

Response to R2

We thank the reviewer for their detailed feedback and interest in our work! We hope that we have managed to correctly convey the reviewer's sentiment in the updated manuscript. Our responses are listed below in green. R2 is in black.

I thank the authors for responding in full to all of my questions and comments. However, in doing so they have introduced errors within their manuscript, and in some cases made incorrect statements and/or misunderstood concepts. These errors and misunderstandings need to be corrected before the manuscript is published. One error that the authors have made (in relation to uncertainties in the GLODAPv2 data) may change some results. So I recommend that the paper is revised before acceptance.

Errors to be corrected

- 1) The authors have attempted to address my questions about temperature handling, but they have misreported and/or misunderstood multiple aspects. These are:
 - a) **lines 159 – 164.** These statements are misleading and they confuse the issue. The re-analysis is needed to ensure that all of the data are valid for a consistent depth. Ship and in situ measurements within a large collated dataset will have been collected from multiple different depths, and these variations in depth will have an unknown impact on the calculated air-sea gas fluxes. This problem is mentioned in Goddijn-Murphy et al, (2015) and explained in detail within Woolf et al., (2016). The use of a remote sensing temperature dataset for the re-analysis means that the data all become valid for a fixed and known depth. The Reynolds OISST dataset that the authors have used is a long term consistent climate quality dataset of remotely sensed infrared data (measurements of the thermal signal from the top few mm of the water) that are calibrated to fixed depth using quality-controlled and calibrated buoy in situ data. The statements on lines 159 to 164 (and the equivalent sentences in the supplementary section, and similar sentences in the discussion – see point 1e below) need to be updated so that they correctly convey this reasoning and explanation.

We have rewritten this paragraph as the reviewer suggested. We have also removed the section in the supplementary materials that addressed this.

The re-analysis to a known and consistent depth then also makes it possible for a more accurate gas flux calculation, as you then have the ability to use or estimate skin and sub-skin temperatures. But this is a separate issue.

The product we use, OISSTv2 AVHRR only, reports bulk temperature - this puts it in an awkward place, where the corrections suggested in Woolf et al. (2016) recommend that skin or foundation temperatures are used. However, in the updated version, we aim to be consistent with the SOCOM study, so that we can compare our study with those independent studies in a meaningful manner. This means that we use the formulation for bulk fluxes where k_w is scaled

to 16 cm/hr.

The version of the Reynolds OISST dataset also needs to be stated (as there are two different methodologies, one uses passive microwave, the other does not).

We have already stated the version of the OISST product (AVHRR only). This has also been updated in Table 1.

Woolf et al., (2016) <https://agupubs.onlinelibrary.wiley.com/doi/full/10.1002/2015JC011427>

Goddijn-Murphy et al, (2015) <https://www.ocean-sci.net/11/519/2015/>

- b) **lines 165 to 167.** This is an unusual scientific justification. It reads like the authors are saying that they are choosing to exclude recent advances, as including them appears to degrade their expected result. It's also not clear if the 'reference' dataset (that they used for the evaluation in the supplementary section) was the original SOCAT data or the re-analysed SOCAT data. You would expect that assessing a re-analysed (interpolated output) result, using the original (not re-analysed) dataset as the reference would indeed give a poorer result (due to the poorer reference value).

We have removed this section from the supplementary materials.

This latter point is an example of why using RMSE, where the E is error, is potentially misleading as it's a difference, not an error – please see my point 3 below.

A better reasoning for their method choice would be that the authors are choosing a method that is consistent with those used by the SOCOM datasets, to which the authors would like to compare/contrast their work.

We have changed our reasoning as suggested by the reviewer. L379-181 now reads: *While this is an important consideration, we do not apply a temperature correction as we aim to be consistent with the $p\text{CO}_2$ estimates from the SOCOM intercomparison (Rödenbeck et al., 2015).*

- c) The methods are confusing and inconsistent with the response to reviewer comments. In the reviewer responses the authors say that they will use a more accurate calculation for the gas fluxes (as given by Woolf et al., 2016 where two solubilities are used), whereas equation 2 says that they are not using a more accurate calculation (as it contains one solubility term), but then on line 315 the authors say that they are using a more accurate calculation (they state they are using the Rapid model, but Equation 2 is not the rapid model). Please, can the authors clarify their methods and ensure that the text matches the equations and the information from the Woolf et al, (2016) reference that they refer to.

We have chosen to use the bulk flux approximation scaled to 16 cm/hr after Rödenbeck et al. (2015). Scaling the fluxes is a recognition that there is uncertainty in the bulk flux approximation. This section (L345-352) now reads: *One of the problems with the bulk estimates of sea-air CO_2 fluxes is that models of gas exchange in the surface layer of the water column are simplified, but there are approaches, such as the rapid equilibrium model, that account for more complex temperature gradients in the upper layer of the surface ocean*

(Wanninkhof et al. 2009; Woolf et al. 2016). However, for the sake of consistency with past studies, we use the bulk approximation of sea-air fluxes (Eq. 2) where k_w is scaled to 16 $\text{cm}\cdot\text{hr}^{-1}$ as in the SOCOM intercomparison (Rödenbeck et al., 2015).

- d) Why is the UEA-SI method an exception? (as stated on line 454 to 455). All of the SOCOM datasets that the authors have used the original SOCAT data and so none of these datasets have re-analysed the SOCAT data. So I don't think that there are any exceptions and they will also contain unknown impacts due to not re-analysing the SOCAT data to a consistent depth.

The text with the exception of UEA-SI has been removed

- e) **lines 655 to 664.** This paragraph is also confused and makes incorrect statements. The same mistakes as described above in point 1a) are again repeated in this paragraph. please refer to the original Woolf et al., 2016 paper where these issues are discussed in depth.

Any difference introduced due to inconsistent sampling depths will vary spatially and dependent upon the ship collecting these measurements, and so the measurements (original versus re-analysed) are unlikely to always shift in unison as suggested. Eg. there are likely to be larger differences between original and re-analysed data within an area of upwelling, in comparison to areas of well-mixed ocean waters.

The paragraph now reads (L736-740): Another source of bias is the mismatch between the temperature at which $p\text{CO}_2$ is measured (i.e. at the depth of a ship's intake) and the temperature to which $p\text{CO}_2$ is predicted (~ 1 m in the case of the dOISSTv2 data; Banzon et al. 2016; Goddijn-Murphy et al. 2015). Goddijn-Murphy et al. (2015) show that this mismatch is considerable in some cases. However, this correction also makes the assumption that temperature is the only factor that influences $p\text{CO}_2$ in the surface layer of the ocean. The correction will thus not account for other processes such as primary production, stratification and gas exchange within the surface layer. This is an issue that should be discussed by the community and tested empirically to assess the impact that processes other than temperature difference may have on $p\text{CO}_2$.

- 2) The authors have used an incorrect combined uncertainty value for the GLODAPv2 dataset. The value taken from the Olsen et al., (2016) paper that the authors have used as a combined uncertainty is actually the bias value, which is only one part of a Type A uncertainty (BIPM, 2008). Both a bias and variance estimate are needed for a Type A uncertainty estimate. The Olsen et al. work does not provide a variance estimate, so I would suggest that the authors use the state of the art uncertainties as provided by Bockmon and Dickson, (2015). Bockmon and Dickson give a combined uncertainty of 0.5% for total alkalinity and for total dissolved carbon (whereas the existing bias estimate that the authors are using equates to $\sim 0.2\%$ for alkalinity, which illustrates that the uncertainty estimate used by the authors is too low). Using a more correct estimate of the combined uncertainty may mean that some results throughout the paper need updating.

We thank the reviewer for highlighting this. The following sentence has been added on line 323:

However, this error may be larger as reported in Table 2, where Bockmon and Dickson (2015) showed

that the uncertainty for DIC and TA is likely closer to $\pm 10 \mu\text{mol.kg}^{-1}$. This error is consistent in the comparisons with the SOCOM products, thus is inconsequential in the context of this study.

- a) BIPM, 2008 https://www.bipm.org/utis/common/documents/jcgm/JCGM_100_2008_E.pdf
 - b) Bockmon and Dickson, (2015)
<https://www.sciencedirect.com/science/article/pii/S0304420315000213>
 - c) Olsen et al., 2016 <https://www.earth-syst-sci-data.net/8/297/2016/>
- 3) Throughout the manuscript, the authors have incorrectly used the word ‘error’ when they actually mean uncertainty or difference (e.g. when using the root mean squared error, RMSE). The use of the word error implies that some ‘truth’ value is known, whereas all measurements and observations contain uncertainties and are not ‘truth’ (including those within GLODAPv2 and SOCAT that are being used as a reference for the RMSE calculations). I would suggest that the authors instead report RMSD (where the D is difference) and check and correct the wording used within sentences throughout the paper (and the supplementary) to refer to differences, uncertainties and errors (as many instances where the word ‘error’ has been used are not strictly ‘errors’).

I have tried to find a reference that substantiates the reviewers terminology, but I could not find any. RMSE is also a widely used term in machine learning literature. We have thus chosen to stick with the RMSE terminology. We have also clarified our definition of “error” with: “Note that the term “error” refers here to the error introduced by the gap-filling method relative to the observations.”

Typographical errors

There typographical errors in the main text. (e.g. ‘sea-are’ on line 311) and in the reference list (e.g. the Woolf et al., 2015 paper in the reference list on page 35 has incorrect lead author, incorrect co-author list and incorrect publication year). There are also sentences that are incomplete and/or contain errors (e.g. line 123 in the supplementary ‘...showing that SOCAT temperatures or on average... [whereas it should be ‘temperatures are on average’]).

I would encourage the authors to take some time and slowly go through the manuscript and check their sentences, phrasing, spelling and references.

We apologise for the multiple typographical errors and the incorrect listing of the reference – this has been removed as fluxes are calculated using bulk estimates in the final form. These corrections and others have been made throughout the document.

A comparative assessment of the uncertainties of global surface-ocean CO₂ estimates using a machine learning ensemble (CSIR-ML6 version 2019a) – have we hit the wall?

Luke Gregor^{1,2,3}, Alice D. Lebehoh^{1,2}, Schalk Kok⁴, Pedro M. Scheel Monteiro¹

¹SOCCO, Council for Scientific and Industrial Research, Cape Town, 7700, South Africa

²MaRe, Marine Research Institute, University of Cape Town, Cape Town, 7700, South Africa

³Environmental Physics, Institute of Biogeochemistry and Pollutant Dynamics, ETH Zürich, 8092 Zürich, Switzerland

⁴Department of Mechanical & Aeronautical Engineering, University of Pretoria, Pretoria, 0028, South Africa

Correspondence to: Luke Gregor (luke.gregor@usys.ethz.ch)

Abstract. Over the last decade, advanced statistical inference and machine learning have been used to fill the gaps in sparse surface ocean CO₂ measurements (Rödenbeck et al. 2015). The estimates from these methods have been used to constrain seasonal, interannual and decadal variability in sea-air CO₂ fluxes and the drivers of these changes (Landschützer et al. 2015, 2016, Gregor et al. 2018). However, it is also becoming clear that these methods are converging towards a common bias and RMSE boundary: *the wall*, which suggests that *p*CO₂ estimates are now limited by both data gaps and scale-sensitive observations. Here, we analyse this problem by introducing a new gap-filling method, an ensemble average of six machine learning models (CSIR-ML6 version 2019a), where each model is constructed with a two-step clustering-regression approach. The ensemble average is then statistically compared to well-established methods. The ensemble average, CSIR-ML6, has an RMSE of 17.16 µatm and bias of 0.89 µatm when compared to a test-dataset kept separate from training procedures. However, when validating our estimates with independent datasets, we find that our method improves only incrementally on other gap-filling methods. We investigate the differences between the methods to understand the extent of the limitations of gap-filling estimates of *p*CO₂. We show that disagreement between methods in the South Atlantic, southeastern Pacific and parts of the Southern Ocean are too large to interpret the interannual variability with confidence. We conclude that improvements in surface ocean *p*CO₂ estimates will likely be incremental with the optimisation of gap-filling methods by (1) the inclusion of additional clustering and regression variables (*e.g.* eddy kinetic energy), (2) increasing the sampling resolution, (3) successfully incorporating *p*CO₂ estimates from alternate platforms (*e.g.* floats, gliders) into existing machine learning approaches.

1 Introduction

The ocean plays a crucial role in mitigating against climate change by taking up about a third of anthropogenic carbon dioxide (CO₂) emissions (Sabine et al. 2004; Khatiwala et al., 2013; McKinley et al. 2016). While the mean state in the global contemporary marine CO₂ uptake is a widely-used benchmark (Le Quéré et al., 2018),
40 underlying assumptions and limited confidence regarding the variability and long-term evolution of this sink persist. Sparse observations of surface ocean CO₂ during winter and in large inaccessible regions has been the biggest barrier in constraining the seasonal and interannual variability of global contemporary sea-air exchange (Monteiro et al. 2010; Rödenbeck et al. 2015; Bakker et al. 2016; Ritter et al. 2017). The increasing ship-based
45 sampling effort and the ongoing development of autonomous observational platforms (e.g. biogeochemical Argo floats and Wave Gliders) have improved confidence of interannual estimates of ocean CO₂ uptake in more recent years (Monteiro et al. 2015; Bakker et al. 2016; Gray et al., 2018).

The community has turned to models and data-based approaches to improve estimates of CO₂ uptake by the oceans for periods and regions with poor or no observational coverage (Wanninkhof et al. 2013a; Rödenbeck et al. 2015; Verdy and Mazloff, 2017). Ocean biogeochemical models are able to capture the general global trend
50 in increasing oceanic CO₂ uptake shown by observations but suffer from significant regional and interannual (~1 PgC yr⁻¹) differences in their estimates because these models cannot yet accurately parameterise the marine carbonate system at computationally feasible resolutions (Wanninkhof et al. 2013a). In recent years, data-based approaches, e.g. statistical interpolations and regression methods, have become a popular alternative to
55 biogeochemical models (Lefèvre et al. 2005; Telszewski et al. 2009; Landschützer et al. 2014; Rödenbeck et al. 2014; Jones et al. 2015; Iida et al. 2015). The regression methods try to maximise the utility of existing ship-based observations by extrapolating CO₂ using proxy variables (observable from space or interpolated). Extrapolating with proxy variables is possible due to the non-linear relationship between the partial pressure of CO₂ (*p*CO₂) in the surface ocean and proxies that may drive changes in surface ocean *p*CO₂. Improved access to
60 quality-controlled ship-based measurements of surface ocean CO₂ through the Surface Ocean CO₂ Atlas (SOCAT) database, and satellite and reanalysis products as proxy variables have aided the development of the data-based methods (Rödenbeck et al. 2015; Bakker et al. 2016).

The current state of machine learning in ocean CO₂ estimates

With the increase in the number of statistical estimates of surface-ocean CO₂, the Surface Ocean CO₂ Mapping (SOCOM) community collated fourteen of these methods in an intercomparison of “gap-filling” methods
65 (Rödenbeck et al. 2015). The intercomparison gives an overview of the SOCOM landscape, with regression and statistical interpolation approaches making up eight and four of the fourteen methods respectively (Rödenbeck et al. 2015). Two model-based approaches were also compared.

70 While SOCOM intercomparison did not seek to identify an optimal mapping method, it assessed members according to how well they represented interannual variability (IAV) relative to climatological surface ocean

$p\text{CO}_2$ increasing at the rate of atmospheric CO_2 concentrations (R^{iaV}). Two methods, the Jena-MLS (Mixed-Layer Scheme) and MPI-SOMFFN (Self-Organising Map Feed-Forward Neural-Network), achieved lower R^{iaV} scores compared to other members of the comparison. The MPI-SOMFFN is a global implementation of a two-step clustering-regression approach and has been widely adopted in the literature (Landschützer et al. 2015, 2016, 2018, Ritter et al. 2017). The elegance of the clustering-regression approach, particularly the clustering step, is that it reduces the problem into smaller parts with more coherent variability and reduces the computational size of the problem per cluster – a beneficial attribute when using regression methods that do not scale well to big datasets.

The SOCOM intercomparison found that the gap-filling methods were in agreement in regions with a large number of seasonally-resolving persistent measurements, but the different methods did not agree in regions where data were sparse (e.g. the Southern Ocean). Similarly, Ritter et al. (2017) found little agreement in the Southern Ocean on seasonal timescales, yet on decadal time-scales, there was agreement on the direction of trends between gap-filling methods.

1.2 Measuring the uncertainty of estimates?

~~The biggest limitation in assessing gap-filling methods is the paucity of data in the Southern Hemisphere (Rödenbeck et al. 2015; Bakker et al. 2016).~~ The assessment of gap-filling methods is largely limited by the distribution of the observational coverage, which is particularly true for the Southern Hemisphere where data is sparse (Rödenbeck et al. 2015; Bakker et al. 2016). The standard use of root-mean-squared error (RMSE) and bias as measures of uncertainty give larger weighting to the observation-heavy regions or periods compared with data-sparse regions and periods, potentially leading to underestimates of uncertainty (Lebehot et al. 2019). Note that the term “error” refers here to the error introduced by the gap-filling method relative to the observations. The R^{iaV} score improves on the standard implementation of RMSE and bias by weighting the uncertainties annually, thus giving a less temporally biased estimate of uncertainty. ~~However, the method is still limited to the regions where there are observations of $p\text{CO}_2$.~~

Previous studies have compared their methods’ estimates to independent datasets, where measurements of $p\text{CO}_2$ are not included in the SOCAT datasets (Landschützer et al. 2013, 2014; Jones et al. 2015; Denvil-Sommer et al. 2018). These data serve as good validation data, particularly with the inclusion of derivations of $p\text{CO}_2$ from autonomous platforms in the Southern Ocean, a historically undersampled area especially during winter (Boutin and Merlivat, 2013; Gray et al. 2018).

