Overall I found this to be an interesting, well-written and illuminating paper that I think will help spur improvements in model development. The separation of the TA biases into preformed TA, remineralization TA and CaCO3 TA is also very useful and points to concretely implementable improvements, especially in the treatment of CaCO3.

We thank the reviewer for their thoughtful und thorough review and helpful suggestions. We addressed the reviewer comments below in blue.

I recommend publication with some minor revisions (see below).

My two major comments are:

Line 254 & Figure 7

As the authors point out the biases in Revelle Factor are of great importance to mCDR. An additional metric of this that would be straightforward to add using CO2SYS and of great value to folks investigating the feasibility and cost of ocean alkalinity enhancement is the uptake efficiency factor (In our work we like to call this $\eta CO2 = \partial DIC/\partial Alk$ at constant pCO2, see https://doi.org/10.1039/D1EE01532J and https://bg.copernicus.org/articles/20/27/2023/). The metric simply indicates the number of moles of CO2 taken up per mol of Alkalinity added after full equilibration (for an infinitesimal increase in Alk) and is generally ~0.8 though it is quite dependent on location (see for example He et al., 2023, https://bg.copernicus.org/articles/20/27/2023/).

I think this number is very practical because it directly represents an efficiency loss going from some alkaline substance to actual CO2 drawdown and thus enters any cost estimates. Thus I would be very curious to see how model biases affect this metric, even if just expressed as a global average or surface average.

We calculated the instantaneous uptake efficiency μCO_2 also with CO2SYS and added the results to Figure 7:

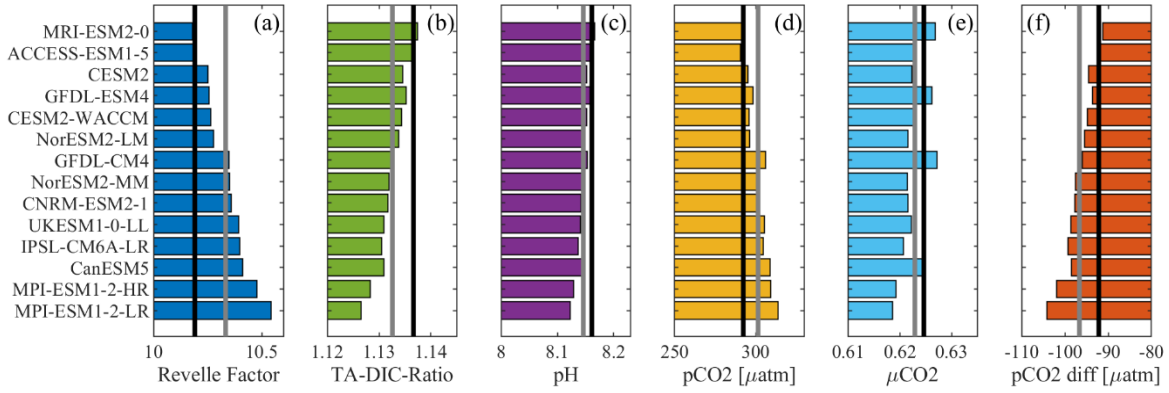
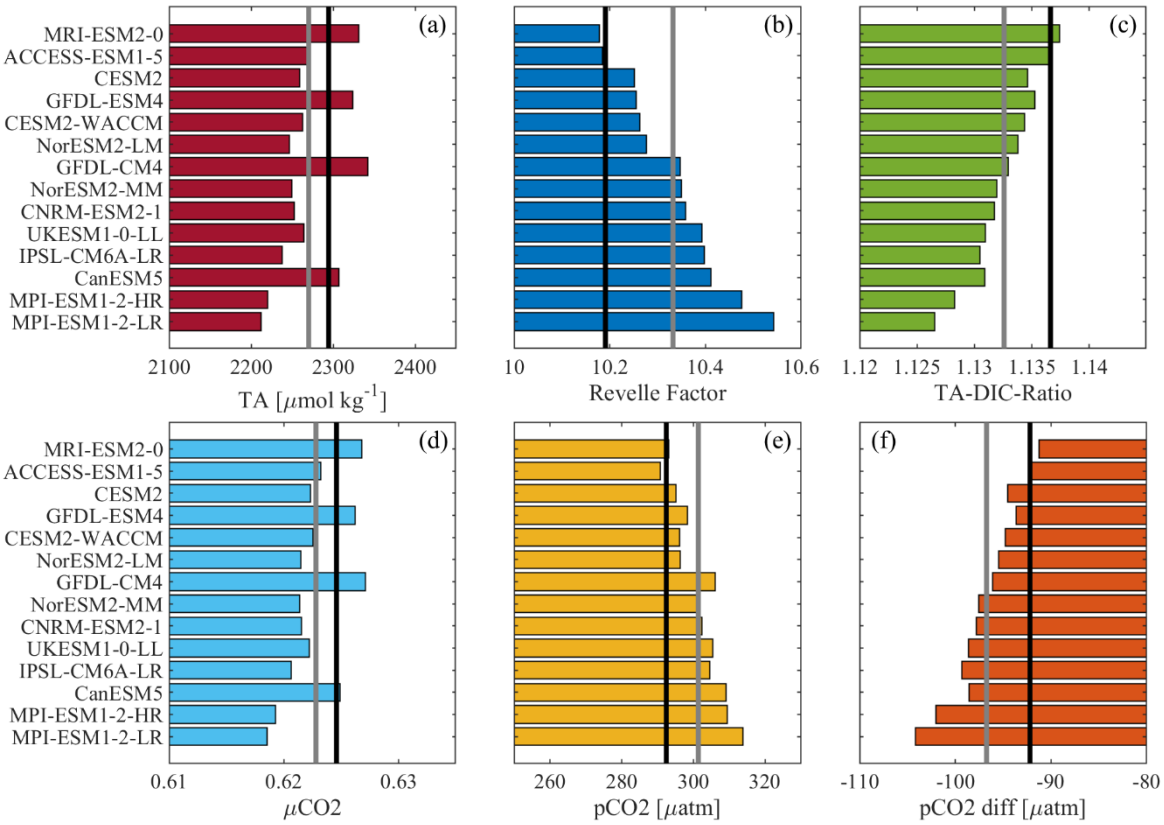


