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

Review of Hinrichs et al. (2023)
Overall comment:
In overall, this work is interesting and innovative in its approach to bias assessment in OAE modelling studies using ESMs. Indeed, the study pushes a CMIP6 bias analysis (essentially focused on alkalinity, but also on DIC) towards a concrete, but idealized, CDR experiment. The approach used to decompose the alkalinity bias is broadly consistent with Planchat et al. (2023), thus reinforcing the message to the modelling community around the representation of alkalinity and the carbonate pump in ESMS. The writing of the manuscript could nevertheless be partly improved, in particular some parts could be restructured to gain clarity. Many technical corrections should be made.

Specific comments
A. Main comments
1. In the decomposition of the alkalinity bias, it is repeatedly mentioned that "preformed alkalinity"
refers to physical biases for the ocean, but it also contains biases of both organic matter and
CaCO3 production (e.g., l. 17 ; l. 121-125 ; l. 218 ; l. 291 ; l. 294)

We amended L 17 with:

The decomposition of the global mean alkalinity biases into contributions from physical processes (preformed alkalinity) which includes i) the physical redistribution of biased alkalinity originating from the soft tissue und carbonates pumps, ii) remineralization, and iii) carbonate formation and dissolution showed that the bias stemming from the physical redistribution of alkalinity is dominant.

We amended the method part L121-125 on preformed alkalinity with:

Preformed Alkalinity also refers to the physical redistribution of alkalinity biases stemming originally from soft tissue and carbonate pumps and the upwelling of water masses with biased alkalinity.

L221 was amended with the sentence:
The first part, preformed alkalinity, includes the advection and upwelling of already biased water masses.

2. The direct link made with the CO2 uptake in the OAE idealized experiment is confusing for me
(e.g., 274, l. 327): a relative pCO2 difference should not directly be a relative CO2 uptake difference, or should it?
According to me, we have:
$dDIC$
$DIC = 1$
$Re \cdot dpCO$!

pCO !

An ESM with a greater relative variation of pCO2 resulting from alkalinity addition, also has a greater Revelle factor. So, the product of the right member of the above equation should at least partly compensate. To access a first approximation of the CO2 uptake associated with alkalinity addition, the relative variation of DIC should be expressed in function of the relative variation of alkalinity. I think that with what is currently offered, it is possible to assess the bias resulting from OAE in ESMs concerning acidification, but I'm confused by the direct link made with the CO 2 uptake.

Thank you for bringing up this point. We have now added the parameter uptake efficiency (μCO_2) to our back of the envelope calculation. The uptake efficiency depends on the ratio of pCO_2 change over pCO2 state with respect to total DIC over alkalinity change (or addition) and the Revelle factor:

$\mu CO_2 = \partial DIC/\partial Alk$ at constant pCO_2,

rearranged to

$\mu CO_2 = -1/Revelle * \partial pCO2/pCO2 * DIC / \partial Alk$.

The inherent model states of pCO2, TA and DIC lead to slightly different efficiencies for a constant alkalinity addition. On your suggestion, we added the TA plot into Figure 7 as well. What becomes more obvious this way, is that the initial TA strongly influences the total uptake efficiency after alkalinity addition, while for the pCO2 change the Revelle factor and the initial pCO2 play a role as well and partly compensate.

