data points the final result using a one-step framework would be very heavily based on the initial ensemble."

I find most of this text framed in a very misleading way. First, the fact that large areas of the globe are largely unaffected by the assimilation is not a surprise at all: it's an inevitable consequence of the fact that the assimilation is based on a mere 14 to 23 regionally clustered data points, each with a rather small 2500km localization radius. As noted earlier, this means that a substantial portion of the globe isn't even being updated in the assimilation. In fact, if the color scheme used by the authors in Fig. 3c were centered with, for example, a white color value atop $\Delta T = 0^{\circ}C$, I'd wager most of the globe would show in white rather than (the somewhat misleading) red or blue it currently is. Second, the authors' suggestion that their two-step process (involving an initial re-centering of their multi-model mean atop the proxy estimates prior to assimilation) is somehow "improving" upon the one-step process (assimilation only) is similarly misleading. Given that updates in the two-step process are even smaller than those without it (as expected given the reduced model-proxy offset, c.f., Fig. 3b), the two-step data assimilation could reasonably be viewed as being *less* of an improvement over the one-step method in the absence of validation efforts, since the two-step posterior remains closer to its prior state.

f) Pg. 7, L28-33 "*We take a uniform uncertainty of 2◦C on all of our data points. With so few data points, this estimate is necessarily itself uncertain, but we consider it reasonable for the following arguments. The RMS difference between the original PlioMIP ensemble members and the data points is around 2.6˚C averaged across the ensemble members, or 2.2˚C relative to the ensemble mean, which precludes a much higher error value since the data should not be closer to the models than they are to reality under the assumption that model errors and data errors are independent. Conversely, our posterior mean estimate after fitting to the data only achieves a residual RMS difference of 1.7˚C.*"

First, rationalizing the choice of proxy uncertainty by comparing proxy-inferred offsets to the same models you wish to assimilate appears to be a circular argument, since the true temperature state is unknown. Second, the RMS difference of the Mg/Ca-inferred temperatures to the posterior is also a meaningless comparison, given that any updated RMS difference will simply be an uncertainty-weighted reflection of the same Mg/Ca data that went into the assimilation! Indeed, had you increased the uncertainty to infinity, the RMS difference would remain 2.2˚C; for increasingly small uncertainty, the RMS difference would converge toward 0˚C. Please see Point 1ii., above.

g) Pg. 8, L8, L1-4 "*We adopt the same parameters for the algorithm that were shown to work well in AHM22, of 4 EOFs, and a localisation length scale of 2500km. While changing these values altered the regional patterns somewhat (for example, using a larger number of EOFs introduced more spatial variability) they made little difference to the large scale results such as global mean temperature anomalies.*"

The authors should either illustrate this visually or via statistical validation testing.

h) *Table 1*. More information on each model would be appreciated. At minimum, the degree of mPWP warming / cooling (SAT's and SST's) relative to their PI-reference would be useful to reference for each model.

Citations:

Annan, J. D. and Hargreaves, J. C. A new global reconstruction of temperature changes at the Last Glacial Maximum, *Clim. Past*, 9, 367–376, 2013. https://doi.org/10.5194/cp-9-367-2013

Annan, J. D., Hargreaves, J. C., and Mauritsen, T. A new global surface temperature reconstruction for the Last Glacial Maximum, *Clim. Past*, 18, 1883–1896 (2022). https://doi.org/10.5194/cp-18-1883-2022

Erb, M. P., McKay, N. P., Steiger, N., Dee, S., Hancock, C., Ivanovic, R. F., Gregoire, L. J., and Valdes, P.: Reconstructing Holocene temperatures in time and space using paleoclimate data assimilation, *Clim. Past*, 18, 2599–2629 (2022). https://doi.org/10.5194/cp-18-2599-2022

Gray, W. R., & Evans, D. Nonthermal influences on Mg/Ca in planktonic foraminifera: A review of culture studies and application to the last glacial maximum. *Paleoceanography and Paleoclimatology*, 34, 306–315 (2019). https://doi.org/10.1029/2018PA003517

Osman, M.B., Tierney, J.E., Zhu, J. et al. Globally resolved surface temperatures since the Last Glacial Maximum. *Nature* 599, 239–244 (2021). https://doi.org/10.1038/s41586-021-03984-4

Rosenthal, Y., Bova, S., & Zhou, X. A user guide for choosing planktic foraminiferal Mg/Ca-temperature calibrations. *Paleoceanography and Paleoclimatology*, 37, e2022PA004413 (2022). https://doi.org/10.1029/2022PA004413

Tardif, R., Hakim, G. J., Perkins, W. A., Horlick, K. A., Erb, M. P., Emile-Geay, J., Anderson, D. M., Steig, E. J., and Noone, D. Last Millennium Reanalysis with an expanded proxy database and seasonal proxy modeling, *Clim. Past*, 15, 1251–1273 (2019). https://doi.org/10.5194/cp-15-1251-2019

Tierney, J.E., Zhu, J., King, J. et al. Glacial cooling and climate sensitivity revisited. *Nature* 584, 569–573 (2020). https://doi.org/10.1038/s41586-020-2617-x

Tierney, J. E., Malevich, S. B., Gray, W., Vetter, L., & Thirumalai, K. Bayesian calibration of the Mg/Ca paleothermometer in planktic foraminifera. *Paleoceanography and Paleoclimatology*, 34, 2005–2030 (2019). https://doi.org/10.1029/2019PA003744

Tierney, J. E., & Tingley, M. P. BAYSPLINE: A new calibration for the alkenone paleothermometer. *Paleoceanography and Paleoclimatology*, 33, 281–301 (2018). https://doi.org/10.1002/2017PA003201