One of the concluding statements in the SOCOM intercomparison is that pseudo- or synthetic data (deterministic model output) experiments should be used to test and compare methods. Gregor et al. (2017) did just this, but their study was limited to the Southern Ocean, and the synthetic data did not fully capture the variability represented by observations, in part due to coarse synthetic data resolution (5-daily mean and $\frac{1}{2}^\circ$ spatially). The authors found that the ensemble average performed slightly better than ensemble members, in

agreement with ensemble averaging approaches previously used in ocean CO₂ studies (Khatiwala et al. 2013).

110 ~~On the other hand~~ Similarly, Lebehot et al. (2019) investigated the performance of an interpolation method in the North Atlantic using an ensemble of model outputs. Their approach offered a unique way of assessing a gap-filling method at places and times where no observations were made.

1.3 Aims

115 The main aim of this study is to present and evaluate a new machine learning approach to estimate surface ocean pCO₂. We propose the use of an ensemble average, where we hypothesise that the “whole is greater than the sum of its parts” as the strengths of the ensemble members are often complementary in such a way to overcome the weaknesses (Khatiwala et al. 2013; Gregor et al. 2017). Further, we aim to evaluate the method for a selection of existing gap-filling methods. From this comparison we aim not only to gain a sense of our method’s performance but also the state of gap-filling based estimates; i.e. where would we be able to improve in future work?

2 Methods

125 There are two ~~main~~major components to this study: surface pCO₂ mapping with multiple methods, and robust error estimation from SOCAT v5 gridded product and independent data sources. This study takes a similar two-step approach used in the JMA-MLR and MPI-SOMFFN approaches, where data is grouped or clustered first, and then a regression algorithm is applied separately to each group or cluster. We use the ocean CO₂ biomes by Fay and McKinley (2014) as an option for grouping. Alongside this grouping, we use an optimal K-means clustering configuration. Next, four non-linear regression methods are applied to each of the groupings. The regression methods are Support Vector Regression (SVR), Feed-Forward Neural Network (FFN), Extremely Randomised Trees (ERT) and Gradient Boosting Machine (GBM). The latter two approaches are new to the application. These methods are then compared to independent data sources. This is outlined in more detail in the Experimental Overview below.

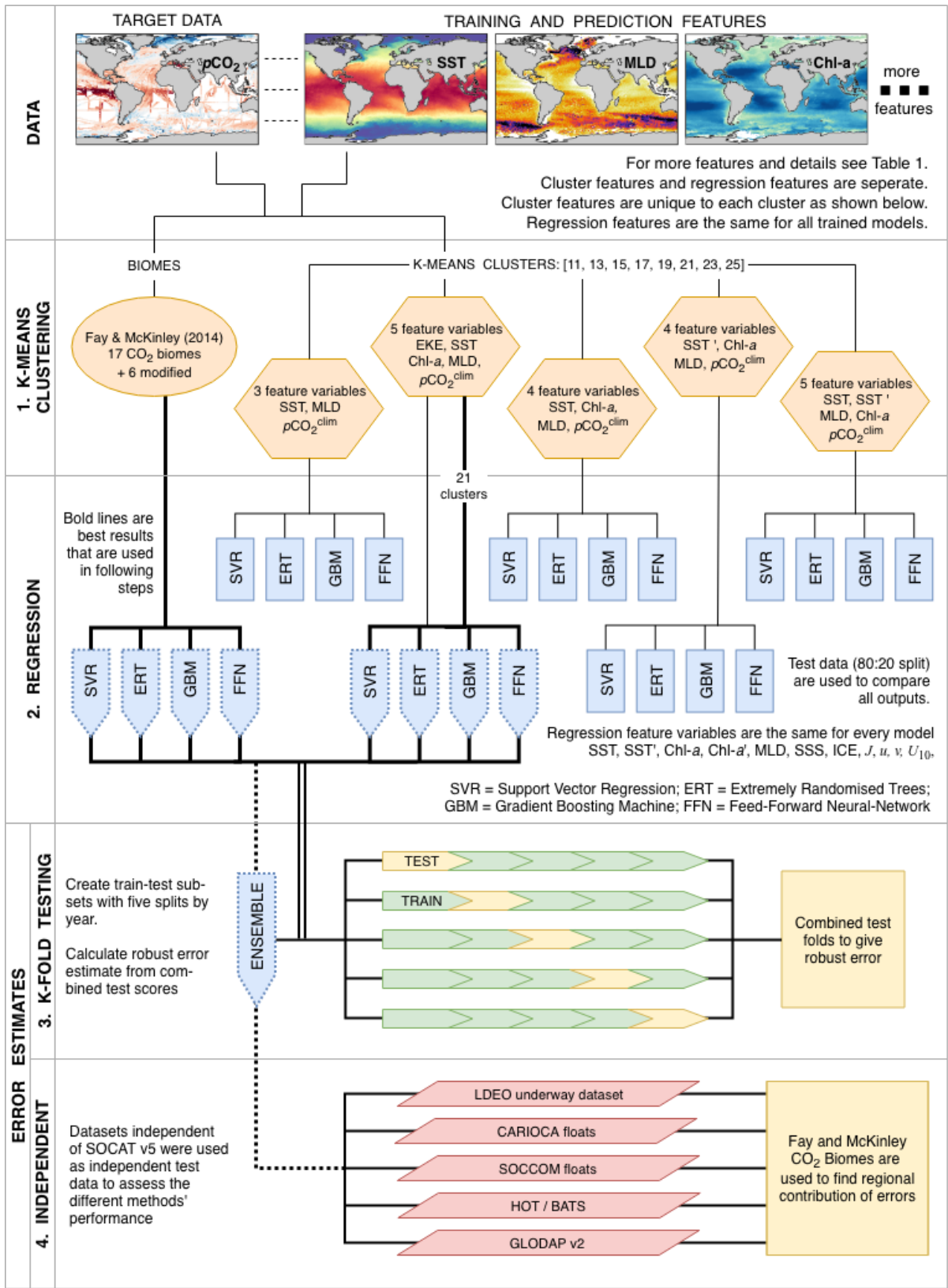
2.1 Experimental Overview

The experimental design, outlined below, is summarised in Figure 1:

- 135 1. In the first step (denoted as “K-means clustering” in Figure 1), we generate climatological biomes using the oceanic CO₂ biomes by Fay and McKinley (2014), and a selection of features variables (five combinations) and number of clusters (a range of 11 to 25 clusters, stepping by two) resulting in a total of 41 clustering configurations.
- 140 2. Four regression algorithms are applied to each clustering configuration, resulting in 164 models (described by the “Regression” section in Figure 1). The test data (isolated from the model training procedure) is used to identify the best performing clustering configuration with annually weighted bias, ~~root-mean-squared error (RMSE)~~ and R^{ia}. The four regression models for CO₂ biomes and the four models from the best performing clustering configuration (as indicated by the bold lines in Figure 1)

are used in the steps that follow. The selected eight models are averaged to create an ensemble average that is included with the eight members for further evaluation.

- 145
3. The third step (as represented by the “K-fold testing” section in Figure 1 and Section 2.5) provides a robust uncertainty evaluation based on the training data (SOCAT v5). An iterative test-train approach is applied to estimate the bias, RMSE and R^{lav} for the complete SOCAT v5 dataset (rather than just one test split).
 4. The fourth step compares the ensemble average estimates of surface ocean pCO_2 with independent test data (that is not in SOCATv5, as represented by the “Independent” section in Figure 1), which allows testing the predictive ability of the ensemble method (Section 2.6). Four methods from the SOCOM gap-filling intercomparison study are included for reference.
 5. Lastly, all gap-filling methods are compared to identify regions where there is a divergence in the trend and seasonal cycle.
- 150



155 **Figure 1:** A flow diagram that shows the experimental procedure used in this study. Abbreviations for feature-variables in the orange hexagons can be found in Table 1. All other abbreviations are given in the diagram. Details of each step are given in the text (Section 2.1).

2.2 Data: clustering, training and prediction

160 Standard machine learning implementation requires a training- and a predictive dataset. The training dataset consists of a target variable that is being predicted (in this case $p\text{CO}_2$) and one or more feature-variables that have samples that correspond with target samples (e.g. SST, Chl- a , MLD co-located in space and time), where feature-variables may directly or indirectly influence the target variable. Features variables are used to predict once a machine learning model has been trained and must thus be available for the full prediction domain.

165

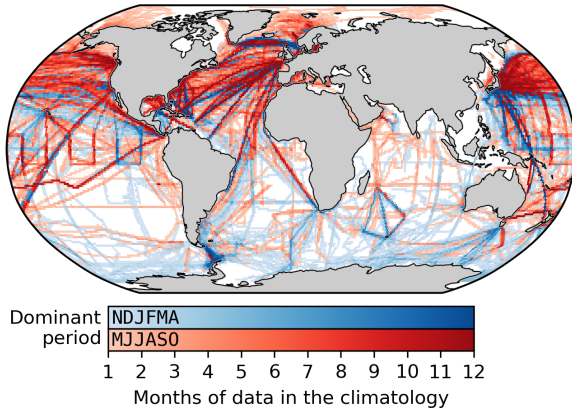


Figure 2: Map showing the distribution of the SOCAT v5 monthly gridded product (1982 to 2016) as a monthly climatology to show how well the seasonal cycle is represented (regardless of the year). The red shading shows grid-points where the majority of data occur from May to October and the blue shading shows grid-points where the majority of data occur from November to April.

170

Here we use surface ocean $p\text{CO}_2$ calculated from the SOCAT v5 monthly gridded $f\text{CO}_2$ (fugacity of CO_2) product (hereinafter SOCAT v5 as shown in Figure 2) as the target variable (Sabine et al. 2013; Bakker et al. 2016). SOCAT v5 is a quality controlled dataset that contains observations of surface ocean $f\text{CO}_2$, which is converted to $p\text{CO}_2$ with:

175

$$p\text{CO}_2 = f\text{CO}_2 \cdot \exp\left(P_{\text{atm}}^{\text{surf}} \cdot \frac{B + 2\delta}{R \cdot T}\right)^{-1} \quad \text{Eq. 1}$$

where $P_{\text{atm}}^{\text{surf}}$ is the atmospheric pressure at the surface of the ocean, T is the sea surface temperature (SST) in $^{\circ}\text{K}$, B and δ are virial coefficients, and R is the gas constant (Dickson et al. 2007). We used ERA-interim $P_{\text{atm}}^{\text{surf}}$ (Dee et al., 2011) and NOAA daily optimally interpolated SST version 2 (dOISSTv2) that uses only Advanced Very-High-Resolution Radiometer data (AVHRR; Reynolds et al. 2007; Banzon et al. 2016).

180

An important consideration in the use of the SOCAT database is that in-situ measurements (i.e. ship measurements) are not collected at the surface. The in-situ temperatures that coincide with $p\text{CO}_2$ in the SOCAT database are thus different from surface temperature product used to estimate $p\text{CO}_2$ and calculate fluxes (Goddijn-Murphy et al. 2015; Bakker et al., 2016). The discrepancy in in-situ and remotely sensed temperature results in a theoretical difference between $p\text{CO}_2$ measured at the ship intake depth and the surface due to warming or cooling (Takahashi et al., 1993). Goddijn-Murphy et al. (2015) suggest that a correction for the

185

theoretical difference in $p\text{CO}_2$ should be made using the empirical relationship between $p\text{CO}_2$ and temperature (Takahashi et al. 1993). While this merits further coordinated consideration by the marine CO_2 observations community, we do not apply such a temperature correction in this study as we aim to be consistent with the earlier $p\text{CO}_2$ estimates from the SOCOM intercomparison (Rödenbeck et al., 2015). However, we do present the potential impact of this discrepancy in Section S2.4. ~~in this application of SOCAT data is the mismatch between the temperature at which $p\text{CO}_2$ is measured ($\text{SST}^{\text{SOCAT}}$) and the OISST (Goddijn-Murphy et al. 2015). dOISST product reports the bulk temperature at -1 m, whereas $\text{SST}^{\text{SOCAT}}$ is measured at the depth of a ship's water intake (Banzon et al., 2016; Bakker et al. 2016). A comparison of $\text{SST}^{\text{SOCAT}}$ and OISST shows that the former has a global mean warm bias of 0.13°C (Figure S3a), the same as that reported by Banzon et al. (2016), which they attribute to warming in the ships intake. Further investigation shows that correcting $p\text{CO}_2$ for the temperature bias reduces the accuracy of the machine learning estimates relative to the training data (Section S2.4) and does not improve spatial biases. We thus do not apply the correction applied in Goddijn-Murphy et al. (2015).~~

Feature-variables in both the training and predictive datasets are globally gridded products, including satellite observations, *in-situ* measurements and reanalysis products (Table 1, see Section S1 for details). All feature-variables are gridded to a monthly frequency onto a global $1^\circ \times 1^\circ$ resolution grid. Thereafter, data processing steps are applied as shown in Table 1 and described in detail in Supplementary Materials (Section S1) with the final output being a complete dataset ranging from 1982 to 2016. Note that the clustering and regression steps use different subsets of the feature-variables as indicated in Table 1.

Table 1: Summary of the products, variables and data processing steps used for feature-variables. The column “Usage” indicates the features that are used for the clustering step (identified by C) and for the regression step (identified by R). Abbreviations are used in Figure 1 and throughout the text. Basic data processing is described in the text with details in the Supplementary Materials (Section S1).

Group: Product	Variable	Abbrev	Usage	Processing	Reference
NOAA: dOISSTv2 (AVHRR only)	Sea surface temperature	SST	C R	-	Reynolds et al. (2007) Banzon et al. (2016)
	SST seasonal anom.	SST'	C R	$\text{SST} - \text{annual average}$	
	Sea ice fraction	ICE	R	-	
MetOffice: EN4	Salinity	SSS	R	-	Good et al. (2013)
CDIAC: ObsPack v3	Atmospheric $p\text{CO}_2$	$p\text{CO}_2^{\text{atm}}$	R	$x\text{CO}_2^{\text{atm}} \times \text{sea level pressure}$	Masarie et al. (2014)
UCSD: Argo Mixed Layers	Mixed Layer Depth	MLD	C R	$\log_{10}(\text{climatology})$	Holte et al. (2017)
ESA: Globcolour	Chlorophyll- <i>a</i>	Chl- <i>a</i>	C R	$\log_{10}(\text{climatology filled}_{1982-1997}^{\text{cloud gaps}})$	Maritorena et al. (2010)
	Chl- <i>a</i> seasonal anom.	Chl- <i>a</i> '	R	$\text{Chl-}a - \text{annual average}$	
ECMWF: ERA-Interim 2	<i>u</i> -wind	<i>u</i>	R	-	Dee et al. (2011)
	<i>v</i> -wind	<i>v</i>	R	-	
	Wind speed	U_{10}	R	$\sqrt{u^2 + v^2}$	
ESA: Globcurrent	Eddy kinetic energy	EKE^{clim}	C	$\log_{10}(\frac{1}{2} \cdot (u'^2 + v'^2))$	Rio et al. (2014)
-	Day of the year	<i>J</i>	R	$\sin(\frac{j}{365}), \cos(\frac{j}{365})$	-
LDEO: $p\text{CO}_2$ climatology	Surface ocean $p\text{CO}_2$	$p\text{CO}_2^{\text{clim}}$	C	Data smoothing	Takahashi et al. (2009)

In this paragraph, we briefly describe the data processing steps shown in Table 1 - detailed product descriptions and in-depth processing steps are in Section S1. We derive an additional SST feature, SST', by subtracting the annual mean of SST from each respective year, leaving the annual mean anomalies (Reynolds et al. 2007; Banzon et al. 2016). We use the \log_{10} transformation of the Globcolour Chl-*a* global product (Maritorena et al. 2010). Cloud gaps and the period before the start of the product (1982 to 1997) are filled with the climatology (1998 – 2016), and high-latitude winter regions (where there is no climatology for Chl-*a*) is filled with low concentration random noise to be consistent with regions of low concentration Chl-*a* (Gregor et al. 2017). We derive an additional Chl-*a* feature, Chl-*a*' using the same procedure as described for the SST annual mean anomalies. We use a \log_{10} transformation of mixed layer depth (MLD) from Argo float density profiles (Holte et al. 2017) to create a monthly climatology, thus imposing the assumption that there is no interannual variability. Wind speed is calculated from 6-hourly data using the equation in Table 1 before taking the monthly average. Atmospheric $p\text{CO}_2$ is calculated with: $p\text{CO}_2 = x\text{CO}_2^{atm} \times P^{atm}$, where $x\text{CO}_2^{atm}$ is the mole fraction of atmospheric CO_2 (from ObsPack v3 by Masarie et al. 2014) and P^{atm} is the reanalysed mean sea-level pressure (from ERA-interim 2; Dee et al. 2011) – further details for the procedure are in Section S1 of the Supplementary Materials. The climatology of ~~e~~Eddy ~~k~~Kinetic ~~e~~Energy (EKE^{clim}) is calculated from u and v surface current components (integrated for depth < 15 m) from the Globcurrent product (Rio et al., 2014), where u' is calculated as $\bar{u} - u$ and similarly with v (Table 1).

