Geosci. Model Dev., 12, 5113–5136, 2019
https://doi.org/10.5194/gmd-12-5113-2019

Geoscientific
Model Development
Open Access

A comparative assessment of the uncertainties of global surface ocean CO_2 estimates using a machine-learning ensemble (CSIR-ML6 version 2019a) – have we hit the wall?

Luke Gregor[1,2,3], **Alice D. Lebehot**[1,2], **Schalk Kok**[4], **and Pedro M. Scheel Monteiro**[1]

[1]SOCCO, Council for Scientific and Industrial Research, Cape Town, 7700, South Africa
[2]MaRe, Marine Research Institute, University of Cape Town, Cape Town, 7700, South Africa
[3]Environmental Physics, Institute of Biogeochemistry and Pollutant Dynamics, ETH Zürich, 8092 Zürich, Switzerland
[4]Department of Mechanical & Aeronautical Engineering, University of Pretoria, Pretoria, 0028, South Africa

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

Received: 14 February 2019 – Discussion started: 5 April 2019
Revised: 5 November 2019 – Accepted: 8 November 2019 – Published: 10 December 2019

Abstract. Over the last decade, advanced statistical inference and machine learning have been used to fill the gaps in sparse surface ocean CO_2 measurements (Rödenbeck et al., 2015). The estimates from these methods have been used to constrain seasonal, interannual and decadal variability in sea–air CO_2 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 root mean square error (RMSE) boundary: "the wall", which suggests that $p\mathrm{CO}_2$ 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, Council for Scientific and Industrial Research – Machine Learning ensemble with Six members), 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 $p\mathrm{CO}_2$. We show that disagreement between methods in the South Atlantic, southeastern Pacific and parts of the Southern Ocean is too large to interpret the interannual variability with confidence. We conclude that improvements in surface ocean $p\mathrm{CO}_2$ 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 and (3) successfully incorporating $p\mathrm{CO}_2$ 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 (CO_2) emissions (Sabine et al., 2004; Khatiwala et al., 2013; McKinley et al., 2016). While the mean state in the global contemporary marine CO_2 uptake is a widely used benchmark (Le Quéré et al., 2018), underlying assumptions and limited confidence regarding the variability and long-term evolution of this sink persist. Sparse observations of surface ocean CO_2 during winter and in large inaccessible regions have 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 sampling effort and the ongoing development of autonomous observational platforms (e.g. biogeochemical Argo floats and Wavegliders) have improved confidence of interannual estimates of ocean CO_2 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 CO_2 uptake by the oceans for periods and regions with poor or no observational coverage (Wanninkhof et al., 2013b; Rödenbeck et al., 2015; Verdy and Mazloff, 2017). Ocean biogeochemical models are able to capture the general global trend in increasing oceanic CO_2 uptake shown by observations but suffer from significant regional and interannual ($\sim 1\,\mathrm{Pg\,C\,yr^{-1}}$) differences in their estimates because these models cannot yet accurately parameterise the marine carbonate system at computationally feasible resolutions (Wanninkhof et al., 2013b). In recent years, data-based approaches, e.g. statistical interpolations and regression methods, have become a popular alternative to 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 CO_2 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 CO_2 ($p\mathrm{CO_2}$) in the surface ocean and proxies that may drive changes in surface ocean $p\mathrm{CO_2}$. Improved access to quality-controlled ship-based measurements of surface ocean CO_2 through the Surface Ocean CO_2 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).

1.1 The current state of machine learning in ocean CO_2 estimates

With the increase in the number of statistical estimates of surface ocean CO_2, the Surface Ocean CO_2 Mapping (SOCOM) community collated 14 of these methods in an intercomparison of "gap-filling" methods (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 14 methods, respectively (Rödenbeck et al., 2015). Two model-based approaches were also compared.

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 $p\mathrm{CO_2}$ 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. 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 timescales, there was agreement on the direction of trends between gap-filling methods.

1.2 Measuring the uncertainty of estimates?

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 are sparse (Rödenbeck et al., 2015; Bakker et al., 2016). The standard use of root mean squared error (RMSE) and bias as measures of uncertainty gives larger weighting to 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.

Previous studies have compared their methods' estimates to independent datasets, where measurements of $p\mathrm{CO_2}$ are not included in the SOCAT datasets (Landschützer et al., 2013, 2014; Jones et al., 2015; Denvil-Sommer et al., 2019). These data serve as good validation data, particularly with the inclusion of derivations of $p\mathrm{CO_2}$ 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 d mean and 1/2° spatially). The authors found that the ensemble average performed slightly better than ensemble members, in agreement with ensemble averaging approaches previously used in ocean CO_2 studies (Khatiwala et al., 2013). On the other hand, 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

The main aim of this study is to present and evaluate a new machine-learning approach to estimate surface ocean $p\mathrm{CO_2}$. 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

There are two main components to this study: surface $p\mathrm{CO_2}$ 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 Japanese Meteorological Agency – multi-linear regression (JMA-MLR) and MPI-SOMFFN approaches, where data are grouped or clustered first, and then a regression algorithm is applied separately to each group or cluster. We use the ocean $\mathrm{CO_2}$ biomes by Fay and McKinley (2014) as an option for grouping. Alongside this grouping, we use an optimal K-means clustering configuration. Next, four nonlinear 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 Fig. 1:

1. In the first step (denoted as "K-means clustering" in Fig. 1), we generate climatological biomes using the oceanic $\mathrm{CO_2}$ 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.

2. Four regression algorithms are applied to each clustering configuration, resulting in 164 models (described by the "regression" section in Fig. 1). The test data (isolated from the model training procedure) are used to identify the best-performing clustering configuration with annually weighted bias, RMSE and R^{iav}. The four regression models for $\mathrm{CO_2}$ biomes and the four models from the best-performing clustering configuration (as indicated by the bold lines in Fig. 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.

3. The third step (as represented by the "K-fold testing" section in Fig. 1 and Sect. 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^{iav} for the complete SOCAT v5 dataset (rather than just one test split).

4. The fourth step compares the ensemble average estimates of surface ocean $p\mathrm{CO_2}$ with independent test data (that are not in SOCATv5, as represented by the "independent" section in Fig. 1), which allows testing the predictive ability of the ensemble method (Sect. 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.

2.2 Data: clustering, training and prediction

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, $p\mathrm{CO_2}$) and one or more feature variables that have samples that correspond with target samples, e.g. sea surface temperature (SST), Chl a and mixed-layer depth (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.

Here, we use surface ocean $p\mathrm{CO_2}$ calculated from the SOCAT v5 monthly gridded $f\mathrm{CO_2}$ (fugacity of $\mathrm{CO_2}$) product (hereinafter SOCAT v5, as shown in Fig. 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 $f\mathrm{CO_2}$, which is converted to $p\mathrm{CO_2}$ with

$$p\mathrm{CO_2} = f\mathrm{CO_2} \cdot \exp\left(P_{\mathrm{atm}}^{\mathrm{surf}} \cdot \frac{B + 2 \cdot \delta}{R \cdot T}\right)^{-1}, \tag{1}$$

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

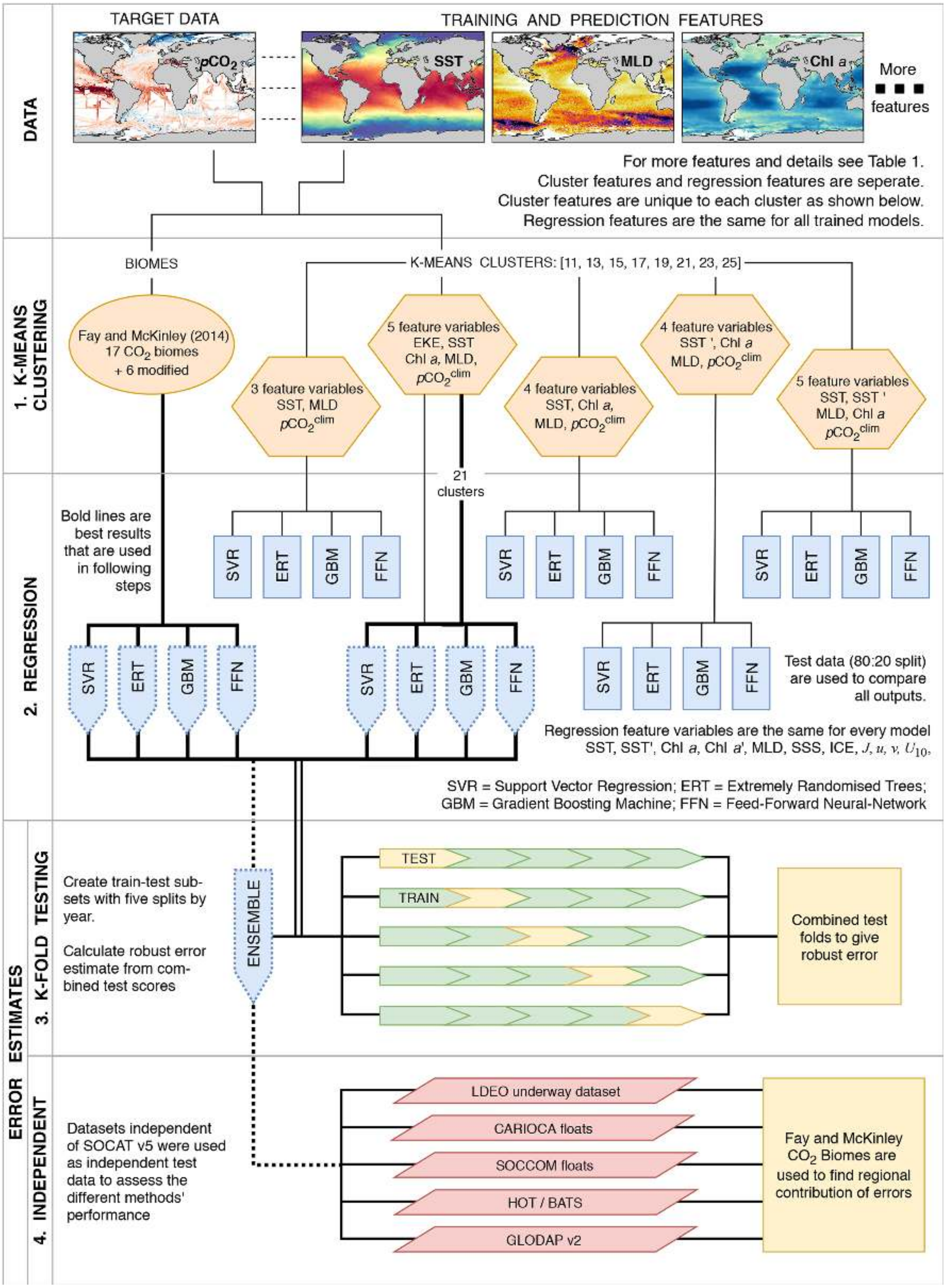
DATA
TARGET DATA
TRAINING AND PREDICTION FEATURES
pCO2
SST
MLD
Chl a
More ■ ■ ■ features
For more features and details see Table 1.
Cluster features and regression features are seperate.
Cluster features are unique to each cluster as shown below.
Regression features are the same for all trained models.
1. K-MEANS CLUSTERING
BIOMES
K-MEANS CLUSTERS: [11, 13, 15, 17, 19, 21, 23, 25]
Fay and McKinley (2014) 17 CO2 biomes + 6 modified
3 feature variables SST, MLD pCO2clim
5 feature variables EKE, SST Chl a, MLD, pCO2clim
4 feature variables SST, Chl a, MLD, pCO2clim
4 feature variables SST ', Chl a MLD, pCO2clim
5 feature variables SST, SST ' MLD, Chl a pCO2clim
21 clusters
2. REGRESSION
Bold lines are best results that are used in following steps
SVR
ERT
GBM
FFN
Test data (80:20 split) are used to compare all outputs.
Regression feature variables are the same for every model SST, SST', Chl a, Chl a', MLD, SSS, ICE, J, u, v, U10,
SVR = Support Vector Regression; ERT = Extremely Randomised Trees; GBM = Gradient Boosting Machine; FFN = Feed-Forward Neural-Network
ERROR ESTIMATES
3. K-FOLD TESTING
Create train-test sub-sets with five splits by year.
Calculate robust error estimate from com-bined test scores
ENSEMBLE
TEST
TRAIN
Combined test folds to give robust error
4. INDEPENDENT
Datasets independent of SOCAT v5 were used as independent test data to assess the different methods' performance
LDEO underway dataset
CARIOCA floats
SOCCOM floats
HOT / BATS
GLODAP v2
Fay and McKinley CO2 Biomes are used to find regional contribution of errors

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 (Sect. 2.1).

