Authors Comments

We would like to thank the reviewers for their thoughtful and constructive feedback, which we think will strengthen the paper and its ideas. Note that we only produce our comments to the reviewers' comments here and do not show the actual changes made to the manuscript. We include some of the reviewers' concerns and statements which are shown in ***blue*** to distinguish them from our text and suggested corrections.

Response to R1 (Peter Lanschützer)

R1's comments were minor and no major changes were suggested. We will implement all the suggestions made by the reviewer. The reviewer made one point which we feel is important to address here:

Page 24 and onward: The authors argue for the inclusion of EKE to move forward, but I was not fully convinced, given the small improvement in Figure 5. Another limiting factor that is not discussed is availability. In an ideal case, one would use time-varying fields of SST, SSS, DIC, TALK, etc. however, in most cases they are not available. As the authors mention, EKE only exists as a climatology, hence one cannot expect directly improved IAV signals from it (as e.g. visible in Figure 5c). I do however agree with the authors on their conclusions regarding the addition of novel proxies.

An important point to make is that we use EKE only as a clustering variable. It will thus only impact pCO2 estimates through the impact that EKE has on clustering. In other words, EKE could capture regions where pCO2 might respond differently to drivers. However, the reviewer is correct in saying that it will not improve the IAV as we use a climatology of EKE.

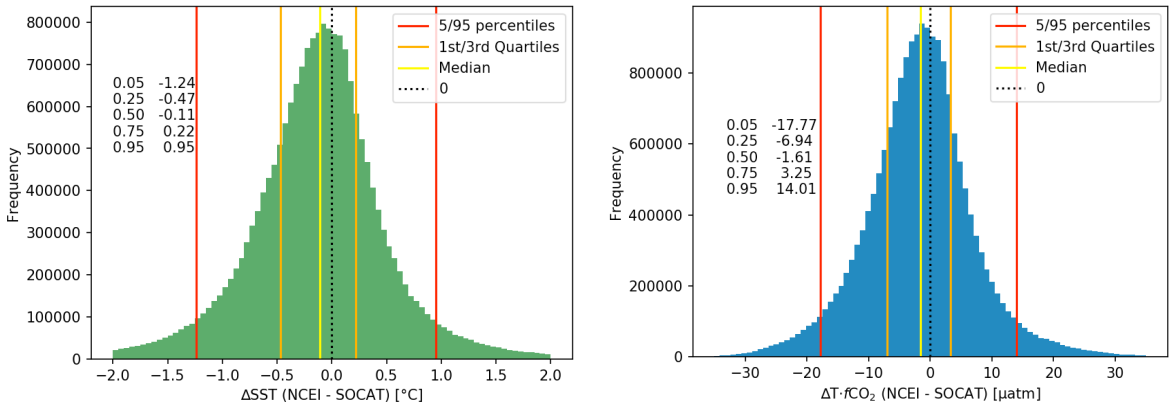
Response to R2 (Jamie Shutler)

R2 had several suggestions in ways to improve the manuscript and were themed primarily around the way in which we deal with SST in our study. The reviewer makes three major suggestions: 1) temperature correction to the pCO2 based on the difference between SOCAT temperatures (ship based) and the optimally interpolated SST; 2) correct handling of SST in the calculation of air-sea CO_2 fluxes; 3) both machine learning methods and validation data suffer from the same temperature uncertainties as highlighted in points 1 and 2. An important point to note is that we have in fact used the optimally interpolated SST by Reynolds et al. (2007) which is estimated only from AVHRR data.

1. *The SOCAT gridded dataset is based on gridded data that have all originated from different ships, systems and times. Hence the gridded values are likely to contain unknown biases due to inconsistent depths and thus temperatures that were originally captured/linked to each pCO2/SST pair, but which was lost as a result of gridding (as SST and pCO2 are gridded individually). Furthermore, the differing depths of each sample means that multiple measurements within each box could be from different depths and so they are not part of the same (statistical) population. This issue can be overcome by first re-analysing the original SOCAT cruise data to a common temperature dataset (and thus a common sampling depth) and then re-gridding them (through the use of a satellite-observed sea surface temperature dataset).*