230

2.3 Clustering and biomes

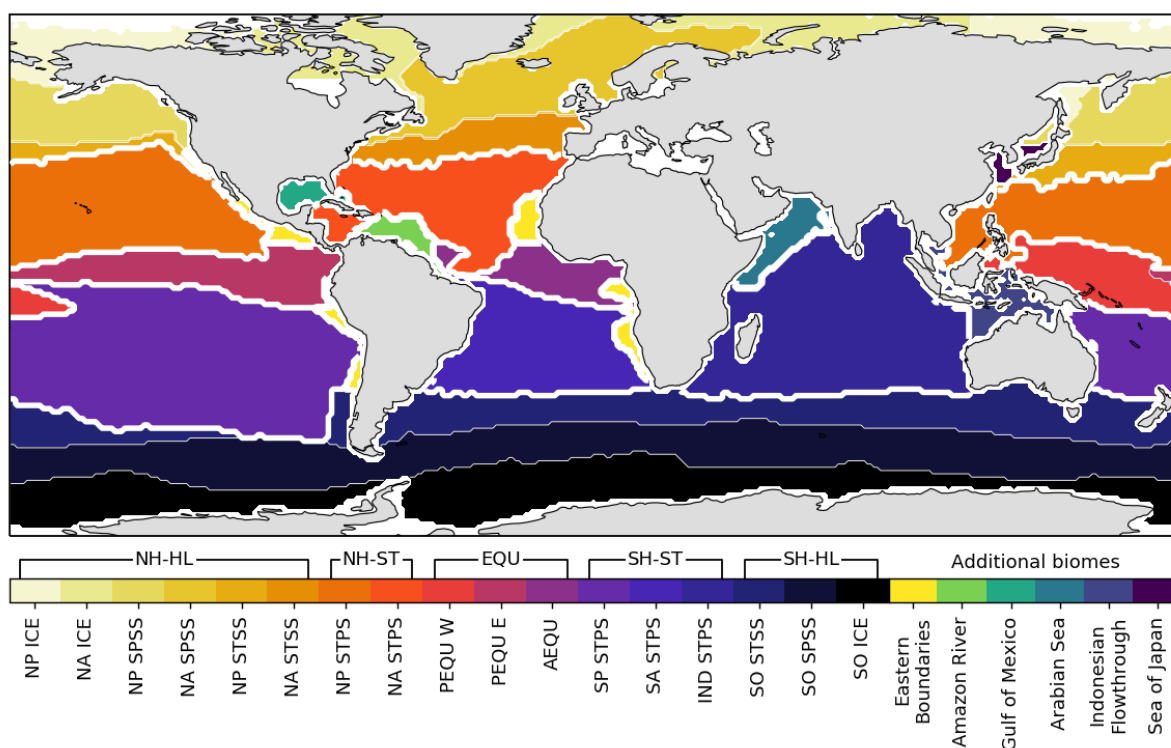
The seasonal and interannual variability of global surface ocean $p\text{CO}_2$ is complex due to interactions of various driver variables acting on the surface ocean at different space and time scales (Lenton et al. 2012; Landschützer et al. 2015; Gregor et al. 2018). Machine learning algorithms applied globally struggle to represent the $p\text{CO}_2$ accurately unless spatial coordinates are included as feature-variables (Gregor et al. 2017). This is due to the fact that $p\text{CO}_2$ may respond inconsistently to observable feature-variables in different regions as it is not possible to observe all feature-variables that drive $p\text{CO}_2$. A common practice to avoid the inclusion of coordinates is to separate the ocean into regions where processes that drive $p\text{CO}_2$ are coherent and then apply individual regressions to each region – five of the eight regression methods in Rödenbeck et al. (2015) apply this approach. We adopt two such approaches to develop regions of internal coherence in respect of CO_2 variability, namely regions defined by biogeochemical properties and clusters defined by a clustering algorithm.

240

Our first “clustering” approach uses the oceanic CO_2 biomes by Fay and McKinley (2014) that divide the ocean into 17 biomes. Fay and McKinley (2014) define their biomes by establishing thresholds for SST, Chl-*a*, sea-ice extent and maximum MLD ~~depth~~. Unclassified regions from the original biomes are manually assigned based on their geographical extent resulting in six additional regions (Figure 3). We maintain these as separate regions from the original Fay and McKinley (2014) **biomes**. Their study originally did not classify these regions in the core biomes because the physical and biogeochemical properties were not accounted for by the set thresholds from their study. This would suggest that drivers of CO_2 in these regions could be quite different from the

245

250 adjacent open ocean biomes. Note that we may refer to the modified Fay and McKinley (2014) ocean CO₂ biomes as “CO₂ biomes” or as “BIO23” from here on (Figure 3). For later analyses, we group certain biomes together as shown by the brackets above the colour-bar in Figure (3).



255 **Figure 3:** Regions or biomes as defined by Fay and McKinley (2014). Unclassified regions from the original data have been assigned manually in this study and are shown by the separate colours. This modified configuration of the CO₂ biomes is referred to as BIO23 in this study. The sea-mask used in Lanschützer et al. (2014) has been applied. For the biome abbreviations (below the colour-bar) see Fay and McKinley (2014). The abbreviations above the colour-bar are used in this study, where selected biomes are grouped together. Thick white lines show the boundaries of the grouped regions. Prefixes are: NH = Northern Hemisphere, SH=Southern hemisphere; suffixes are HL = high latitudes, ST = subtropics, and EQU = equatorial.

260

We also use K-means clustering, which groups data based on Euclidean distances. More specifically, we implement mini-batch K-means from Python's Scikit-Learn package (Sculley 2010; Pedregosa et al. 2012), which is described in the [Supplementary Materials](#) (Section S2.2; Figure S2). We apply clustering with various feature combinations and the number of clusters (shown by orange hexagons in Figure 1). We tested a range of 11 to 25 clusters (stepping by two). The performance of each clustering configuration is not tested with a clustering metric; instead, we test the performance based on the test scores of the regressions in the next step as a more complete indicator of performance. We find optimal results in respect of RMSE and biases with 21 and 23 clusters. We selected 21 clusters (Figure S2). Each method of defining regional coherence in respect of $p\text{CO}_2$ variability has its methodological weaknesses so in this study we adopted the approach of incorporating both K-means and CO₂ biomes into the ensemble average (Figure 1). Although this likely weakens the geophysical meaning of the ensembled domains we show that it strengthens the overall performance of the ensemble average.

265

270

2.4 Regression

275 Here we describe the underlying machine learning principles of regression. The co-located data (*i.e.* SOCAT v5) are split into training and test-subsets with a roughly 80:20 split. The test-subset is isolated from the training

process to attain a reliable estimate of uncertainty. We make the split between training and test-subsets based on a random subset of years in the time series (1982 to 2016): 1984, 1990, 1995, 2000, 2005, 2010 and 2014. We avoid using a shuffled train–test split (completely random) as this leads to artificially low uncertainties in machine learning algorithms that are prone to overfitting (see the experiment in S2.1), where the models can reproduce the shuffled test data better as these data are adjacent to samples of the same ship track.

We further reduce the possibility of overfitting by tuning the hyper-parameters for each model to be more generalised, *i.e.* able to fit the data that the model has not been exposed to. The search for the optimal hyper-parameters is achieved with grid-search cross-validation, where a portion of the training subset is iteratively kept separate from the training process for a certain set of hyper-parameters (Hastie et al. 2009). The hyper-parameters that result in the best score from the grid-search are used for the fit with the full training subset (see S2.3 for more details). We use a variation of K-fold cross-validation called *group K-fold* in Scikit-Learn (Pedregosa et al. 2012). Rather than having arbitrary splits for each fold, a given grouping variable is used to split the data – in this case, years. Using years as the grouping variable reduces bias towards the second half of the time series where data is less sparse.

The train-test split and cross-validation are applied identically to each of the four machine learning algorithms for each clustering configuration. We use the following machine learning algorithms: Extremely Randomised Trees (ERT – Geurts 2006); Gradient Boosting Machines (GBM – Friedman 2001); Support Vector Regression (SVR – Drucker et al. 1997); and Feed-Forward Neural Networks (FFN). The details of these methods and how they were tuned are explained in the supplementary materials (Section S2.3). The first two methods, ERT and GBM, are new to this application. SVR has been implemented as a single global domain by Zeng et al. (2017), and FFN is used by several different methods, some of which are in the SOCOM intercomparison (Landschützer et al. 2014; Zeng et al. 2014; Sasse et al. 2013).

Regression performance is tested using RMSE primarily but also bias (Equations 3 and 4 below) and R^{iaV} (Equation 5) with only the models from the best averaged clustering configuration used for the rest of the study.

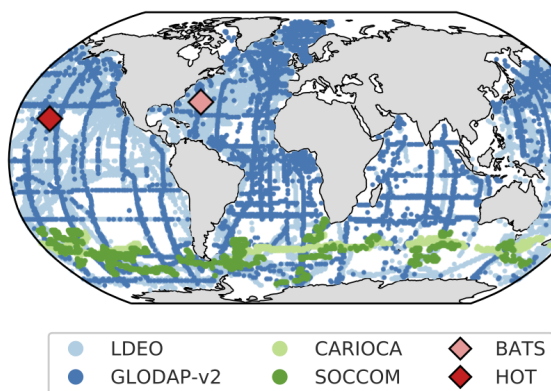
2.5 Robust biases and root-mean-square errors

Standard practice in machine learning is to set aside a test-subset of the data as described in Section 2.4. We use this standard approach in the second step of our experiment (regression comparison) as an estimate of the performance for each of the machine learning models (164 in total). However, this grouped train-test split gives a bias and RMSE estimate limited to the random test years of test-subset (see Section 2.4). To overcome this limitation, we iteratively apply the train-test split method with multiple selections of years. The splits in the test fold are based on a subset of years spaced five years apart. We then refactor the five test-fold estimates into a complete test-estimate (with the same structure as the original SOCAT v5), thus giving a complete estimate of bias and RMSE (Figure 1 step 3). This robust test-estimate method ensures that correct biases and RMSE scores are reported even if methods are prone to overfitting (see Section S2.1 and Figure S1). We limit this procedure

315 to only the CO₂ biome and best clustered regressions as it has five times the computational cost of a single train-test split.

2.6 Method validation data

320 For method validation we use observation data that are not used in SOCAT (Figure 4 and Table 2) as they are either: 1) included in the Lamont-Doherty Earth Observatory (LDEO) database, but not in SOCAT; 2) not measured with an infrared analyser; 3) derived from two other variables in the marine carbonate system, where these include dissolved inorganic carbon (DIC), pH and total alkalinity (TA) – where the Southern Ocean Carbon and Climate Observation and Modeling (SOCCOM) floats use empirically calculated TA.



325 **Figure 4:** The distribution of the validation data. Details of these datasets are given in Table 2. The Hawaii Ocean Time-series (HOT) and the Bermuda Atlantic Time-series (BATS) are marked as diamonds to distinguish them as time series stations.

330 **Table 2:** Details for the validation datasets. The measured variables are shown (DIC = dissolved inorganic carbon; TA = total alkalinity) along with the estimated accuracy of $p\text{CO}_2$. This includes the propagated uncertainty in the conversion from DIC and TA to $p\text{CO}_2$ as defined by Lueker et al. (2000), where the estimates marked with * are an extrapolation of the estimates as the DIC and TA uncertainties do not match or exceed those listed in the publication. Note that the error estimates for GLODAP v2 are larger than shown in the table as measurement uncertainty is defined as $\pm 10 \mu\text{mol.kg}^{-1}$ in Bockmon and Dickson (2015). Grid points show the number of data at the same resolution as the feature-variables.

Platform	Project	Measured variable	Accuracy (μatm)	Reference	Grid points
Ship	LDEO	$p\text{CO}_2$ Equilibrator	$\pm 2.5 \mu\text{atm}$	Takahashi et al. (2016)	16161
	GLODAP v2	DIC + TA	$>= 12 \mu\text{atm @ } 400 \mu\text{atm}^*$	Olsen et al. (2016); Bockmon and Dickson (2015)	5976
Surface floats	CARIOCA	$p\text{CO}_2$ Colourimetric	$\pm 3.0 \mu\text{atm}$	Boutin and Merlivat (2013)	613
Profiling floats	SOCCOM	pH + TA (LIAR)	$\sim 11 \mu\text{atm @ } 400 \mu\text{atm}$	Carter et al. (2016)	1037
Mooring	BATS	DIC + TA	$\sim 4 \mu\text{atm @ } 400 \mu\text{atm}$	Bates (2007)	246
	HOT	DIC + TA	$< 7.6 \mu\text{atm @ } 400 \mu\text{atm}^*$	Dore et al. (2009)	214

335 The uncertainty of $p\text{CO}_2$ that is calculated from DIC and TA is dependent on the accuracy of these two measurements, as well as the derivation of $p\text{CO}_2$ with dissociation constants, for which we use the CBSys package in Python (Hain et al. 2015). CBSys implements the constants from Lueker et al. (2000) that reports an uncertainty of 1.9% standard deviation of the calculated $p\text{CO}_2$ where DIC and TA uncertainties are $2.0 \mu\text{mol.kg}^{-1}$ and $4.0 \mu\text{mol.kg}^{-1}$ respectively. The measurements in GLODAP v2 are slightly larger than this at 4

and $6 \mu\text{mol.kg}^{-1}$, which would result in an error larger than 1.9% – this is $12 \mu\text{atm}$ for a $400 \mu\text{atm}$ estimate at a hypothetical 3% error. However, this error may be larger as reported in Table 2, where Bockmon and Dickson (2015) showed that the uncertainty for DIC and TA is likely closer to $\pm 10 \mu\text{mol.kg}^{-1}$. While this potentially large error range may seem concerning, we argue that the inclusion of these data in data-sparse regions is more valuable than their omission. Additionally, GLODAP v2 data has been adjusted on a per-profile basis to minimise the biases through the comparison of deep slow-changing ocean properties (Olsen et al. 2016).
 Moreover, the errors from the previous gap-filling products are on the order of $20 \mu\text{atm}$, below the potential uncertainty from the DIC/TA conversion to $p\text{CO}_2$ (Landschützer et al. 2014, Rödenbeck et al. 2014). Williams et al. (2017) estimated the error for $p\text{CO}_2$ calculated empirically to be 2.7%, where TA was calculated empirically with the Locally Interpolated Alkalinity Regression (LIAR) algorithm (Carter et al. 2016). Note that the datasets in Table 2 likely suffer from biases unaccounted for due to temperature mismatches as discussed in Section 2.2 (Goddijn-Murphy et al. 2015). It is important to note that each of the validation datasets are compared independently of each other, thus avoiding the complications of accounting for the biases between datasets. All $p\text{CO}_2$ data are then gridded to the same time and space resolution as the feature-variables (monthly $\times 1^\circ$) using *xarray* and *pandas* packages in Python (McKinney, 2010; Hoyer and Hamman, 2017).

2.7 Sea-air CO_2 flux calculation

Bulk sea-air CO_2 flux ($F\text{CO}_2$) is calculated with:

$$F\text{CO}_2 = k_w \cdot K_0 \cdot (p\text{CO}_2^{\text{sea}} - p\text{CO}_2^{\text{atm}}) \quad \text{Eq. 2}$$

where K_0 is the solubility of CO_2 in seawater (Weiss 1974) and k_w is the gas-transfer velocity calculated from wind speed using formulation by Nightingale et al. (2000) as this parameterisation was the closest match to *in-situ* observations of CO_2 fluxes (Goddijn-Murphy et al. 2016). The ERA-interim v2 wind product is used to calculate k_w . $p\text{CO}_2^{\text{sea}}$ is from the gap-filling methods, and $p\text{CO}_2^{\text{atm}}$ is atmospheric $p\text{CO}_2$. All ancillary variables required in these calculations are the same as those listed in Table 1, except for $p\text{CO}_2^{\text{atm}}$, which is the CarboScope atmospheric $p\text{CO}_2$ product from Rödenbeck et al. (2014). One of the problems with the bulk estimates of sea-air CO_2 fluxes is that models of gas exchange in the surface layer of the water column are simplified, but there are approaches, such as the rapid equilibrium model, that account for more complex temperature gradients in the upper layer of the surface ocean (Wanninkhof et al. 2009; Woolf et al. 2016).
 However, for the sake of consistency with past studies, we use the bulk approximation of sea-air fluxes (Eq. 2) where k_w is scaled to 16 cm.hr^{-1} as in the SOCOM intercomparison (Rödenbeck et al., 2015). Other approaches, such as the rapid equilibrium model maintain a degree of complexity (Woolf et al. 2016). This requires the separate calculation of atmospheric and oceanic CO_2 concentrations in the foundation (*fund*) and interfacial (*skin*) layers. We thus include the variables required to calculate sea-air fluxes using the rapid equilibrium model in the data repository.

2.8 Relative interannual variability and interquartile range metrics

2.8.1 Regression metrics

We use bias and root-mean-square error (RMSE) as first-order metrics of model performance.

Bias is the mean difference between the target variable and the estimates thereof:

$$Bias = \sum_{i=1}^n \frac{\hat{y}_i - y}{n} \quad \text{Eq. 3}$$

375 where n is the number of training samples, y is the array of target data and \hat{y} is the corresponding array of estimates. Similarly, RMSE is a measure of the difference between the target variable and the estimates thereof:

$$RMSE = \sqrt{\sum_{i=1}^n \frac{(y_i - \hat{y}_i)^2}{n}} \quad \text{Eq. 4}$$

In our study, these metrics are calculated for each year and then the mean of the annual bias or RMSE scores is taken as a more robust measure of performance in the context of temporally imbalanced data. This is typically done for the global domain unless otherwise stated.

380

The relative interannual variability metric (R^{iav}) was used in the SOCOM intercomparison by Rödenbeck et al. (2015) to measure how well a method represents the interannual variability of the SOCAT data. The metric furthers the idea of RMSE calculated by year (and region if stated, otherwise global) by normalising annually weighted RMSE to a benchmark with interannual variability driven only by atmospheric pCO_2 :

$$R^{iav} = \frac{\sigma_{1982-2015}(M^{iav}(t))}{\sigma_{1982-2015}(M_{bench}^{iav}(t))} \quad \text{Eq. 5.1}$$

$$M^{iav}(t) = \sqrt{\frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n-1}} \quad \text{Eq. 5.2}$$

$$M_{bench}^{iav}(t) = \sqrt{\frac{\sum_{i=0}^n (y_i - \hat{y}_i^b)^2}{n-1}} \quad \text{Eq. 5.3}$$

385 Here σ is the standard deviation of M^{iav} and M_{bench}^{iav} respectively, which are both represented as yearly time series. Equations 5.2 and 5.3 show the formulation for $M^{iav}(t)$ and $M_{bench}^{iav}(t)$, which represent these metrics for a single year (t). The symbol i represents individual data points in a particular year t , y is the observation-based data for that year, \hat{y} is the predicted data and n is the number of points in the year and region. The benchmarked M_{bench}^{iav} is calculated to normalise M^{iav} . The \hat{y}^b represents the data where IAV has been removed by summing
390 the climatology of the mapped surface ocean pCO_2 and the annual trend of atmospheric pCO_2 .