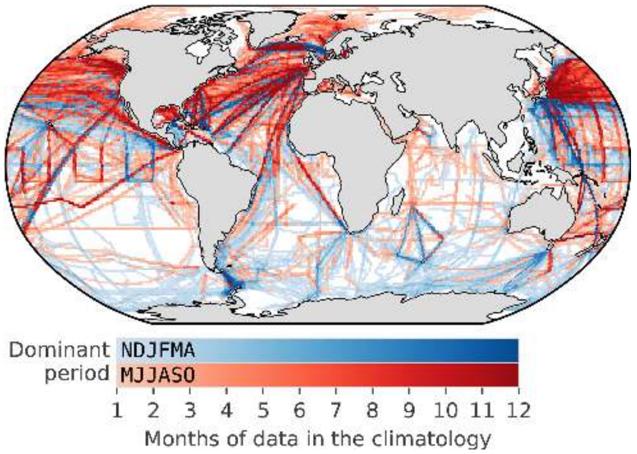

Figure 2. Map showing the distribution of the SOCAT v5 monthly gridded product (1982–2016) as a monthly climatology to show how well the seasonal cycle is represented (regardless of the year). The red shading shows grid points where the majority of data occur from May to October, and the blue shading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

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 $p\mathrm{CO}_2$ in the SOCAT database are thus different from surface temperature products used to estimate $p\mathrm{CO}_2$ 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 $p\mathrm{CO}_2$ 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 theoretical difference in $p\mathrm{CO}_2$ should be made using the empirical relationship between $p\mathrm{CO}_2$ and temperature (Takahashi et al., 1993). While this merits further coordinated consideration by the marine CO_2 observation community, we do not apply such a temperature correction in this study, as we aim to be consistent with the earlier $p\mathrm{CO}_2$ estimates from the SOCOM intercomparison (Rödenbeck et al., 2015). However, we do present the potential impact of this discrepancy in Sect. S2.4.

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 Sect. S1 for details). All feature variables are gridded to a monthly frequency onto a global 1° × 1° resolution grid. Thereafter, data processing steps are applied as shown in Table 1 and described in detail in the Supplement (Sect. 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.

In this paragraph, we briefly describe the data processing steps shown in Table 1; detailed product descriptions and in-depth processing steps are in Sect. 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 (1982–1997) are filled with the climatology (1998–2016), and high-latitude winter regions (where there is no climatology for Chl a) are 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 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 $p\mathrm{CO}_2$ is calculated with $p\mathrm{CO}_2 = x\mathrm{CO}_2^{\mathrm{atm}} \times P^{\mathrm{atm}}$, where $x\mathrm{CO}_2^{\mathrm{atm}}$ 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 Sect. S1. The climatology of eddy kinetic energy (EKE$^{\mathrm{clim}}$) 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 $\underline{u} - u$ and similarly with v (Table 1).

2.3 Clustering and biomes

The seasonal and interannual variability of global surface ocean $p\mathrm{CO}_2$ is complex due to interactions of various driver variables acting on the surface ocean at different space scales and timescales (Lenton et al., 2012; Landschützer et al., 2015; Gregor et al., 2018). Machine-learning algorithms applied globally struggle to represent the $p\mathrm{CO}_2$ accurately unless spatial coordinates are included as feature variables (Gregor et al., 2017). This is due to the fact that $p\mathrm{CO}_2$ may respond inconsistently to observable feature variables in different regions as it is not possible to observe all feature variables that drive $p\mathrm{CO}_2$. A common practice to avoid the inclusion of coordinates is to separate the ocean into regions where processes that drive $p\mathrm{CO}_2$ 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 with respect to CO_2 variability, namely regions defined by biogeochemical properties and clusters defined by a clustering algorithm.

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

Table 1. Summary of the products, variables and data processing steps used for feature variables. The "usage" column 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 Fig. 1 and throughout the text. Basic data processing is described in the text with details in the Supplement (Sect. S1).

Group: product	Variable	Abbreviation	Usage	Processing	Reference
NOAA: dOISSTv2 (AVHRR only)	Sea surface temperature SST seasonal anomaly Sea ice fraction	SST SST′ ICE	C, R C, R R	– SST – annual average –	Reynolds et al. (2007); Banzon et al. (2016)
Met Office: EN4	Salinity	SSS	R	–	Good et al. (2013)
CDIAC: ObsPack v3	Atmospheric $p\mathrm{CO_2}$	$p\mathrm{CO_2}^{\mathrm{atm}}$	R	$x\mathrm{CO}_2^{\mathrm{atm}}$ × sea-level pressure	Masarie et al. (2014)
UCSD: Argo mixed layers	Mixed-layer depth	MLD	C, R	$\log_{10}$(climatology)	Holte et al. (2017)
ESA: Globcolour	Chl *a* Chl *a* seasonal anomaly	Chl *a* Chl *a*′	C, R R	$\log_{10}\left(\text{climatology filled}^{\text{cloud gaps}}_{1982–1997}\right)$ Chl *a* – annual average	Maritorena et al. (2010)
ECMWF: ERA-Interim 2	*u* wind *v* wind Wind speed	u v U_{10}	R R R	– – $\sqrt{u^2+v^2}$	Dee et al. (2011)
ESA: Globcurrent	Eddy kinetic energy	$\mathrm{EKE}^{\mathrm{clim}}$	C	$\log_{10}\left(1/2\cdot\left(u'^2+v'^2\right)\right)$	Rio et al. (2014)
–	Day of the year	J	R	$\sin\left(\frac{j}{365}\right)$, $\cos\left(\frac{j}{365}\right)$	–
LDEO: $p\mathrm{CO_2}$ climatology	Surface ocean $p\mathrm{CO_2}$	$p\mathrm{CO}_2^{\mathrm{clim}}$	C	Data smoothing	Takahashi et al. (2009)

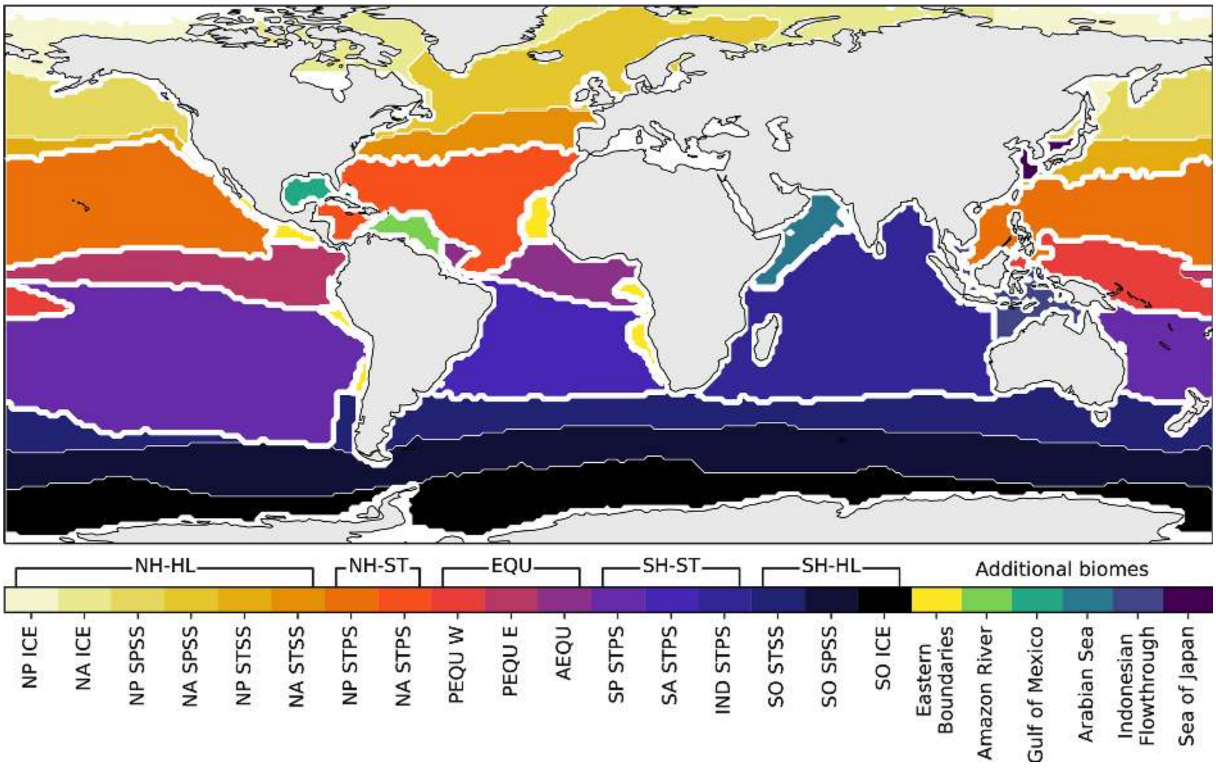

Figure 3. Regions or biomes as defined by Fay and McKinley (2014). Unclassified regions from the original data have been assigned manually in this study and are shown by the separate colours. This modified configuration of the $\mathrm{CO_2}$ biomes is referred to as BIO23 in this study. The sea mask used in Landschützer 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 as follows: NH is Northern Hemisphere and SH is Southern Hemisphere. Suffixes are as follows: HL is high latitudes, ST is subtropics, and EQU is equatorial.

biomes by establishing thresholds for SST, Chl a, sea-ice extent and maximum MLD. Unclassified regions from the original biomes are manually assigned based on their geographical extent, resulting in six additional regions (Fig. 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 adjacent open-ocean biomes. Note that we may refer to the modified Fay and McKinley (2014) ocean CO_2 biomes as "CO_2 biomes" or as "BIO23" from here on (Fig. 3). For later analyses, we group certain biomes together, as shown by the brackets above the colour bar in Fig. 3.