We will include this additional result and the new based on the uptake efficiency in the revised manuscript.
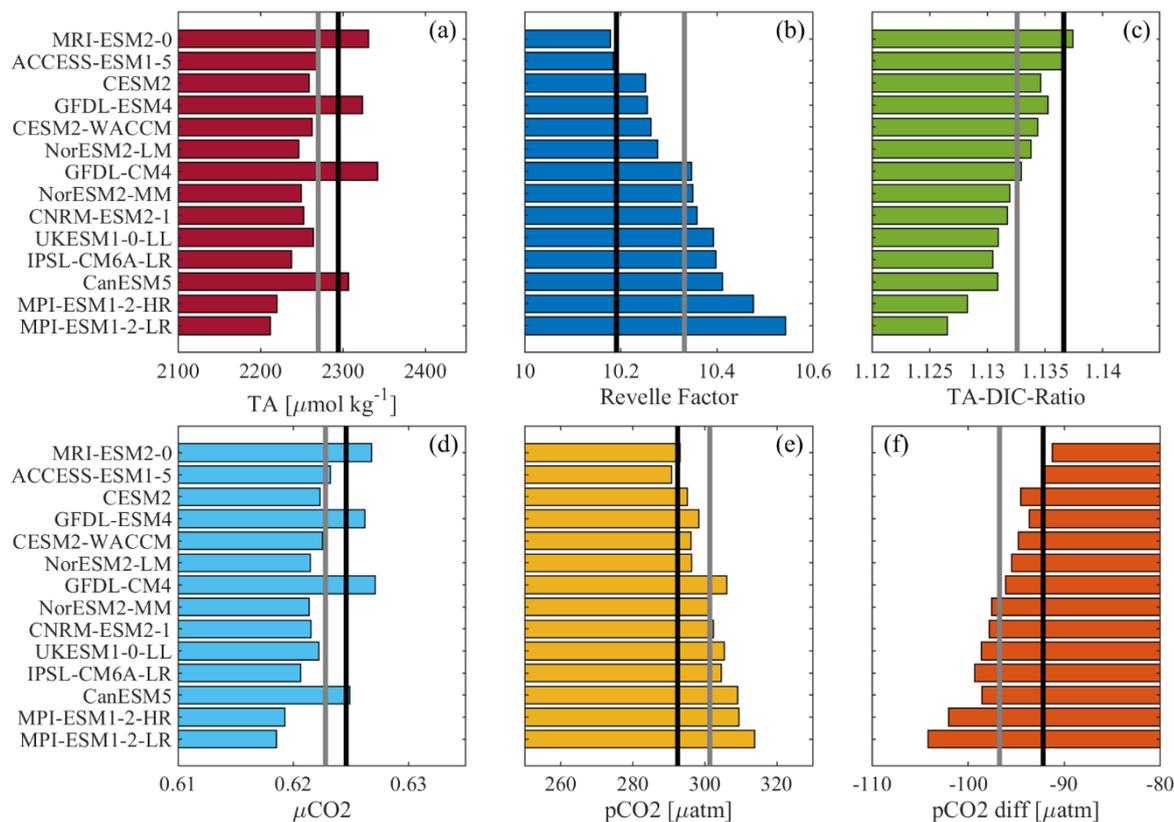
Figure 7 NEW: Carbonate system parameters were computed for all models, the MMM (grey line) and the GLODAP data (black line) with the CO2SYS toolbox, based on the two input parameters global mean alkalinity and DIC. The results are sorted by Revelle Factor in ascending order for all panels. Shown are TA [μmol kg-1] (a), the Revelle factor (b), the TA-DIC ratio (c), μCO₂ uptake efficiency (d), pCO₂ [μatm] (d), and the difference in pCO₂ after a 100 μmol kg-1 addition of TA (e)

B. Minor comments

1)
The second part of the abstract could be improved and made more readable. In particular, I do not understand: "We find that the degree of compensation of DIC and alkalinity biases at the surface is more important for the marine CO2 uptake capacity than the alkalinity biases themselves."

We adapted the paragraph, which now reads:
"The relative biases of alkalinity versus DIC at the surface affect the Revelle factor and therefore the marine CO_2 uptake capacity. The global mean surface alkalinity bias relative to GLODAPv2 in the different models ranges from -85 mmol m^{-3} (-3.6%) to +50 mmol m^{-3}(+2.1%) (mean: -25 mmol m^{-3} or -1.1%). For DIC the relative bias ranges from -55 mmol m^{-3} (-2.6%) to 53 mmol m^{-3} (+2.5%) (mean: -13 mmol m^{-3} or -0.6%). All but two of the CMIP6 models evaluated here overestimate the Revelle factor at the surface and thus overestimate the CO_2-draw-down after alkalinity addition by up to 13%. This overestimate has to be taken into account when reporting on efficiencies of ocean alkalinity enhancement experiments using CMIP6 models. „

2)
(l. 25-26); and l. 27-29 are too heavy,
Broke this up into two sentences.
perhaps misplaced, and with a unit error (mmol/kg instead of mmol/m3 as l. 209-211).

