The manuscript “Reconstructing ocean carbon storage with CMIP6 models and synthetic Argo observations” by Turner et al. is well written and easy to follow. The figures are well chosen, good to read, and facilitate the understanding of the study. The methods are sound and well presented, although some precisions at a few points would be needed, I think.

However, I am somewhat concerned about the novelty and significance of this work for this journal, although this might well be due to a misunderstanding of the methods. To be sure that no misunderstandings have occurred, I am first summarizing the paper very(!) briefly in my own words. If the authors detect significant flaws, please consider them when reading my following comments.

As far as I understand, the here presented method builds on the assumption that the DIC (averaged over the first 100 m in the ocean) at a point x,y can be expressed as the average DIC from 1955 to 2014 at this point x,y plus a delta DIC. This delta DIC can, for any moment in time, be approximated by a linear combination of delta pCO2, delta T, and delta S at locations more or less close to point delta x,y.

My comments are:

- Long-term trend
To me, it seems somewhat trivial that the long-term trajectory of DIC in the first 100 m, a depth that is more or less representative of the average mixed layer depth, is mainly determined by the atmospheric pCO2. The provided BATS example is an excellent example (blue dashed line in Figure 7b). By knowing also temperature, which affects solubility, and salinity, which is closely related to alkalinity, at only one point close to the point of interest, slight deviations of this long-term trend can be represented, as well as inter-annual and decadal variability that is not due to changes in primary production or remineralization can be represented. This is nicely represented by the HOT example (blue dashed line in Figure 7c). As the authors describe in their paper, the underlying mechanisms are well-understood: air-sea CO2 flux due to increasing atmospheric pCO2, decreased solubility with increasing temperatures, and more DIC with increasing alkalinity. Therefore, I think that the predictability of DIC in the mixed layer from pCO2, T, and S as well as a background DIC, is not surprising.

- Variability (detrended signal)

When the detrended response is looked at, pCO2, which is responsible for the largest part of the trend, should not play any role anymore as annual globally averaged pCO2 has little variability. At this point, the RMSE improvement seems to drop (Figures 5d-f and 6d-f). It remains mainly high in the relatively slow and calm subtropical gyres and becomes smaller in the more dynamical regions like the tropical Pacific, eastern upwelling regions, or the Southern Ocean. In these regions, variability in close-to-surface DIC is likely influenced by a variability in the circulation and hence the upwelling of nutrients and carbon. The negative correlation between DIC and T in the tropical eastern Pacific (2d) due to El Nino is a good example. It would be nice to discuss and analyze a little bit more, why these regions are less predictable. As these regions are usually also the regions with the highest variability in surface ocean DIC, I was wondering how the statement that the RMSE of the detrended signal was reduced by 60% was calculated. Did you calculate the area-averaged eta from Figures 5e and 6e? Or did you compare the detrended true signal of globally averaged DIC to the predicted signal of globally averaged DIC? I am not sure which way is better as both have potential pitfalls. Taking the area-weighted average of eta gives potentially to strong emphasize to regions with little variability, as the subtropical gyres, that contribute only little to the global ocean DIC signal. However, by taking the area-averaged DIC first, errors in different directions may globally compensate. Maybe it would be best to use the local eta at each point x,y and weight them by the true DIC variability at that point x,y?

- Significance of the first 100 m for the long-term ocean carbon sink
Another point that I am not sure about is the importance of this analysis for the global ocean carbon sink. How much of the additional DIC is in the first 100 m? The Global Carbon Budget 2021 (https://essd.copernicus.org/articles/14/1917/2022/essd-14-1917-2022.html) estimates that from 1960 to 2020, 115 Pg C has entered the global ocean. Recent estimates from observationally constrained ESM output suggests that it might be a bit more. Thus from 1960 to 2014, the ocean has taken up very approximately ~100 Pg C. The upper ocean carbon inventory shows an uptake of ~14 Pg C (tried to read that number from Figure 7a). Thus, the upper ocean is ‘only’ responsible for 14% of the global ocean carbon sink. Wouldn’t it be more important to quantify how much is transferred from the surface ocean to the interior ocean, i.e., subducted below the mixed layer (https://agupubs.onlinelibrary.wiley.com/doi/full/10.1002/gbc.20092, https://agupubs.onlinelibrary.wiley.com/doi/full/10.1002/2015GL065073). Recent studies indicate that salinity and temperature might indeed be able to reproduce the interior ocean Cant (https://www.science.org/doi/10.1126/sciadv.abd5964, https://www.nature.com/articles/s41586-020-2360-3, https://bg.copernicus.org/articles/18/2221/2021/bg-18-2221-2021.html#&gid=1&pid=1, and https://bg.copernicus.org/preprints/bg-2022-134/). However, the interior ocean transport might make this much harder. The authors mention that this will happen in the next step, but I think it could/should also be incorporated in here if(!) the long-term trend remains a focus of this study.