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., 2011), which is described in the Supplement (Sect. S2.2; Fig. S2). We apply clustering with various feature combinations and the number of clusters (shown by orange hexagons in Fig. 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 with respect to RMSE and biases with 21 and 23 clusters. We selected 21 clusters (Fig. S2). Each method of defining regional coherence with respect to $p\mathrm{CO}_2$ variability has its methodological weaknesses so in this study, we adopted the approach of incorporating both K means and CO_2 biomes into the ensemble average (Fig. 1). Although this likely weakens the geophysical meaning of the ensemble domains, we show that it strengthens the overall performance of the ensemble average.

2.4 Regression

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 (1982–2016): 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 Sect. 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 hyperparameters 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 hyperparameters 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 hyperparameters (Hastie et al., 2009). The hyperparameters that result in the best score from the grid search are used for the fit with the full training subset (see Sect. S2.3 for more details). We use a variation of K-fold cross validation called "group K fold" in Scikit-learn (Pedregosa et al., 2011). 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 are 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: ERT – Geurts et al. (2006); GBM – Friedman (2001); SVR – Drucker et al. (1997); and FFNs. The details of these methods and how they were tuned are explained in the Supplement (Sect. 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 (Eqs. 3 and 4 below) and R^{iav} (Eq. 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 Sect. 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 Sect. 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 5 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 (Fig. 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 Sect. S2.1 and Fig. S1). We limit this procedure to only the CO_2 biome and best clustered regressions as it has 5 times the computational cost of a single train-test split.

2.6 Method validation data

For method validation, we use observation data that are not used in SOCAT (Fig. 4 and Table 2) as they are either (1) in-

Table 2. Details for the validation datasets. The measured variables are shown (DIC is dissolved inorganic carbon; TA is 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 those shown in the table, as measurement uncertainty is defined as ±10 µmol 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 GLODAP v2	$p\text{CO}_2$ equilibrator DIC + TA	±2.5 µatm > 12 µatm at 400 µatm*	Takahashi et al. (2017) Olsen et al. (2016); Bockmon and Dickson (2015)	16 161 5976
Surface floats	CARIOCA	$p\text{CO}_2$ colourimetry	±3.0 µatm	Boutin and Merlivat (2013)	613
Profiling floats	SOCCOM	pH + TA (LIAR)	∼ 11 µatm at 400 µatm	Carter et al. (2016)	1037
Mooring	BATS HOT	DIC + TA DIC + TA	∼ 4 µatm at 400 µatm < 7.6 µatm at 400 µatm*	Bates (2007) Dore et al. (2009)	246 214

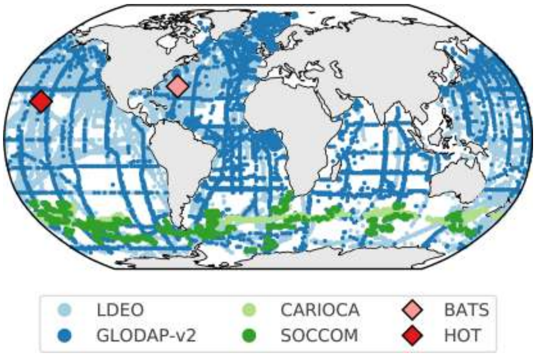

Figure 4. The distribution of the validation data. Details of these datasets are given in Table 2. The Hawaii ocean time series (HOT) and the Bermuda Atlantic time series (BATS) are marked as diamonds to distinguish them as time series stations.

cluded in the Lamont–Doherty Earth Observatory (LDEO) database but not in SOCAT; (2) not measured with an infrared analyser; or (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.

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 and 4.0 µmol kg^{-1}, respectively. The measurements in GLODAP v2 are slightly larger than this at 4 and 6 µmol kg^{-1}, which would result in an error larger than 1.9 % – this is 12 µatm for a 400 µatm 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 ±10 µmol 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 have been adjusted on a per-profile basis to minimise the biases through the comparison of deep slow-changing ocean properties (Olsen et al., 2016). 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 Sect. 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 × 1°) 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

$$F\text{CO}_2 = k_\text{w} \cdot K_0 \cdot \left(p\text{CO}_2^\text{sea} - p\text{CO}_2^\text{atm}\right), \qquad (2)$$

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 $16\,\mathrm{cm\,h^{-1}}$ as in the SOCOM intercomparison (Rödenbeck et al., 2015).

2.8 Relative interannual variability and interquartile range metrics

2.8.1 Regression metrics

We use bias and RMSE as first-order metrics of model performance.

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

$$\mathrm{Bias} = \sum\nolimits_{i=1}^{n} \frac{\hat{y}_i - y}{n}, \tag{3}$$

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:

$$\mathrm{RMSE} = \sqrt{\sum_{i=1}^{n} \frac{\left(y_i - \hat{y}_i\right)^2}{n}}. \tag{4}$$

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.

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 $p\mathrm{CO}_2$:

$$R^\mathrm{iav} = \frac{\sigma_{1982\text{–}2015}\left(M^{\mathrm{iav}(t)}\right)}{\sigma_{1982\text{–}2015}\left(M^{\mathrm{iav}(t)}_\mathrm{bench}\right)} \tag{5a}$$

$$M^{\mathrm{iav}(t)} = \sqrt{\frac{\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)}{n-1}} \tag{5b}$$

$$M^{\mathrm{iav}(t)}_\mathrm{bench} = \sqrt{\frac{\Sigma_{i=0}^{n}\left(y_i - \hat{y}_i^{b}\right)}{n-1}}. \tag{5c}$$

Here, σ is the standard deviation of M^iav and $M^\mathrm{iav}_\mathrm{bench}$, respectively, which are both represented as yearly time series. Equation (5b) and (5c) show the formulation for $M^{\mathrm{iav}(t)}$ and $M^{\mathrm{iav}(t)}_\mathrm{bench}$, 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 $M^\mathrm{iav}_\mathrm{bench}$ is calculated to normalise M^iav. $\hat{y}^b$ represents the data where IAV has been removed by summing the climatology of the mapped surface ocean $p\mathrm{CO}_2$ and the annual trend of atmospheric $p\mathrm{CO}_2$.

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 2 in Fig. 1) are depicted in Fig. 5a–c, which plot the matrix of the (a) average bias, (b) RMSE and (c) R^iav for each combination of the experimental number of clusters and clustering features.

Results show that the configuration that includes $\mathrm{EKE}^\mathrm{clim}$ (column E in Fig. 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 Fig. 5a, b). The increased dynamics associated with high-EKE regions might change the way $p\mathrm{CO}_2$ behaves compared to low-EKE regions (Boutin and Merlivat, 2013; 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 Fig. 5), while we do not weight R^iav strongly in this assessment as a R^iav 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 (Figs. S5 and S6, respectively) do not match the distributions exactly, the two selected clustering configurations (black boxes in Fig. 5) score consistently low for both metrics (with the exception of ERT – discussed in greater detail further on). We are motivated 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 reduce the possible complexity at little cost. The selected clustering configuration with 21 clusters has the following fea-

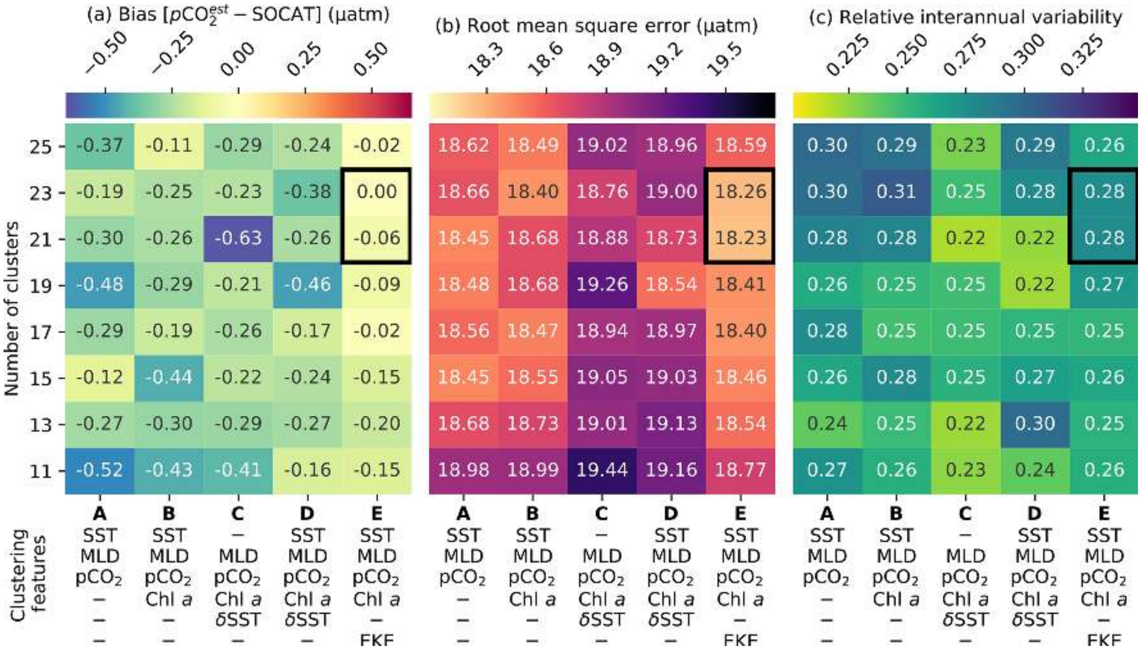

Figure 5. Heat maps showing the average cluster **(a)** bias, **(b)** RMSE and **(c)** relative interannual variability (R^{iav}) for different cluster configurations, where smaller scores are better for all metrics. The rows show the number of clusters, and the columns show clustering feature-variable 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 value of $R^{\text{iav}} < 0.3$ falls within the best category of performance in Rödenbeck et al. (2015).

tures: 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 Fig. S2 for the distribution of the climatology for these clusters).

Comparatively, the Fay and McKinley (2014) CO_2 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 CO_2 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 pCO_2 estimates (CSIR-ML8, Council for Scientific and Industrial Research – Machine Learning ensemble with Eight members).

All regression methods have lower RMSE scores for K21E than for BIO23, but R^{iav} and bias do not indicate that either 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 and 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 betters its members, with RMSE and bias scores of 17.25 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).