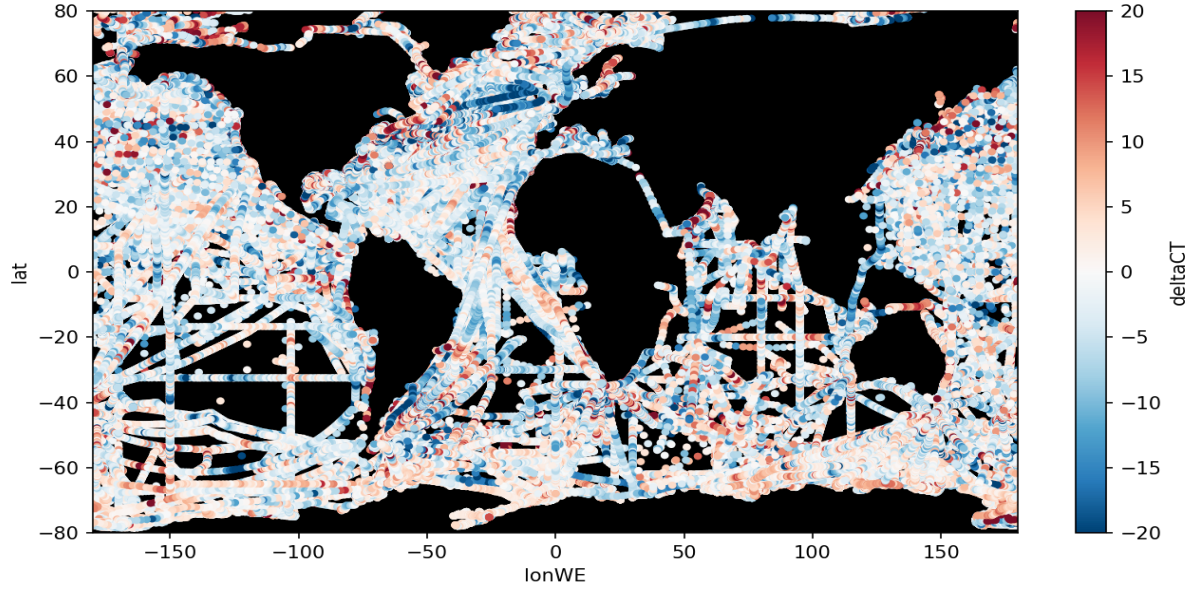
We find the reviewer's first point interesting and potentially important — it is something that we had not considered. We have made a preliminary investigation on this discrepancy between SOCAT SST and optimally interpolated AVHRR SST and the impact it has on pCO_2.

We used the raw (and not gridded) SOCAT data for this experiment. This discrepancy is surprisingly large with the 25th and 75th percentiles being ~ -18 and ~14 µatm. However, the median bias of pCO_2 is -1.59 µatm and the error is pseudo-normally distributed. The spatial distribution of these "biases" are shown on the map on the following page. There seems to be little to no spatial pattern (latitudinal or basin) in the data (at least at our superficial level of investigation). However, in some cases it seems that the along-track decorrelation length scale is longer than completely random noise would suggest, indicating that biases may be driven by the structure/stratification of the upper layer of the water column. There are other factors that could also contribute to these biases such as ship specific biases (warming of intake water before thermosalinograph).



(left) A histogram of temperature residuals (NCEI - SOCAT);

(right) A histogram of fCO2 residuals when fCO2 is corrected (NCEI - SOCAT)



Map of fCO2 difference between SOCAT SST and AVHRR SST (µatm) using correction as suggested by the reviewers: $pCO_2^{AVHRR} = pCO_2^{SOCAT} \times e^{(0.0423 * \Delta T)}$. Note that all SOCAT flags were included.

We performed a second preliminary investigation of the impact that correcting pCO2 would have on the machine learning estimates. The experimental setup for this experiment is somewhat simplified (for the sake of time). We use only gradient boosting and the same testing years used in the manuscript training. We also do not use clusters for this first order approach. We perform three regressions to compare:

1) a control where SOCAT SST is used without correcting pCO_2;
2) OISST AVHRR temperature is used but we do correct pCO_2 for temperature;
3) OISST AVHRR temperature is used and we correct pCO_2 for temperature.

We show these results in the table below

Test results based on gridded 1° SOCAT data	**MAE (µatm)**	**RMSE (µatm)**	**r^2**
SOCAT temperatures no pCO_2 correction	12.93	20.65	0.73
AVHRR temperatures no pCO_2 correction	13.35	21.07	0.72
AVHRR temperatures pCO_2 corrected to AVHRR	14.42	22.74	0.68

The results show relatively large RMSE values compared to the results in the experiment and this is likely because data is not clustered. The magnitude of these results relative to each other is the important part here. The SOCAT SST regression scores the lowest errors, but surprisingly, the OISST corrected pCO_2 does not perform as well as the uncorrected pCO_2.

This result complicates whether or not to implement the corrections suggested by R2. We thus suggest that we will discuss the importance of the way that temperature is handled in machine learning applications of pCO_2 – this is part of the problems (the wall) we face. However, we will also include the results from a more complete analysis of the effect of correcting pCO_2 to the temperature discrepancy in the supplementary materials.