- **Significance of the first 100 m for the variability of the ocean carbon sink**

Figure 7a suggests, however, that the first 100 m might be very interesting for the inter-annual or decadal variability. In Figure 7a, you show the variability of the upper ocean 100 m DIC. How much of this is caused by the air-sea CO2 flux and how much by transport to the interior ocean. Could the variability in the air-sea CO2 flux maybe largely be derived from variability in the ocean T and S? Or in the variability of subduction on these time-scales? Can your method provide an estimate of the subduction of DIC on an annual basis using the surface ocean air-sea CO2 flux estimates in combination with your estimates? The example of HOT in NorESM is striking as it suggests that the upper-ocean has taken up almost no carbon between 1990 and 2014 despite an increase in atmospheric pCO2.

- **Proof-of-concept without application to real data**

After having read through the manuscript, I was really disappointed that this new method is not yet applied to observations right now. There is much potential, and I understand the
idea of making two publications, one for 'proof-of-concept' and one for the application, but now it seems as if something is missing in this paper. If no institutional, or PhD-related restrictions exists, I would recommend to also present the application here.

My overall recommendation would therefore be to de-emphasize the long-term trend, focus on the variability, apply the method to observations, compare it to estimates of the air-sea CO2 flux, and try to address the subject of the decadal and inter-annual variability of the ocean carbon sink.

As you already have the models, you might even be able to draw conclusions why and where models do not capture the variability. Are the regions with the largest variability in the air-sea CO2 flux also the regions with the largest variability in the upper ocean DIC?

If the focus is more on the detrended signal, the subsurface ocean would not fit anymore and could be left for another study.

While this sounds somewhat negative, I am in strong favor of publishing this manuscript in Biogeoscience but with an application to observations. If not applied, maybe GMD is more suited for 'proof-of-concept' studies? But that is an editorial decision.

**Minor comments:**


Line 25, page 2: Gruber et al. is about compound extreme events and rather summarizes literature. I would recommend to replace it by this earlier study on global ocean acidity extremes (https://bg.copernicus.org/articles/17/4633/2020/) and this regional study (https://www.nature.com/articles/s43247-021-00254-z) that provides one example in detail.
Line 57: I think this is the first time that you use the term ‘upper ocean’. Maybe define it here.

Lines 60-66 and lines 99-100: Would suggest to add the Southern Ocean studies (https://www.science.org/doi/10.1126/sciadv.abd5964, https://www.nature.com/articles/s41467-022-27979-5), and one study from the Arctic Ocean (https://www.nature.com/articles/s41586-020-2360-3) that show how surface observations (and variations) of S and T can alter the CO2 uptake. Or maybe keep that for later in the Discussion when you describe the potential to go below the surface.

Line 112: Does the method account for inter-dependencies between the predictors? Sorry about the question, but I am not very familiar with this kind of method.

Line 132-133: Just wanted to say that I really like that you provided the regridding method here in such detail. Often it is not possible to reproduce data because the regridding software is not named.

Line 141: Something is wrong in the sentence, I think.

Lines 147-154: Just to be sure that I understand it right. When you use the ARGO 2002 coverage, you assume that you have data at each of these points for all years from 1955-2014? Is that right? If yes, how do changes in the temporal coverage effect time series as shown in Figure 7? If you have little data for the first 30 years and then more, you should get a better representation of the variability over time, right? Figure 7c seems to suggest that there is only 1 observation for the blue line, but that means one observation per year in the vicinity of the HOT station, right?

Section 3: I found this section a little bit difficult to follow. What could help (maybe it doesn’t), is to look only at the long-term signal and only at the detrended for the correlations. Than the long-term should give you a correlation between pCO2, T, and DIC. T and DIC because the ocean is warming and taking up carbon (but one does not cause the other) and the other between pCO2 in the atmosphere that causes DIC in the ocean to increase. For the detrended signal, I would expect no correlation between pCO2 and DIC as the pCO2 trend is gone. T and DIC should be negatively correlated (like shown in Figure 3b). And salinity should be positively correlated. A little bit along the lines of Figure 2 in https://www.nature.com/articles/s41467-022-32120-7

General: Did you think about including the area of sea ice in each cell as well? As this blocks the air-sea gas exchange to some extent, it might be a powerful predictor in the high-latitude oceans. Just a guess.
Section 6: While I completely understand the reason to only choose Nor-ESM to make these tests, I would be curious how other models perform.