Table 3. Regression scores for the CO_2 biomes (BIO23), the clustering configuration from column E in Fig. 5 (K21E) and the ensemble average (CSIR-ML8). Abbreviations are as follows: RMSE is the root mean square error; R^{iav} is the relative interannual variability (Eq. 5). Regression methods are as follows: SVR is support vector regression; ERT is extremely randomised trees; GBM is the gradient-boosting machine; FFN is the feed-forward neural network. Bold values are significantly lower than the mean for that column ($p < 0.05$ for the two-tailed Z test; absolute values are used for the 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**

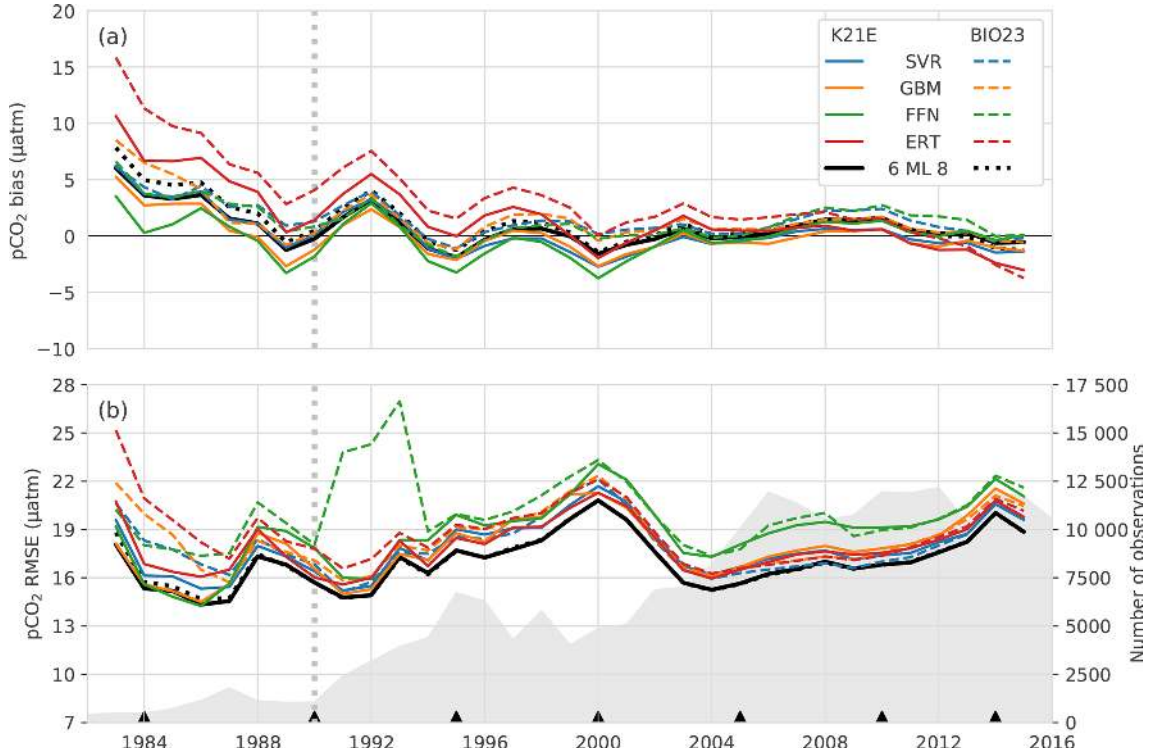

Figure 6. Annually averaged **(a)** bias and **(b)** RMSE for the eight individual regression methods in Table 3: BIO23 (dashed lines) and K21E (solid lines). The dotted black lines show the ensemble averages for all eight models (CSIR-ML8), and the solid black line shows metrics for the ensemble average of the SVR, GBM and FFN (CSIR-ML6) from BIO23 and K21E. The grey-filled area in panel **(b)** shows the number of observations per year, and black triangles show 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.

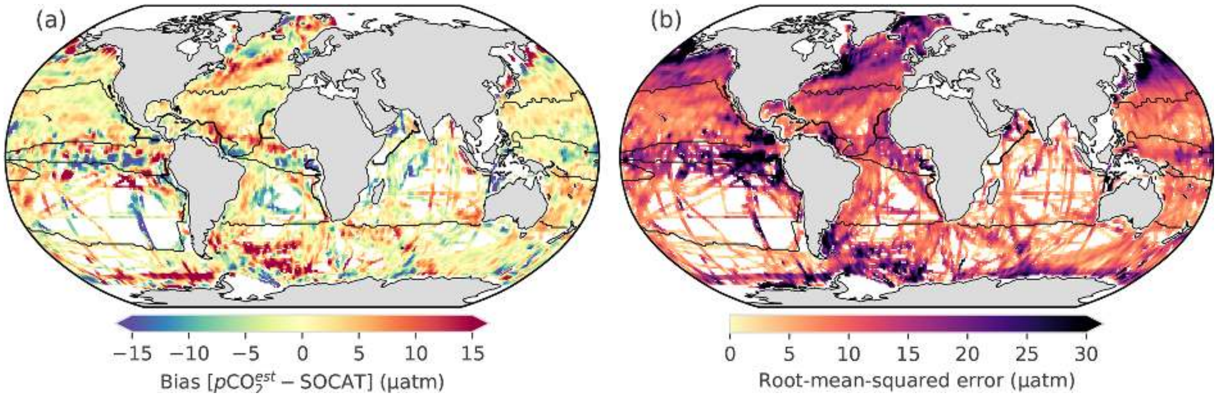
(a)
(b)
−15
−10
−5
0
5
10
15
Bias [pCO$_2^{est}$ − SOCAT] (µatm)
0
5
10
15
20
25
30
Root-mean-squared error (µatm)

Figure 7. Panel **(a)** shows the biases from the robust test estimates; panel **(b)** shows the RMSEs for CSIR-ML6. Convolution has been applied to panels **(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 Fig. 3.

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 (Fig. 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 and 2.24 µatm, respectively). These biases are larger than those reported in Table 3 (with averages of absolute biases of 0.48 and 0.41 µatm for K21E and BIO23, respectively),

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 the two-tailed Z test; absolute values are used for the 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

but this is likely since selected test years (black triangles in Fig. 6b) fall on years of low bias. While FFN has the largest RMSE (18.93 and 20.24 µatm for K21E and BIO23), it has a smaller bias compared to other regression methods (0.04 and 1.60 µatm, respectively), motivating 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 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 and 1.48 µatm, respectively; Table 4).

Similar to the biases, RMSEs for all models (Fig. 6b) have similar interannual tendencies and variability, with a sharp peak in the year 2000 (> 20 µatm, where the mean RMSE is 18.61 µatm). The increased RMSE scores are likely due to the spatial distribution of sampling density (see Fig. 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 (< 19 µatm 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 and 17.25 µatm for CSIR-ML6 and CSIR-ML8, respectively (see Table S1 comparisons of ensemble averages with different members).

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 (Eq. 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).

The spatial distribution of the bias and RMSE is now studied for CSIR-ML6 (Fig. 7a 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 (|bias| < 5 µatm and RMSE < 10 µatm). The equatorial regions (EQU), especially the eastern Pacific, contrasts this with large uncertainties in both bias and RMSE (> |10 µatm| and 30 µatm, respectively). The high-latitude oceans (NH-HL and SH-HL) have considerable uncertainties due to the large interannual variability of surface ocean $p\mathrm{CO}_2$ caused by the 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).

In summary, the robust test estimates show that there is a positive bias in $p\mathrm{CO}_2$ predictions before 1990 for all models, but it is largest for ERT, and excluding these models from the ensemble results in better $p\mathrm{CO}_2$ 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.

3.3 Validation with independent datasets

Here, we validate the accuracy of $p\mathrm{CO}_2$ estimates from CSIR-ML6 with independent data (that are 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); JMA-MLR, updated on 2 December 2018 (Iida et al. 2015); MPI-SOMFFN v2016 (Landschützer et al., 2017); and University of East Anglia – Statistical Interpo-

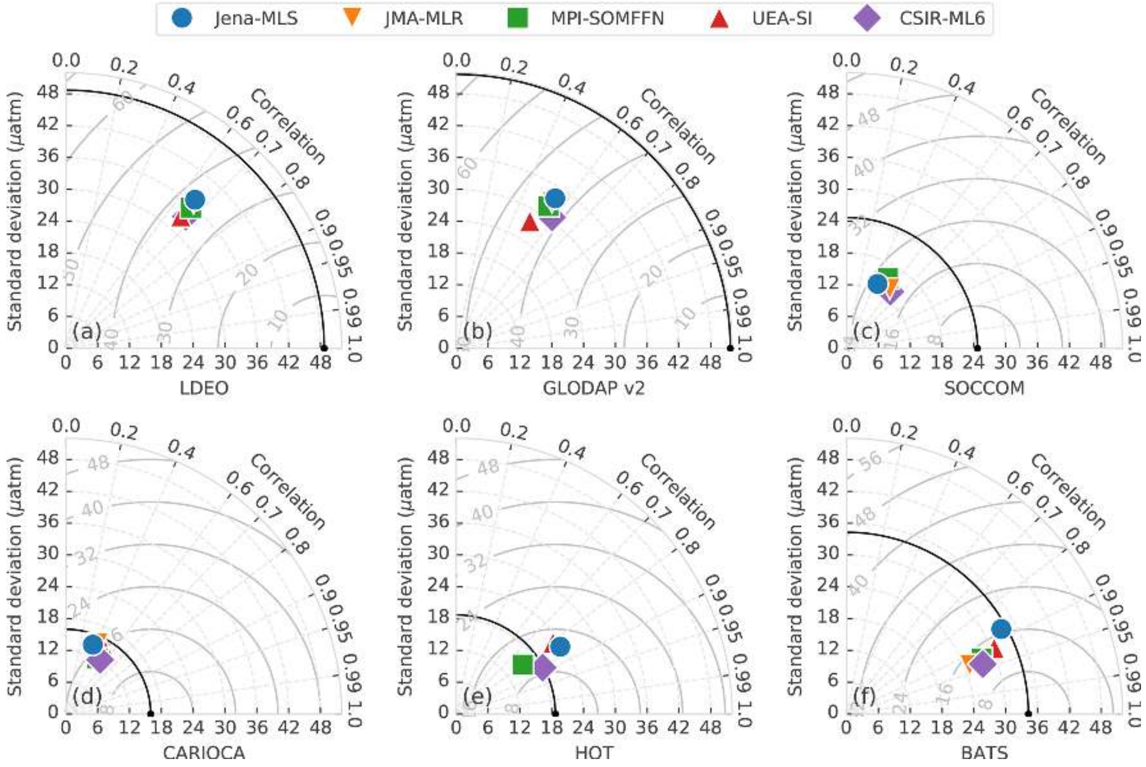

Figure 8. Taylor diagrams comparing the $p\mathrm{CO_2}$ estimates of five gap-filling methods (represented by the different markers) with validation datasets (Table 2) for the period 1990–2015. Each validation dataset has its own Taylor diagram, as labelled on the bottom axes. The black marker on the 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). A description of the gap-filling methods from independent studies is provided in the text (Sect. 3.3).

lation (UEA-SI version 1.0; Jones et al., 2015). $p\mathrm{CO_2}$ estimates by the Jena-MLS were resampled to monthly temporal resolution and interpolated to a 1° grid using Python's *xarray* package. Note that these datasets will also suffer from the same temperature biases discussed in Sect. S2.4.

The performance of each gap-filling method is represented with a Taylor diagram for each independent validation dataset (Fig. 8; Taylor, 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 Fig. 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 arched line for all but the station HOT (Fig. 8e).

All methods fail to represent the standard deviation of the two global validation datasets, LDEO and GLODAP v2 (Fig. 8a, b), with centred RMSE scores greater than 35 µatm. However, calculating RMSE annually results in scores of ~ 27 µatm for LDEO and ~ 35 µatm for GLODAP v2, much lower than shown in Fig. 8a–b, due to high RMSE scores (> 40 µatm) for a small subset of years (Sect. S3.4 and Fig. S7). Estimates of the Southern Ocean datasets (Fig. 8c, d), SOCCOM and CARIOCA, have lower RMSE scores (~ 16 and ~ 23 µatm, 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 underestimated by at 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 is largely underestimated (Williams et al., 2017). In contrast, all methods estimate the two time series stations, HOT and BATS (Fig. 8e, f and Table 5), relatively well with correlation scores > 0.8 and low average bias ~ 4.5 µatm.

