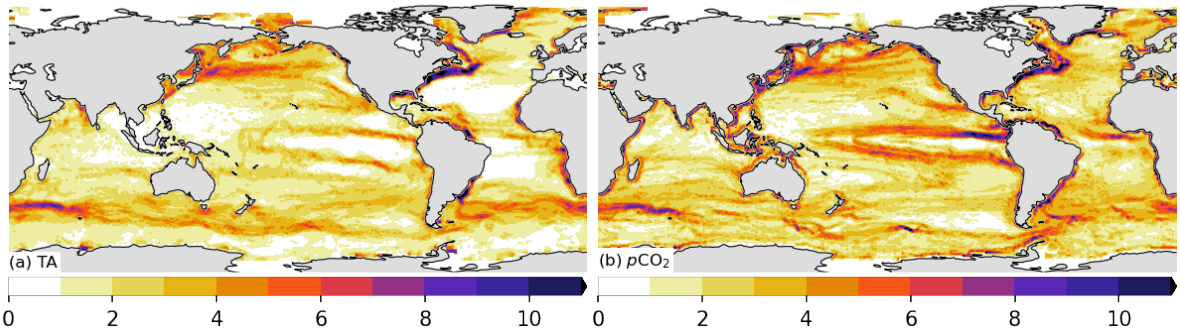
Response to Reviewer 1 (R1)

We thank reviewer 1 for their prompt and positive feedback. Their review was detailed and was a great help in preparing the revised manuscript. Below, we outline how we will address the major points raised : 1) recommendation on how modelers should use the data set; 2) discussion of coastal estimates , particularly considering the new merged Landschützer et al. (2020) product; 3) missing appendix. We included below the reviewer's comments in blue and our responses in black. Italic green font indicates text that will be added to the manuscript. The location of the text in the manuscript is indicated if applicable. The line specific grammar issues will be addressed and shown in the track-changed document. Note that we have also refined our definition of error and uncertainties. These changes have also been marked in the tack changes file.

1. Inclusion of the coastal ocean

It would be good to add 1–2 paragraphs on potential issues in the coastal zone compared to open ocean and recommendations for future efforts. It would also be useful to connect to this recent paper: Landschützer, P., Laruelle, G. G., Roobaert, A., and Regnier, P.: A uniform pCO2 climatology combining open and coastal oceans, Earth Syst. Sci. Data, 12, 2537–2553, https://doi.org/10.5194/essd-12-2537-2020, 2020.

We will add a paragraph in the discussion about the validity of our coastal estimates. In particular, we will outline why we have some confidence in these estimates (with the figure below supporting our point), even though there also some clear limits. Further we will make reference to the MPI-ULB-SOMFFN merged product. Lastly, we will emphasize that users should consult the original SOCAT data for comparison if they would like to use the product beyond the climatological scale. We plan to include the following text in the discussion.



Caption (Appendix): *Map of the position of cluster boundaries across all ensemble members and months for (a) total alkalinity and (b) pCO2. The white regions indicate locations that belong almost exclusively to the same cluster. Dark regions show where cluster boundaries are persistent.*

Discussion: *The OceanSODA-ETHZ product extends further into the coastal margin than most previous studies (Iida et al., 2015; Land-schützer et al., 2016; Denvil-Sommer et al., 2019). This is achieved i) by including coastal observations during the training, and ii) by using a larger number of clusters compared to other clustering approaches (Landschützer et al., 2016; Watson et al., 2020). This permits to better separate open ocean and coastal variability through the inclusion of suitable variables in the clustering step (e.g. Chl-a for pCO_2, and see Figure A3 to see a representation of cluster boundaries). This gives us some*

confidence in the coastal estimates, at least on a climatological scale with regard to the seasonal cycle. Our product is therefore comparable to that of Landschützer et al. (2020) who blended separate coastal and open ocean pCO_2 products into a single climatological product with monthly resolution (Landschützer et al., 2016; Laruelle et al., 2017).

The total uncertainties of our estimates in the coastal ocean are considerably larger compared to the open ocean estimates (Figure 7). This reflects the much higher spatio-temporal variability of the physical and chemical environment in the coastal ocean, leading to much higher variations in the marine carbonate system (Laruelle et al., 2017). Since our predictor variables are only partially reflecting this variability, a large portion of the high total uncertainty is due to a high representation error (Table 3). Increasing the resolution of the products may improved coastal estimates as done by Laruelle et al. (2017). Until we arrive at this point, the OceanSODA-ETHZ data should be used with care in the coastal ocean. Further, we recommend that researchers interested in the investigation of interannual variability and trends in the coastal ocean using the OceanSODA-ETHZ product should also look a the underlying in situ data to gain a better understanding of the variability, trends, and uncertainties for the coastal region of interest.

2. Recommendation to modelers

You mention ocean models briefly in the introduction (L45). It would be helpful for the community if you could make some recommendations (based on what you have learned in the development of this paper) for accurately simulating and benchmarking ocean acidification in numerical ocean models.