2.8.2 Ensemble metrics

We use the interquartile range (IQR) between different gap-filling methods as a robust metric of disagreement, in contrast to the standard deviation which is sensitive to outliers. IQR is calculated as the third quartile (75th

percentile) minus the first quartile (25th percentile). The disagreement between methods is calculated with annually-averaged data with the resulting difference averaged over the time series to arrive at the interannual disagreement (IQR^{IA}). This is calculated per pixel if the representation of the data is spatial (maps) and per time step of a time series.

3 Results

3.1 Regression results

The results from the regression comparisons (step two in Figure 1) are depicted in Figure (5a-c) which plots the matrix of the (a) average bias, (b) RMSE and (c) R^{ia^v} for each combination of the experimental number of clusters and clustering features.

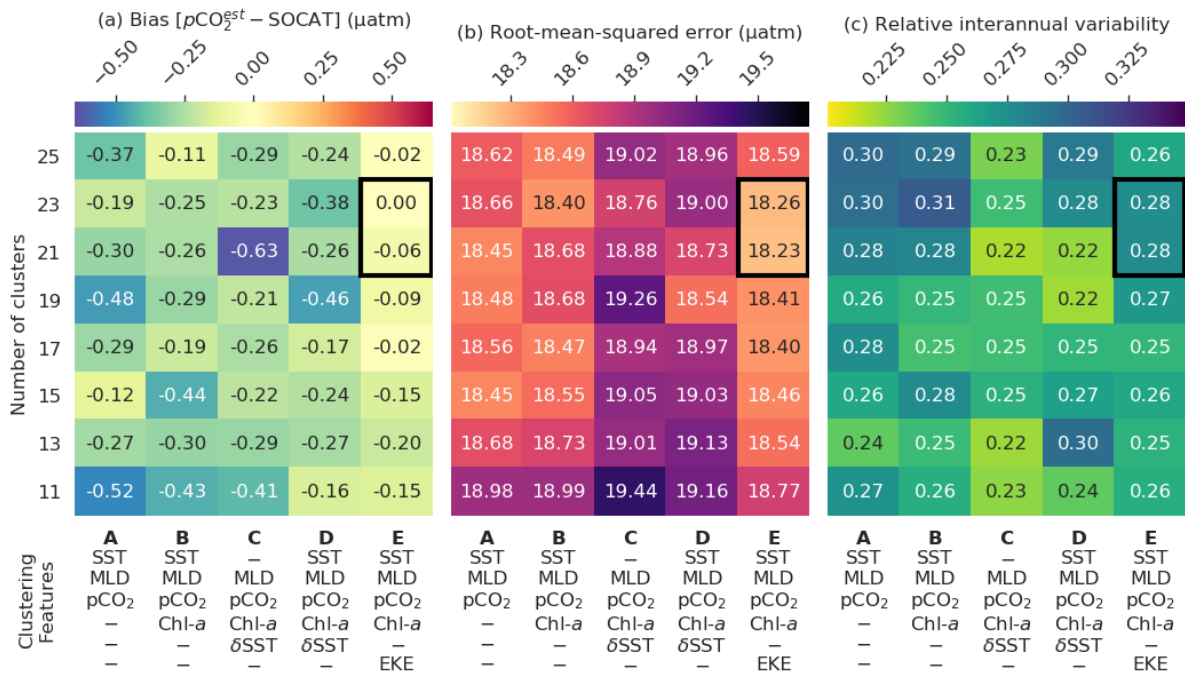


Figure 5: Heatmaps showing the average cluster (a) bias, (b) root-mean-squared error (RMSE) and (c) relative interannual variability (R^{ia^v}) for different cluster configurations, where smaller scores are better for all metrics. The rows show the number of clusters, and the columns show clustering feature-variable configurations. Each cluster contains the average of the scores for four regression methods: support vector regression, extremely randomised trees, gradient boosting machine, and feed-forward neural network. The black box indicates clustering configurations that perform well across all metrics – note that a R^{ia^v} < 0.3 falls within the best category of performance in Rödenbeck et al. (2015).

Results show that the configuration that includes EKE^{clim} (column E in Figure 5a-c) as a clustering feature has the lowest average RMSE and absolute bias for nearly all clustering configurations, regardless of the number of clusters (rows in Figure 5a,b). The increased dynamics associated with high EKE regions might change the way pCO₂ behaves compared to low EKE regions (Boutin et al., 2008; Monteiro et al. 2015; du Plessis et al., 2017, 2019). The optimal number of clusters within this configuration is either 21 or 23, based on the smallest bias and RMSE scores (as indicated by the black box in Figure 5), while we do not weight R^{ia^v} strongly in this assessment as a R^{ia^v} score of less than 0.3 is in the top-performing category in the SOCOM intercomparison

(Rödenbeck et al. 2015). While the individual regression methods' bias and RMSE scores (Figures S5 and S6 respectively) do not match the distributions exactly, the two selected clustering configurations (black boxes in Figure 5) score consistently low for both metrics (with the exception of ERT – discussed in greater detail further on). We motivate to select only one clustering configuration for the sake of simplicity. Furthermore, we select the configuration with 21 clusters (rather than 23), as fewer clusters further reduces the possible complexity at little cost. The selected clustering configuration with 21 clusters has the following features: SST, $\log_{10}(\text{MLD}^{\text{clim}})$, $p\text{CO}_2^{\text{clim}}$, $\log_{10}(\text{Chl-}a^{\text{clim}})$, and $\log_{10}(\text{EKE}^{\text{clim}})$; and is hereinafter abbreviated as K21E (see Figure S2 for the distribution of the climatology for these clusters).

Comparatively, the Fay and McKinley (2014) CO₂ biomes have an average RMSE score of 18.98 μatm (Table 3) but have a lower mean R^{iaV} (0.26) and smaller bias (0.03 μatm) than the K21E configuration. Given that the CO₂ biomes perform well and provide an alternate clustering approach, we include the regression estimates. The eight machine learning models from K21E and BIO23 (four each) were used to create an ensemble average by averaging $p\text{CO}_2$ estimates (CSIR-ML8).

Table 3: Regression scores for the CO₂ biomes (BIO23), the clustering configuration from column E in Figure 5 (K21E) and the ensemble average (CSIR-ML8). Abbreviations are: RMSE = root-mean-square error; R^{iaV} = relative interannual variability (Equation 5). Regression methods are: SVR = support vector regression; ERT = extremely randomised trees; GBM = gradient boosting machine; FFN = feed-forward neural network. Bold values are significantly lower than the mean for that column ($p < 0.05$ for two-tailed Z-test; absolute values used for bias column).

Clustering	Regression	Bias (μatm)	RMSE (μatm)	R^{iaV}
CSIR-ML8		0.04	17.25	0.25
K21E	SVR	-0.45	17.95	0.24
	ERT	0.84	17.96	0.36
	GBM	-0.32	18.21	0.24
	FFN	-0.30	18.82	0.27
BIO23	SVR	-0.19	18.47	0.15
	ERT	0.85	18.76	0.38
	GBM	0.02	19.05	0.28
	FFN	-0.58	19.65	0.21

All regression methods have lower RMSE scores for K21E than for BIO23, but R^{iaV} and bias do not indicate that any of the two clustering approaches is preferable (Table 3). Comparing the RMSE scores of the individual regression methods, we see that the model scores are ranked the same in each cluster from first to last: SVR, ERT, GBM, FFN. However, it is important to note that this ranking does not apply to bias or R^{iaV} , where ERT has low RMSE, but the largest bias and R^{iaV} in each clustering approach. CSIR-ML8 only slightly better its members with RMSE and bias scores of 17.25 μatm and 0.04 μatm respectively. However, the ensemble average R^{iaV} (0.25) is only just less than the average of the ensemble members' average (0.26).

3.2 Robust RMSE, bias and R^{iav}

Here, we study the change in the bias and RMSE for all selected methods (i.e. K21E, BIO23 and CSIR-ML8; Table 3) across 1982-2016 (Figure 6). Most notable is that bias scores for all models have the same interannual tendencies, with a positive bias at the beginning of the time series (1982 to 1993) that is strongest before 1990, strongly influencing the mean bias (Table 4). Secondly, the biases for K21E (solid lines) are, on average, smaller than for BIO23 (dashed lines) as shown for the annually-averaged results in Table 4 (0.73 μatm and 2.24 μatm respectively). These biases are larger than those reported in Table 3 (with averages of absolute biases of 0.48 μatm and 0.41 μatm for K21E and BIO23 respectively), but this is likely since selected test years (black triangles in Figure 6b) fall on years of low bias. While FFN has the largest RMSE (18.93 μatm and 20.24 μatm for K21E and BIO23), it has a smaller bias compared to other regression methods (0.04 μatm and 1.60 μatm respectively), motivating for including FFN regressions in the ensemble average (Table 4). Conversely, the ERT approach has a significant positive bias likely due to the method's resilience to outliers, where sparse measurements could be treated as outliers (2.08 μatm and 3.88 μatm for K21E and BIO23 respectively, with $p > 0.95$ for both values; Table 4; Gregor et al. 2017). A second ensemble average without ERT regressions, thus with six members (CSIR-MLR6 version 2019a, hereafter called CSIR-ML6), has lower biases compared to CSIR-ML8 (0.98 μatm and 1.48 μatm respectively; Table 4).

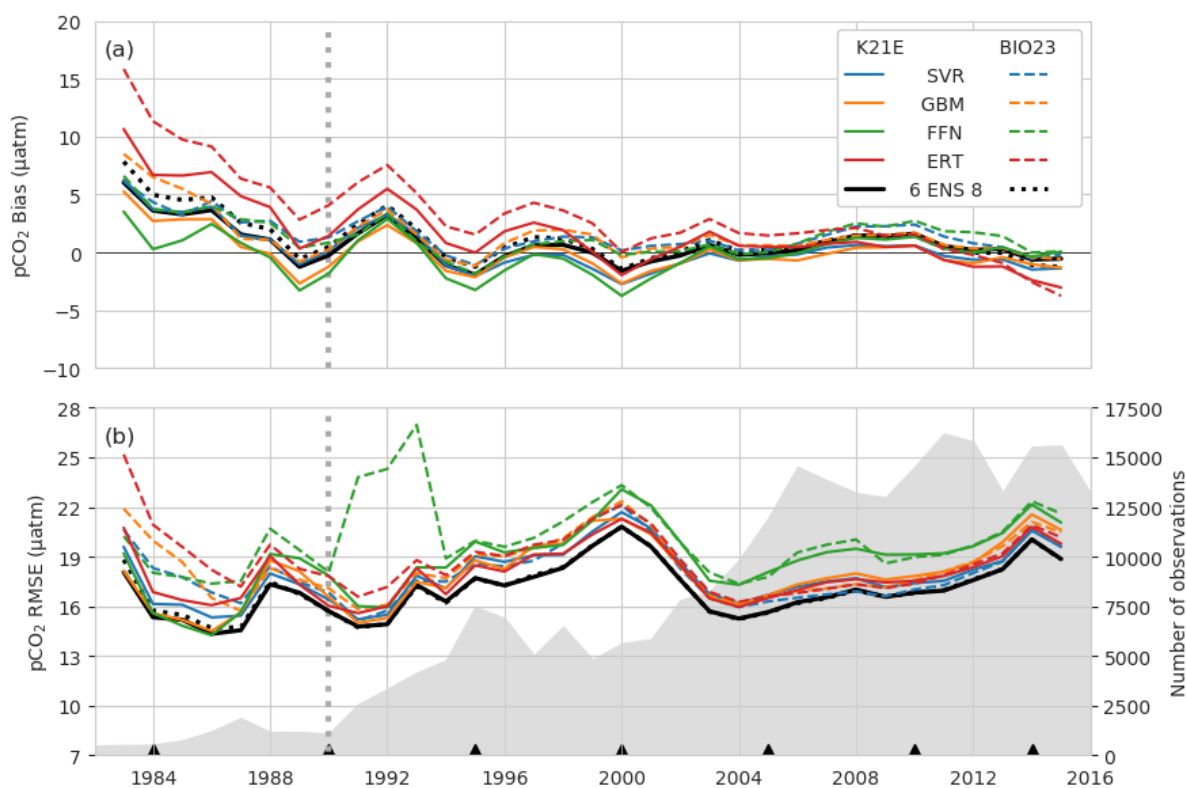


Figure 6: Annually averaged (a) bias and (b) RMSE for the eight individual regression methods in Table 3: BIO23 (dashed lines) and K21E (solid lines). The dotted black lines show the ensemble averages for all eight models (CSIR-ML8), and the solid black line shows metrics for the ensemble average of the SVR, GBM and FFN (CSIR-ML6) from BIO23 and K21E. The grey filled area in (b) shows the number of observations per year and black triangles shows the years that are isolated as the test subset. The vertical dashed grey line demarks 1990 prior to which there is a large positive bias.

470 Similarly to the biases, RMSE for all models (Figure 6b) have similar interannual tendencies and variability,
 with a sharp peak in the year 2000 ($> 20 \mu\text{atm}$ where the mean RMSE is $18.61 \mu\text{atm}$). The increased RMSE
 scores are likely due to the spatial distribution of sampling density (see Figure S7), *e.g.* an increase in sampling
 in the high latitudes during spring and summer, a region and period of high variability and biogeochemical
 complexity, would increase the weight of these data in the final RMSE calculation, thus resulting in larger
 RMSE scores. The increase in the number of samples from 2002 to 2016 results in a sharp decrease in RMSE (
 475 $< 19 \mu\text{atm}$ for the majority of this period). Both ensemble averages perform slightly better than all other
 methods for the majority of the time series with RMSE scores of $17.16 \mu\text{atm}$ and $17.25 \mu\text{atm}$ for CSIR-ML6 and
 CSIR-ML8 respectively (see Table S1 comparisons of ensemble averages with different members).

480 The R^{iaV} scores for the robust errors (Table 4) are lower than train-test results with a single split reported in Table
 3, likely due to an increase of standard deviation for the IAV benchmark (Equation 5). The lowest score is held
 by CSIR-ML6 (0.20) and is lower (better) than the average for its members (0.21). These R^{iaV} estimates compare
 well to the Jena-MLS and SOM-FFN, which both scored < 0.3 (Rödenbeck et al. 2015).

485 **Table 4:** The robust estimates of bias, RMSE and R^{iaV} from 1982 to 2016 for BIO23, K21E and the ensemble averages,
 CSIR-ML6 and CSIR-ML8, where the first excludes the ERT method. Bold values are significantly lower than the mean for
 that column ($p < 0.05$ for two-tailed Z-test; absolute values used for bias column). See Table S1 for further comparisons
 between different ensemble average configurations.

Clustering	Regression	Bias (μatm)	RMSE (μatm)	R^{iaV}
CSIR	ML6	0.98	17.16	0.20
	ML8	1.48	17.25	0.22
K21E	SVR	0.58	18.04	0.21
	ERT	2.08	18.20	0.27
	GBM	0.21	18.05	0.21
	FFN	0.04	18.93	0.22
BIO23	SVR	1.76	18.17	0.21
	ERT	3.88	19.16	0.32
	GBM	1.72	18.59	0.21
	FFN	1.60	20.24	0.21

490 The spatial distribution of the bias and RMSE is now studied for CSIR-ML6 (Figure 7 a and b, respectively),
 particularly focusing on the regional patterns emerging from the data. CSIR-ML6 clearly represents the
 subtropical regions (NH-ST and SH-ST) with relatively low biases and RMSE scores ($|\text{bias}| < 5 \mu\text{atm}$ and RMSE
 $< 10 \mu\text{atm}$). The equatorial regions (EQU), especially the eastern Pacific, contrasts this with large uncertainties
 in both bias and RMSE ($> |10 \mu\text{atm}|$ and $30 \mu\text{atm}$ respectively). The high-latitude oceans (NH-HL and SH-HL)
 have considerable uncertainties due to the large interannual variability of surface ocean $p\text{CO}_2$ caused by the
 495 formation and retreat of sea-ice (around Antarctica; Ishii et al. 1998; Bakker et al. 2008) and phytoplankton
 spring blooms (Atlantic sector of the Southern Ocean, North Pacific and Arctic Atlantic; Thomalla et al. 2011;

Lenton et al. 2013; Gregor et al. 2018). There are two bands of overestimates on the southern and northern boundaries of the North Atlantic Gyre, where the latter coincides with the Gulf Stream. Regression approaches may be prone to a positive bias in the North Atlantic as this was also shown by Landschützer et al. (2013; 2014).

500

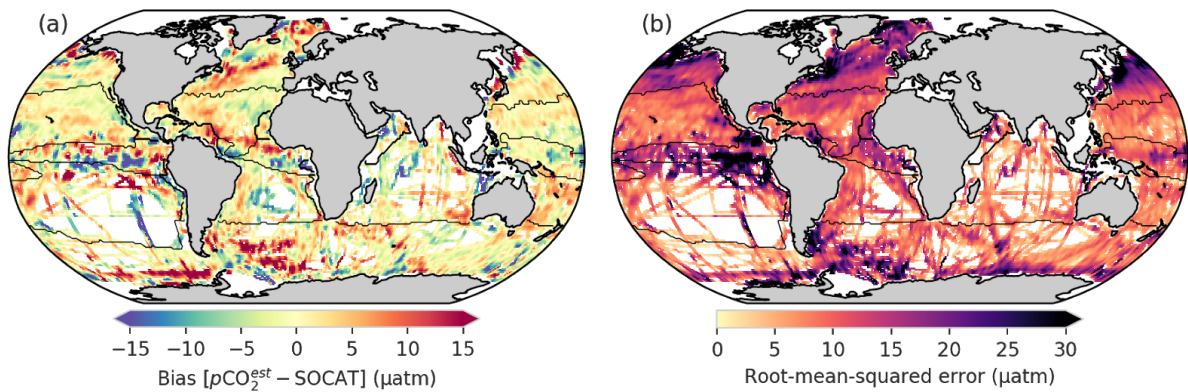


Figure 7: (a) shows the biases from the robust test-estimates; (b) shows the root-mean-squared errors for CSIR-ML6.