Despite all scores being closely grouped (Fig. 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 (Fig. 8) is beneficial with regard to achiev-

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 the two-tailed Z test; absolute values are used for the biases). The UEA-SI method does not have error estimates for SOCCOM floats as these two time series do not overlap.

Metric	Method	LDEO	GLODAP-v2	SOCCOM	CARIOCA	BATS	HOT
Count	Count	16 161	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

ing 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.

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 Sect. 3.3) of the sea–air CO_2 flux ($F\mathrm{CO}_2$) over the period 1990 to 2016. $F\mathrm{CO}_2$ was calculated using the same gas-transfer velocity and solubility for each gap-filling method (Sect. 2.7). Differences in $F\mathrm{CO}_2$ are thus driven by variations in $p\mathrm{CO}_2$ from each gap-filling method.

The average $F\mathrm{CO}_2$ for 1990–2016 by CSIR-ML6 (Fig. 9a) contextualises the regional distribution of fluxes: strong outgassing in the equatorial Pacific, strong sink in the midlatitudes, 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 $F\mathrm{CO}_2$ as simulated by CSIR-ML6 (Fig. 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 $F\mathrm{CO}_2$ between 2016 and 2000 (Fig. 9b), since those are the two years where the difference in global $F\mathrm{CO}_2$ is greatest for CSIR-ML6 (Fig. 10a). Note that Fig. 9b serves as a snapshot for the change in $F\mathrm{CO}_2$ between those two years, whose interpretation cannot be linked to an overall anthropogenically forced change as the comparison between the two years could reflect interannual, decadal or multi-decadal variability. The differences in $F\mathrm{CO}_2$ between 2016 and 2000 are 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 $F\mathrm{CO}_2$ ($> 10\,\mathrm{g\,C\,m^{-2}\,yr^{-1}}$) between the two specific years.

The annual change in $F\mathrm{CO}_2$ is also studied for the different regions. The Southern Hemisphere high-latitude (SH-HL) region is the strongest contributor to the trend (Fig. 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 (Fig. S10c). The equatorial (EQU) region is the only persistent source of CO_2 to the atmosphere (also seen in Fig. 9a). The subtropical regions (Fig. 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.

We use the average interquartile range between the 1-year rolling mean estimates ($\mathrm{IQR^{IA}}$) as a measure of agreement or divergence between gap-filling methods, where large values indicate a divergence (Sect. 2.8.2). We also show the $\mathrm{IQR^{IA}}$ scaled to the range of the regional interannual variability (max–min) as a percentage (relative $\mathrm{IQR^{IA}}$), 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 (Fig. 10e), with diverging $F\mathrm{CO}_2$ throughout the period with an $\mathrm{IQR^{IA}}$ of $0.11\,\mathrm{Pg\,C\,yr^{-1}}$ and a large relative $\mathrm{IQR^{IA}}$ of 28 %. Similarly, the $\mathrm{IQR^{IA}}$ for the

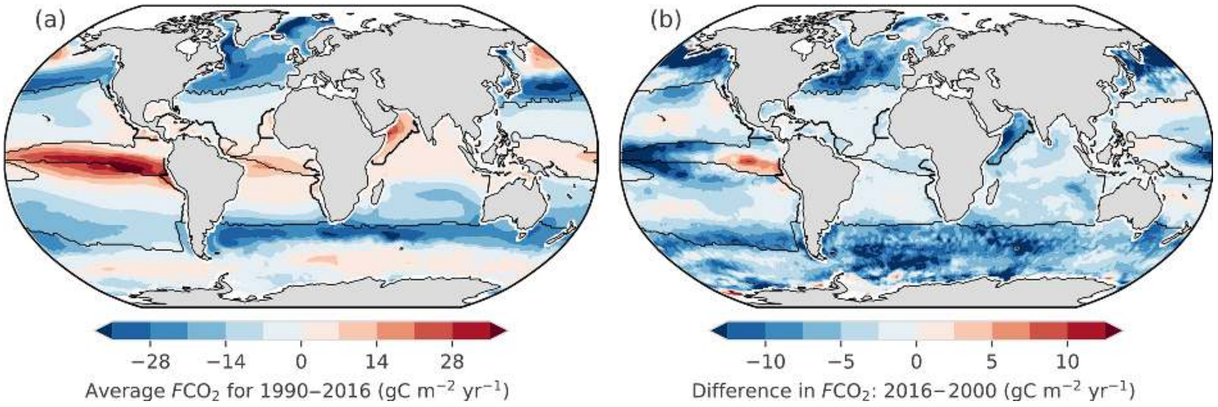
(a)
(b)
−28
−14
0
14
28
Average FCO_2 for 1990–2016 (gC m^{-2} yr^{-1})
−10
−5
0
5
10
Difference in FCO_2: 2016–2000 (gC m^{-2} yr^{-1})

Figure 9. (a) Average sea–air CO_2 fluxes ($F\mathrm{CO}_2$) of CSIR-ML6 for 1990 to 2016, where $F\mathrm{CO}_2$ is calculated as shown in Eq. (2). Negative $F\mathrm{CO}_2$ (blue) indicates regions of atmospheric CO_2 uptake. **(b)** The differences between $F\mathrm{CO}_2$ in 2016 and 2000, which are the minimum and maximum of global ocean uptake flux ($F\mathrm{CO}_2$) estimates, respectively (for CSIR-ML6 in Fig. 10a). Black lines show the regions as defined in Fig. 3.

SH-HL region (Fig. 10f) is 0.08 Pg C yr^{-1}, but the relative IQR$^{\mathrm{IA}}$ 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$^{\mathrm{IA}}$ estimates of 0.04 and 0.05 Pg C yr^{-1} for the NH-ST and NH-HL regions, respectively (Fig. 10c, d). However, a larger relative IQR$^{\mathrm{IA}}$ of 20 % suggests that the interannual $F\mathrm{CO}_2$ 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 $F\mathrm{CO}_2$. The EQU region (Fig. 10b) has an IQR$^{\mathrm{IA}}$ and relative score at 0.03 Pg C yr^{-1} and 14 %.

The CSIR-ML8 method is not included in the IQR$^{\mathrm{IA}}$ calculations but is included in Fig. 10 to show the impact of the ERT models' positive bias in $p\mathrm{CO}_2$ on $F\mathrm{CO}_2$ (Fig. 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 Pg C yr^{-1}. The positive biases have the strongest impact on the SH-ST that occupies 36 % total area (Fig. S10c), with only 11 % of the total observations in SOCAT, suggesting that this method is sensitive to imbalanced datasets.

3.5 Regional disagreement between methods

In order to better understand the regional distribution of the uncertainties in $F\mathrm{CO}_2$, we assess the level of agreement between independent gap-filling methods in their interannual surface ocean $p\mathrm{CO}_2$ estimates (Fig. 11). We use $p\mathrm{CO}_2$ for this representation as no spatial integration occurs – only time averaging.

The interannual estimates of IQR$^{\mathrm{IA}}$ (Fig. 11a) show the disagreement between methods is relatively small in the majority of the ocean (<6 µatm), with the exceptions being the Southern Ocean, South Atlantic, southeastern Pacific and eastern equatorial Pacific with differences of >10 µatm, where these regions coincide with regions of low sampling density (Fig. 2). The IQR$^{\mathrm{IA}}$ scaled to the maximum–minimum range of interannual $p\mathrm{CO}_2$ suggests that the NH-ST trend is relatively well constrained (<10 %), which is in conflict with the IQR$^{\mathrm{IA}}$ for $F\mathrm{CO}_2$ in Fig. 10c (where the relative IQR$^{\mathrm{IA}}$ is 20 %). The disagreement may stem from the magnifying impact that wind speed has on $F\mathrm{CO}_2$; i.e. small differences in $p\mathrm{CO}_2$ may become large when fluxes are calculated. The same principle may apply to the EQU in Fig. 11b, where relative IQR$^{\mathrm{IA}}$ is large (>10 %) for $p\mathrm{CO}_2$, but low wind speeds result in a low relative IQR$^{\mathrm{IA}}$ for $F\mathrm{CO}_2$ (7 % in Fig. 10b). The largest relative IQR$^{\mathrm{IA}}$ scores occur in the SH-ST (>10 % in Fig. 11c) where data are sparse, specifically the South Atlantic and southeastern Pacific (Fig. 2a). The relative IQR$^{\mathrm{IA}}$ scores suggest that the gap-filling methods agree on $p\mathrm{CO}_2$ in the SH-HL east of the Greenwich meridian ($>0°$ E).

In summary, we show that there is an agreement between gap-filling methods in the Northern Hemisphere for interannual $p\mathrm{CO}_2$, but the methods show considerable disagreement in the Southern Hemisphere, particularly in 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 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