Figure 7 NEW: Carbonate system parameters were computed for all models, the MMM (grey line) and the GLODAP data (black line) with the CO2SYS toolbox, based on the two input parameters global mean alkalinity and DIC. The results are sorted by Revelle Factor in ascending order for all panels. Shown are the Revelle factor (a), the TA-DIC ratio (b) pH (c), pCO_2 (d), μCO_2 uptake efficiency (e), and difference in pCO_2 after a 100 μmol kg^{-1} addition of TA (f).

As one can see the uptake efficiency in this case hovers around ~0.62 for the models and GLODAP. Figure 5 in https://bg.copernicus.org/articles/20/27/2023/ suggests that it would

take some time to get the expected values of ~ 0.8. Still, the models exhibit some differences in this instantenous uptake efficiency that reflects their initial state in TA, DIC and pCO2.

In response to Reviewer 2, we added TA in the same order to Figure 7 (and removed pH for now), this shows more clearly how the initial state of TA directly influences the uptake efficiency after alkalinity addition, while the effect on the pCO_2 difference is further modified by the Revelle factor and the PCO_2 initial state :



L317ff and Figure 6, panel (d). There is clearly a large amount of difference in TA* between models and also in some models these biases are clearly depth-dependent while in others they are less so. This is one of the major insights of this paper. Not being familiar with the details of each of the models tested, I am very curious about whether there is any pattern or correlation between the sophistication of the CaCO3-cycle-model in each GCM and the amount and type of bias observed ?

E.g. The blue-ish models mostly overestimate TA at depth - do they have something in common in the way they treat the CaCO3 dissolution?

The two models that show the highest bias in the CaCO3 cycle in Figure 6d, CNRM-ESM2-1 and IPSL-CM6A-LR, have in common that their ocean model is NEMO and the biogeochemical model is PISCESv2. Dissolution in PISCESv2 is treated explicitly and is dependent on omega and the sinking speed for PIC is depth-dependent, while for other models it is constant. More details on the model equations for all CMIP6 models can be found in Planchat et al. (2023) (https://bg.copernicus.org/articles/20/1195/2023/bg-20-1195-2023.pdf).