505 Convolution has been applied to (a) and (b) to make it easier to see the regional nature of the biases and RMSE. Figure S8 shows the bias for every ensemble member. **Black lines show the regions as defined in Figure 3.**

In summary, the robust test-estimates show that there is a positive bias in $p\text{CO}_2$ predictions before 1990 for all models, but it is largest for ERT, and excluding these models from the ensemble results in better $p\text{CO}_2$ predictions. The spatial evaluation of the performance metrics for CSIR-ML6 shows that regions with specific oceanic features (e.g. western boundary currents) mostly have positive biases. However, it is important to note that these uncertainty assessments are limited as the characteristics and biases of the dataset are intrinsic to the models. Validation with independent data is thus a more reliable estimate of the performance of these methods.

510

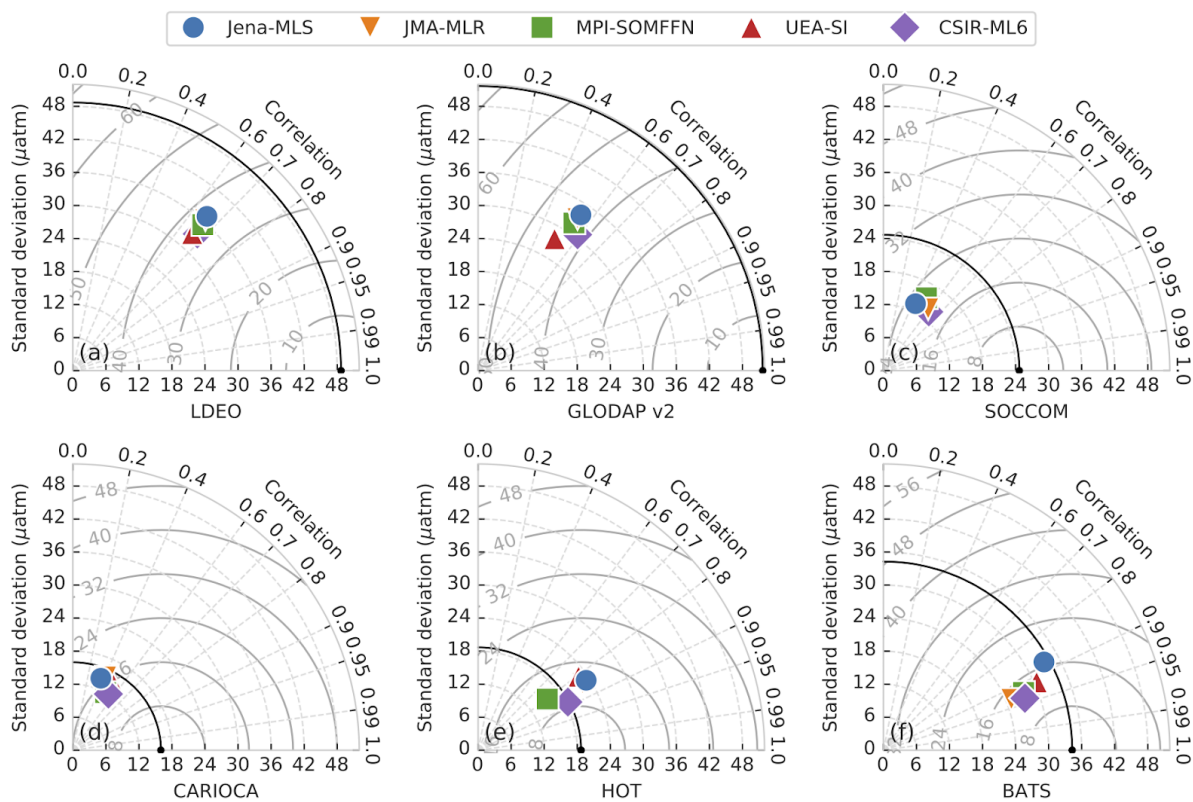
3.3 Validation with independent datasets

515 Here, we validate the accuracy of $p\text{CO}_2$ estimates from CSIR-ML6 with independent data (that is not in SOCAT v5 as described in Table 2). To further study the behaviour of our ensemble average estimates relative to previous studies, we compare the results from four independent methods of the SOCOM intercomparison project against the independent data calculated over individual data points (Rödenbeck et al. 2015). Those four independent methods are: the Jena mixed-layer scheme (Jena-MLS version *oc_v1.6*, Rödenbeck et al. 2014); Japanese Meteorological Agency – multi-linear regression (JMA-MLR updated on 2018-12-2, Iida et al. 2015); Max Planck Institute – Self-organising Map Feed-forward Neural-network (MPI-SOMFFN *v2016*, Landschützer et al. 2017); and University of East Anglia – Statistical Interpolation (UEA-SI version 1.0, Jones et al. 2015). $p\text{CO}_2$ estimates by the Jena-MLS were resampled to monthly temporal resolution and interpolated to a one-degree grid using Python’s *xarray* package. Note that these datasets ~~(with the exception of the UEA-SI)~~ will also suffer from the same temperature biases discussed in S2.4.

525

The performance of each gap-filling method is represented with a Taylor diagram for each independent validation dataset (Figure 8; Taylor et al. 2001). The most important characteristic learnt from these plots is that the gap-filling methods are tightly bunched for nearly all validation datasets, indicating a similar RMSE, correlation and standard deviation relative to the reference datasets. Poor estimates in Figures 8a-d may indicate that the training data for gap-filling methods is the limiting factor. Secondly, the gap-filling methods almost always underestimate the standard deviation of the validation datasets, being below the black arced line for all but the station=HOT (Figure 8e).

530



535 **Figure 8:** Taylor diagrams comparing the $p\text{CO}_2$ estimates of five gap-filling methods (represented by the different markers) with validation datasets (Table 2), for the period 1990-2015. Each validation dataset has its own Taylor diagram as labelled on the bottom axes. The black marker on the bottom axis in each subplot represents the validation dataset and the black arc shows the standard deviation thereof. The closer the gap-filling estimates are to this point, the better the model's performance, in terms of variance, centred RMSE and correlation (for bias information, see Table 5). The solid grey arcs show the centred RMSE for the datasets (with bias removed). Description of the gap-filling methods from independent studies is provided in the text, Section 3.3.

540

All methods fail to represent the standard deviation of the two global validation datasets, LDEO and GLODAP v2 (Figures 8a,b), with centred RMSE scores greater than $35 \mu\text{atm}$. However, calculating RMSE annually results in scores of $\sim 27 \mu\text{atm}$ for LDEO and $\sim 35 \mu\text{atm}$ for GLODAP v2, much lower than shown in Figure 8a,b due to high RMSE scores ($> 40 \mu\text{atm}$) for a small subset of years (Section S3.4 and Figure S7). Estimates of the Southern Ocean datasets (Figures 8c, d), SOCCOM and CARIOCA, have lower RMSE scores ($\sim 16 \mu\text{atm}$ and $\sim 23 \mu\text{atm}$ respectively) relative to LDEO and GLODAP v2. However, for standard deviation scores of similar magnitude and low correlation coefficients, the datasets are not well constrained (Table 5). The SOCCOM dataset also has the largest average absolute bias for estimates, with gap-filling methods underestimating by at

550

least 11 μatm (Table 5). This large bias may be because SOCCOM floats have a proportionately large number of winter samples – suggesting that our knowledge of Southern Ocean winter fluxes are largely underestimated (Williams et al. 2017). In contrast, all methods estimate the two time-series stations, HOT and BATS (Figures 8e,f and Table 5) relatively well with correlation scores $r > 0.8$ and low average bias $\sim 4.5 \mu\text{atm}$.

555

Table 5: The RMSE and bias for each gap-filling method compared to the validation datasets. For more information on the validation-datasets see Table 2. The first row of data (count) shows the number of gridded samples in the dataset during the period 1990-2015 (that are not in the SOCAT v5 gridded product). Values shown in bold are significantly different from the mean for the column ($p < 0.05$ for two-tailed Z-test; absolute values used for biases). **The UEA-SI method does not have error estimates for SOCCOM floats as these two time series do not overlap.**

560

Metric	Method	LDEO	GLODAP-v2	SOCCOM	CARIOCA	BATS	HOT
Count	Count	16161	5976	1037	613	246	214
RMSE	CSIR-ML6	26.55	32.84	23.15	14.26	12.53	8.62
	MPI-SOMFFN	27.43	35.96	25.21	15.08	13.39	10.40
	JMA-MLR	29.11	34.53	22.32	16.05	14.29	11.64
	Jena-MLS	27.61	35.52	26.83	18.24	16.14	12.28
	UEA-SI	27.35	35.07		15.73	13.35	18.52
Bias	CSIR-ML6	-1.18	8.48	-13.12	4.28	0.32	0.46
	MPI-SOMFFN	-0.19	9.16	-13.79	4.00	-1.41	-0.12
	JMA-MLR	-1.86	6.62	-11.25	2.85	-3.98	2.22
	Jena-MLS	-0.14	8.48	-14.68	7.18	4.09	6.15
	UEA-SI	-0.71	9.20		0.79	-2.02	16.27

Despite all scores being closely grouped (Figure 8), Table 5 shows that the CSIR-ML6 method scores significantly lower RMSE scores (using a two-tailed Z-test with $p < 0.05$) for all but one of the datasets (SOCCOM). However, bunching of the RMSE scores (Figure 8) is beneficial with regard to achieving low p -values. No single method dominates the biases, with JMA-MLR and MPI-SOMFFN each scoring the lowest bias on two occasions. To summarise, all gap-filling methods underperform when validated against independent observational products. Tight bunching of gap-filling method scores per validation dataset shows that training data may limit all methods in the same manner.

565

570

3.4 The effect of uncertainties on the sea-air CO_2 flux interannual variability

In this section, we assess the regional implications of the differences in gap-filling methods' estimates (within CSIR-ML6 and the four independent methods described in Section 3.3) of the sea-air CO_2 flux (FCO_2) over the period 1990 to 2016. FCO_2 was calculated using the same gas transfer velocity and solubility for each gap-filling method (Section 2.7). Differences in FCO_2 are thus driven by variations in $p\text{CO}_2$ from each gap-filling method.

575

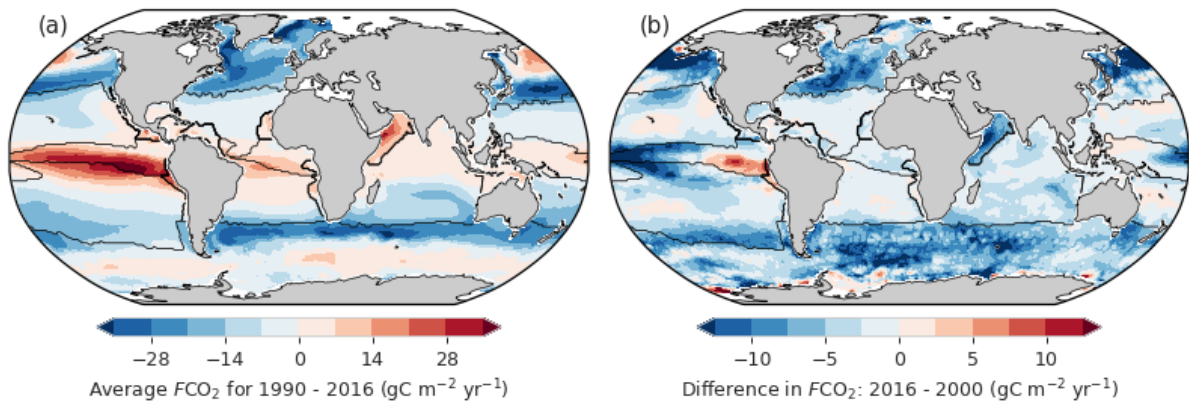


Figure 9: (a) Average sea-air CO_2 fluxes (FCO_2) of CSIR-ML6 for 1990 to 2016, where FCO_2 is calculated as shown in Equation 2. Negative FCO_2 (blue) indicates regions of atmospheric CO_2 uptake. (b) The difference between FCO_2 in 2016 and 2000, which are the minimum and maximum of global ocean uptake flux (FCO_2) estimates respectively (for CSIR-ML6 in Figure 10a). Black lines show the regions as defined in Figure 3.

580

The average FCO_2 for 1990-2016 by CSIR-ML6 (Figure 9a) contextualises the regional distribution of fluxes: strong outgassing in the Equatorial Pacific, strong sink in the mid-latitudes, a moderate uptake for the most part of the subtropics, and weak source in the majority of the Southern Ocean (in agreement with e.g. Takahashi et al., 2009). The global annual time-series for FCO_2 as simulated by CSIR-ML6 (Figure 10a) indicates a strengthening for 2000 to 2016 (as for the other methods). To give spatial context to this strengthening, we display the differences in FCO_2 between 2016 and 2000 (Figure 9b), since those are the two years where the difference in global FCO_2 is greatest for CSIR-ML6 (Figure 10a). Note that Figure 9b serves as a snapshot for the change in FCO_2 between those two years, whose interpretation cannot be linked to an overall anthropogenically-forced change as the comparison between two years could reflect interannual, decadal or multi-decadal variability. The differences in FCO_2 between 2016 and 2000 is negative in the high latitudes and moderately positive in the subtropics, indicating a respective increase and decrease in the CO_2 ocean uptake between the two years. The Eastern Equatorial Pacific is the only region that shows a considerable increase in FCO_2 ($> 10 \text{ gC m}^{-2} \text{ yr}^{-1}$) between the two specific years.

585

590

595

The annual change in FCO_2 is also studied for the different regions. The Southern Hemisphere high-latitude (SH-HL) region is the strongest contributor to the trend (Figure S10b), where there is a steady increase in the uptake of CO_2 since the 2000s for all methods (Landschützer et al. 2015; Gregor et al. 2018). On average, the Northern Hemisphere high latitudes (NH-HL) are a weaker sink relative to the SH-HL, because the SH-HL is more than double the area of the NH-HL (Figure S10c). The equatorial (EQU) region is the only persistent source of CO_2 to the atmosphere (also seen in Figure 9a). The subtropical regions (Figure 10c, e) contribute to global flux on similar orders of magnitude; however, there is a large divergence between gap-filling methods in the SH-HL.

600

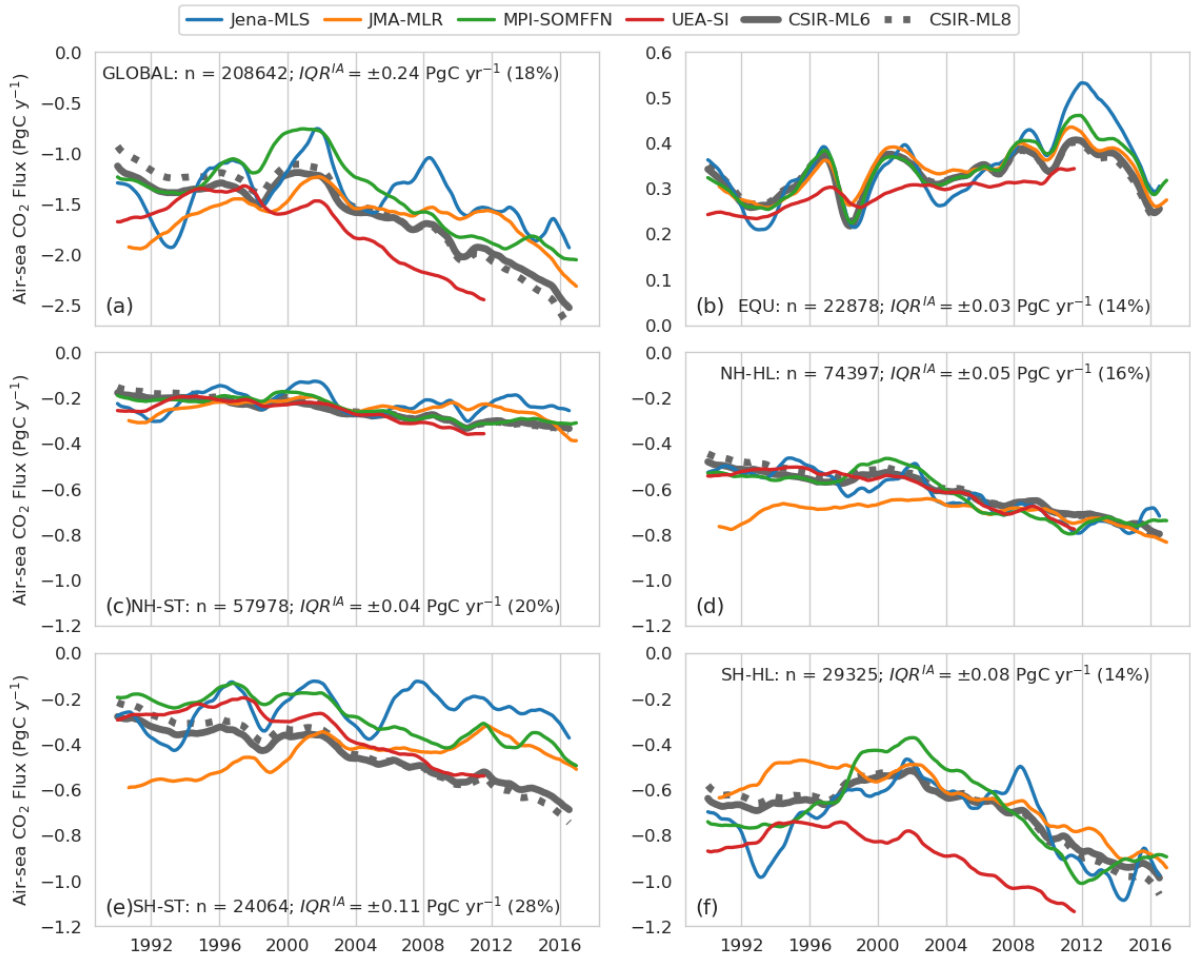


Figure 10: Sea-air CO₂ fluxes averaged for regions as shown in Figure 2: (a) global domain, (b) Equatorial regions, (c) Northern Hemisphere Subtropical, (d) Northern Hemisphere High Latitude, (e) Southern Hemisphere Subtropical. (f) Southern Hemisphere High Latitude. The coloured lines show the four SOCOM products. The thick and dotted grey lines show the results for CSIR-ML6 and CSIR-ML8, respectively. A moving average of 12 months has been applied to smooth the data. Note that the y-axes' scales differ for the top (a) and (b). Note that the uncertainties of each model (e.g. bias and RMSE from Figure 6) are not shown here. The text at the right of each figure shows the number of SOCAT v5 gridded data points for each region (*n*) and the inter-annual interquartile range (IQR^{IA}).

