Biogeosciences

Research article

Assessing improvements in global ocean pCO_2 machine learning reconstructions with Southern Ocean autonomous sampling

Thea H. Heimdal[1], Galen A. McKinley[1], Adrienne J. Sutton[2], Amanda R. Fay[1], and Lucas Gloege[3]

[1]Lamont-Doherty Earth Observatory, Columbia University, Palisades, NY, USA
[2]Pacific Marine Environmental Laboratory, National Oceanic and Atmospheric Administration, Seattle, WA, USA
[3]Department of Earth and Planetary Sciences, Yale University, New Haven, CT, USA

Correspondence: Thea H. Heimdal (theimdal@ldeo.columbia.edu)

Abstract. The Southern Ocean plays an important role in the exchange of carbon between the atmosphere and oceans and is a critical region for the ocean uptake of anthropogenic CO_2. However, estimates of the Southern Ocean air–sea CO_2 flux are highly uncertain due to limited data coverage. Increased sampling in winter and across meridional gradients in the Southern Ocean may improve machine learning (ML) reconstructions of global surface ocean pCO_2. Here, we use a large ensemble test bed (LET) of Earth system models and the "pCO_2-Residual" reconstruction method to assess improvements in pCO_2 reconstruction fidelity that could be achieved with additional autonomous sampling in the Southern Ocean added to existing Surface Ocean CO_2 Atlas (SOCAT) observations. The LET allows for a robust evaluation of the skill of pCO_2 reconstructions in space and time through comparison to "model truth". With only SOCAT sampling, Southern Ocean and global pCO_2 are overestimated, and thus the ocean carbon sink is underestimated. Incorporating uncrewed surface vehicle (USV) sampling increases the spatial and seasonal coverage of observations within the Southern Ocean, leading to a decrease in the overestimation of pCO_2. A modest number of additional observations in Southern Hemisphere winter and across meridional gradients in the Southern Ocean leads to an improvement in reconstruction bias and root-mean-squared error (RMSE) of as much as 86 % and 16 %, respectively, as compared to SOCAT sampling alone. Lastly, the large decadal variability of air–sea CO_2 fluxes shown by SOCAT-only sampling may be partially attributable to undersampling of the Southern Ocean.

1 Introduction

The ocean plays an important role in mitigating climate change by sequestering anthropogenic carbon emissions. From 1850 to 2023, the oceans removed a total of 180 ± 35 Gt of carbon (Friedlingstein et al., 2023). In order to fully understand the climate impacts from rising emissions, it is essential to accurately quantify the air–sea CO_2 flux and the global ocean carbon sink in space and time. The Surface Ocean CO_2 ATlas (SOCAT; Bakker et al., 2016) is the largest global database of surface ocean CO_2 observations, with data starting in 1957. The main synthesis and gridded products contain over 33 million high-quality direct shipboard measurements of fCO_2 (fugacity of CO_2), with an uncertainty of < 5 µatm (Bakker et al., 2022). However, due to the limited resources for ocean observing, limited number of ships and/or routes, inaccessible regions and unsafe waters, the database covers only about 1 % of the global ocean at a monthly $1° \times 1°$ spatial resolution over the period 1982–2023 and is highly biased towards the Northern Hemisphere.

Mapping methods have been developed to estimate full-coverage surface ocean pCO_2 across space and time by extrapolating to global coverage from these sparse SOCAT observations (e.g., Landschützer et al., 2014; Rödenbeck et al., 2015; Gloege et al., 2022; Bennington et al., 2022a, b). Most of these data products utilize machine learning (ML) algorithms to estimate a non-linear function between a suite of driver variables (i.e., sea surface temperature, SST; sea surface salinity, SSS; mixed layer depth, MLD; Chlorophyll a, Chl a; xCO_2, atmospheric CO_2) and surface ocean pCO_2 (the target variable) where these are co-located. The driver

variables are proxies for processes influencing ocean $p\mathrm{CO_2}$. Full-coverage driver variable datasets are then processed through these ML algorithms to produce estimated global full-coverage surface ocean $p\mathrm{CO_2}$. Since the data products rely on $p\mathrm{CO_2}$ observations to estimate functions between the target and driver variables, data sparsity remains a fundamental limitation of this technique.