Do any of these models treat the natural occurrence and distribution of CaCO3 sediments

explicitly (see work by Sulpis et al and others for maps of this) ? or do they only account for precipitation and redissolution ? If not, then perhaps there is a spatial correlation between TA biases and occurence of CaCO3 sediments? I'd love to see more discussion of this phenomenon - it's very interesting! The discussion on L294-330 is in very general terms rather than looking at algorithm differences between the specific models that could explain the differences.

In two of the models (of Figure 6d), MRI-ESM2-0 and UKESM1-0-LL, CaCO3 is dissolved without a sediment, while the other models do have explicit sediment treatments where CaCO3 is buried or dissolved, either depend on omega or a set rate (Planchat et al. 2023). A direct link to the bias at depths is not obvious in this case. Since our study is more focused on biases at the surface and the subsequent implications for ocean alkalinity enhancement, we did not go into more detail here.

Figure Style comments:

As I was parsing the figures I felt some improvements in the plots could make better use of the space, aid visual parsing and generally make the paper even easier to follow. Please take these as suggestions, perhaps try them out and see if you like them.

Figure 1: Maybe expressing the MMM as a (say, dashed) line rather than an additional row would be more intuitive and allow visual comparison of each model vs the MMM ?

Figure 1: does the thickness of the GLODAP line have meaning (e.g. a standard error) or is it incidental ? If a standard error for the GLODAP measurement is known or can be computed it would be neat to use the thickness (using a semi-tranparent color) this way (unless the certainty is so high that it reduces to a thin line of course). I think this is important as a large GLODAP uncertainty could change or weaken the conclusions.

Figure 1, Line 264 As you note Alk and DIC are highly correlated, and they are compensating variables in the carbonate system, with respect to pH, pCO2 etc.

An alternative for the two panels in Figure 1 would be to plot both together as a scatter graph with DIC on the x-axis and Alk on the y axis (or vice versa). This way the exact same information is displayed but the extent of the correlation is visually immediately apparent as well. The scatter points could be labelled directly on the graph with a floating text for example. Error bars on each pint could indicate the variance of these values over the surface average.

Thank you for the suggestion. In response, we replaced Figure 1 with a scatter plot of TA versus DIC, as we agree that this might be more intuitive. This figure also contains an error estimate for the global mean GLODAP data. While the MMM is almost within the estimated GLODAP range, we see that the difference for individual models can be quite large. We also

see that TA and DIC biases are highly correlated.