We will add a section with specific recommendations to users of the product, containing a paragraph aimed at the modeling community. Specifically, we will add an additional uncertainty estimate that provides climatologically mapped errors based on test data that has not been seen by the trained model. This will permit modelers to assess model-data misfits on a local basis, thereby making model-data comparisons more quantitative.

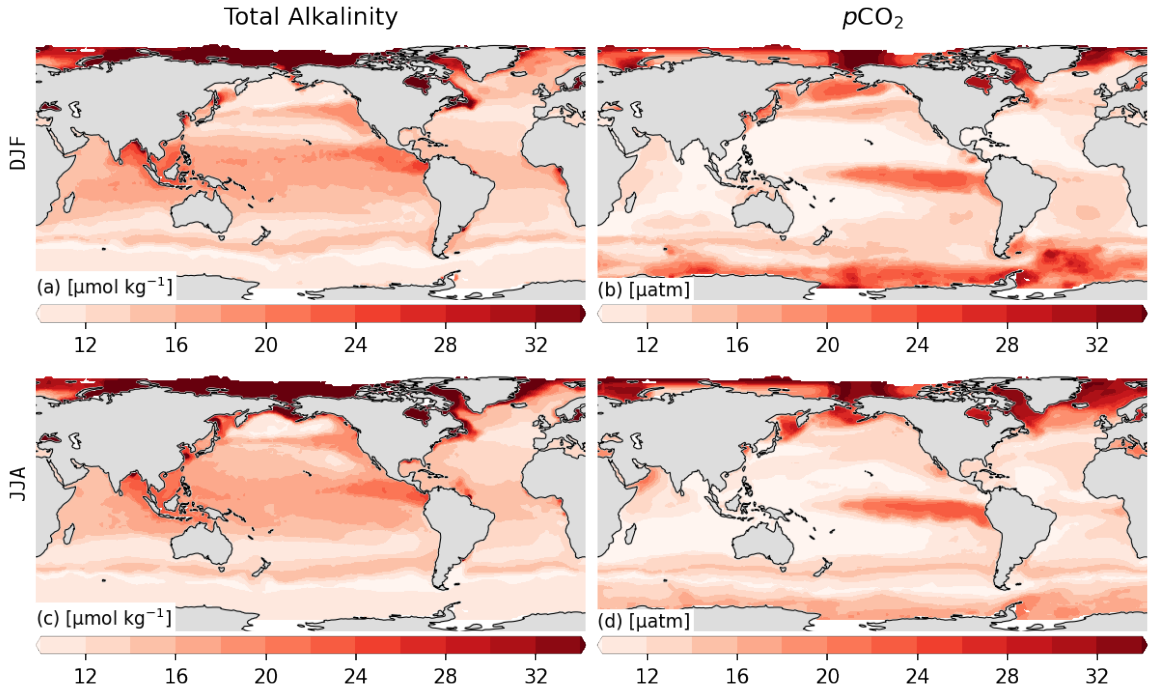


Figure A2: *The Huber test scores mapped to the ensemble clusters for Total Alkalinity (TA) and pCO_2. The top row shows Huber scores averaged for December, January, and February (DJF) and the bottom row June, July, and August (JJA). The Huber score is a blend between root mean squared error (RMSE) and mean absolute error (MAE), where MAE is applied to values that are considered outliers. Only test data is used to calculate these climatological scores, meaning that the scores are based on GLODAP2 and SOCAT data for TA and pCO_2 respectively.*

Discussion: *The OceanSODA-ETHZ product provides a useful comparison for numerical models with the full marine carbonate system. The spatially resolved climatological estimates of uncertainty for TA and pCO2, based on in situ data, provide useful context for ocean modelers on a climatological time-frame (Figure A2). In the same way that previous studies have used pCO2, the OceanSODA-ETHZ data set can also be used to compare interannual trends and variability of the marine carbonate system (Landschützer et al., 2015, 2016; Gregor et al., 2018; Keppler and Landschützer, 2019).*

(Appendix) *One of the advantages of using the GRaCER approach is that any metric can be mapped from the results to the appropriate clusters, resulting in an ensemble of metric scores. The possible metrics that can be applied include bias, root mean squared error, and mean absolute error. Further, these metrics can be applied to test data, meaning that the resulting scores can be based on test scores — that is data that are unseen by the model during the training process, thus giving a true representation of the uncertainty. Given that the cluster used in this study are climatological, we can get fully mapped climatological estimates of uncertainty.*

Missing appendix

Appendix A3.4 seems to be missing, please correct this

This will be corrected to include text on how the variable importance figure was created. The main text now will refer to the figure instead of the section in the appendix.