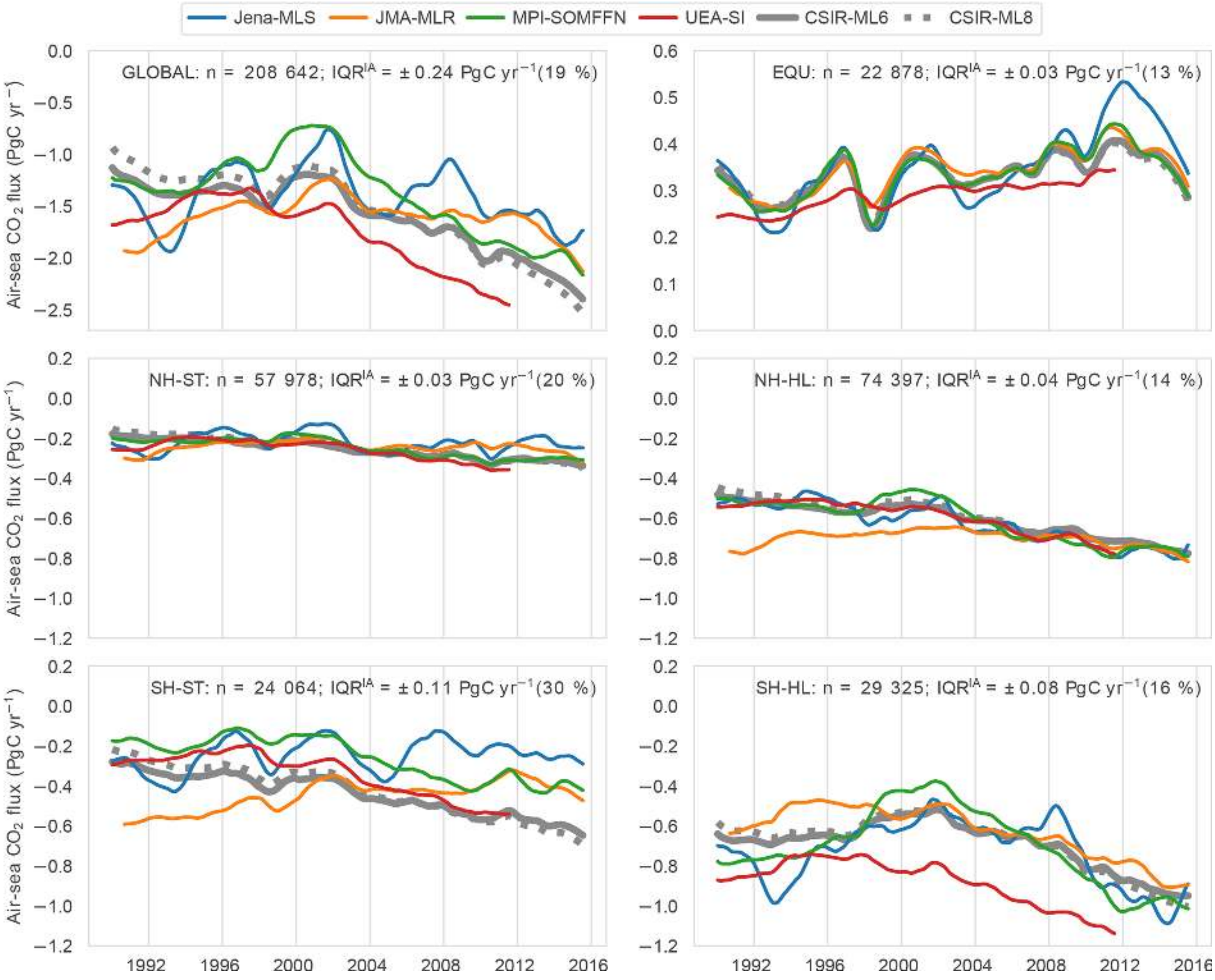

Figure 10. Sea-air CO_2 fluxes averaged for regions as shown in Fig. 2: **(a)** global domain, **(b)** equatorial regions, **(c)** Northern Hemisphere subtropical, **(d)** Northern Hemisphere high latitude, **(e)** Southern Hemisphere subtropical. **(f)** Southern Hemisphere high latitude. The coloured 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-axis scales differ for the top **(a, b)**. Note that the uncertainties of each model (e.g. bias and RMSE from Fig. 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 interannual interquartile range (IQR^{IA}).

an ensemble average of two-step machine-learning models that estimate global surface ocean $p\mathrm{CO}_2$. 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 clustering and regression, we create the ensemble average with a good understanding of each model's biases (Figs. 6 and 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 nonlinear models, with the exception of basic diagnostics such as feature-variable importance (see Fig. S11) from decision-tree-based approaches (Pedregosa et al., 2011; Castelvecchi, 2016). Our results thus show that there is, in fact, a benefit in creating an ensemble average of models (Table 5), and if carefully implemented, it is an additional tool that can be used to reduce the uncertainties in gap-filling estimates of $p\mathrm{CO}_2$.

It could be argued that an exhaustive search for the optimal configuration (Fig. 5) for CSIR-ML6 may result in poorly trained individual models. However, we think that the merit of introducing and assessing regression 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 ensemble averages stronger, as the CSIR-ML6 performs well relative to other gap-filling methods.

In the search for the optimal clustering configuration (Fig. 5a, b), we show that including EKE (along with SST)

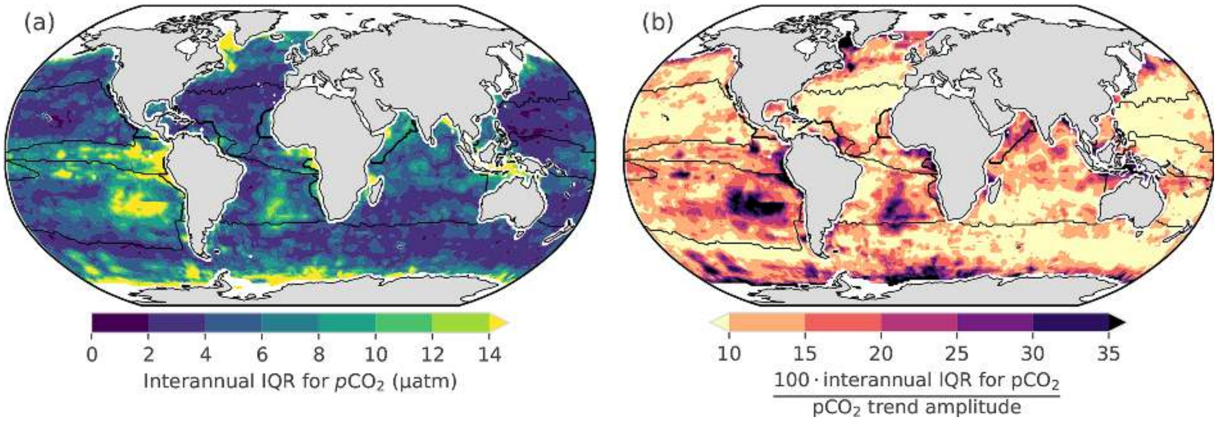
(a)
0
2
4
6
8
10
12
14
Interannual IQR for pCO$_2$ (µatm)
(b)
10
15
20
25
30
35
$\frac{100 \cdot \text{interannual IQR for pCO}_2}{\text{pCO}_2 \text{ trend amplitude}}$

Figure 11. (a) The magnitude of the interannual disagreement between independent gap-filling methods (IQR$^{\mathrm{IA}}$) as shown in Fig. 10; hence, low IQR$^{\mathrm{IA}}$ indicates good agreement amongst the different methods. **(b)** Level of agreement on the interannual variability across methods (in %), more specifically IQR$^{\mathrm{IA}}$ scaled by the difference between the maximum and minimum values for interannual pCO$_2$ (the range).

as a clustering feature variable leads to an improvement in bias and RMSE for nearly all numbers of clusters, albeit a small improvement. Increased intraseasonal variability of pCO$_2$ appears to be associated with regions of high EKE compared to low EKE regions (Monteiro et al., 2015; du Plessis et al., 2017, 2019). Moreover, the importance of EKE as a part of the clustering constraints also shows that more thought should be given to how we sample pCO$_2$ 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 (Fig. 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 pCO$_2$. The bias in ERT may be due to its sensitivity to imbalanced datasets (Crone and Finlay, 2012), where the data in SOCAT v5 are sparse 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 pCO$_2$ and FCO$_2$ vary significantly in some regions, particularly in regions where data are sparse, such as in the Southern Hemisphere oceans (Fig. 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 increases, the biases are reduced, and the methods converge.

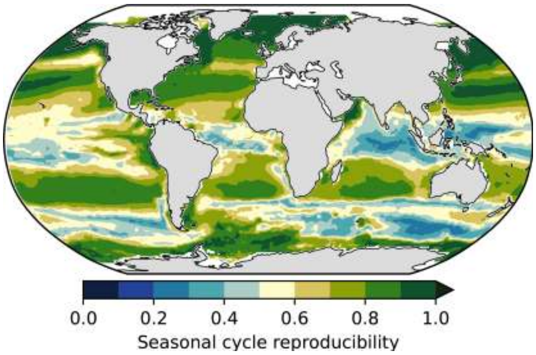

Figure 12. The seasonal cycle reproducibility of CSIR-ML6 pCO$_2$, which is a correlation of detrended pCO$_2$ with its own climatology – the larger the correlation, the stronger the reproducibility of the seasonal cycle (method from Thomalla et al., 2011).

The Northern Hemisphere subtropical regions are a good example of a region where the gap-filling methods converge (Fig. 11b), as also shown by the low RMSE scores and high correlation for the two mooring stations, HOT and BATS (Fig. 8e, f). One of the reasons that the methods predict the variability well in the subtropics (Fig. 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 (Fig. 12).

The gap-filling methods' divergences also serve as a metric to inform where there are not enough data to constrain the pCO$_2$ or FCO$_2$ estimates; i.e. the divergences inform us where estimates should be treated with caution. The IQR$^{\mathrm{IA}}$,

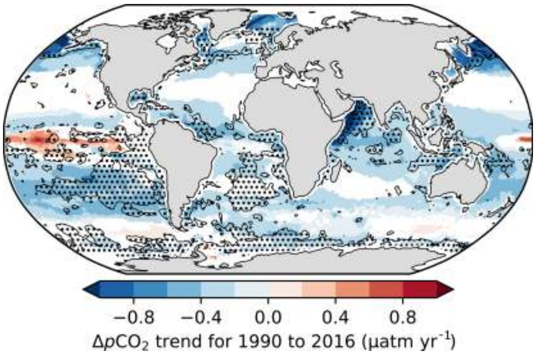

Figure 13. $\Delta p\mathrm{CO}_2$ trends ($p < 0.05$), where $\Delta p\mathrm{CO}_2$ is calculated as the estimated surface ocean $p\mathrm{CO}_2$ from the CSIR-ML6 method minus atmospheric $p\mathrm{CO}_2$ from the CarboScope project (Rödenbeck et al., 2014). The shaded areas show the regions where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

when scaled to the range of interannual variability (Fig. 11b), should be taken into account when analysing interannual trends of $\Delta p\mathrm{CO}_2$ (Fig. 13). For instance, significant trend estimates in $\Delta p\mathrm{CO}_2$ for CSIR-ML6 ($p < 0.05$) are negative for the majority of the global ocean, even in regions where method estimates are too disparate to resolve interannual variability (relative $\mathrm{IQR}^{\mathrm{IA}} > 15\,\%$; dotted regions in Fig. 13). However, the relative $\mathrm{IQR}^{\mathrm{IA}}$ 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.

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 (Fig. 8), and result in divergences between estimates when there are insufficient data to constrain the methods (Fig. 11b). From these points, it may seem that we may have in fact "hit the wall" in terms of better resolving surface ocean $p\mathrm{CO}_2$. 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.

4.3.1 Reducing existing biases

The robust test estimates show that there are regions where training data are not sparse, yet estimates still suffer from large uncertainties (e.g. northern and southern boundaries of the North Atlantic Gyre in Figs. 7a, b and 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 $p\mathrm{CO}_2$ variability at current resolutions (monthly × 1° or coarser) or with the current regression feature variables (Gregor et al., 2017; Denvil-Sommer et al., 2018). It may be possible to reduce these uncertainties with consideration about the drivers of CO_2 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 midlatitude oceans (Monteiro et al., 2015). It is worthwhile to note that increasing the resolution may not be the panacea for poor estimates. For example, the Jena-MLS method is able to estimate $p\mathrm{CO}_2$ with relative accuracy (Fig. 8) at a low spatial resolution ($\approx 4° \times 5°$; 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 temperature at which $p\mathrm{CO}_2$ is measured (i.e. at the depth of a ship's intake) and the temperature to which $p\mathrm{CO}_2$ 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 (> 5 µatm for large regions, as shown in Fig. S3b). However, the correction of the intake temperature to the remotely sensed surface temperature also makes the assumption that temperature is the only factor that influences $p\mathrm{CO}_2$ 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 experimentally to assess the impact that these processes may have on $p\mathrm{CO}_2$.

4.3.2 Improving estimates in data-poor regions

All gap-filling methods suffer from similar biases and uncertainties (Fig. 8, Table 5) when compared to 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 are the limiting factor (Rödenbeck et al., 2015; Landschützer et al., 2016; Ritter et al., 2017; Denvil-Sommer et al., 2018).