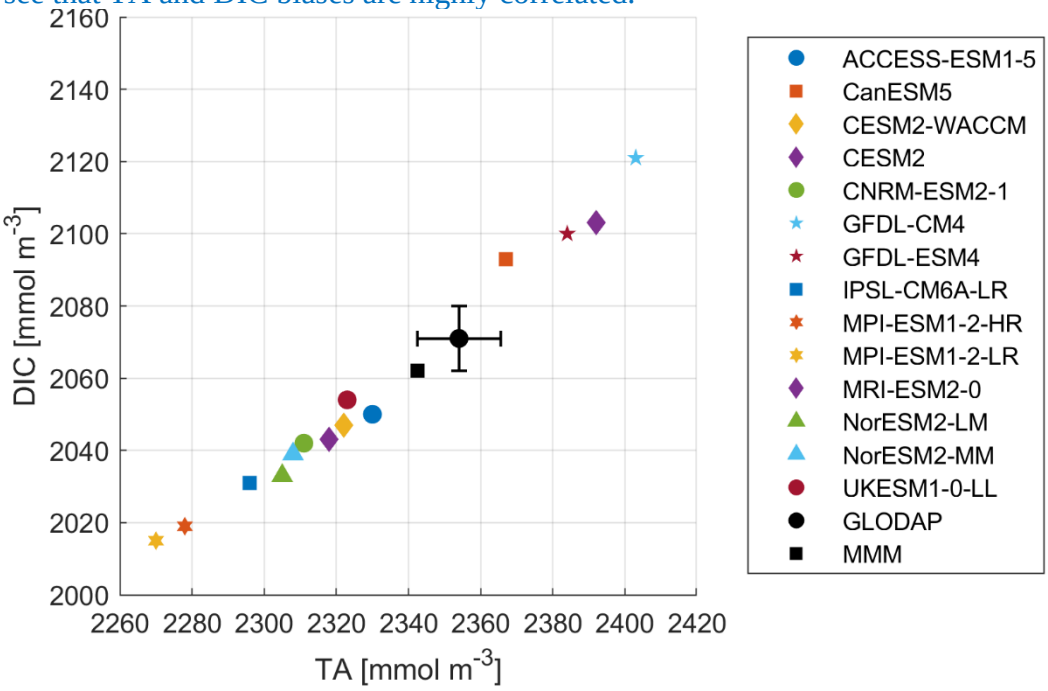


Figure 1 NEW: Global mean surface total alkalinity (TA) [mmol m^{-3}] of the 14 CMIP6 models, the multi-model-mean (MMM), and GLODAP including its error estimate versus dissolved inorganic carbon (DIC) [mmol m^{-3}].

Figure 2&5: "Absolut error" → "Absolute error"

This has been corrected.

Figure 2&5: If the whitespace between globes could be reduced at all, that would make everything bigger and easier to parse (It's already tricky without looking at the PDF on a large screen).

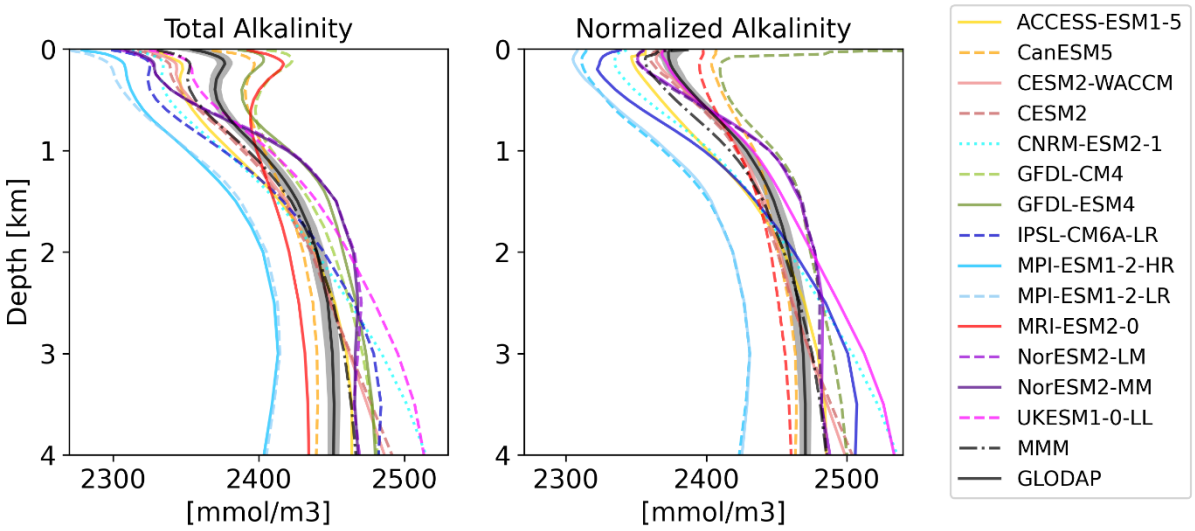
Space between the subplots has been reduced as much as possible.

Figure 3: Ah, I see here there is a GLODAP error estimate. Great! Could this be added to Figure 1 also ?

Yes, this has been added to Figure 1. Thank you for the suggestion.

Figure 3&4: Style considerations: For a colour-blind person (like myself) it is nigh impossible to know which line is which, among similar shades/hues. I would recommend blending color with different dash/dot patterns to help with this. Perhaps the error (currently dashed lines) can be given simply by a transparent shaded area ?