Response to Reviewer 2 (R2)

We thank reviewer 2 for their prompt and positive feedback. R2 provided an in-depth critique of the method and of the data product itself. The major concerns raised include: 1) The use of a gap-filled pCO2 climatology to perform clustering; 2) the use of climatological nutrients in solving the marine carbonate system; 3) uncertainties of the dataset, particularly over time. We have included the reviewer's comments in italicized blue font, while our answers are given in black. We grouped all comments and answers into these three topics as there is significant overlap in many of the reviewer's comments. We have also refined our definition of error and uncertainties. These changes have also been marked in the tack changes file.

1. Use of gap-filled pCO2 to cluster

I believe the novelty of the GRaCER method is [it] that produces an ensemble of clusters. The ensemble is produced because the clustering process randomly assigns the first cluster center in the predictor-variable space. Thus, each member of the ensemble has a different center, and therefore the ensemble mean does not have discrete boundaries. Based on this, it would be helpful to clarify a couple of details:

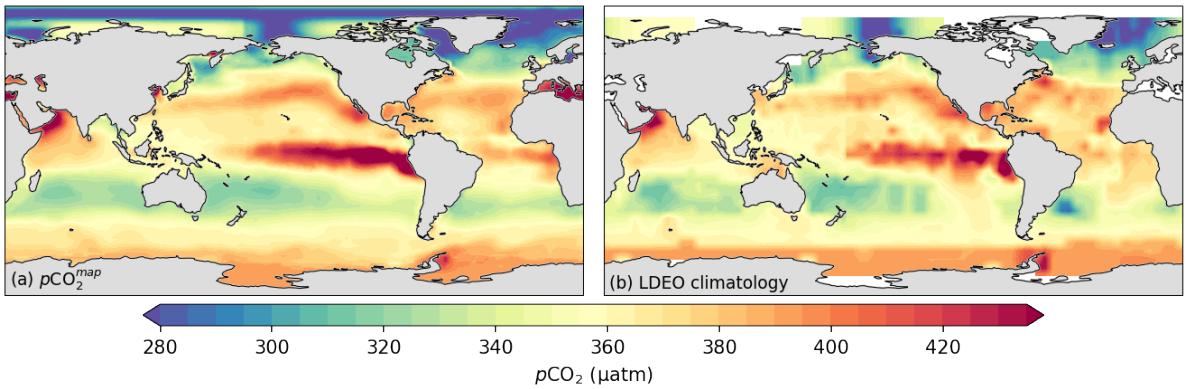
- *Lines 245-248: "It may seem tautological to use other machine learning estimates, but these data are just used to create regional clusters, i.e., they are not used in the regression step. " How do the results vary if you do not use the previous machine learning estimates for clustering? Previous methods (Landschutzer, Rodenbeck etc), only use the observations from SOCAT for clustering. In this manuscript these datasets are not used in the regression step, but I believe that the results are affected by which data is used for the clustering.*
- *Lines 172-179: How does the result vary, if you use monthly data instead of climatologies for the clustering process?*
- *The cluster seems to collect data by climatological month. How does he method change if monthly data is used instead?*

The reviewer raised a question about our using of pCO2 climatologies (called pCO2map) that are based on similar machine-learning methods. This indeed could come across as tautological, especially since two of the products we are using were derived using related methods, i.e., the the SOM-FFN estimate of Landschützer et al. (2013) and the LSCE-FFNN estimate of Denvil-Sommer et al. (2019). We consider this issue as very minor, since use these climatologies just for the clustering step, and not for the final estimation of pCO2. This is done through our regression approach. Extensive testing has shown that our results are very robust with regard to the choice of pCO2map. This is because, all four mapping methods are able to predict the seasonal cycle of pCO2 well (Sommer-Denvil et al. 2018), which is the main dimension of variability that is captured by the clustering step.

Thus one could argue that given this lack of sensitivity, one should restrict pCO2map to just the truly independent methods, such as the LDEO climatology. We feel that the benefit of using multiple products in pCO2map outweighs this alternative. The use of an ensemble avoids overfitting in regions of where the LDEO climatological distribution is noisy (see

Figure). It also permits us to cluster regions that are not covered by the LDEO climatology. The following change was made to the manuscript with underlined text being inserted.

Clustering is performed on climatological values of pCO2, SST, mixed layer depth and Chlorophyll-a, with additional weighting given to pCO2. As with TA, all variables are standardized prior to clustering with (x - μ) / σ, after which pCO2 is multiplied by 3 to give it stronger weighting. The larger weight given to pCO2 means that monthly clustering would result in very similar results to climatological clustering, as only SST and Chl-a would vary over time and not pCO2. Details of the regression method, and of the hyper-parameter selection are given in section.



Will not be added to manuscript: Owing to the substantial amount of “”noise” present in the LDEO climatology stemming from the way the measurements are interpolated (b) it is prone to create spurious clusters. Further, the coverage of pCO2map (a) allows for clustering and thus predictions in regions not mapped by the LDEO climatology, such as the Mediterranean.