Strides have been made in closing these data-sparse gaps with the deployment of autonomous sampling platforms. The 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 have shown that we have previously underestimated winter outgassing of CO_2 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 anthropogenic CO_2 uptake (Gruber et al., 2019). Incorporating these data successfully into existing models may not be straightforward due to the strong temporal bias of these data toward the end of the time series. For instance, the inclusion of atmospheric $p\mathrm{CO}_2$ could result in temporally skewed estimates due to the "memory" effect that including the annually increasing atmospheric $p\mathrm{CO}_2$ could have on estimates.

The complex machine-learning models often used to estimate $p\mathrm{CO}_2$ are prone to overfitting the data, particularly in regions where data are sparse. Using less complex models, e.g. multi-linear regression, in such regions would reduce the risk of overfitting 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 $p\mathrm{CO}_2$ has a strong seasonal signal driven primarily by temperature (Fig. 12; Taylor, 2001).

One of the weaknesses of our study is that our approach is similar to other regression methods (e.g. MPI-SOMFFN by Landschützer et al., 2014, and JMA-MLR and LSCE-FFNN by Denvil-Sommer et al., 2019) that predict $p\mathrm{CO}_2$ 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 $p\mathrm{CO}_2$ (Verdy and Mazloff, 2017). These methods may be able to provide better constraints on $p\mathrm{CO}_2$ 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 CO_2 community can improve estimates within the bounds of the current observations and make recommendations for future observations.

We introduce a new surface ocean $p\mathrm{CO}_2$ 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, result 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 resolve $p\mathrm{CO}_2$ 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 $p\mathrm{CO}_2$ 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 $p\mathrm{CO}_2$ 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 $p\mathrm{CO}_2$ derived from pH measured by biogeochemical Argo floats and new platforms such as $p\mathrm{CO}_2$-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 the Supplement. Data (global surface ocean $p\mathrm{CO}_2$ from CSIR-ML6 version 2019a) are available from the Ocean Carbon Data System (OCADS; https://doi.org/10.25921/z682-mn47, Gregor et al., 2019).

Supplement. The supplement related to this article is available online at: https://doi.org/10.5194/gmd-12-5113-2019-supplement.

Author contributions. LG is the lead author and developed the method and wrote the manuscript. ADL contributed to the model assessment and contributed to 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.

Competing interests. The authors declare that they have no conflict of interest.

Acknowledgements. We acknowledge the support and computational hours from the Centre for High-Performance Computing (CSIR-CHPC). The Surface Ocean CO_2 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 CO_2 database. The many researchers and funding agencies responsible for the collection of data and quality control are thanked for their contributions to SOCAT.

Financial support. This work is part of a post-doctoral research fellowship funded by the CSIR 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.

This work received support from the European Space Agency (ESA)'s OCEANSODA – Ocean Acidification project (contract no. 4000125955/18/I-BG).

Review statement. This paper was edited by Andrew Yool and reviewed by Peter Landschützer, Jamie Shutler, and one anonymous referee.

References

Bakker, D. C. E., Hoppema, M., Schröder, M., Geibert, W., and de Baar, H. J. W.: A rapid transition from ice covered CO_2-rich waters to a biologically mediated CO_2 sink in the eastern Weddell Gyre, Biogeosciences, 5, 1373–1386, https://doi.org/10.5194/bg-5-1373-2008, 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., 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., Ibánhez, 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 multidecade record of high-quality $f\mathrm{CO}_2$ data in version 3 of the Surface Ocean CO2 Atlas (SOCAT), Earth Syst. Sci. Data, 8, 383–413, https://doi.org/10.5194/essd-8-383-2016, 2016.

Banzon, V., Smith, T. M., Chin, T. M., Liu, C., and Hankins, W.: A long-term record of blended satellite and in situ sea-surface temperature for climate monitoring, modeling and environmental studies, Earth Syst. Sci. Data, 8, 165–176, https://doi.org/10.5194/essd-8-165-2016, 2016.

Bates, N. R.: Interannual variability of oceanic CO_2 and biogeochemical properties in the Western North Atlantic subtropical gyre, Deep-Res. Pt. II, 48, 1507–1528, https://doi.org/10.1016/S0967-0645(00)00151-X, 2001.

Bockmon, E. E. and Dickson, A. G.: An inter-laboratory comparison assessing the quality of seawater carbon dioxide measurements, Mar. Chem., 171, 36–43, https://doi.org/10.1016/j.marchem.2015.02.002, 2015.

Boutin, J. and Merlivat, L.: Sea surface $f\mathrm{CO}_2$ measurements in the Southern Ocean from CARIOCA Drifters, Carbon Dioxide Information Analysis Center, Oak Ridge National Laboratory, US Department of Energy, Oak Ridge, Tennessee, https://doi.org/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. Meth., 16, 119–131, https://doi.org/10.1002/lom3.10232, 2018.

Castelvecchi, D.: Can we open the black box of AI?, Nature, 538, 20–23, https://doi.org/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, 224–238, https://doi.org/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, I., Biblot, J., Bormann, N., Delsol, C., Dragani, R., Fuentes, M., Greer, A. J., Haimberger, L., Healy, S. B., Hersbach, H., Holm, E. V., Isaksen, L., Kallberg, P., Kohler, M., Matricardi, M., McNally, A. P., Mong-Sanz, B. M., Morcette, J.-J., Park, B.-K., Peubey, C., de Rosnay, P., Tavolato, C., Thepaut, J. N., and Vitart, F.: The ERA-Interim reanalysis: Configuration and performance of the data assimilation system, Q. J. Roy. Meteorol. Soc., 137, 553–597, https://doi.org/10.1002/qj.828, 2011.

Denvil-Sommer, A., Gehlen, M., Vrac, M., and Mejia, C.: LSCE-FFNN-v1: a two-step neural network model for the reconstruction of surface ocean pCO2 over the global ocean, Geosci. Model Dev., 12, 2091–2105, https://doi.org/10.5194/gmd-12-2091-2019, 2019.

Dickson, A. G.: Guide to best practices for ocean CO_2 measurements, edited by: Andrew, G., Sabine, C. L., and Christian, J. R., North Pacific Marine Science Organization, available at: https://www.oceanbestpractices.net/handle/11329/249 (last access: 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, P. Natl. Acad. Sci. USA, 106, 12235–12240, https://doi.org/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, 155–161, 1997.

du Plessis, M., Swart, S., Ansorge, I. J., and Mahadevan, A.: Submesoscale processes promote seasonal restratification in the Subantarctic Ocean, J. Geophys. Res., 122, 2960–2975, https://doi.org/10.1002/2016JC012494, 2017.

du Plessis, M., Swart, S., Ansorge, I. J., Mahadevan, A., and Thompson, A. F.: Southern Ocean seasonal restratification delayed by submesoscale wind-front interactions, J. Phys. Oceanogr., 49, 1035–1053, https://doi.org/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, 273–284, https://doi.org/10.5194/essd-6-273-2014, 2014.

Friedman, J. H.: Greedy function approximation: A gradient boosting machine, Ann. Stat., 29, 1189–1232, https://doi.org/10.1214/aos/1013203451, 2001.

Geurts, P., Ernst, D., and Wehenkel, L.: Extremely randomized trees, Mach. Learn., 63, 3–42, https://doi.org/10.1007/s10994-006-6226-1, 2006.

Goddijn-Murphy, L. M., Woolf, D. K., Land, P. E., Shutler, J. D., and Donlon, C.: The OceanFlux Greenhouse Gases methodology for deriving a sea surface climatology of CO_2 fugacity in support of air–sea gas flux studies, Ocean Sci., 11, 519–541, https://doi.org/10.5194/os-11-519-2015, 2015.

Goddijn-Murphy, L., Woolf, D. K., Callaghan, A. H., Nightingale, P. D., Shutler, J. D.: A reconciliation of empirical and mechanistic models of the air-sea gas transfer velocity, J. Geophys. Res.-Oceans, 121, 818–835, https://doi.org/10.1002/2015JC011096, 2016.

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.-Oceans, 118, 6704–6716, https://doi.org/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, 9049–9057, https://doi.org/10.1029/2018GL078013, 2018.

Gregor, L., Kok, S., and Monteiro, P. M. S.: Empirical methods for the estimation of Southern Ocean CO_2: support vector and random forest regression, Biogeosciences, 14, 5551–5569, https://doi.org/10.5194/bg-14-5551-2017, 2017.

Gregor, L., Kok, S., and Monteiro, P. M. S.: Interannual drivers of the seasonal cycle of CO_2 in the Southern Ocean, Biogeosciences, 15, 2361–2378, https://doi.org/10.5194/bg-15-2361-2018, 2018.

Gregor, L., Lebehot, A. D., Kok, S., and Monteiro, P. M. S.: Global surfaceocean partial pressure of carbon dioxide (pCO2) estimates from a machine learning ensemble: CSIR-ML6 v2019a (NCEI Accession 0206205), Version 1.1, NOAA National Centers for Environmental Information Dataset, https://doi.org/10.25921/z682-mn47, last access: 15 November, 2019.

Gruber, N., Clement, D., Carter, B. R., Feely, R. A., van Heuven, S., Hoppema, M., Ishii, M., Key, R. M., Kozyr, A., Lauvset, S. K., Lo Monaco, C., Mathis, J. T., Murata, A., Olsen, A., Perez, F. F., Sabine, C. L., Tanhua, T., and Wanninkhof, R.: The oceanic sink for anthropogenic CO_2 from 1994 to 2007, Science, 363, 1193–1199, https://doi.org/10.1126/science.aau5153, 2019.

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. Cy., 29, 517–533, https://doi.org/10.1002/2014GB004986, 2015.

Hastie, T., Tibshirani, R., and Friedman, J. H.: The Elements of Statistical Learning: Data mining, Inference, and Prediction, 2nd edn., Springer, 2009.

Holte, J., Talley, L. D., Gilson, J., and Roemmich, D.: An Argo mixed layer climatology and database, Geophys. Res. Lett., 44, 5618–5626, https://doi.org/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, https://doi.org/10.5334/jors.148, 2017.

Iida, Y., Kojima, A., Takatani, Y., Nakano, T., Sugimoto, H., Midorikawa, T., and Ishii, M.: Trends in $p\mathrm{CO}_2$ and sea–air CO_2 flux over the global open oceans for the last two decades, J. Oceanogr., 71, 637–661, https://doi.org/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.-Sea Res. Pt. I, 45, 1187–1209, https://doi.org/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, 1554–1575, https://doi.org/10.1002/2014MS000416, 2015.

Khatiwala, S., Tanhua, T., Mikaloff Fletcher, S., 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, 2169–2191, https://doi.org/10.5194/bg-10-2169-2013, 2013.

Landschützer, P., Gruber, N., Bakker, D. C. E., and Schuster, U.: Recent variability of the global ocean carbon sink, Glob. Planet. Change, 28, 927–949, https://doi.org/10.1002/2014GB004853, 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, 349, 1221–1224, https://doi.org/10.1126/science.aab2620, 2015.

Landschützer, P., Gruber, N., and Bakker, D. C. E.: Decadal variations and trends of the global ocean carbon sink, Global Biogeochem. Cy., 30, 1396–1417, https://doi.org/10.1002/2015GB005359, 2016.

