This manuscript is likely to be of great interest to the paleoceanography community because it makes significant progress in finding a self-consistent decomposition of global mean benthic d18O into temperature and seawater (ice volume) components in a way which is consistent with independent estimates of global mean sea surface temperature (GMSST) and sea level constraints. Overall, it is well written and well supported by evidence. However, the manuscript could be significantly improved with some additional clarification.

Thank you for this positive assessment and very helpful comments which have improved the manuscript.

We first point out that, at the suggestion of Reviewer 2, we downloaded the 15 model results from the PLIOMIP2 experiment (e.g., Haywood et al., 2020, *Climate of the Past*) to obtain ΔGMSST and ΔMOT for each model run. We added these to the model results shown in Fig. 2A and reassessed the HSE using several statistical models. As explained in the revised text, two statistical models (LOESS and segmented regression with two breakpoints) provide equivalent fits to the data that are superior to the linear regression used in the original Fig. 2A. These model results now suggest that HSE is 1 for ΔGMSST <0°C, 0.6 for ΔGMSST 0°C to 5°C, and 1.2 for ΔGMSST >5°C. These results are thus consistent with what we derived from the proxy data for the last 4.5 Ma (i.e.,), providing two independent lines of evidence for an increase in HSE during the MPT from ~0.5 to 1. These new results have been incorporated into the revision.

Major points:

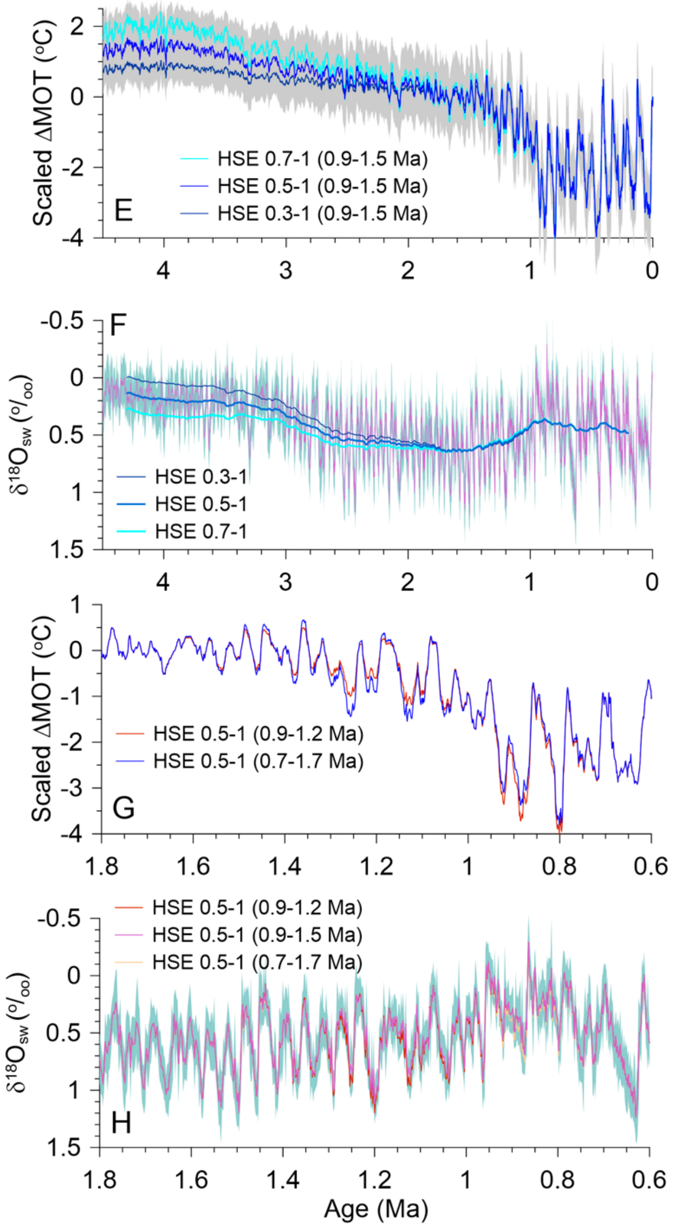
1. The calculations of mean ocean temperature (MOT) change relies on a transition in the ocean heat storage efficiency (HSE) from ~0.5 before the MPT to ~1 after the MPT. While the need for such a transition is well justified by comparison with BWT measurements, the available proxy data before the MPT (particularly in the Pacific) are quite sparse with large scatter and uncertainties. Although the authors appropriately provide a large uncertainty estimate for HSE, they provide calculations for the decomposition of d18Osw using only one scenario, in which HSE changes linearly between 1.5-0.9 Ma. It would be enormously helpful for the interpretation of the d18Osw estimate if the authors also provided the d18Osw results of a few sensitivity tests in which the timing and amplitude of HSE change are varied within the range consistent with BWT estimates.