2. *The gas flux calculation itself as used by the authors (equation 2) is likely to add temperature related errors into the analysis. This bulk formulation using DpCO2 ignores vertical temperature gradients and so is likely to introduce additional (and unknown) errors into the analysis. Woolf et al., (2016) also discuss the shortfalls of using an inaccurate gas flux calculation. The work would benefit from using a version of the equation that accounts for differing solubilities at the top and bottom of the mass boundary layer. Section 2 of Woolf et al., (2016) provides the information needed to achieve this.*

We will use the suggested flux calculation suggested by the reviewer. We will also do this for the comparison datasets (SOMFFN, MLS, MLR) as pCO_2 for each product is available and the comparison will only be fair if the same procedure is applied to calculate fluxes.

3. *Both machine learning methods and validation data suffer from the same temperature uncertainties as highlighted in points 1 and 2.*

Based on the results above, the issue of temperature corrections in SOCAT and the validation datasets are clearly an important but complex matter and will thus be discussed.

Other minor concerns by the reviewer will be addressed and corrected in the manuscript.

Response to reviewer 3 (anonymous)

R3 provided a very thorough and well thought out critique of the manuscript. The reviewer had several concerns about the manuscript, particularly with the discussion. The reviewer tentatively suggested separating some of the ideas in the discussion into a second paper. We would like to keep the manuscript as one and we hope that the suggestions that we make below will address their major concerns sufficiently.

The first point is that we perhaps claim a little too strongly that "*we have hit the wall*" and by making this claim we assert our model as an exhaustive approach without regard for methods beyond regressive approaches:

> *The paper also discusses more general aspects of pCO2 interpolation, such as the potential "wall" mentioned in the title, which is definitely an interesting and relevant question. However, I'm a bit surprised by some formulations, such as L677-678 or L578 ("stagnant"), which seem to suggest that "there must be intrinsic limits if not even our method performs better than other methods". Why should we expect your particular method to exhaust all that's achievable?*

We will address the reviewer's concern by rephrasing some of the bold statements (e.g. "stagnant") and clarifying the scope and framework of the assessment where we compare only within the statistical gap-filling framework. At the same time, we will broaden the discussion with reference to projects such as the Southern Ocean State Estimate (SOSE) which uses the assimilation of observations to correct toward the truth (Verdy and Mazloff, 2017). Further, we will make it clear that this study strengthens the case for models such as SOSE as there is seemingly (from our work) a limit to what regressive models can achieve. Where the latter point is made quite clear by the Taylor diagrams in which regression methods are tightly grouped for each of the respective validation datasets.

The second major point is that the discussion should be simplified, particularly around the subsection "scale-sensitive sampling strategies":

> *In order to make the discussion more interesting in the revised paper, therefore, I feel that it should be done in a wider context of the existing literature and make more concrete statements on how to go forward … While autonomous sampling devices are presented as "a new way", there are papers already, e.g., from the SOCCOM project, which are not even cited in the discussion. These papers already discuss possibilities and limits on a higher level.*

We agree that the discussion should be shorter. Our suggestion of "scale-sensitive sampling" is perhaps a little premature and is thus not constructive in "inching over the wall". Particularly since this has not yet been tested at even the regional scale. This will thus be removed as the primary focus of the section. Further, we recognise, as the reviewer points out, that major strides are being made by the observational community (e.g. SOCCOM) and the discussion will emphasise this more (e.g. Gray et al. 2018; Bushinsky et al. 2019). Moreover, the successful inclusion

of these data in machine learning estimates of pCO_2 should be a priority at this stage; in other words, regressive or other models that can incorporate unbalanced data (i.e. new winter data in the Southern Ocean at the end of the time-series) should be explored.

We have removed the quote from the Rödenbeck et al (2015) study from the manuscript as the point made below was also made by R1.

When comparing the presented ensemble of pCO2 interpolations with e.g. the SOCOM ensemble, however, there seem to be distinct differences in how statistically homogeneous these ensembles are: The presented ensemble of regressions against the same explanatory variables likely spreads in the finer details (see the rather similar behaviour in Fig 6) such that averaging may reduce noise, while there are large systematic differences (including members with limited ability to fit the data) in the SOCOM ensemble or other ensembles from the literature.

Removed (L544-546): "*We also discourage any ensemble averaging (or medians, etc.) of full spatiotemporal fields or time series, as this would result in variations that are not self-consistent any more and fit the data less well than individual products.*"

There were many minor corrections that the reviewer identified that will be corrected as required.

Perhaps important to mention from the reviewer's minor comments is that we described the RIAV incorrectly and the IQR^IA unclearly; this will be cleared up. The RIAV was in fact calculated in the same was as Rödenbeck et al. (2015).

Corrections to be made, but not requested by the reviewers

- Figure 10: the key for JMA-MLR and Jena-MLS were incorrectly swapped in the submitted manuscript. This will be corrected.
- Correction of SST product: we used the AVHRR product by Reynolds et al (2007) and not the OSTIA product by Donlon et al (2012). This is discussed in more detail in Reviewer 2's comments.
- Data availability: we will make the data available on OCADS (https://www.nodc.noaa.gov/ocads/oceans/) instead of figshare.