We also prefer to stick with our original choice of using the climatological distribution of pCO2 (and Alk) for the clustering. First, as also pointed out by the reviewer, there is much less confidence in the interannual (monthly) estimates. Using these data for clustering rather than the climatology of these products would run the risk of creating problems given that such a step would also further increase the weight of the products. This is because pCO2 is weighted three times more than the other variables in the clustering step. Second, and from a more fundamental perspective, we argue that such a step would undermine a strength of the two step approach, i.e., its separation of variations on different timescales. The clustering step is meant to isolate primarily regions with the same seasonal cycle. The regression step is meant to explain the variability within each region. This is based on the assumption that, to first order, interannual variability can be considered as modifications of the seasonal cycle. If we were to allow the clusters to vary inter annually, we would lose this fundamental distinction, and we would ask the SOM step to take over a bigger burden of the total variance. Given the discrete nature of the mapping, this can actually lead to worse results.

In summary, we consider our choice to be well justified. At the same time, our experience indicates that it is unlikely that the outcome of the regression is impacted much by the details of the clustering step. We added some small comment on this issue to the method section. The inserted text is underlined:

The main advantage of such a two-step approach is that the first clustering step organizes the variability regionally and temporally. This greatly enhances then the fidelity of the second step, i.e., the regression, as the size of the regression problem is reduced from the global

domain to smaller, more homogeneous regions. A second advantage is that this clustering brings together regions with similar seasonality and similar co-variability with potential predictors, irrespective of the number of observations. The regression step explains the variability within each region over time and space dimensions, including interannual variability. Further, the clustering permits the regression to transfer information from spatially distant, but geochemically similar regions, making the inter and extrapolation more robust in data poor regions.

…

For the clustering step, we use monthly climatological data of pCO2 and TA and related parameters (Figure 2a-c), to deter-mine the main patterns of variability of the target variable and its co-variability with potential predictor variables. Concretely, the clustering step is meant to isolate primarily regions with the same seasonal cycle.

Figure 8 (d) shows the mean of the monthly differences between the SOMFFN and GRaCER pCO2 datasets. When I plot the time-series of the monthly differences between SOMFFN and GRaCER averaged over the eastern equatorial Pacific, I see a continuous decrease in the difference from 1985-2018. That could mean that for the beginning of the time period there is a larger difference between the two methods than by the end. Could this be a consequence of using the SOMFFN and other products for the clustering process? Since at the beginning of the period there is less observational points compared to the end.

No, various tests showed that this trend in difference is not a consequence of our use of SOMFFN estimats in our clustering. The first part of the difference is largely due to the fact that data are very sparse. This makes the mapped estimates more sensitive to the specifics of the methods, especially in the regression step (not in the clustering step). In the regression step, SOMFFN and GRaCER are actually quite different, which explains the divergence of the estimates. In response, we emphasize in our new section on data use that the first part of the timeseries should be used with great caution.

The MLD sub-annual and inter annual variability is removed.

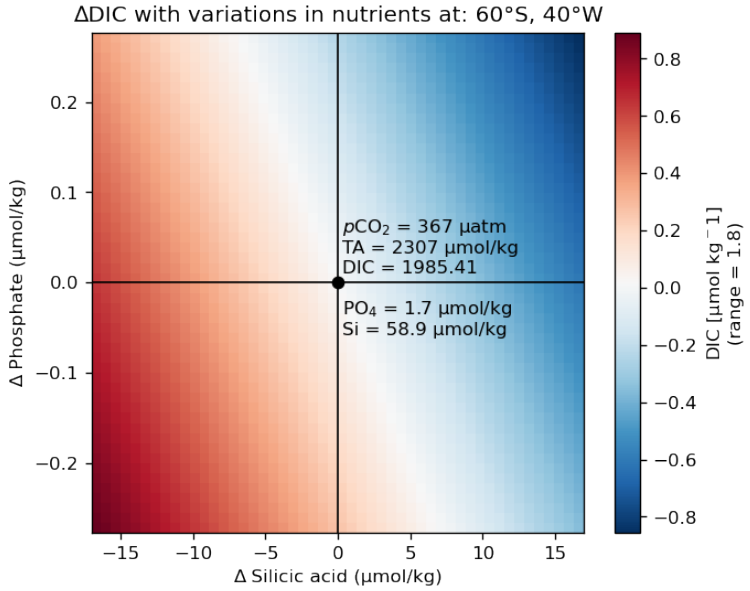
We use an observationally based mixed-layer depth product that is only available as a climatology (Holte et al. 2018). This product reports the mixed layer depth for Argo profiles globally and is not normalized to a specific year. Hence, the interannual variability of the MLD is still present in this data product. In a climatological context, this interannual variability acts as noise rather than signal. We thus remove this interannual signal using a Gaussian smoother.

2. Climatological nutrients for marine carbonate system calculations

To calculate DIC and pH they use climatologies of silicic acts and phosphate instead of monthly data.