It has been suggested that targeted sampling from autonomous platforms combined with ships, filling in the state space of $p\mathrm{CO_2}$, represents a path forward to improve surface ocean $p\mathrm{CO_2}$ reconstructions (Bushinsky et al., 2019; Gregor et al., 2019; Gloege et al., 2021; Djeutchouang et al., 2022; Landschützer et al., 2023; Hauck et al., 2023). One major obstacle, however, is that the indirect $p\mathrm{CO_2}$ estimates from floats have high uncertainties ($\pm 11.4\,\mu$atm) and may be biased by as much as $\sim 4\,\mu$atm (Bakker et al., 2016; Williams et al., 2017; Fay et al., 2018; Gray et al., 2018; Sutton et al., 2021; Mackay and Watson, 2021; Wu et al., 2022). These large uncertainties and biases arise when $p\mathrm{CO_2}$ is not measured directly as in the observations included in SOCAT but is rather estimated using measurements of pH combined with a regression-derived alkalinity estimate (Williams et al., 2017; Gray et al., 2018). SOCAT includes only direct $p\mathrm{CO_2}$ observations. Biases and uncertainties may have large impacts on global air–sea $\mathrm{CO_2}$ flux estimates given that the global mean air–sea disequilibrium is only 5–8 μatm (McKinley et al., 2020). It is therefore critical that bias and uncertainty corrections are well constrained over different oceanic conditions and over time.

Uncrewed surface vehicles (USVs), such as those manufactured and maintained by Saildrone, Inc., represent a new type of autonomous platform that can obtain direct $p\mathrm{CO_2}$ observations with significantly lower uncertainties compared to other autonomous methods and equivalent to the highest-quality shipboard measurements contained in SOCAT ($\pm 2\,\mu$atm; Sabine et al., 2020; Sutton et al., 2021). Such improvements in sampling are critically important in the undersampled Southern Ocean. This region is fundamental in terms of the ocean's ability to remove carbon from the atmosphere, being responsible for $\sim 40\,\%$ of the global ocean uptake of anthropogenic $\mathrm{CO_2}$ (Khatiwala et al., 2009). Improved data coverage in the Southern Ocean thus represents a major opportunity to advance our understanding of the global ocean carbon sink (Lenton et al., 2006, 2013; Takahashi et al., 2009; Monteiro et al., 2015; Gregor et al., 2019; Gray et al., 2018; Mongwe et al., 2018; Bushinsky et al., 2019; Sutton et al., 2021; Long et al., 2021; Mackay et al., 2022; Wu et al., 2022; Landschützer et al., 2023; Hauck et al., 2023). A combination of SOCAT and Saildrone USV observations would include high-accuracy data from both the long-record and global coverage of ship tracks and the expanded finer resolution of spatial and seasonal coverage of the poorly sampled Southern Ocean. Importantly, Saildrone USVs are also able to cover the spatial extent and seasonal cycle of the meridional gradients, which has been shown to be criti-

cal in order to reduce errors in reconstructing surface ocean $p\mathrm{CO_2}$ (Djeutchouang et al., 2022). A combined approach, with autonomous samples such as those obtained from Saildrone USVs and high-quality observations collected from ships, thus represents a promising solution to improve surface ocean $p\mathrm{CO_2}$ ML reconstructions.