610

615

620

625

We use the average interquartile range between the one-year rolling mean estimates (IQR^{IA}) as a measure of agreement or divergence between gap-filling methods, where large values indicate a divergence (Section 2.8.2). We also show the IQR^{IA} scaled to the range of the regional interannual variability (max – min) as a percentage (relative IQR^{IA}), which shows if the trend for a particular region is agreed on by all methods (the smaller the percentage, the better the agreement across methods). The disagreement between methods in the SH-ST is substantial (Figure 10e), with diverging *FCO*₂ throughout the period with an IQR^{IA} of 0.11 PgC yr⁻¹ and a large relative IQR^{IA} of 28%. Similarly, the IQR^{IA} for the SH-HL region (Figure 10f) is 0.08 PgC yr⁻¹, but the relative IQR^{IA} is lower at 14%, indicating that all methods agree on the observed strong trend. Compared to the Southern Hemisphere, the Northern Hemisphere regions are both relatively well constrained, with IQR^{IA} estimates of 0.04 PgC yr⁻¹ and 0.05 PgC yr⁻¹ for the NH-ST and NH-HL regions respectively (Figure 10c,d). However, a larger relative IQR^{IA} of 20% suggests that the interannual *FCO*₂ estimates in the NH-ST region are potentially not resolving the trend, or more likely that there is a weak trend with a small difference between the minimum and

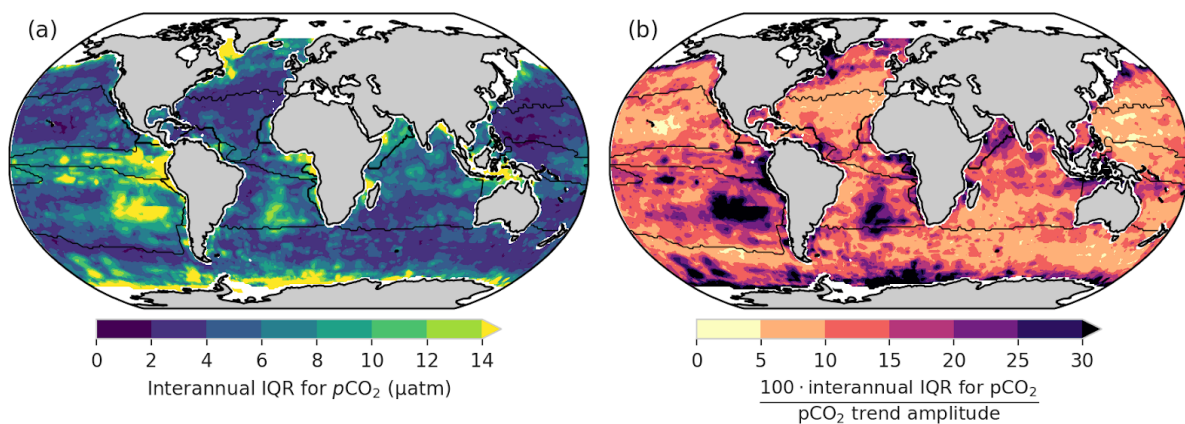
maximum interannual estimates of FCO_2 . The equatorial region (EQU - Figure 10b) has an IQR^{IA} and relative score at 0.03 PgC yr^{-1} and 14%.

630 The CSIR-ML8 method is not included in the IQR^{IA} calculations but is included in Figure 10 to show the impact of the ERT models' positive bias in pCO_2 on FCO_2 (Figure 6a). The biases are positive at the beginning and negative end of the time series, with the average absolute difference between the CSIR methods being 0.08 PgC yr^{-1} . The positive biases have the strongest impact on the SH-ST that occupies 36% total area (Figure S10c), with only 11% of the total observations in SOCAT, suggesting that this method is sensitive to imbalanced datasets.

635

3.5 Regional disagreement between methods

In order to better understand the regional distribution of the uncertainties in FCO_2 , we assess the level of agreement between independent gap-filling methods in their interannual surface ocean pCO_2 estimates (Figure 11). We use pCO_2 for this representation as no spatial integration occurs – only time averaging.



640 **Figure 11:** (a) The magnitude of the interannual disagreement between independent gap-filling methods (IQR^{IA}) as shown in Figure 10; hence low IQR^{IA} indicates good agreement amongst the different methods. (b) Level of agreement on the interannual variability across methods (in %), more specifically IQR^{IA} scaled by the difference between the maximum and minimum values for interannual pCO_2 (the range).

645 The interannual estimates of interquartile range (IQR^{IA} ; Figure 11a) show the disagreement between methods is relatively small in the majority of the ocean ($\approx 5 \mu\text{atm}$). The exceptions being the Southern Ocean, South Atlantic, southeastern Pacific and eastern equatorial Pacific with differences of $> 10 \mu\text{atm}$, where these regions coincide with regions of low sampling density (Figure 2). The IQR^{IA} scaled to the maximum-minimum range of interannual pCO_2 suggests that the NH-ST trend is relatively well constrained ($< 10\%$), which is in conflict with the IQR^{IA} for FCO_2 in Figure 10c (where the relative IQR^{IA} is 20%). The disagreement may stem from the magnifying impact that wind speed has on FCO_2 , *i.e.* small differences in pCO_2 may become large when fluxes are calculated. The same principle may apply to the EQU in Figure 11b, where relative IQR^{IA} is large ($> 10\%$) for pCO_2 , but low wind speeds result in a low relative IQR^{IA} for FCO_2 (7% in Figure 10b). The largest relative IQR^{IA} scores occur in the SH-ST ($> 10\%$ in Figure 11c) where data is sparse, specifically the South Atlantic and

650

655 south eastern Pacific (Figure 2a). The relative IQR^{IA} scores suggest that the gap-filling methods agree on $p\text{CO}_2$
in the SH-HL east of the Greenwich meridian ($> 0^\circ \text{ E}$).

In summary, we show that there is an agreement between gap-filling methods in the Northern Hemisphere for
interannual $p\text{CO}_2$, but the methods show considerable disagreement in the Southern Hemisphere, particularly in
660 the subtropics. Disagreements in the Equatorial and Southern Hemisphere high-latitude regions are large ($>$
10%) and should be treated with caution when considering trends in these regions.

4 Discussion

4.1 Not all models are equal

In their study, Khatiwala et al. (2013) stated that: “*our comparison of different methods suggests, that multiple*
665 *approaches, each with its own strengths and weaknesses, remain necessary to quantify the ocean sink of*
anthropogenic CO_2 ”. In our study, we embrace this philosophy by creating an ensemble average of two-step
machine learning models that estimate global surface ocean $p\text{CO}_2$. We show robustly that the CSIR-ML6
method reproduces the available data with greater accuracy than previous methods, albeit in an incremental way.
Our method is methodologically consistent with regard to feature-variables. Though there is variability in the
670 clustering and regression, we create the ensemble average with a good understanding of each model’s biases
(Figure 6 and Figure S8). The argument that ensemble averages reduce transparency is also somewhat
diminished by the fact that little additional information that can be gained from highly non-linear models, with
the exception of basic diagnostics such as feature-variable importance (see Figure S11) from decision-tree-based
approaches (Pedregosa et al. 2012; Castellevecchi, 2016). Our results thus show that there is, in fact, a benefit in
675 creating an ensemble average of models (Table 5), and if carefully implemented is an additional tool that can be
used to reduce the uncertainties in gap-filling estimates of $p\text{CO}_2$.

It could be argued that an exhaustive search for the optimal configuration (Figure 5) for CSIR-ML6 may result
in poorly trained individual models. However, we think that the merit of introducing and assessing regression
680 algorithms new to the application (for gradient boosting machines and extremely randomised trees) outweighs
the marginal loss in potential performance for individual methods. Moreover, lessons learnt from our study can
be used to improve on future iterations. It also makes the case for ensembles averages stronger as the
CSIR-ML6 performs well relative to other gap-filling methods.

685 In the search for the optimal clustering configuration (Figure 5a,b), we show that including EKE (along with
SST) as a clustering feature-variable leads to an improvement in bias and RMSE for nearly all number of
clusters, albeit a small improvement. Increased intra-seasonal variability of $p\text{CO}_2$ appears to be associated with
regions of high EKE compared to low EKE regions (Monteiro et al. 2015; du Plessis, 2017, 2019). Moreover,
the importance of EKE as a part of the clustering constraints also shows that more thought should be given to
690 how we sample $p\text{CO}_2$ in high-EKE regions and at what resolution regression methods are run at.

Our findings suggest the following about the individual regression methods: the SVR and GBM algorithms produce good estimates with lower RMSE scores and biases, the FFN approach has larger RMSE scores yet low biases than the other methods, and the ERT approach has low RMSE scores but large biases in the estimates (Figure 6a,b; Table 4). We do not include the ERT approach in the ensemble average (CSIR-ML6) due to the large time-evolving biases, suggesting that ERT (with our tuning) is not suitable for estimating surface ocean $p\text{CO}_2$. The bias in ERT may be due to its sensitivity to imbalanced datasets (Crone and Finlay, 2012), where the data in SOCAT v5 are few before 2000. Returning to the above quote by Khatiwala et al. (2013), we thus find that the weaknesses of ERT outweigh its strengths.

4.2 Divergent gap-filling estimates

While we see that the improvements in the performance of gap-filling methods are relatively stagnant (relative to the training and validation data), the differences between the methods' estimates of $p\text{CO}_2$ and $F\text{CO}_2$ vary significantly in some regions, particularly in regions where data is sparse, such as in the Southern Hemisphere oceans (Figure 2). We also find that training the gap-filling methods with limited training data exposes the intrinsic biases of the algorithms, or in the words of Ritter et al. (2017): “the difference [between gap-filling methods] is a result of how the spatial and seasonal heterogeneity and the sparseness of the data is dealt with”. Conversely, as the number of training data increase, the biases are reduced, and the methods converge.

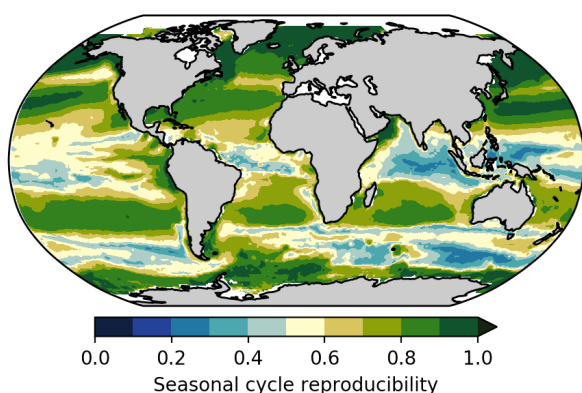


Figure 12: The seasonal cycle reproducibility of CSIR-ML6 $p\text{CO}_2$, which is a correlation of detrended $p\text{CO}_2$ with its own climatology – the larger the correlation the stronger the reproducibility of the seasonal cycle (method from Thomalla et al. 2011).

The Northern Hemisphere subtropical regions are a good example of a region where the gap-filling methods converge (Figure 11b), as also shown by the low RMSE scores and high correlation for the two mooring stations, HOT and BATS (Figure 8e,f). One of the reasons that the methods predict the variability well in the subtropics (Figure 8e,f) is that these regions are less biogeochemically complex and driven primarily by seasonal changes in SST (Bates 2001; Dore et al. 2009). This strong SST-driven seasonality in the subtropics is shown by the high seasonal cycle reproducibility (Figure 12).

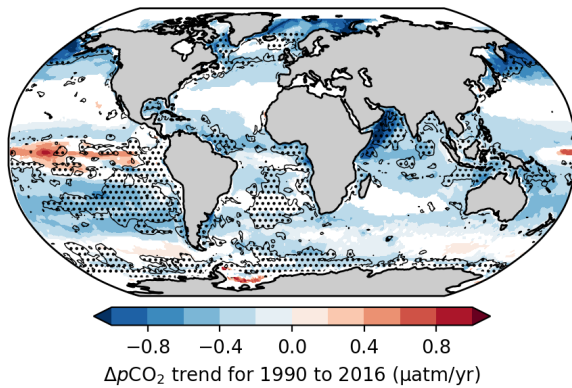


Figure 13: $\Delta p\text{CO}_2$ trends ($p < 0.05$), where $\Delta p\text{CO}_2$ is calculated as the estimated surface ocean $p\text{CO}_2$ from the CSIR-ML6 method minus atmospheric $p\text{CO}_2$ from the CarboScope project (Rödenbeck et al. 2014). The shaded areas show the regions where IQR^{IA} is $> 15\%$, thus indicating regions where trends should be interpreted with caution.

725

The gap-filling methods' divergences also serve as a metric to inform where there is not enough data to constrain the $p\text{CO}_2$ or $f\text{CO}_2$ estimates, *i.e.* the divergences inform us where estimates should be treated with caution. The IQR^{IA} , when scaled to the range of interannual variability (Figure 11b), should be taken into account when analysing interannual trends of $\Delta p\text{CO}_2$ (Figure 13). For instance, significant trend estimates in

730

$\Delta p\text{CO}_2$ for CSIR-ML6 ($p < 0.05$) are negative ($p < 0.05$) for the majority of the global ocean, even in regions where method estimates are too disparate to resolve interannual variability (relative $\text{IQR}^{\text{IA}} > 15\%$; dotted regions in Figure 13). However, the relative IQR^{IA} is not without its limits, as there may be regions where methods are in agreement but share the same biases, thus reporting false confidence in the estimates. Regions of false confidence would most likely occur in data sparse areas, but could only truly be identified with better data coverage in these regions.

735

4.3 Inching up and over the wall: incremental improvements

In our study, we show that all gap-filling methods suffer from the same uncertainties where there are data to test and validate the estimates (Figure 8), and divergences between estimates when there are insufficient data to constrain the methods (Figure 11b). From these points, it may seem that we may have in fact “hit the wall” in terms of better resolving surface ocean $p\text{CO}_2$. In this section, we discuss how we might overcome this proverbial wall. First, by addressing the existing uncertainty and biases, and then discussing how we could improve on estimates in data-poor regions. ▢

740

4.3.1 Reducing existing biases

The robust test-estimates show that there are regions where training data is not sparse, yet estimates still suffer from large uncertainties (*e.g.* northern and southern boundaries of the North Atlantic gyre in Figure 7a,b and Figure S8). These errors are spatially consistent with those reported by Landschützer et al. (2014). Such regional mismatches between gridded observations and estimates are likely systematic – meaning that gap-filling methods are not able to resolve the more complex $p\text{CO}_2$ variability at current resolutions (monthly $\times 1^\circ$ or coarser) or with the current regression feature-variables (Gregor et al. 2017; Denvil-Sommer et al. 2018). It may

745

750 be possible to reduce these uncertainties with consideration about the drivers of CO₂ in a specific region. Including appropriate additional feature-variables (if available), such as reanalysis mixed-layer depth products, may improve the uncertainties of gap-filling methods (Gregor et al. 2017). Similarly, increasing the temporal and spatial resolution may be able to improve estimates where aliasing occurs in regions of high dynamic variability such as the mid-latitude oceans (Monteiro et al. 2015). It is worthwhile noting that increasing the
755 resolution may not be the panacea for poor estimates. For example, the Jena-MLS method is able to estimate pCO₂ with relative accuracy (Figure 8) at a low spatial resolution ($\approx 4^\circ \times 5^\circ$; Rödenbeck et al. 2014); however, with the trade-off in spatial resolution, the method is able to increase the temporal resolution to daily estimates.

Another source of bias is the mismatch between the ~~depth and~~ temperature at which pCO₂ is measured (i.e. at
760 ~~the depth of a ship's intake~~) and the temperature to which pCO₂ is predicted (~ 1 m in the case of the dOISSTv2 data; Banzon et al. 2016; Goddijn-Murphy et al. 2015). Goddijn-Murphy et al. (2015) show that this mismatch is considerable in some cases ($> 5 \mu\text{atm}$ for large regions as shown in Figure S3b). However, ~~the correction of the intake temperature to the remotely sensed surface temperature this correction~~ also makes the assumption that temperature is the only factor that influences pCO₂ in the surface layer of the ocean. The correction will thus not
765 account for other processes such as primary production, stratification and gas exchange within the surface layer. This is an issue that should be discussed by the community and tested experimentally to assess the impact that these processes may have on pCO₂. ~~As we show in the supplementary material (S2.4), compensating for this mismatch by adjusting pCO₂ measurements to OISST results in larger errors. There is a warm bias for the ship temperatures (0.13°C), which Banzon et al. (2016) attribute to ship-based warming. However, this general bias would not contribute to the increase in RMSE as the target variable, pCO₂, shifts accordingly when the correction is applied. Rather we suggest that these biases are a complex interaction of ship-specific biases and water column processes (other than temperature) that may lead to different CO₂ concentrations. The sampling configuration and set up is unique to each vessel, which suggests that ship-specific corrections should be applied to reported SST—this would require a community wide effort. The second potential contribution to the~~
770 ~~mismatch could be considerable in cases where primary production near the surface is highly stratified.~~ ¶
775 ¶

4.3.2 Improving estimates in data-poor regions

All gap-filling methods suffer from similar biases and uncertainties (Figure 8, Table 5) when compared to
780 independent validation data, yet the same methods show vastly different results in data-sparse regions. These shared uncertainties and regionally consistent divergences between methods are in agreement with past studies, which find that insufficient training data is the limiting factor (Rödenbeck et al. 2015; Landschützer et al. 2016; Ritter et al. 2017; Denvil-Sommer et al. 2018).