Thanks for pointing this out, it should be mmol/m3. This has been corrected.

3)
For the sake of clarity and consistency throughout the manuscript, I would suggest that alkalinity and DIC should always be shared in mmol/m3 rather than mixing mmol/m3 and umol/kg.
Changed to mmol/m3 in Figure 1 and text where applicable.
It would simply be necessary to clarify how the conversions are done for GLODAPv2 and for using CO2SYS in the methodology (see l. 28-29, l. 149-152, Fig. 1 etc.).
For the model evaluation the GLODAP TA and DIC data is converted from units of μmol kg^{-1} to mmol m^{-3} using the potential density computed from GLODAP salinity and temperature data. We will add this to the methods section.

4)
Table 1: It would be great also to share the ensemble member for each ESM, rather than writing "the
first available ensemble member" l. 109, which is slightly confusing.
The ensemble member has been added to Table 1.

5)
The marine biogeochemical model of CNRM-ESM2-1 (not CNRM-ESM-2-1) is PISCESv2-gas. The one
for IPSL-CM6A-LR is PISCESv2. The one for NorESM2-LM/MM is iHAMOCC.
This has been corrected.

Why do you precise the grid only for a few ocean models? If you wish to precise it, add it in the "ocean horizontal resolution" column and just precise maybe whether it is a tripolar grid or not.
Added information about tripolar grids to grid column.
6)
l. 109-119: These lines would benefit from restructuring.
The TA* method: are there limitations to this decomposition of the biases?
Koeve et al. (2014) discuss the limitations as follows: although the computation of TA* according to equations (2) to (4) reproduces tracer-based simulated TA* robustly in most of the global ocean, higher uncertainties occur in the Atlantic and in the 500m to a 1,000m layer in the Pacific and Indian ocean.
In our study we focus on the TA* results for the global mean ocean.

Another caveat that was mentioned by Koeve et al. is related to the issue that AOU overestimates true oxygen utilization by 20–25 %. Hence TAr computed from AOU is probably also overestimated by this percentage. TAr is small and, in our study, we focus on the implication of TA* biases in ESMs and potential remedies for these biases.

What are the advantages of it rather than the one suggested in Sarmiento and Gruber (2006)? Not using DIC?
In some ways the methods are similar. Both are partly based on Feely et al. (2002). E.g. the TA_r computed in Koeve et al. (2014) is similar to δC_{soft}. Koeve then aims to further separate the influence of ocean physics from the impact of carbonate dissolution. Although, as the reviewer correctly points out the physical distribution of TA includes the distribution of TA biases from the biological and the carbonate pump

8)
l. 139, you explicitly mention the fact that TA* "is computed as residual after rearranging Eq. (2)". This is questioning, since it is finally on the dissolution of CaCO3 that we have the more certainty over its affects on alkalinity (+2 eq for 1 eq of CaCO3 dissolved). On the contrary, the effect of remineralisation, for example, is more complicated to extract from ESMs because of the N-reactions whether or not they are taken into account in the biogeochemical models. It is therefore perhaps a pity not to take advantage of the ease with which the carbonate pump can be used in terms of its effect on alkalinity with this method.

Yes, as the reviewer points out, there are uncertainties related to N-reactions, i.e., of reaction of PON to ammonium to nitrate. However, on the one hand, this is a fast reaction and on the other hand most models don't simulate ammonium explicitly. We thus think that the uncertainty that goes with neglecting the effect from N-reactions is small at present. Koeve et al. (2014) states that the "global mean ratio of the alkalinity effect stemming from organic matter remineralization to TA background

is only 0.005. Contributions from N2-fixation, denitrification, sulfate reduction, and shelf alkalinity fluxes are even smaller or of local importance only.
Koeve et al. (2014) also tested the a posteriori TA* method used here against explicitly modeled tracer of TA0, TAr and TA* with good results on a global scale.