Here, we assess to what extent surface ocean $p\mathrm{CO_2}$ reconstructions can improve by implementing the $p\mathrm{CO_2}$-Residual machine learning (ML) reconstruction (Bennington et al., 2022a) with the combined inputs of SOCAT and Saildrone USV coverage. However, instead of using real-world observations, we sample the target (i.e., surface ocean $p\mathrm{CO_2}$) and driver variables (i.e., SST, SSS, MLD, Chl a and $x\mathrm{CO_2}$) from our large ensemble test bed (LET) of Earth system models (ESMs) (e.g., Stamell et al., 2020; Gloege et al., 2021; Bennington et al., 2022a). There are two major benefits of using a test bed compared to actual observations. First, in an ESM, the surface ocean $p\mathrm{CO_2}$ field is provided precisely at all model times and $1° \times 1°$ points. Therefore, the $p\mathrm{CO_2}$ reconstructed by the ML algorithm can be robustly evaluated in space and time against a known "truth" (i.e., "model truth"). The reconstruction evaluation is thus not limited to the availability of sparse real-world ocean observations. Secondly, a test bed can be used to plan and evaluate the impact of different sampling strategies on the reconstructed $p\mathrm{CO_2}$. It is important to stress that, by using a model test bed, we do not predict real-world surface ocean $p\mathrm{CO_2}$ and air–sea $\mathrm{CO_2}$ fluxes. The goal here is to assess the accuracy with which an ML algorithm can reconstruct the model truth when given inputs of samples consistent with real-world data coverage from the SOCAT database and Saildrone USVs.

By utilizing the observational coverage of SOCAT and Saildrone USV transects, we assess to what extent the $p\mathrm{CO_2}$-Residual method accurately reconstructs model surface ocean $p\mathrm{CO_2}$ in space and time. We test the impact of two different USV Southern Ocean sampling schemes – the first based on a sampling campaign completed in 2019 (Sutton et al., 2021) and the second on logistically feasible potential future meridional sampling. Additionally, we explore the timing, magnitude, duration and spatial extent of Southern Ocean USV sample additions that most significantly improve the $p\mathrm{CO_2}$ predictions. Combined, the sampling patterns tested here complement previous studies exploring the impact of additional sampling in the Southern Ocean based on idealized full global coverage of floats and float observations from recent deployments, including the Southern Ocean Carbon and Climate Observations and Modeling (SOCCOM) project, moorings and sailboats (Bushinsky et al., 2019; Denvil-Sommer et al., 2021; Djeutchouang et al., 2022; Hauck et al., 2023; Behncke et al., 2024; Landschützer et al., 2023).

2 Methods

2.1 The large ensemble test bed (LET)

In this study, the large ensemble test bed (LET) includes 25 members from three independent initial-condition ensemble models (i.e., CanESM2, CESM-LENS and GFDL-ESM2M; Kay et al., 2015; Rodgers et al., 2015; Fyfe et al., 2017), giving a total of 75 members within the test bed. We do not use the MPI-GE model that was included in the past LET studies because its Southern Ocean pCO$_2$ seasonality and decadal variability appear to be anomalously large (Gloege et al., 2021; Fay and McKinley, 2021; Bennington et al., 2022a). Each individual Earth system model (ESM) is an imperfect representation of the actual Earth system, so the multiple large ensembles are used to span different model structures and their representation of internal variability. Each ensemble member undergoes the same external forcing (i.e., historical atmospheric CO$_2$ before 2005 and Representative Concentration Pathway 8.5 through 2016, as well as solar and volcanic forcing), but the spread across the ensemble members gives a unique trajectory of the ocean–atmosphere state over time (i.e., a different state of internal variability and the difference across models).

The LET used in this study includes monthly $1° \times 1°$ model output from 1982–2016 (Gloege et al., 2021). For each individual ensemble member of the LET, surface ocean pCO$_2$ and co-located driver variables (i.e., SST, SSS, Chl a, MLD and xCO$_2$) were sampled monthly at a $1° \times 1°$ resolution at times and locations equivalent to SOCAT and Saildrone USV observations (Fig. 1; Step 1). While the SOCAT observations were sampled from the test bed matching the actual years of sampling, the USV observations were sampled from the test bed starting in 2007 (for 10-year sampling) or 2012 (for 5-year sampling) (see Sect. 2.4). As our focus is on reconstruction for the open ocean, test bed output for coastal areas, the Arctic Ocean ($> 79°$ N) and marginal seas (Hudson Bay, Caspian Sea, Black Sea, Mediterranean Sea, Baltic Sea, Java Sea, Red Sea and Sea of Okhotsk) was removed prior to algorithm processing.