Using climatological concentrations of PO_4 and SiO_4 in the calculations of the marine carbonate system instead of the interannually varying concentrations has a very small impact on the computed values. We base this conclusion on the following worst-case scenario, i.e., that the year to year variability is larger than the seasonal cycle in these nutrients. To illustrate this, we take a location in the Southern Ocean (60S, 40W) characterized by a very large seasonal cycle, and vary silicic acid (59 ± 15 µmol/kg) and phosphate (1.7 ± 0.3

µmol/kg) over this seasonal range. As shown by the figure below, the maximal impact of using climatological nutrients in this calculation is 1.8 µmol/kg. In reality, the range of interannual variations will be much smaller. Thus, we consider the potential implication of our using climatological nutrient concentrations instead of interannually varying ones as neglibile. In response, we will add to the text that this assumption has very little impact.



Not added to manuscript: The range of DIC when using a range of phosphate (PO4) and silicic acid (Si) to solve the marine carbonate system from pCO2 and TA. The input ranges for PO4 and Si were determined from the magnitude of the seasonal cycle from a region where the variability is large.

3. Uncertainty related points

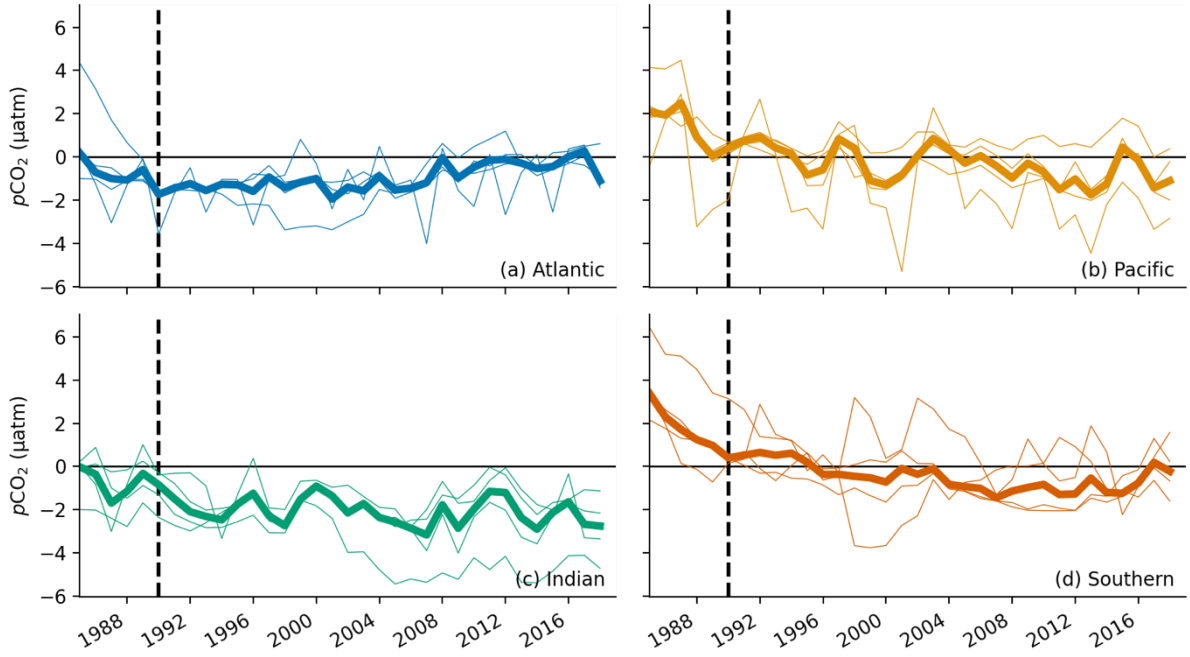
Moreover, some methodologies used in the manuscript should be discussed in the context of interannual to decadal variability

We will add a section in the discussion of the manuscript that gives recommendations for the use of the OceanSODA-ETHZ data (as per request of R1). In this section we caution users of the product to treat data prior to 1990 with care as sparse data results in substantial uncertainties in the estimates, and also larger differences between the different methods as shown in Watson et al. (2020). This will be accompanied by the figure shown below.

Discussion – Recommendations for use: *However, users of the OceanSODA-ETHZ product should be aware of the fact that that data prior to the 1990's should be treated with care due to the paucity of SOCAT pCO2 training data during this period (Rödenbeck et al., 2015; Watson et al., 2020). This was rescently demonstrated by Watson et al. (2020) who used an ensemble of various regression approaches to show that the spread of pCO2 estimates prior to the 1990's is large due to the paucity of data. Similarly, Gregor et al. (2019) showed that pCO2 estimates prior to 1990 tend to have aslightly positive bias. Hence, the trends shown in Table 5 are calculated for the years after 1990, i.e., covering the period 1990–2018.*

If the authors consider that their dataset is good to estimate internal and decadal variability, then they should add some analysis and comparison with other existent pCO_2 observation-based datasets.