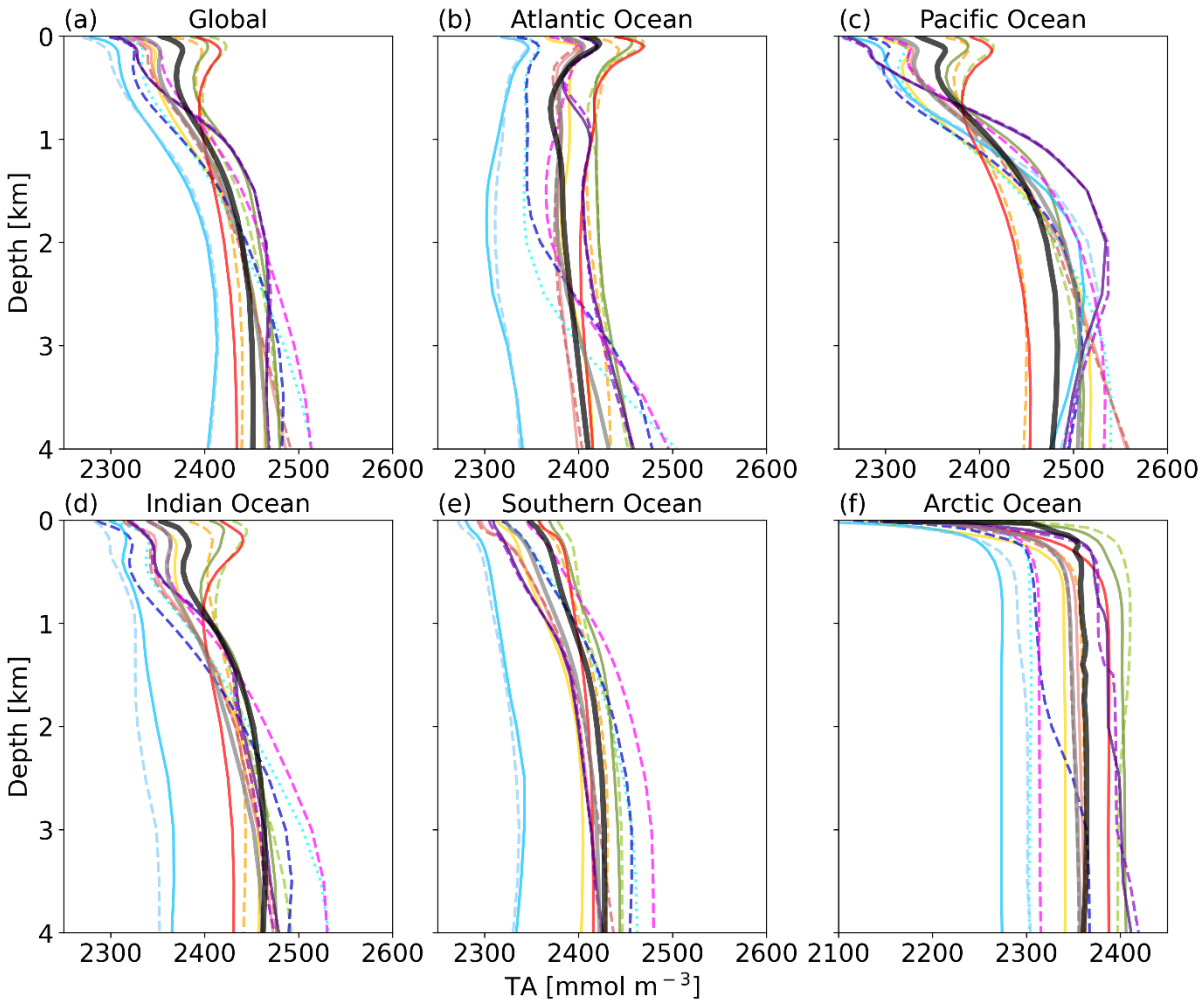
We added linestyles to the profiles and indicate the GLODAP error estimate with shading.



Also, a lot of features of this graph occur in the upper 0.5km and are visually cramped in a very small area. For the same reason that most models have non-uniform vertical z slices, perhaps it would be possible to plot the vertical axis on a log scale or a mixed log-linear scale. Or split the graph into two linear regions, one for 0-500m and one 500-4000m ?