9)
l. 156-157: Could you explain why you have chosen these values rather than the mean surface ESM values for instance?
We chose to apply the same typical values for all models. We now did a test and applied the actual MMM surface salinity (34.5), temperature (18.4) and PO (0.5) values for the calculation and found that the differences in the results are very, very small.

Model	% difference in PCO2 after alkalinity addition using S=34, T=15, PO=2	% difference in PCO2 after alkalinity addition using S=34.5, T=18.4, PO=0.5
'MRI-ESM2-0'	-1.02	-0.99
'ACCESS-ESM1-5'	0.16	0.14
'CESM2'	2.53	2.47
'GFDL-ESM4'	1.54	1.54
'CESM2-WACCM'	2.83	2.77
'NorESM2-LM'	3.55	3.47
'MMM'	4.94	4.87
'NorESM2-MM'	4.22	4.21
'GFDL-CM4'	5.83	5.73
'CNRM-ESM2-1'	6.06	5.96
'UKESM1-0-LL'	6.99	6.88
'IPSL-CM6A-LR'	7.70	7.57
'CanESM5'	6.90	6.83
'MPI-ESM1-2-HR'	10.63	10.44
'MPI-ESM1-2-LR'	12.98	12.76

10)
l. 166-168: This sentence should be part of the methodology.
Sentence was moved to methods part.

11)
l. 190 and Fig. 3b: The very high TAn value observed at the ocean surface for GFDL-ESM4 is linked to very low salinity values (even possibly null) in the Baltic Sea if I remember well. You might be able to neglect this closed sea for this ESM when averaging to avoid this issue. You could also use a salinity threshold maybe.
Thank you for the suggestion, this issue has been fixed by using a salinity threshold of 10, which we noted in the methods section.

12)
Fig. 3 and 4: I would combine both figures so that the first panel is not repeated.
We would like to keep the comparison of TA and TAn separate from the comparison of global to regional profiles.

13)

l. 201: It could be mentioned that the bias observed for CNRM, IPSL and UK is associated with the ocean model NEMO.

Thanks for the suggestion, we will check.

14)

l. 211: "TA biases likely lead DIC biases, as DIC can adjust through gas-exchange of CO_2". I suggest to add at the end "to maintain a surface chemical equilibrium with the atmospheric CO_2 concentration".

The suggested sentence has been added.

Besides, this aspect could be slightly more highlighted to understand why you decompose the alkalinity biases and not the DIC biases in addition, whereas you consider both surface alkalinity and DIC in the following with CO2SYS.

We focused on TA because of its conservative nature and because we came from the background of alkalinity enhancement studies. One result then became that DIC biases can act in a compensating way.

We will aim to highlight the importance of DIC more in the revised manuscript.

15)

l. 221-230: This paragraph should be rewritten.

It should be readable and meaningful independently of Fig. 6 (avoiding "are shown in Figure 6" instance).

The paragraph has been rewritten:

The decomposition of the TA biases (Figure 6a) shows that in the upper 1 km most of the models' alkalinity biases are due to their preformed TA (Figure 6b). Per definition, models with a negative surface TA bias have a negative bias in preformed TA. Below about 1,000 m depth preformed TA stays constant with depth. TA biases from the representation of organic matter remineralization processes are in the order of 5 to 10 mmol m^{-3} and play only a negligible role in absolute terms (Figure 6c). The bias in TA from calcium carbonate dissolution in the interior ocean (Figure 6d) can be in absolute terms comparable to or even larger than the bias in preformed TA. The MRI model and the GFDL models have a small negativebias in TA* in the order of ~10-20 mmol m^{-3}, relatively constant with depth. The MPI and NorESM models have a slightly positive TA* bias in about the same order of magnitude, also relatively constant with depth, while the UKESM, the CNRM-ESM and the IPSL ESM exhibit TA* biases that increase with depth. The CNRM model has the largest TA* bias with about 100 mmol m^{-3} at 4,000 m depth.

The first sentences are too repetitive compared to what was already described in the previous sections.

Removed the repetitive sentence.

l. 226-228: This sentence is out of topic as you are talking about biases in this section. It would be great also to share some values.

The sentence has been removed and some values are added, see new paragraph above.