Agreed. In response, we will add the figure below to the manuscript showing the basin-mean difference between OceanSODA-ETHZ pCO2 and that of the other methods: MPI-SOMFFN, Jena-MLS, CMEMS-FFNN, and CSIR-ML6. The thin lines show the individual method differences, while the thick line shows the mean difference of the four methods. In the Indian ocean, the OceanSODA-ETHZ pCO2 estimate is persistently lower than the pCO2 estimated by the ensemble of the four other methods. A similar negative difference is found in the Atlantic, but the difference diminishes from 2008 onward. In the Pacific and Southern Ocean, there are positive differences prior to 1990 that diminish thereafter. The spread of the relative differences is larger in the Pacific and Southern Ocean, not unexpected given the much larger data gaps in these ocean basins.

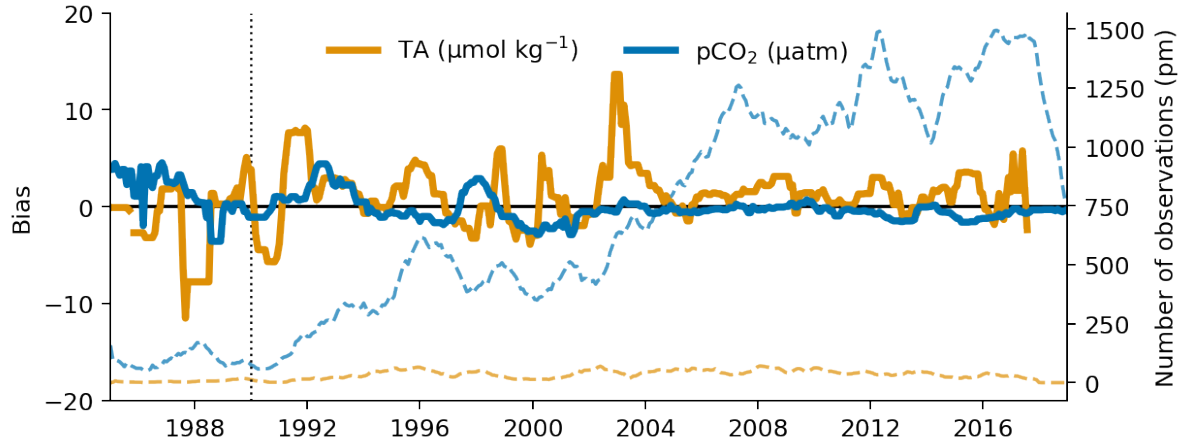


Results – Comparison with other products: *basin-mean difference between OceanSODA-ETHZ pCO2 and other methods: MPI-SOMFFN, Jena-MLS, CMEMS-FFNN, and CSIR-ML6. The thin lines show the differences to the individual methods, while the thick line shows the mean difference across the four methods.*

Results – Comparison with other products: *We also show the basin-mean temporal differences between OceanSODA-ETHZ pCO2 and other gap-filling methods (Figure 9). In the Atlantic (Figure 9a), OceanSODA-ETHZ pCO2 is < 2 µatm lower than the mean of the other gap-filling methods for the period 1990 to 2008. Thereafter, the difference is < 1 µatm. In the Indian ocean, our pCO2 estimates have a persistent negative difference of ~ 2 µatm (Figure 9c). The comparison in the Pacific (Figure 9b) is the most consistent with the other methods, with a slight positive difference in the beginning of the period (pre-1990). The OceanSODA-ETHZ estimates of pCO2 in the Southern Ocean (Figure 9d) have a large positive difference prior to 1990 – up to 6 µatm for one of the ensemble members. This difference quickly diminishes and is near zero by 1990. There is also a negative difference later in the period (2004 to 2015); however, the ensemble spread over this period is large.*

The comparison with other methods illustrates that while gap-filling methods are converging on a global scale, there are regionally differences. Further, large differences in pCO_2 between methods prior to 1990 indicates high uncertainty for this period.

To further address R2's point, we will add the figure of the bias of the TA and pCO2 relative to the training data sets (GLODAPv2 and SOCAT respectively) in the supplementary material. The biases for pCO_2 are larger (~5 µatm) at the beginning of the time series when there is less data. Gregor et al. (2019) found similar biases for pCO_2 in the pre-1990 period using gradient boosted trees. TA biases are erratic over time due to the highly uneven sampling distribution (*e.g.*, not only in time, but also in space, e.g., by sampling in river plume areas where the uncertainty is large). The low number of samples exacerbates this effect.