785 Strides have been made in closing these data-sparse gaps with the deployment of autonomous sampling platforms. The Southern Ocean Carbon and Climate Observations and Modelling (SOCCOM) project, in

particular, has been influential in closing the gap in the Southern Ocean with the deployment of ~200 pH-capable biogeochemical Argo floats in the region since 2015 (Williams et al., 2017; Gray et al., 2018). The data collected by these floats during winter has shown that we have previously underestimated winter outgassing of CO₂ in the Southern Ocean (Gray et al. 2018). Incorporating these new estimates into machine learning estimates should be a priority for the community as the Southern Ocean plays an important role in the uptake of anthropogenic CO₂ uptake (Gruber et al. 2019). Incorporating this data successfully into existing models may not be straight-forward due to the strong temporal bias of these data toward the end of the time-series. For instance, the inclusion of atmospheric *p*CO₂ could result in temporally skewed estimates due to the “memory” effect that including the annually increasing atmospheric *p*CO₂ could have on estimates.

The complex machine learning models often used to estimate *p*CO₂ are prone to overfitting to the data, particularly in regions where data is sparse. Using less complex models, e.g. multi-linear regression, in such regions would reduce the risk of overfitting to the data. A regionally weighted ensemble approach may be an eloquent way to address this problem. In regions with sparse data coverage, simpler models could be favoured, while more complex models could be weighted more in regions with more data. However, the user would have to apply a potentially subjective model-complexity ranking for each approach. This may work well in the subtropical gyres where *p*CO₂ has a strong seasonal signal driven primarily by temperature (Figure 12; Lefèvre and Taylor, 2002).

One of the weaknesses of our study is that our approach is similar to other regression methods, namely (e.g. MPI-SOMFFN by Landschützer et al. 2014, and JMA-MLR and LSCE-FFNN by Denvil-Sommer et al. 2019) that predict *p*CO₂ based on the instantaneous physical and biological variables without regard for past states. There is thus a need to explore methods that incorporate the past state into future state estimates. This includes assimilative modelling approaches, such as B-SOSE (Biogeochemical Southern Ocean State Estimate), which would also provide greater understanding of the driver for changes in surface *p*CO₂ (Verdy and Mazloff, 2017). These methods may be able to provide better constraints on *p*CO₂ in data-poor regions. However, these assimilative models are not yet in a stage to fit the data closely (Verdy and Mazloff, 2017).

5 Summary

Our study suggests that we may be reaching the limits of gap-filling methods’ abilities to reduce uncertainties, as shown by the limited incremental improvement in errors by the ensemble method we compare with established methods. Significant uncertainties still prevail across all gap-filling methods, most likely limited by the extent of basin-scale observational gaps in the Southern Hemisphere as well as sampling aliases in mesoscale intensive ocean regions. We propose ways in which the surface ocean CO₂ community can improve estimates within the bounds of the current observations, and make recommendations for future observations.

We introduce a new surface ocean $p\text{CO}_2$ gap-filling method that is a machine learning ensemble average of six two-step clustering-regression models (CSIR-ML6 version 2019a). An exhaustive search process was used to find the best K-means clustering configuration which was used alongside the Fay and McKinley (2014) oceanic CO_2 biomes. The regression models applied to each clustering method are support vector regression, feed-forward neural-networks and gradient boosting machines. We show that the ensemble average of the six methods marginally outperforms each of its members, thus promoting the idea that averaging model estimates, each with different strengths and weaknesses, results in an improvement in the overall estimates.

The CSIR-ML6 (version 2019a) approach was compared to validation data alongside four other methods from the SOCOM intercomparison study (Rödenbeck et al. 2015). Our new method marginally outperformed the SOCOM methods when comparing RMSE scores for the validation data but fared equally on biases. Despite this improvement, all methods had errors of roughly the same magnitude, suggesting that the methods are resolving $p\text{CO}_2$ equally outside the bounds of the training data.

Closer assessment of the spatial distribution of errors shows that there is spatial coherence between regression approaches for the Northern Hemisphere. Some of these errors coincide with regions of high dynamic variability or complex biogeochemistry, suggesting that increasing the spatial and temporal resolution of gap-filling methods could improve estimates. Moreover, introducing additional feature-variables for regression, such as eddy kinetic energy, may improve estimates in these regions.

A comparison of the distribution of mismatches in $p\text{CO}_2$ between gap-filling methods shows that there are regions (primarily in the Southern Hemisphere) where the compared methods, as an ensemble, cannot resolve interannual variability of $p\text{CO}_2$ and as such, trends analyses in those regions should be interpreted with caution. These large mismatches likely occur due to amplification of algorithm specific biases in data-sparse areas. We suggest that an ensemble with data density-driven weighting for model complexity could be a way to reduce potential overfitting in data-sparse regions. We also urge the community to focus on incorporating new measurements from autonomous platforms such as the $p\text{CO}_2$ derived from pH measured by biogeochemical Argo floats, and new platforms such as $p\text{CO}_2$ capable Wavegliders.

In closing, we suggest that it is time to consider another SOCOM-like intercomparison. Several new methods have been developed since the last intercomparison and the addition of these would improve the robustness of ensemble average flux estimates. Further, the authors of the SOCOM intercomparison suggest that a future intercomparison should include a comparison of methods using simulated data, a method to overcome the limitation of the lack of data to test the estimates.

Code and data availability

Supporting code is available in Supplementary Materials. Data (global surface ocean $p\text{CO}_2$ from CSIR-ML6 version 2019a) is available at OCADS (link specified once the manuscript is accepted).

Author contributions

LG is the lead author and developed the method and wrote the manuscript. ADL contributed to the model assessment and contributed in editing the manuscript. SK contributed to the initial conceptualisation of the methods and proofread the manuscript. PMSM contributed to the development of the manuscript and its reviews.

Acknowledgements

This work is part of a Post-doctoral research fellowship funded by the CSIR's Southern Ocean Carbon - Climate Observatory (SOCCO) through financial support from the Department of Science and Technology (DST) and the National Research Foundation (NRF) and hosted at the MaRe Institute at UCT. We acknowledge the support from the Centre for High-Performance Computing (CSIR-CHPC). The Surface Ocean CO₂ Atlas (SOCAT) is an international effort, endorsed by the International Ocean Carbon Coordination Project (IOCCP), the Surface Ocean Lower Atmosphere Study (SOLAS) and the Integrated Marine Biogeochemistry and Ecosystem Research program (IMBER), to deliver a uniformly quality-controlled surface ocean CO₂ database. The many researchers and funding agencies responsible for the collection of data and quality control are thanked for their contributions to SOCAT. The lead author would also like to thank Nicolas Gruber (ETH) for allowing LG to correct the manuscript while under his employment.

References

- Bakker, D. C. E., Hoppema, M., Schr, M., Geibert, W. and Baar, H. J. W. De: A rapid transition from ice covered CO₂-rich waters to a biologically mediated CO₂ sink in the eastern Weddell Gyre, *Biogeosciences*, 5, 1373–1386, 2008.
- Bakker, D. C. E., Pfeil, B., Landa, C. S., Metzl, N., O'Brien, K. M., Olsen, A., Smith, K., Cosca, C., Harasawa, S., Jones, S. D., Nakaoka, S., Nojiri, Y., Schuster, U., Steinhoff, T., Sweeney, C., Takahashi, T., Tilbrook, B., Wada, C., Wanninkhof, R. H., Alin, S. R., Balestrini, C. F., Barbero, L., Bates, N. R., Bianchi, A. A., Bonou, F., Boutin, J., Bozec, Y., Burger, E. F., Cai, W.-J., Castle, R. D., Chen, L., Chierici, M., Currie, K., Evans, W., Featherstone, C., Feely, R. A., Fransson, A., Goyet, C., Greenwood, N., Gregor, L., Hankin, S., Hardman-Mountford, N. J., Harlay, J., Hauck, J., Hoppema, M., Humphreys, M. P., Hunt, C. W., Huss, B., Ibáñez, J. S. P., Johannessen, T., Keeling, R., Kitidis, V., Körtzinger, A., Kozyr, A., Krasakopoulou, E., Kuwata, A., Landschützer, P., Lauvset, S. K., Lefèvre, N., Lo Monaco, C., Manke, A., Mathis, J. T., Merlivat, L., Millero, F. J., Monteiro, P. M. S., Munro, D. R., Murata, A., Newberger, T., Omar, A. M., Ono, T., Paterson, K., Pearce, D., Pierrot, D., Robbins, L. L., Saito, S., Salisbury, J., Schlitzer, R., Schneider, B., Schweitzer, R., Sieger, R., Skjelvan, I., Sullivan, K. F., Sutherland, S. C., Sutton, A. J., Tadokoro, K., Telszewski, M., Tuma, M., Van Heuven, S. M. A. C., Vandemark, D., Ward, B., Watson, A. J. and Xu, S.: A multi-decade record of high-quality fCO₂ data in version 3 of the Surface Ocean CO₂ Atlas (SOCAT), *Earth Syst. Sci. Data*, 8(2), 383–413, doi:10.5194/essd-8-383-2016, 2016.
- Banzon, V., Smith, T. M., Mike Chin, T., Liu, C., & Hankins, W. (2016). A long-term record of blended satellite and in situ sea-surface temperature for climate monitoring, modeling and environmental studies. *Earth System Science Data*, 8(1), 165–176. <https://doi.org/10.5194/essd-8-165-2016>
- Bates, N. R.: Interannual variability of oceanic CO₂ and biogeochemical properties in the Western North Atlantic subtropical gyre, *Deep. Res. Part II Top. Stud. Oceanogr.*, 48(8–9), 1507–1528, doi:10.1016/S0967-0645(00)00151-X, 2001.
- Boutin, J. and Merlivat, L.: Sea surface fCO₂ measurements in the Southern Ocean from CARIOCA Drifters, ,

doi:10.3334/CDIAC/OTG.CARIOCA, 2013.

- Carter, B. R., Feely, R. A., Williams, N. L., Dickson, A. G., Fong, M. B. and Takeshita, Y.: Updated methods for global locally interpolated estimation of alkalinity, pH, and nitrate, *Limnol. Oceanogr. Methods*, 16(2), 119–131, doi:10.1002/lom3.10232, 2018.
- Castelvecchi, D.: Can we open the black box of AI?, *Nature*, 538(7623), 20–23, doi:10.1038/538020a, 2016.
- Crone, S. F. and Finlay, S.: Instance sampling in credit scoring: An empirical study of sample size and balancing, *Int. J. Forecast.*, 28(1), 224–238, doi:10.1016/j.ijforecast.2011.07.006, 2012.
- Dee, D. P., Uppala, S. M., Simmons, A. J., Berrisford, P., Poli, P., Kobayashi, S., Andrae, U., Balmaseda, M. A., Balsamo, G., Bauer, P., Bechtold, P., Beljaars, A. C. M., van de Berg, L., Bidlot, J., Bormann, N., Delsol, C., Dragani, R., Fuentes, M., Geer, A. J., Haimberger, L., Healy, S. B., Hersbach, H., Holm, E. V., Isaksen, L., Kallberg, P., Köhler, M., Matricardi, M., McNally, A. P., Monge-Sanz, B. M., Morcrette, J. J., Park, B. K., Peubey, C., de Rosnay, P., Tavolato, C., Thepaut, J. N., Vitart, F., Hólm, E. V., Kållberg, P. and Thépaut, J. N.: The ERA-Interim reanalysis: Configuration and performance of the data assimilation system, *Q. J. R. Meteorol. Soc.*, 137(656), 553–597, doi:10.1002/qj.828, 2011.
- Denvil-Sommer, A., Gehlen, M., Vrac, M. and Mejia, C.: FFNN-LSCE: A two-step neural network model for the reconstruction of surface ocean pCO₂ over the Global Ocean, *Geosci. Model Dev. Discuss.*, (November), 1–27, doi:10.5194/gmd-2018-247, 2018.
- Dickson, A. G. (Andrew G., Sabine, C. L., Christian, J. R. and North Pacific Marine Science Organization.: Guide to best practices for ocean CO₂ measurements, North Pacific Marine Science Organization. [online] Available from: <https://www.oceanbestpractices.net/handle/11329/249> (Accessed 16 February 2019), 2007.
- Dore, J. E., Lukas, R., Sadler, D. W., Church, M. J. and Karl, D. M.: Physical and biogeochemical modulation of ocean acidification in the central North Pacific, *Proc. Natl. Acad. Sci.*, 106(30), 12235–12240, doi:10.1073/pnas.0906044106, 2009.
- Drucker, H., Burges, C. J. C., Kaufman, L., Smola, A. and Vapnik, V. N.: Support vector regression machines, *Adv. Neural Inf. Process. Syst.* 9, 1, 155–161, doi:10.1.1.10.4845, 1997.
- du Plessis, M., Swart, S., Anson, I. J. and Mahadevan, A.: Submesoscale processes promote seasonal restratification in the Subantarctic Ocean, *J. Geophys. Res.*, 122(4), 2960–2975, doi:10.1002/2016JC012494, 2017.
- du Plessis, M., Swart, S., Anson, I. J., Mahadevan, A. and Thompson, A. F.: Southern Ocean seasonal restratification delayed by submesoscale wind-front interactions, *J. Phys. Oceanogr.*, JPO-D-18-0136.1, doi:10.1175/JPO-D-18-0136.1, 2019.
- Fay, A. R. and McKinley, G. A.: Global open-ocean biomes: Mean and temporal variability, *Earth Syst. Sci. Data*, 6(2), 273–284, doi:10.5194/essd-6-273-2014, 2014.
- Friedman, J. H.: Greedy function approximation: A gradient boosting machine, *Ann. Stat.*, 29(5), 1189–1232, doi:10.1214/aos/1013203451, 2001.
- Geurts, P., Ernst, D. and Wehenkel, L.: Extremely randomized trees, *Mach. Learn.*, 63(1), 3–42, doi:10.1007/s10994-006-6226-1, 2006.
- Goddijn–Murphy, L., Woolf, D. K., Land, P. E., Shutler, J. D., & Donlon, C. J. (2015). The OceanFlux Greenhouse Gases methodology for deriving a sea surface climatology of CO₂ fugacity in support of air-sea gas flux studies. *Ocean Science*, 11(4), 519–541. <https://doi.org/10.5194/os-11-519-2015>
- Goddijn–Murphy, L., Woolf, D. K., Callaghan, A. H., Nightingale, P. D., & Shutler, J. D. (2016). A reconciliation of empirical and mechanistic models of the air–sea gas transfer velocity. *Journal of Geophysical Research: Oceans*, 121(1), 818–835. <https://doi.org/10.1002/2015JC011096>
- Good, S. A., Martin, M. J. and Rayner, N. A.: EN4: Quality controlled ocean temperature and salinity profiles and monthly objective analyses with uncertainty estimates, *J. Geophys. Res. Ocean.*, 118(12), 6704–6716, doi:10.1002/2013JC009067, 2013.
- Gray, A. R., Johnson, K. S., Bushinsky, S. M., Riser, S. C., Russell, J. L., Talley, L. D., Wanninkhof, R. H., Williams, N. L. and Sarmiento, J. L.: Autonomous Biogeochemical Floats Detect Significant Carbon Dioxide Outgassing in the High-Latitude Southern Ocean, *Geophys. Res. Lett.*, 45(17), 9049–9057, doi:10.1029/2018GL078013, 2018.
- Gregor, L., Kok, S. and Monteiro, P. M. S.: Empirical methods for the estimation of Southern Ocean CO₂: support vector