This is an excellent suggestion. Some of this can be inferred from existing information in the paper. For example, Figure 6B shows ΔMOT reconstructions based on HSE = 0.7 and HSE = 0.3, which closely encompass the 1σ uncertainty on our ΔMOT reconstruction based on HSE = 0.5, suggesting that $\delta^{18}O_{sw}$ based on HSE = 0.3 and 0.7 will similarly fall within the uncertainty of our $\delta^{18}O_{sw}$ based on HSE = 0.5. Figures 9A and 9B similarly indicate the sensitivity of $\delta^{18}O_{sw}$ based on HSE = 1 and 0.5.

To address the reviewer's comment more thoroughly, we have added four panels to Figure 9 (see below). Panels E and F address the amplitude question. Panel E (reproducing Figure 6B) shows that ΔMOT reconstructions based on HSE = 0.7 and HSE = 0.3 fall within the 1σ uncertainty on our ΔMOT reconstruction based on HSE = 0.5. Panel F shows our high-resolution $\delta^{18}O_{sw}$ (violet) with 1σ uncertainty compared to long-term (401-kyr running average) $\delta^{18}O_{sw}$ for the three HSE scenarios. The differences during the early Pleistocene are small (<0.1 per mil), and the high-resolution $\delta^{18}O_{sw}$ for the two bracketing HSE scenarios fall within the 1σ uncertainty of high-resolution $\delta^{18}O_{sw}$ based on 0.5-1 HSE scenario.

Panels G and H address the timing question by comparing our preferred scenario (HSE increased from 0.5 to 1 between 1.5 Ma and 0.9 Ma) to one scenario where increase occurred more rapidly (1.2-0.9 Ma) and another where it increased more gradually (1.7-0.7 Ma). Panels G and H show that the differences in ΔMOT and $\delta^{18}O_{sw}$, respectively, are negligible.

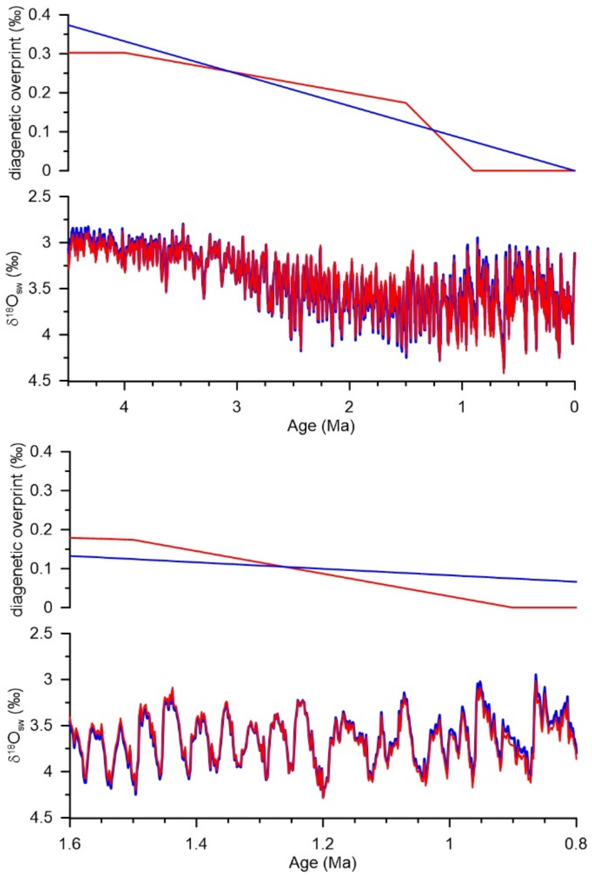
We thus conclude that our main findings regarding $\delta^{18}O_{sw}$ are robust to the range of ΔMOT suggested by the models and data.