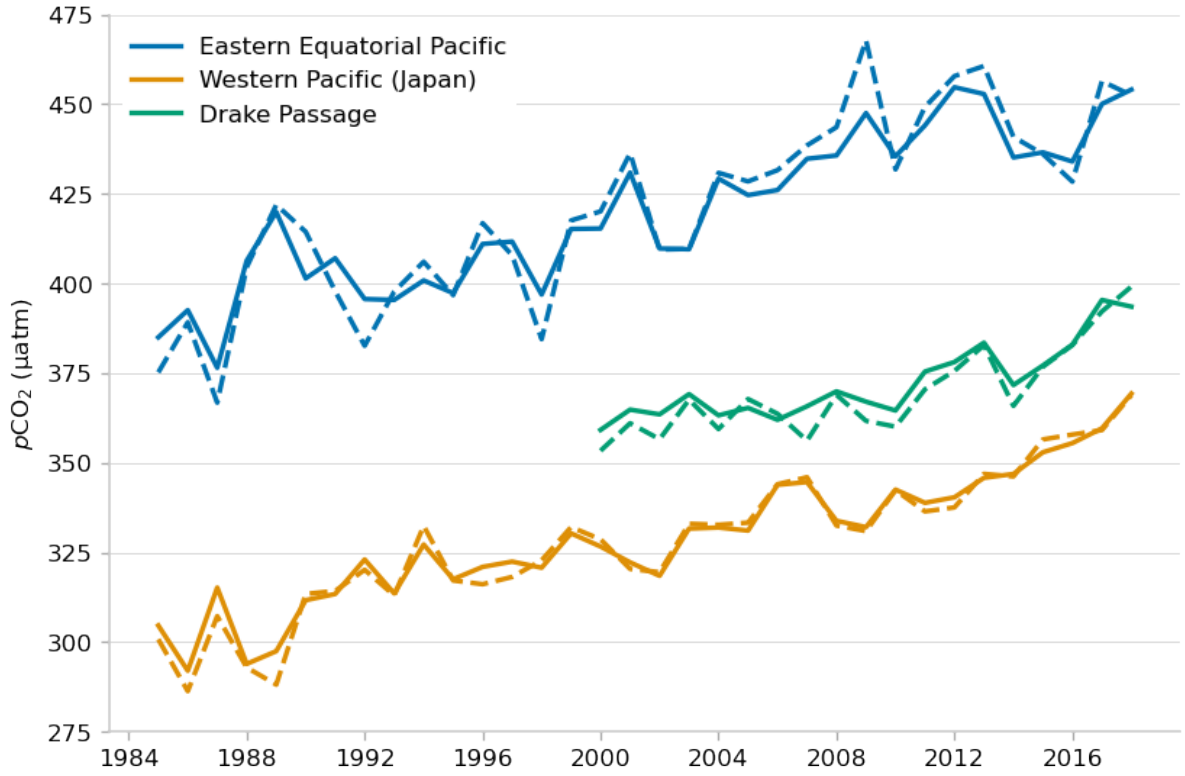


Appendix: *Timeseries of the median bias (solid lines) of TA (orange) and pCO_2 (blue) relative to the GLODAP and SOCAT datasets, respectively. The dashed lines show the number of observations for each of the data (right axis).*

A pCO_2 RMSE of 12 muatm is larger than inter annual variability for many locations.

The RMSE represents a distribution of the uncertainty theoretically centered around a zero mean. It provides an estimate of the expected uncertainty of a particular instance. In our case, this is an estimate for a single grid cell for a single month. When assessing variations in pCO2 (or any of the other quantities), we usually analyze the temporal variations averaged over a larger region or averaged over an entire year. Even though the data are to a certain degree autocorrelated in time and space which reduces the number of degrees of freedom somewhat, the uncertainty of the mean estimate is reduced by the square-root of the (reduced) number of degrees for freedom. For example, when averages are formed over 10°x10° regions in the annual mean, we expect at least a factor of ten reduction in the uncertainty of this mean. Thus, we expect an uncertainty of around 1µatm of this mean, which is much smaller than the signal one is interested.

When looking at long-term trends, potential biases matter as well. The biases in our product are in general $\ll 10$ µatm, with the exception of parts of the Southern Ocean, and the Eastern Tropical Pacific (see paragraph below). Fortunately, there are well sampled areas (through time) in these regions allowing us to compare the OceanSODA-ETHZ pCO_2 estimates with SOCAT directly. In the figure below we show direct comparisons of pCO_2 in the Eastern Equatorial Pacific, Western Pacific, and the Drake Passage. The interannual variability is well captured in these regions, confirming our hypothesis. Still, the large interannual variability in the Eastern Equatorial Pacific is occasionally underestimated.



Not added to manuscript: Comparison of OceanSODA-ETHZ (solid) and SOCAT (dashed) pCO2 for open ocean regions. The three chosen regions have persistent occupancy for the selected regions and periods: 62% for the Eastern Equatorial Pacific (< 5°N/S < 180°W), 84% for the Western Pacific (25° to 40° N, 128° to 145°E), and 76% for the Drake Passage (> 50°S, between 73° and 65°W over the period 2000 to 2018).

Lines 412-415: "The highest biases on pCO_2 when comparing with observations are located in the eastern tropical Pacific where inter annual variability is higher". This may suggest that the dataset does not represent well inter annual variability

The analyses above demonstrate that the Ocean-SODA-ETHZ product is able to capture most of the large variability observed in the Eastern Tropical Pacific. The large RMSE and biases are found at the edges of the tropical Pacific, where the mean lateral gradients are high. These lateral gradients shift strongly during El Niños and La Niñas, posing a challenge to any interpolation method. Any small deviation in the specific location of the gradient leads to large local biases, although the large-scale spatial mean is well captured. The fact that the interannual variability of the Eastern Equatorial Pacific is still well captured (as shown above) reinforces this.