- and random forest regression, *Biogeosciences*, 14(23), 5551–5569, doi:10.5194/bg-14-5551-2017, 2017.
- Gregor, L., Kok, S. and Monteiro, P. M. S.: Interannual drivers of the seasonal cycle of CO₂ in the Southern Ocean, *Biogeosciences*, 15(8), 2361–2378, doi:10.5194/bg-15-2361-2018, 2018.
- Hain, M. P., Sigman, D. M., Higgins, J. A. and Haug, G. H.: The effects of secular calcium and magnesium concentration changes on the thermodynamics of seawater acid/base chemistry: Implications for Eocene and Cretaceous ocean carbon chemistry and buffering, *Global Biogeochem. Cycles*, 29(5), 517–533, doi:10.1002/2014GB004986, 2015.
- Hastie, T., Tibshirani, R. and Friedman, J. H.: *The Elements of Statistical Learning: Data mining, Inference, and Prediction*, Second Edi., Springer., 2009.
- Holte, J., Talley, L. D., Gilson, J. and Roemmich, D.: An Argo mixed layer climatology and database, *Geophys. Res. Lett.*, 44(11), 5618–5626, doi:10.1002/2017GL073426, 2017.
- Hoyer, S. and Hamman, J. J.: xarray: N-D labeled Arrays and Datasets in Python, *J. Open Res. Softw.*, 5, 1–6, doi:10.5334/jors.148, 2017.
- Iida, Y., Kojima, A., Takatani, Y., Nakano, T., Sugimoto, H., Midorikawa, T. and Ishii, M.: Trends in pCO₂ and sea–air CO₂ flux over the global open oceans for the last two decades, *J. Oceanogr.*, 71(6), 637–661, doi:10.1007/s10872-015-0306-4, 2015.
- Ishii, M., Inoue, H. Y., Matsueda, H. and Tanoue, E.: Close coupling between seasonal biological production and dynamics of dissolved inorganic carbon in the Indian Ocean sector and the western Pacific Ocean sector of the Antarctic Ocean, *Deep. Res. Part I Oceanogr. Res. Pap.*, 45(7), 1187–1209, doi:10.1016/S0967-0637(98)00010-7, 1998.
- Jones, S. D., Le Quéré, C., Rödenbeck, C., Manning, A. C. and Olsen, A.: A statistical gap-filling method to interpolate global monthly surface ocean carbon dioxide data, *J. Adv. Model. Earth Syst.*, 7(4), 1554–1575, doi:10.1002/2014MS000416, 2015.
- Khatiwal, S., Tanhua, T., Mikaloff Fletcher, S. E., Gerber, M., Doney, S. C., Graven, H. D., Gruber, N., McKinley, G. A., Murata, A., Ríos, A. F. and Sabine, C. L.: Global ocean storage of anthropogenic carbon, *Biogeosciences*, 10(4), 2169–2191, doi:10.5194/bg-10-2169-2013, 2013.
- Landschützer, P., Gruber, N. and Bakker, D. C. E.: An updated observation-based global monthly gridded sea surface pCO₂ and air-sea CO₂ flux product from 1982 through 2015 and its monthly climatology (NCEI Accession 0160558). Version 2.2. NOAA National Centers for Environmental Information. Dataset., 2017.
- Landschützer, P., Gruber, N., Bakker, D. C. E., Stemmler, I. and Six, K. D.: Strengthening seasonal marine CO₂ variations due to increasing atmospheric CO₂, *Nat. Clim. Chang.*, 8(2), 146–150, doi:10.1038/s41558-017-0057-x, 2018.
- Landschützer, P., Gruber, N. and Bakker, D. C. E.: Decadal variations and trends of the global ocean carbon sink, *Global Biogeochem. Cycles*, 30(10), 1396–1417, doi:10.1002/2015GB005359, 2016.
- Landschützer, P., Gruber, N., Bakker, D. C. E. and Schuster, U.: Recent variability of the global ocean carbon sink, *Glob. Planet. Change*, 927–949, doi:10.1002/2014GB004853. Received, 2014.
- Landschützer, P., Gruber, N., Haumann, F. A., Rödenbeck, C., Bakker, D. C. E., Van Heuven, S. M. A. C., Hoppema, M., Metzl, N., Sweeney, C., Takahashi, T. T., Tilbrook, B. and Wanninkhof, R. H.: The reinvigoration of the Southern Ocean carbon sink, *Science* (80-.), 349(6253), 1221–1224, doi:10.1126/science.aab2620, 2015.
- Lebehot, A. D., Halloran, P. R., Watson, A. J., McNeill, D. J., Ford, D. A., Landschützer, P., Lauvset, S. K., Schuster, U.: Reconciling observation and model trends in North Atlantic surface CO₂. *Global Biogeochemical Cycles*, 33, doi:10.1029/2019GB006186, 2019
- Le Quéré, C., Andrew, R. M., Friedlingstein, P., Sitch, S., Hauck, J., Pongratz, J., Pickers, P., Korsbakken, J. I., Peters, G. P. and Canadell, J. G.: Global carbon budget 2018, *Earth Syst. Sci. Data*, 10, 2141–2194, doi:10.5194/essd-10-2141-2018, 2018.
- Lefèvre, N., Watson, A. J. and Watson, A. R.: A comparison of multiple regression and neural network techniques for mapping in situ pCO₂ data, *Tellus B Chem. Phys. Meteorol.*, 57(5), 375–384, doi:10.3402/tellusb.v57i5.16565, 2005.
- Lenton, A., Metzl, N., Takahashi, T. T., Kuchinke, M., Matear, R. J., Roy, T., Sutherland, S. C., Sweeney, C. and Tilbrook, B.: The observed evolution of oceanic pCO₂ and its drivers over the last two decades, *Global Biogeochem. Cycles*, 26(2), 1–14, doi:10.1029/2011GB004095, 2012.

- Lenton, A., Tilbrook, B., Law, R. M., Bakker, D. C. E., Doney, S. C., Gruber, N., Ishii, M., Hoppema, M., Lovenduski, N. S., Matear, R. J., McNeil, B. I., Metzl, N., Fletcher, S. E. M., Monteiro, P. M. S., Rödenbeck, C., Sweeney, C. and Takahashi, T. T.: Sea-air CO₂ fluxes in the Southern Ocean for the period 1990-2009, *Biogeosciences*, 10(6), 4037–4054, doi:10.5194/bg-10-4037-2013, 2013.
- Lueker, T. J., Dickson, A. G. and Keeling, C. D.: Ocean pCO₂ calculated from dissolved inorganic carbon, alkalinity, and equations for K₁ and K₂: Validation based on laboratory measurements of CO₂ in gas and seawater at equilibrium, *Mar. Chem.*, 70, 105–119, doi:10.1016/S0304-4203(00)00022-0, 2000.
- Maritorena, S., Fanton D'andon, O. H., Mangin, A. and Siegel, D. A.: Merged satellite ocean color data products using a bio-optical model: Characteristics, benefits and issues, *Remote Sens. Environ.*, 114, 1791–1804, doi:10.1016/j.rse.2010.04.002, 2010.
- Masarie, K. A., Peters, W., Jacobson, A. R. and Tans, P. P.: ObsPack: A framework for the preparation, delivery, and attribution of atmospheric greenhouse gas measurements, *Earth Syst. Sci. Data*, 6(2), 375–384, doi:10.5194/essd-6-375-2014, 2014.
- Mazloff, M. R., Cornuelle, B. D., Gille, S. T. and Verdy, A.: Correlation Lengths for Estimating the Large-Scale Carbon and Heat Content of the Southern Ocean, *J. Geophys. Res. Ocean.*, 1–35, doi:10.1002/2017JC013408, 2018.
- McKinley, G. A., Pilcher, D. J., Fay, A. R., Lindsay, K., Long, M. C. and Lovenduski, N. S.: Timescales for detection of trends in the ocean carbon sink, *Nature*, 530(7591), 469–472, doi:10.1038/nature16958, 2016.
- McKinley, G. A., Pilcher, D. J., Fay, A. R., Lindsay, K., Long, M. C. and Lovenduski, N. S.: Timescales for detection of trends in the ocean carbon sink, *Nature*, 530(7591), 469–472, doi:10.1038/nature16958, 2016.
- Mckinney, W.: Data Structures for Statistical Computing in Python. [online] Available from: <https://conference.scipy.org/proceedings/scipy2010/pdfs/mckinney.pdf> (Accessed 16 February 2019), 2010.
- Monteiro, P. M. S.: A Global Sea Surface Carbon Observing System: Assessment of Changing Sea Surface CO₂ and Air-Sea CO₂ Fluxes, *Proc. Ocean. Sustain. Ocean Obs. Inf. Soc.*, (1), 702–714, doi:10.5270/OceanObs09.cwp.64, 2010.
- Olsen, A., Key, R. M., van Heuven, S., Lauvset, S. K., Velo, A., Lin, X., Schirnack, C., Kozyr, A., Tanhua, T., Hoppema, M., Jutterström, S., Steinfeldt, R., Jeansson, E., Ishii, M., Pérez, F. F. and Suzuki, T.: The Global Ocean Data Analysis Project version 2 (GLODAPv2) – an internally consistent data product for the world ocean, *Earth Syst. Sci. Data*, 8(2), 297–323, doi:10.5194/essd-8-297-2016, 2016.
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, C., Thirion, B., Grisel, O., Blondel, M., Prettenhoffer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A. and Cournapeau, D.: Scikit-learn: Machine learning in Python, *J. Mach. Learn. Res.*, 12, 2825–2830, doi:10.1007/s13398-014-0173-7.2, 2011.
- Reynolds, R. W., Smith, T. M., Liu, C., Chelton, D. B., Casey, K. S., & Schlax, M. G. (2007). Daily high-resolution-blended analyses for sea surface temperature. *Journal of Climate*, 20(22), 5473–5496. <https://doi.org/10.1175/2007JCLI1824.1>
- Rio, M.-H., Mulet, S. and Picot, N.: Beyond GOCE for the ocean circulation estimate: Synergetic use of altimetry, gravimetry, and in situ data provides new insight into geostrophic and Ekman currents, *Geophys. Res. Lett.*, 41(24), 8918–8925, doi:10.1002/2014GL061773, 2014.
- Ritter, R., Landschützer, P., Gruber, N., Fay, A. R., Iida, Y., Jones, S., Nakaoka, S., Park, G.-H., Peylin, P., Rödenbeck, C., Rodgers, K. B., Shutler, J. D. and Zeng, J.: Observation-Based Trends of the Southern Ocean Carbon Sink, *Geophys. Res. Lett.*, 44(24), 12,339-12,348, doi:10.1002/2017GL074837, 2017.
- Rödenbeck, C., Bakker, D. C. E., Gruber, N., Iida, Y., Jacobson, A. R., Jones, S., Landschützer, P., Metzl, N., Nakaoka, S., Olsen, A., Park, G.-H., Peylin, P., Rodgers, K. B., Sasse, T. P., Schuster, U., Shutler, J. D., Valsala, V., Wanninkhof, R. and Zeng, J.: Data-based estimates of the ocean carbon sink variability – first results of the Surface Ocean pCO₂ Mapping intercomparison (SOCOM), *Biogeosciences*, 12(23), 7251–7278, doi:10.5194/bg-12-7251-2015, 2015.
- Rödenbeck, C., Bakker, D. C. E., Metzl, N., Olsen, A., Sabine, C., Cassar, N., Reum, F., Keeling, R. F. and Heimann, M.: Interannual sea–air CO₂ flux variability from an observation-driven ocean mixed-layer scheme, *Biogeosciences*, 11(17), 4599–4613, doi:10.5194/bg-11-4599-2014, 2014.
- Sabine, C. L., Feely, R. A., Gruber, N., Key, R. M., Lee, K., Bullister, J. L., Wanninkhof, R. H., Wong, C. S., Wallace, D.

- W. R., Tilbrook, B., Millero, F. J., Peng, T., Kozyr, A., Ono, T. and Ríos, A. F.: The Oceanic Sink for Anthropogenic CO₂, *Science* (80-.), 305(5682), 367–371, doi:10.1126/science.1097403, 2004.
- Sabine, C. L., Hankin, S., Koyuk, H., Bakker, D. C. E., Pfeil, B., Olsen, A., Metzl, N., Kozyr, A., Fassbender, A. J., Manke, A., Malczyk, J., Akl, J., Alin, S. R., Bellerby, R. G. J., Borges, A. V., Boutin, J., Brown, P. J., Cai, W.-J., Chavez, F. P., Chen, A., Cosca, C. E., Feely, R. A., González-Dávila, M., Goyet, C., Hardman-Mountford, N. J., Heinze, C., Hoppema, M., Hunt, C. W., Hydes, D., Ishii, M., Johannessen, T., Key, R. M., Körtzinger, A., Landschützer, P., Lauvset, S. K., Lefèvre, N., Lenton, A., Lourantou, A., Merlivat, L., Midorikawa, T., Mintrop, L., Miyazaki, C., Murata, A., Nakadate, A., Nakano, Y., Nakaoka, S., Nojiri, Y., Omar, A. M., Padin, X. A., Park, G.-H., Paterson, K., Perez, F. F., Pierrot, D., Poisson, A., Ríos, A. F., Salisbury, J., Santana-Casiano, J. M., Sarma, V. V. S., Schlitzer, R., Schneider, B., Schuster, U., Sieger, R., Skjelvan, I., Steinhoff, T., Suzuki, T., Takahashi, T. T., Tedesco, K., Telszewski, M., Thomas, H., Tilbrook, B., Vandemark, D., Veness, T., Watson, A. J., Weiss, R., Wong, C. S. and Yoshikawa-Inoue, H.: Surface Ocean CO₂ Atlas (SOCAT) gridded data products, *Earth Syst. Sci. Data*, 5(1), 145–153, doi:10.5194/essd-5-145-2013, 2013.
- Sasse, T. P., McNeil, B. I. and Abramowitz, G.: A novel method for diagnosing seasonal to inter-annual surface ocean carbon dynamics from bottle data using neural networks, *Biogeosciences*, 10(6), 4319–4340, doi:10.5194/bg-10-4319-2013, 2013.
- Takahashi, T. T., Olafsson, J., Goddard, J. G., Chipman, D. W. and Sutherland, S. C.: Seasonal variation of CO₂ and nutrients in the high-latitude surface oceans: A comparative study, *Global Biogeochem. Cycles*, 7(4), 843–878, doi:10.1029/93GB02263, 1993.
- Takahashi, T., Sutherland, S. C. and Kozyr, A.: Global Ocean Surface Water Partial Pressure of CO₂ Database: Measurements Performed During 1957-2017 (Version 2017), ORNL/CDIAC-160, NDP-088(V2017). (NCEI Access. 0160492). Version 4.4. NOAA Natl. Centers Environ. Information. Dataset., doi:10.3334/CDIAC/OTG.NDP088(V2015), 2017.
- Takahashi, T. T., Sutherland, S. C., Wanninkhof, R. H., Sweeney, C., Feely, R. A., Chipman, D. W., Hales, B., Friederich, G. E., Chavez, F. P., Sabine, C. L., Watson, A. J., Bakker, D. C. E., Schuster, U., Metzl, N., Yoshikawa-Inoue, H., Ishii, M., Midorikawa, T., Nojiri, Y., Körtzinger, A., Steinhoff, T., Hoppema, M., Olafsson, J., Arnarson, T. S., Tilbrook, B., Johannessen, T., Olsen, A., Bellerby, R. G. J., Wong, C. S., Delille, B., Bates, N. R. and de Baar, H. J. W.: Climatological mean and decadal change in surface ocean pCO₂, and net sea–air CO₂ flux over the global oceans, *Deep. Res. Part II Top. Stud. Oceanogr.*, 56(8–10), 554–577, doi:10.1016/j.dsr2.2008.12.009, 2009.
- Taylor, K. E.: Summarizing multiple aspects of model performance in a single diagram, *J. Geophys. Res. Atmos.*, 106(D7), 7183–7192, doi:10.1029/2000JD900719, 2001.
- Telszewski, M., Chazottes, A., Schuster, U., Watson, A. J., Moulin, C., Bakker, D. C. E., González-Dávila, M., Johannessen, T., Kortzinger, A., Luger, H., Olsen, A., Omar, A. M., Padin, X. A., Ríos, A. F., Steinhoff, T., Santana-Casiano, M., Wallace, D. W. R. and Wanninkhof, R. H.: Estimating the monthly pCO₂ distribution in the North Atlantic using a self-organizing neural network, *Biogeosciences*, 6(8), 1405–1421, doi:10.5194/bg-6-1405-2009, 2009.
- Thomalla, S. J., Fauchereau, N., Swart, S. and Monteiro, P. M. S.: Regional scale characteristics of the seasonal cycle of chlorophyll in the Southern Ocean, *Biogeosciences*, 8(10), 2849–2866, doi:10.5194/bg-8-2849-2011, 2011.
- Verdy, A. and Mazloff, M. R.: A data assimilating model for estimating Southern Ocean biogeochemistry, *J. Geophys. Res. Ocean.*, 122(9), 6968–6988, doi:10.1002/2016JC012650, 2017.
- Wanninkhof, R. H., Bakker, D., Bates, N., Steinhoff, T. and Sutton, A.: Incorporation of Alternative Sensors in the SOCAT Database and Adjustments to Dataset Quality Control Flags, (2009), 1–26, doi:10.3334/CDIAC/OTG.SOCAT, 2013.
- Wanninkhof, R. H., Park, G.-H., Takahashi, T. T., Sweeney, C., Feely, R. A., Nojiri, Y., Gruber, N., Doney, S. C., McKinley, G. A., Lenton, A., Le Quééré, C., Heinze, C., Schwinger, J., Graven, H. D. and Khatiwala, S.: Global ocean carbon uptake: magnitude, variability and trends, *Biogeosciences*, 10(3), 1983–2000, doi:10.5194/bg-10-1983-2013, 2013.
- Weiss, R.: Carbon dioxide in water and seawater: the solubility of a non-ideal gas, *Mar. Chem.*, 2(3), 203–215, doi:10.1016/0304-4203(74)90015-2, 1974.
- Williams, N. L., Juranek, L. W., Feely, R. A., Johnson, K. S., Sarmiento, J. L., Talley, L. D., Dickson, A. G., Gray, A. R., Wanninkhof, R. H., Russell, J. L., Riser, S. C. and Takeshita, Y.: Calculating surface ocean pCO₂ from biogeochemical Argo floats equipped with pH: An uncertainty analysis, *Global Biogeochem. Cycles*, 31(3), 591–604,

doi:10.1002/2016GB005541, 2017.

- Woolf, D. K., Land, P. E., Shutler, J. D., Goddijn–Murphy, L., & Donlon, C. J. (2016). On the calculation of air-sea fluxes of CO₂ in the presence of temperature and salinity gradients. *Journal of Geophysical Research: Oceans*, 121(2), 1229–1248. <https://doi.org/10.1002/2015JC011427>
- Zeng, J., Matsunaga, T., Saigusa, N., Shirai, T., Nakaoka, S. I. and Tan, Z. H.: Technical note: Evaluation of three machine learning models for surface ocean CO₂ mapping, *Ocean Sci.*, 13(2), 303–313, doi:10.5194/os-13-303-2017, 2017.
- Zeng, J., Nojiri, Y., Landschützer, P., Telszewski, M. and Nakaoka, S.: A global surface ocean fCO₂ climatology based on a feed-forward neural network, *J. Atmos. Ocean. Technol.*, 31(8), 1838–1849, doi:10.1175/JTECH-D-13-00137.1, 2014.