16)

Fig. 6: The addition of MMM would be meaningful on these plots.

The MMM has been added to Figure 6.

(a) TA bias (b) Preformed TA bias (c) TA bias from remineralization (d) TA bias from CaCO3 cycle

Legend:
CNRM-ESM2-1
GFDL-CM4
GFDL-ESM4
IPSL-CM6A-LR
MPI-ESM1-2-HR
MPI-ESM1-2-LR
MRI-ESM2-0
NorESM2-LM
NorESM2-MM
UKESM1-0-LL
MMM
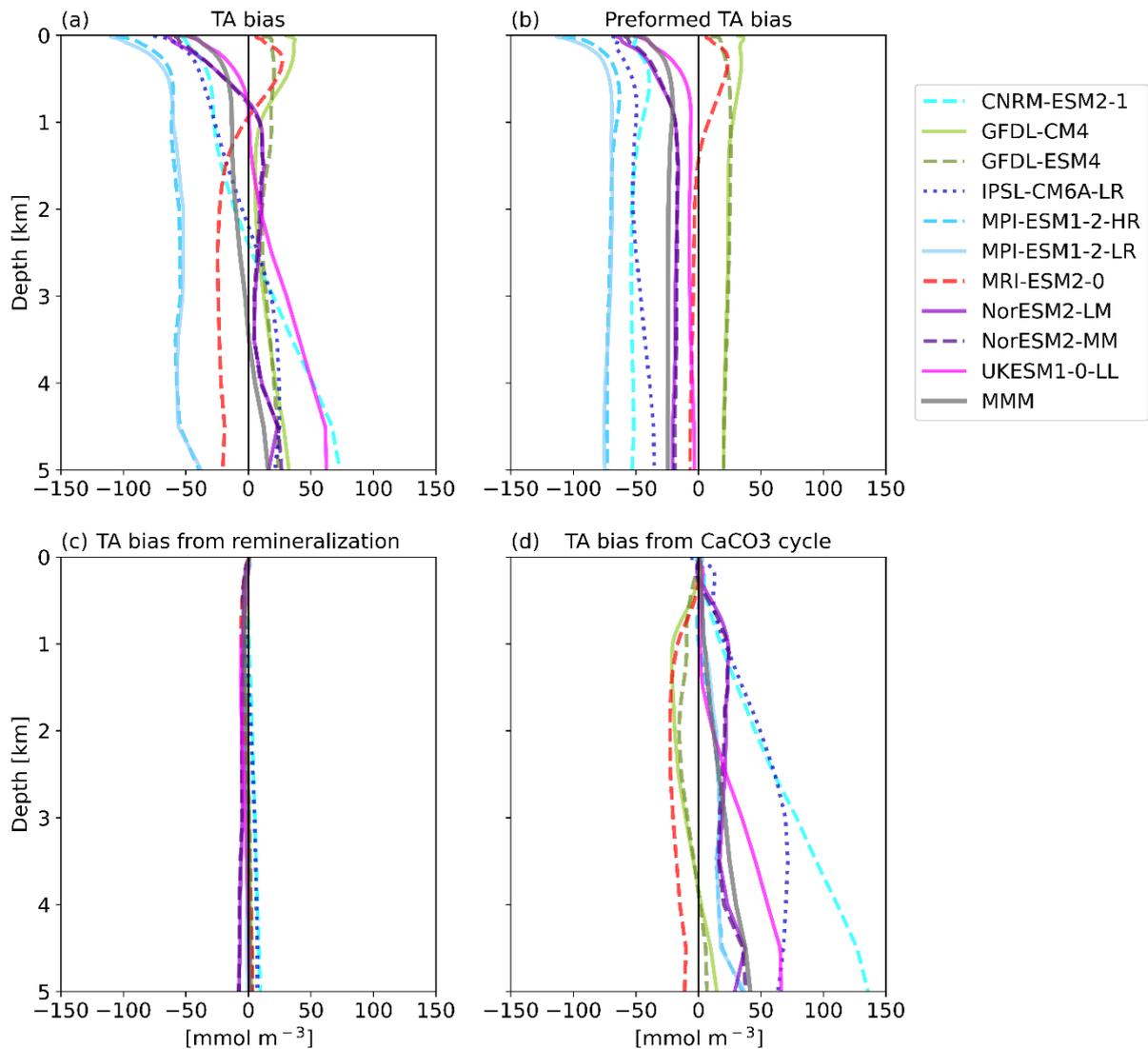
17)
l. 237: I am not convinced of the veracity of the causal sequence (cf. "thus"). Could you detail it?
Biases in simulated surface TA and surface DIC have implications for the individual models' efficiency of OAE in terms of Revelle factor and change in pCO2 and thus in the marine CO2 uptake capacity. This is now shown in the newly added μCO2 metric in Figure 7.