2. Similarly, the timing of the hypothesized diagenetic alteration of benthic d18O is not well constrained by proxy data. Although Raymo et al (2018) proposed a simple linear trend for this effect, one might alternatively hypothesize that the effect would covary with MOT or BWT change if the mechanism responsible for the effect is the cooling of BWT. Because the manuscript estimates that MOT cools most dramatically during the MPT, it would be informative to also show the results of a sensitivity test in which the rate of diagenesis is greater for d18O immediately preceding the MPT (keeping the same estimated total diagenetic contribution at 3 Ma).

We do show sensitivity to different linear trends, ranging from 0.05 to 0.12 ‰ Myr^{-1} (Fig. 9C). These yield small differences for the past 1 to 2 Myr.

We have now also followed the reviewer's suggestion and applied a diagenetic correction of similar overall magnitude but that tracks the MOT reconstruction through time – shown in the figure below. The faster rate of change across the MPT slightly increases the trend in d18Osw over this interval (more depleted before 1 Ma, more enriched after 1 Ma), but the changes are small, ≤ 0.1‰.



3. These two sensitivity tests would be particularly helpful for interpreting the unexpected observation that smoothed d18Osw and glacial maxima d18Osw at ~1.5 Ma are similar to (or possibly more enriched than) the d18Osw of the Late Pleistocene. It's important to clarify whether this finding is relatively robust to the specified timing and amplitude of HSE change and d18O diagenesis, neither of which is well constrained by the available proxy data.

We think our responses show that our findings are robust. In any event, we want to emphasize that our sensitivity results in Figure 9A-9C show that one obtains unrealistic $\delta^{18}O_{sw}$ values in the Pliocene and early Pleistocene without accounting for a change in HSE suggested by data and models and a change in some long-term control such as diagenesis or the carbonate ion effect.

4. An additional surprising result is the relative amplitudes of orbital-scale MOT variability and orbital-scale d18Osw variability in the pre-MPT time period. The pre-MPT MOT record contains very weak glacial-interglacial change compared to relatively large amplitude d18Osw changes from 2.6-1.5 Ma. The authors should add some discussion of the reliability of the amplitudes of the orbital-scale signal in GMSST change and MOT change. Are the resolution and age uncertainty of the SST records sufficient to accurately estimate orbital-scale changes in GMSST and, thus, its application to estimating orbital responses in MOT and d18Osw?

We addressed these issues at length in the Supplementary Material of Clark et al. (2024, Science) (see p. 5-9, Figs. S4-S6), and reached the following conclusions.
(1) Our assessment of resolution on the variability of our SST stack suggests minimal loss of the 100- and 41-kyr signals and that that despite some loss of the 23-kyr signal, it remains readily detectable.
(2) Our assessment of age model uncertainties suggests there is minimal preferential signal loss due to age misalignments.
(3) We used three different statistical models to assess whether the trends in the standard deviations of our global stack and individual-record averages differ from one another, with our results suggesting that the trends parallel one another and would be interpreted similarly in terms of the evolution of variability over the past 4 Ma.
(4) Finally, we compared our stack to several other composite reconstructions and found that all reconstructions show a gradual increase in variability over the Pleistocene similar to the stack.

We address the reviewer's comment by adding the following to our revised text:
We note that Clark et al. (2024) found little loss of variability in the ΔGMSST reconstruction due to age uncertainties and resolutions of individual SST records used in the reconstruction.