4. Extra additions

Lastly, we have added a table showing the trends of a set number of variables of the marine carbonate system. The trends are calculated for the period 1990-2018 due to the reasons explained above.

Results – Regional Trends: *The global and basin-scale trends for pCO_2 are remarkably consistent (~ 16.5 µatm decade^{-1}) when compared with pCO_2^{atm} (~ 18.6 µatm decade^{-1}), with the atmospheric slope being slightly steeper than the oceanic trend (Table 5). The basin-scale consistency holds true for pH (-0.016 units decade^{-1}) and Ω_{ar} (-0.07 units decade^{-1}), where global values are consistent with the regional values. Total alkalinity trends are more variable from basin to basin but are driven almost entirely by salinity with a basin-scale correlation of 0.99. Lastly, DIC shows similar variability to TA trends, which makes sense in terms of TA increasing the buffering capacity of seawater while pCO_2 remains relatively consistent on a basin-scale.*

Results – Regional Trends: *Table showing the slopes and associated standard error for OceanSODA-ETHZ variables for the period 1990-2018. All columns show increases per decade (d). All trends are significant (P > 0.05). We exclude the Arctic as the OceanSODA-ETHZ product only covers 23% of this region and may thus give spurious trends. The Ocean basins are defined by the map shown in Figure A4.*

Region	pH	Ω_{AR}	TA	DIC	pCO_2	pCO_2^{atm}
Units	(units/d)	(units/d)	(µmol/kg/d)	(µmol/kg/d)	(µatm/d)	(µatm/d)
Global	-0.015 ± 0.0	-0.06 ± 0.0	1.4 ± 0.1	8.2 ± 0.1	15.5 ± 0.1	17.8 ± 0.1
Atlantic	-0.015 ± 0.0	-0.06 ± 0.0	3.5 ± 0.1	9.9 ± 0.3	15.7 ± 0.1	17.9 ± 0.1
Pacific	-0.015 ± 0.0	-0.07 ± 0.0	0.6 ± 0.1	7.9 ± 0.2	15.7 ± 0.1	17.8 ± 0.1
Indian	-0.015 ± 0.0	-0.07 ± 0.0	3.0 ± 0.3	9.7 ± 0.4	15.3 ± 0.3	17.6 ± 0.1
Southern	-0.016 ± 0.0	-0.06 ± 0.0	0.2 ± 0.1	6.7 ± 0.5	15.0 ± 0.3	17.7 ± 0.1

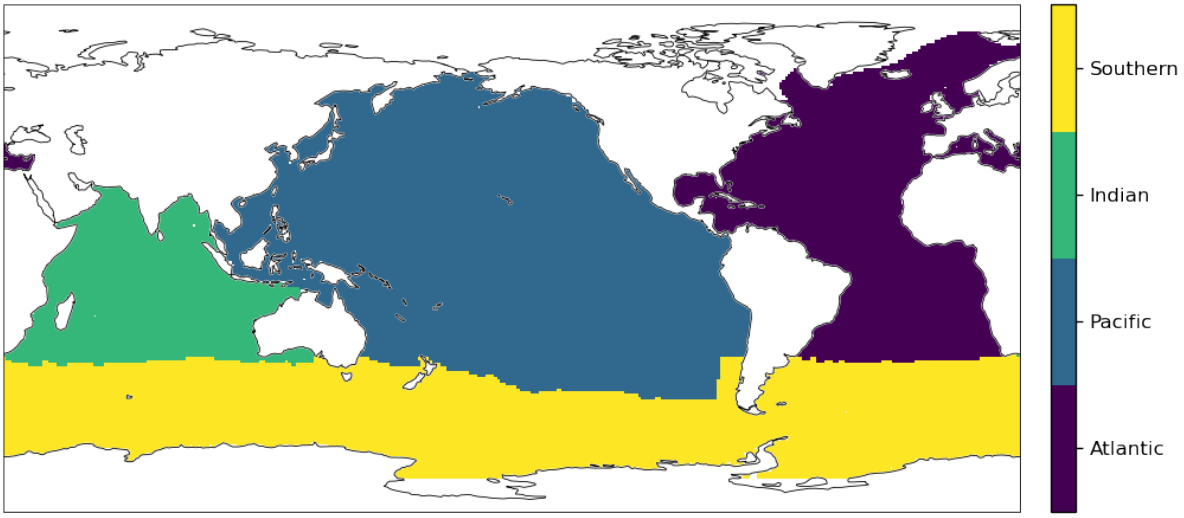


Figure A4: *Ocean basin boundaries used in Table (above) as used by the RECCAP2 project (https://reccap2-ocean.github.io/regions/). The Southern Ocean and North Atlantic boundaries are defined by biome boundaries defined in Fay and McKinley (2014).*