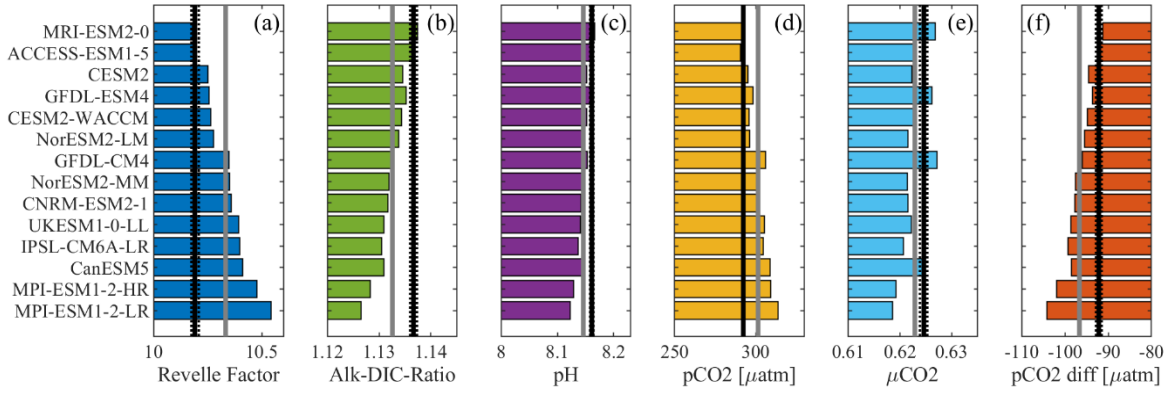
We would like to keep this figure as is since we were interested in how well the profiles overall match.

Figure 4(f) I'd adjust the x axis to not clip the values at shallow depth. Ah - I see all the X axes are coordinated. Hmmm. Not sure how to solve this. How low do the TA values go in the Arctic Ocean (last panel)? Perhaps one could plot all these as Deltas from GLODAP the same way that Figure 6 ? That might help with the dynamic range of the x axis (which IMHO does not necessarily have to be the same for each subpanel). I thought Figure 6 was very nice.



We adjusted the y-axis for the Arctic in Figure 4f and applied the linestyles here as in Figure 3 as well as added a line for MMM.

Figure 7: Again, using the thickness of the GLODAP vertical line to indicate variance would be neat.



The estimated error for the GLODAP values based on the TA and DIC GLODAP errors are very small for the computed parameters. Here, we added dashed vertical lines next to the black GLODAP line. We think the figure would be more clear without the dashed lines though.

Figure 8: I found this figure rather difficult to parse. Because the three different variables have such different dynamic range on the %-scale, especially the second one (TA-DIC ratio) is virtually impossible to read off. Is this figure really necessary ? I feel like most of the information content is already contained in Figure 7.

We will drop Figure 8 and instead refer to the percentage values listed in Supplement Table S2.

L87: Some other recent studies that would be worth including that also aim to be more realistic than the earlier large-scale uniform OAE simulations, i.e. near-coastal or ship-track-based releases or regional assessments.

https://doi.org/10.1002/2017EF000659
Model-Based Assessment of the CO2 Sequestration Potential of Coastal Ocean Alkalinization, Feng, Koeve, Keller, Orschlies 2017

https://doi.org/10.1029/2022EF002816
Simulated Impact of Ocean Alkalinity Enhancement on Atmospheric CO2 Removal in the Bering Sea, Weng et al., 2022

https://doi.org/10.5194/bg-20-27-2023
Limits and CO2 equilibration of near-coast alkalinity enhancement, He and Tyka, 2023

https://doi.org/10.5194/egusphere-egu23-9305
Atmospheric CO2 removal by alkalinity enhancement in the North Sea, Liu et al. 2023

L87 has been appended to include most of the above suggested references and now reads:

Now, more and more projects are underway or in planning that seek to apply more realistic scenarios for OAE e.g., in regional OAE applications (Butenschön et al. (2021), Wang et al. (2023) or coastal applications (Feng et al. (2017), He and Tyka (2023)), which is why a model evaluation is even more important.

L149: Since the carbonate system isn’t linear wrt TA and DIC, does it make sense to first area-average the TA and DIC values and *then* put them through the CO2SYS calculation ? It seems to me it would be more accurate to compute Revelle, pH, pCO2 etc for each surface location and or time and *then* do the area-weighted average of each metric. Perhaps over the range of values encountered the system is linear enough and this doesn't make much of a difference, but I’m not sure.

This study was meant to introduce the issue of alkalinity and DIC biases in ESMs, their implications for assessing model OAE experiments and to suggest some potential areas for model improvements. The suggested approach would certainly be a worthwhile follow-up study.