18)
l. 240-242: "All panels are sorted by Revelle factor in ascending order." This is enough in the legend.
Sentence was removed.

19)
"The Revelle factor ... at the ocean surface." should be part of the Introduction.
Removed here.

20)
Fig. 7: Could you set "MMM" in bold so that we can easily spot it?
Changed the figure to also have vertical MMM bars. Please see above.

21)
I also suggest to keep the x-axis increasing in the last panel, to avoid confusions.

Okay, this has been changed.
22)
l. 260-261: Too repetitive with what was already said for Fig. 1.
23)
l. 262: It is less than 13.0 % in Fig. 8, I think: t should be about 10.7 %.
Figure 8 has now been removed following the suggestion by Reviewer 1. In the initial figure though, the model with largest deviation, MPI-ESM-LR, of ~13% was missing. Thanks for pointing this out.
All values are listed in Table S2. The maximum value is 12.98 %.
24)
l. 263: It would be great to have also the TA bias in Fig. 7 with the ESMs ordered in the same way.
Thank you for the suggestion. TA has been added to Figure 7, which helps to illustrate our findings.
25)
Fig. 8: Could you set "MMM" in bold so that we can easily spot it?
Could you adapt the y-axis scale to avoid the blank space to the top?
Figure 8 has been removed in response to a comment by reviewer 1.

26)
Discussion and conclusions: In overall, this section could be improved in terms of organization and
content. There are some repetitions and it is sometimes difficult to follow you.
In particular, l. 310-323 are quite messy for instance and the points overlap each other sometimes. We sorted the suggestions for improvement into points that relate to tuning and points that relate to implementing or expanding parametrizations.
In the first point, are you pointing towards including aragonite as well as calcite?
The first point was a suggestion to tune the calcification rate.
What would be the effect of reducing the calcification rate in the Southern Ocean, where deep waters are upwelled?
Biases related to primary production and the calcification rate would be distributed, relatively speaking, equally at the ocean surface and in the upwelling regions of the Southern Ocean there is very little calcification anyway.
For almost all ESMs the TA bias is the same sign in the SO as in most of the global ocean (see Figure 2) so that tuning the calcification rate would have the same directional impact.
Hopefully, this answers the question, otherwise we are not quite sure what specifically the reviewer is asking.

27)
l. 320: are you talking about calcite dissolution (effectively sometimes explicitly modelled) or calcite production (always implicitly modelled in that case)?
The carbon cycle formulation could be expanded to also simulate aragonite, i.e. including both aragonite formation and dissolution. Indeed, it is more important to separate aragonite and calcite for dissolution than for $CaCO_3$ formation.

28)
Another important point, which is not mentioned, is the burial and dissolution at the seafloor.
We will add a suggestion to look into implementing a or improving the model formulation for dissolution and burial at the ocean bottom. These processes, although

now all very well understood yet, are important for the TA simulation in the deep ocean. Globally speaking, improving those processes in the ESMs could help improve surface TA slightly on long time scales, while in shallow shelf seas they might be more important.

29)

Fig. S2: It is not compared to the GLODAP climatology here, as opposed to Fig. 6.

Figure S2 shows the results from the TA* method for the models and GLODAP side-by-side. GLODAP is the thick black line.

30)

Table S1: Not useful, I think.

We would like to keep it, just in case.

31)

There might be some issues in the References. There is at least one for Planchat et al., 2023, since some authors are missing.

Corrected the reference.