Landschützer, P., Gruber, N., and Bakker, D. C. E.: An updated observation-based global monthly gridded sea surface $p\mathrm{CO}_2$ and air-sea CO_2 flux product from 1982 through 2015 and its monthly climatology (NCEI Accession 0160558), Version 2.2, NOAA National Centers for Environmental Information, Dataset, available at: https://www.nodc.noaa.gov/ocads/oceans/SPCO2_1982_2015_ETH_SOM_FFN.html (last access: 8 February 2019), 2017.

Landschützer, P., Gruber, N., Bakker, D. C. E., Stemmler, I., and Six, K. D.: Strengthening seasonal marine CO_2 variations due to increasing atmospheric CO_2, Nat. Clim. Chang., 8, 146–150, https://doi.org/10.1038/s41558-017-0057-x, 2018.

Lebehot, A. D, Halloran, P. R., Watson, A. J., McNeall, D. J., Ford, D. A., Landschützer, P., Lauvset, S. K., and Schuster, U.: Reconciling observation and model trends in North Atlantic surface CO_2, Global Biogeochem. Cy., 33, 1204–1222, https://doi.org/10.1029/2019GB006186, 2019.

Lefévre, N., Watson, A. J., and Watson, A. R.: A comparison of multiple regression and neural network techniques for mapping in situ $p\mathrm{CO}_2$ data, Tellus B, 57, 375–384, https://doi.org/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 $p\mathrm{CO}_2$ and its drivers over the last two decades, Global Biogeochem. Cy., 26, GB2021, https://doi.org/10.1029/2011GB004095, 2012.

Lenton, A., Tilbrook, B., Law, R. M., Bakker, D., Doney, S. C., Gruber, N., Ishii, M., Hoppema, M., Lovenduski, N. S., Matear, R. J., McNeil, B. I., Metzl, N., Mikaloff Fletcher, S. E., Monteiro, P. M. S., Rödenbeck, C., Sweeney, C., and Takahashi, T.: Sea–air CO_2 fluxes in the Southern Ocean for the period 1990–2009, Biogeosciences, 10, 4037–4054, https://doi.org/10.5194/bg-10-4037-2013, 2013.

Le Quéré, C., Andrew, R. M., Friedlingstein, P., Sitch, S., Hauck, J., Pongratz, J., Pickers, P. A., Korsbakken, J. I., Peters, G. P., Canadell, J. G., Arneth, A., Arora, V. K., Barbero, L., Bastos, A., Bopp, L., Chevallier, F., Chini, L. P., Ciais, P., Doney, S. C., Gkritzalis, T., Goll, D. S., Harris, I., Haverd, V., Hoffman, F. M., Hoppema, M., Houghton, R. A., Hurtt, G., Ilyina, T., Jain, A. K., Johannessen, T., Jones, C. D., Kato, E., Keeling, R. F., Goldewijk, K. K., Landschützer, P., Lefèvre, N., Lienert, S., Liu, Z., Lombardozzi, D., Metzl, N., Munro, D. R., Nabel, J. E. M. S., Nakaoka, S., Neill, C., Olsen, A., Ono, T., Patra, P., Peregon, A., Peters, W., Peylin, P., Pfeil, B., Pierrot, D., Poulter, B., Rehder, G., Resplandy, L., Robertson, E., Rocher, M., Rödenbeck, C., Schuster, U., Schwinger, J., Séférian, R., Skjelvan, I., Steinhoff, T., Sutton, A., Tans, P. P., Tian, H., Tilbrook, B., Tubiello, F. N., van der Laan-Luijkx, I. T., van der Werf, G. R., Viovy, N., Walker, A. P., Wiltshire, A. J., Wright, R., Zaehle, S., and Zheng, B.: Global Carbon Budget 2018, Earth Syst. Sci. Data, 10, 2141–2194, https://doi.org/10.5194/essd-10-2141-2018, 2018.

Lueker, T. J., Dickson, A. G., and Keeling, C. D.: Ocean $p\mathrm{CO}_2$ 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, https://doi.org/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, https://doi.org/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, 375–384, https://doi.org/10.5194/essd-6-375-2014, 2014.

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, 469–472, https://doi.org/10.1038/nature16958, 2016.

McKinney, W.: Data Structures for Statistical Computing in Python, available at: https://conference.scipy.org/proceedings/scipy2010/pdfs/mckinney.pdf (last access: 16 February 2019), 2010.

Monteiro, P. M. S.: A Global Sea Surface Carbon Observing System: Assessment of Changing Sea Surface CO_2 and Air-Sea CO_2 Fluxes, in: Proceedings of OceanObs'09: Sustained Ocean Observations and Information for Society, Vol. 2, Venice, Italy, 21–25 September 2009, edited by: Hall, J., Harrison, D. E., and Stammer, D., ESA Publication WPP-306, https://doi.org/10.5270/OceanObs09.cwp.64, 2010.

Monteiro, P. M. S., Gregor, L., Lévy, M., Maenner, S., Sabine, C. L., and Swart, S.: Intraseasonal variability linked to sampling alias in air-sea CO_2 fluxes in the Southern Ocean, Geophys. Res. Lett., 42, 8507–8514, https://doi.org/10.1002/2015GL066009, 2015.

Nightingale, P. D., Malin, G., Law, C. S., Watson, A. J., Liss, P. S., Liddicoat, M. I., Boutin, J., and Upstill-Goddard, R. C.: In situ evaluation of air-sea gas exchange parameterizations using novel conservative and volatile tracers, Global Biogeochem. Cy., 14, 373–387, https://doi.org/10.1029/1999GB900091, 2000.

Olsen, A., Key, R. M., van Heuven, S., Lauvset, S. K., Velo, A., Lin, X., Schirnick, 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, 297–323, https://doi.org/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, https://doi.org/10.1007/s13398-014-0173-7.2, 2011.

Reynolds, R. W., Smith, T. M., Liu, C., Chelton, D. B., Casey, K. S., and Schlax, M. G.: Daily high-resolution-blended analyses for sea surface temperature, J. Climate, 20, 5473–5496, https://doi.org/10.1175/2007JCLI1824.1, 2007.

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, 8918–8925, https://doi.org/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, 12339–12348, https://doi.org/10.1002/2017GL074837, 2017.

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 CO_2 flux variability from an observation-driven ocean mixed-layer scheme, Biogeosciences, 11, 4599–4613, https://doi.org/10.5194/bg-11-4599-2014, 2014.

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 $p\mathrm{CO}_2$ Mapping intercomparison (SOCOM), Biogeosciences, 12, 7251–7278, https://doi.org/10.5194/bg-12-7251-2015, 2015.

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 CO_2, Science, 305, 367–371, https://doi.org/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., Manke, A., Malczyk, J., Akl, J., Alin, S. R., Bellerby, R. G. J., Borges, A., Boutin, J., Brown, P. J., Cai, W.-J., Chavez, F. P., Chen, A., Cosca, C., Feely, R. A., González-Dávila, M., Goyet, C., Hardman-Mountford, N., 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. S., Schlitzer, R., Schneider, B., Schuster, U., Sieger, R., Skjelvan, I., Steinhoff, T., Suzuki, T., Takahashi, 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 CO_2 Atlas (SOCAT) gridded data products, Earth Syst. Sci. Data, 5, 145–153, https://doi.org/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, 4319–4340, https://doi.org/10.5194/bg-10-4319-2013, 2013.

Sculley, D.: Web-Scale K-Means Clustering, in: Proceedings of the 19th international conference on World wide web – WWW'10, https://doi.org/10.1145/1772690.1772862, 2010.

Takahashi, T. T., Olafsson, J., Goddard, J. G., Chipman, D. W., and Sutherland, S. C.: Seasonal variation of CO_2 and nutrients in the high-latitude surface oceans: A comparative study, Global Biogeochem. Cy., 7, 843–878, https://doi.org/10.1029/93GB02263, 1993.

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 $p\mathrm{CO}_2$, and net sea-air CO_2 flux over the global oceans, Deep.-Sea Res. Pt. II, 56, 554–577, https://doi.org/10.1016/j.dsr2.2008.12.009, 2009.

Takahashi, T., Sutherland, S. C., and Kozyr, A.: Global Ocean Surface Water Partial Pressure of CO_2 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, https://doi.org/10.3334/CDIAC/OTG.NDP088(V2015), 2017.

Taylor, K. E.: Summarizing multiple aspects of model performance in a single diagram, J. Geophys. Res.-Atmos., 106, 7183–7192, https://doi.org/10.1029/2000JD900719, 2001.

Telszewski, M., Chazottes, A., Schuster, U., Watson, A. J., Moulin, C., Bakker, D. C. E., González-Dávila, M., Johannessen, T., Körtzinger, A., Lüger, H., Olsen, A., Omar, A., Padin, X. A., Ríos, A. F., Steinhoff, T., Santana-Casiano, M., Wallace, D. W. R., and Wanninkhof, R.: Estimating the monthly $p\mathrm{CO}_2$ distribution in the North Atlantic using a self-organizing neural network, Biogeosciences, 6, 1405–1421, https://doi.org/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, 2849–2866, https://doi.org/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.-Oceans, 122, 6968–6988, https://doi.org/10.1002/2016JC012650, 2017.

Wanninkhof, R. H., Asher, W. E., Ho, D. T., Sweeney, C., and McGillis, W. R.: Advances in Quantifying Air-Sea Gas Exchange and Environmental Forcing, Annu. Rev. Mar. Sci., 1, 213–244, https://doi.org/10.1146/annurev.marine.010908.163742, 2009.

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, 1–26, https://doi.org/10.3334/CDIAC/OTG.SOCAT_ADQCF, 2013a.

Wanninkhof, R., Park, G.-H., Takahashi, T., Sweeney, C., Feely, R., Nojiri, Y., Gruber, N., Doney, S. C., McKinley, G. A., Lenton, A., Le Quéré, C., Heinze, C., Schwinger, J., Graven, H., and Khatiwala, S.: Global ocean carbon uptake: magnitude, variability and trends, Biogeosciences, 10, 1983–2000, https://doi.org/10.5194/bg-10-1983-2013, 2013b.

Weiss, R.: Carbon dioxide in water and seawater: the solubility of a non-ideal gas, Mar. Chem., 2, 203–215, https://doi.org/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 $p\mathrm{CO}_2$ from biogeochemical Argo floats equipped with pH: An uncertainty analysis, Global Biogeochem. Cy., 31, 591–604, https://doi.org/10.1002/2016GB005541, 2017.

Woolf, D. K., Land, P. E., Shutler, J. D., Goddijn-Murphy, L., and Donlon, C. J.: On the calculation of air-sea fluxes of CO_2 in the presence of temperature and salinity gradients, J. Geophys. Res.-Oceans, 121, 1229–1248, https://doi.org/10.1002/2015JC011427, 2016.

Zeng, J., Nojiri, Y., Landschützer, P., Telszewski, M., and Nakaoka, S.: A global surface ocean $f\mathrm{CO}_2$ climatology based on a feed-forward neural network, J. Atmos. Ocean. Tech., 31, 1838–1849, https://doi.org/10.1175/JTECH-D-13-00137.1, 2014.

Zeng, J., Matsunaga, T., Saigusa, N., Shirai, T., Nakaoka, S., and Tan, Z.-H.: Technical note: Evaluation of three machine learning models for surface ocean CO_2 mapping, Ocean Sci., 13, 303–313, https://doi.org/10.5194/os-13-303-2017, 2017.