2.2 The pCO$_2$-Residual approach

We used the pCO$_2$-Residual approach following Bennington et al. (2022a), which removes the well-studied direct effect of temperature on pCO$_2$ from the LET model output before algorithm processing. Temperature has both direct and indirect effects on surface ocean pCO$_2$. The direct effect of temperature, due to solubility and chemical equilibrium, is that an increase in temperature directly causes an increase in pCO$_2$ (Takahashi et al., 1993). Indirectly, temperature changes are associated with biological production and wintertime vertical mixing, and these processes tend to result in opposing pCO$_2$ changes. To build reconstruction algorithms through the data-driven training that occurs in ML, the statis-

tics in all other algorithms developed to date must identify a function that disentangles these competing effects of SST on pCO$_2$. Here, the algorithm is assisted by removing this known temperature effect, and it must therefore only learn the pCO$_2$ impacts from biogeochemical drivers. The pCO$_2$-Residual method leads to physically understandable connections between the input data and the output (Bennington et al., 2022a), which mitigates to some degree "black box" concerns typically associated with ML algorithms (Toms et al., 2020). Bennington et al. (2022a) demonstrate higher skill for reconstructions using pCO$_2$-Residual as the target variable as opposed to pCO$_2$ (Fig. S1 in Bennington et al., 2022a), indicating that the removal of the temperature-driven component enhances the performance of the method. Further, the pCO$_2$-Residual method has been shown to perform slightly better against independent observations than other common mapping methods (Bennington et al., 2022a). A brief description is provided here, but for further details, see Bennington et al. (2022a).

The temperature-driven component of pCO$_2$ (pCO$_2$-T) is calculated using the following equation:

$$p\text{CO}_2\text{-}T = p\text{CO}_2^{\text{mean}} \times \exp[0.0423 \times (\text{SST} - \text{SST}^{\text{mean}})],$$

where pCO$_2^{\text{mean}}$ and SST$^{\text{mean}}$ are the long-term means of surface ocean pCO$_2$ and temperature, respectively, using all $1° \times 1°$ grid cells from the test bed. Alternative sources of mean pCO$_2$ were assessed by Bennington et al. (2022a), but they found no significant impact on the test statistics or reconstructed pCO$_2$. Once pCO$_2$-T is determined, pCO$_2$-Residual is calculated as the difference between pCO$_2$ and the calculated pCO$_2$-T:

$$p\text{CO}_2\text{-Residual} = p\text{CO}_2 - p\text{CO}_2\text{-}T.$$

Prior to algorithm processing, pCO$_2$-Residual values > 250 and < -250 µatm from the test bed were filtered out, targeting values that are not representative of the real ocean. The majority of the pCO$_2$-Residual values that were filtered out correspond to high pCO$_2$ values above the maximum value in SOCAT (816 µatm; Stamell et al., 2020). The excluded data points (less than 0.2 % per member) mostly occurred in the output from the CanESM2 model and were restricted geographically, predominantly along the western coastline of South America.