5. In Figure 13, the authors present a very interesting comparison of BWT and d18Osw estimates from two Pacific cores and their global compilation estimates. They make the compelling argument that the estimates from the two cores are unlikely to provide reliable global estimates because they imply that sea level would need to be ~50 m higher than PI for significant amounts of time between 1.4-1 Ma, suggesting that these sites may be affected by local salinity changes. Could the authors slightly expand upon this idea to explain how the locations of those Pacific cores could have significantly different bottom properties than the rest of the deep Pacific?

Unfortunately, having just the two widely spaced records cannot address this question beyond our statement that they "reflect regional hydrographic changes (i.e., salinity) that were perhaps associated with the large changes in ocean circulation during the MPT." We can speculate that there might have been a different δ^{18}O-salinity relationship, which is expected if δ^{18}O of Antarctic ice was not as negative or there may be a problem with the Mg/Ca data. If these records represent the whole Pacific, we need to have saltier water elsewhere to keep the salt and O isotope budget of the ocean. In any event, at this point it is hard to tell, which is why we are using our approach.

6. I really appreciated the section of the paper using model results to explore the mechanisms responsible for scaling between MOT and GMSST and why it might differ before the MPT. However, one question I have is about the authors' apparent conclusion that AABW's contribution to MOT was constant (and approximately equal to pre-industrial) from 4-1.5 Ma (Figure 16D). How can this be consistent with the PlioMIP2 findings that the deep Southern Ocean was 1.5-2.5 C warmer than pre-industrial

As noted above, we have now downloaded the PLIOMIP2 data which allows us to use their constraints on the volume of the two ocean heat reservoirs and the relationship of the temperature of the upper reservoir to GMSST. Using these improved constraints, we now find a ~1ºC decrease of the temperature of our deeper reservoir ($\Delta T_{>2000}$) from 4-1.5 Ma which, with the ±2ºC uncertainty of our approach, can readily accommodate the warmer Pliocene deep Southern Ocean. More importantly, however, we emphasize that our $\Delta T_{>2000}$ is for the entire ocean >2000 m, not just the Southern Ocean. Our new analysis of PLIOMIP2 ΔMOT (now included in Figure 14) shows that much of the deep ocean warming is less than the deep Southern Ocean, as is expected as AABW moves northward.

and that increased stratification caused decreased AABW formation?

We refer to the PlioMIP2 findings in Weiffenbach et al. (2024) that increased surface stratification (because of warmer SSTs and less sea ice) caused a decrease in AABW formation (which leads to warming):
We thus conclude that it was the persistence of a highly stratified Southern Ocean that caused a smaller AABW formation rate and persistently warmer Td than present until ~1.5 Ma, when the gradual decay of stratification and increase in sea-ice extent and variability then enhanced conditions for AABW formation.

Minor points:
Line 546: The statement that 1123 records large ice sheets pre-MPT is unclear because most of the pre-MPT d18Osw record is significantly lighter than the post-MPT glacial values. I think the authors might be referring to one particularly large glacial maximum at ~1.5 Ma. Please clarify exactly what is referred to here and how it provides support for the new d18Osw record.

We have revised as:
This implies that site 1123 is recording large ice sheets at ~1.5 Ma, or before the MPT.

Lines 605-606: The same text is repeated on these two lines.

Thank you – now corrected.

Lines 647-648: The meaning of this sentence isn't clear. Ice sheets have enhanced the warming relative to what? How is this visible in Figure 14C?

Thank you. Clarified as:
The effect of lowering of Northern Hemisphere ice sheets as they retreat induces surface warming…

Figure 1: Many of the individual records are partially/mostly hidden behind other data in this figure. Also, the caption suggests that there are two different orange lines in the figure, which seems like a problem.

Now only one orange line. We don't think it's necessary to completely see every record – the main point is that there is a large spread in the reconstructions.

Figure 10B: It's very hard to see the light blue line (which is an important result to be able to see) due to overlap with the gray line. Maybe make the shade of blue darker or leave off the gray line.

We have darkened the blue line.

Figure 16F: The caption doesn't provide the color information for all the different records shown.

Now added.