The eXtreme Gradient Boosting method (XGB; Chen and Guestrin, 2016) is used to develop an algorithm that allows driver variables (i.e., SST, SSS, Chl a, MLD and xCO$_2$) to predict the pCO$_2$-Residual (Fig. 1; Step 2). The pCO$_2$-Residual and associated feature variables are split into validation, training and testing sets. The test and validation sets each account for 20 % of the data, leaving 60 % for training. The validation set is used to optimize the algorithm hyperparameters, which define the architecture of decision trees used in the model. The training set is used to build the decision trees in XGB, while the test set is used to evaluate
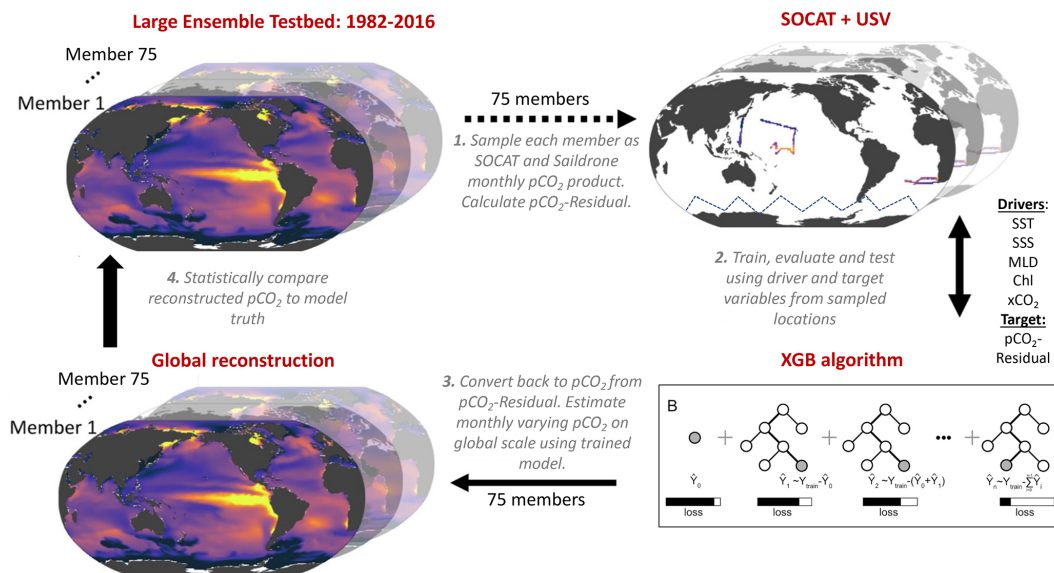
Figure 1. Schematic of the large ensemble test bed (LET; modified from Gloege et al., 2021). (1) Surface ocean pCO_2 from each of the 75 model members is sampled in space and time mimicking real-world SOCAT and Saildrone USV observations (see Fig. 2; Table 1; Sect. 2.4). Prior to algorithm processing, pCO_2-Residual is calculated (Sect. 2.2). (2) The pCO_2-Residual (target variable) and co-located driver variables (i.e., SST, SSS, MLD, Chl and xCO_2) sampled from the test bed are processed by the XGBoost (XGB) algorithm (Sect. 2.2). (3) Based on the full coverage of driver variables, pCO_2-Residual is reconstructed globally. This process is repeated 75 times, individually for every single test bed model member. The temperature component (pCO_2-T) is then added back to the pCO_2-Residual for each value. (4) The globally reconstructed pCO_2 is evaluated against the model truth at all $1° \times 1°$ grid cells. SST: sea surface temperature. SSS: sea surface salinity. MLD: mixed layer depth. Chl: chlorophyll. xCO_2: atmospheric concentration of CO_2.

the performance of the final algorithm. The XGB algorithm for this study used 4000 decision trees with a maximum depth of six levels, and this was fixed for all experiments (see Sect. S1 in the Supplement). For the final reconstruction of surface ocean pCO_2 across all space and time points, the previously calculated pCO_2-T values are added back to the reconstructed pCO_2-Residual (Fig. 1; Step 3).

The full XGB process, including (1) training, evaluating and testing and (2) reconstructing globally at a monthly resolution, was repeated individually for each LET member. This process therefore provided a total of 75 unique reconstruction vs. model truth pairs, which can be statistically compared (Fig. 1; Step 4).

2.3 Statistical analysis in the test bed

The statistical comparisons between the test set and the reconstructions are equivalent to what would be derived using real-world data ("seen" values). Here, we calculate error statistics based on the full reconstruction (pCO_2 from all $1° \times 1°$ grid cells of the test bed, except for those masked or filtered out). In the full reconstruction, $\sim 99\%$ of the data do not correspond to SOCAT or Saildrone USV observations used to train the algorithm (Fig. S1 in the Supplement). Training data would ideally be removed before performance evaluation, but since the training data represent only $\sim 1\%$, the impact of not removing them is negligible (Fig. S2). A

suite of statistical metrics can be used to compare the reconstruction to the model truth in order to assess how well the algorithm can extrapolate from sparse data to full-field coverage (Fig. 1; Step 4). In this study, we focus on the bias and root-mean-squared error (RMSE). Bias is calculated as mean prediction minus mean observation (i.e., pCO_2 predicted by XGB subtracted by the pCO_2 model truth) and is a measure of over- or underestimation in the reconstructions. RMSE measures the magnitude of the predicted error and is calculated as the square root of the mean of the squared errors. We focus our discussion on the mean across 75 members of the test bed for bias and the RMSE. The spread across test bed ensemble members is non-negligible and will be the focus of future work; here, we present the test bed spread primarily in the Supplement.

2.4 Overview of sampling patterns and model runs

First, we sampled target and driver variables from the LET based on sampling distributions equivalent to that of the SOCAT database ("SOCAT baseline"). Then, we combined the SOCAT baseline with test bed output representing additional Saildrone USV coverage in the Southern Ocean. The additional Southern Ocean coverage was based on (1) the Sutton et al. (2021) sampling campaign from 2019 ("one-latitude" track) and (2) realistic potential future meridional USV observations ("zigzag" track) (see Sect. 2.4.2; Fig. 2). We per-

formed a total of 10 experimental runs (Table 1). These represent different sampling approaches, including (1) repeating USV sampling over a period of 5 or 10 years, (2) varying the number of USVs and thus the total number of monthly $1° \times 1°$ observations, and (3) restricting all observations to Southern Hemisphere winter months. By comparing the different runs, we can assess whether or not certain targeted sampling strategies in the Southern Ocean can improve surface ocean $p\mathrm{CO_2}$ ML reconstructions. As discussed above, the LET runs until 2016 only (Gloege et al., 2021). Saildrone USV observations were therefore sampled from the test bed starting in year 2006 or 2007 (for the 10-year sampling) or 2012 (for the 5-year sampling) and lasting until 2016, i.e., the final year of the test bed.

2.4.1 One-latitude runs

Out of the 10 experimental runs, 6 include the one-latitude track (Table 1). The 2019 Saildrone USV journey (Sutton et al., 2021) covered an 8-month period, from January to August. Since the USV was recovered in early August, it did not cover the entire Southern Hemisphere winter (Fig. S3). We repeated this one-latitude 8-month sampling pattern for 5 years (5Y_J-A; 2075 observations) and 10 years (10Y_J-A; 4150 observations). To evaluate year-round (YR) coverage, the 8-month sampling period (January–August) was shifted by 1 month each year for 10 years (10Y_YR; 4150 observations). To evaluate the impact of increased sampling, the 2019 Saildrone USV track was repeated 12 times with incremental offsets of $1°$ from the original track, covering an additional $6°$ north and south (Fig. S4). This "high-sampling" run (x13_10Y_J-A; 44 250 observations) represents a total of 13 USVs. We also performed an additional 13 USV runs but including observations from Southern Hemisphere winter (W) months only (x13_10Y_W; 25 395 observations). Finally, considering the cost of deploying 13 USVs, a downscaled multiple-USV winter-only run was tested, including five USVs sampling over a period of 5 years (x5_5Y_W; 5022 observations). This run covers an additional $2°$ north and south from the original USV track.

2.4.2 Zigzag runs

Of the 10 experimental runs, 4 represent realistic potential meridional sampling in the Southern Ocean (zigzag tracks; Table 1), as suggested by Djeutchouang et al. (2022). Saildrone USVs can operate at a speed capable of covering the spatial extent of meridional gradients in the Southern Ocean (Djeutchouang et al., 2022). However, Saildrone USVs are solar-powered, and thus, their range is restricted by the availability of solar radiation. To account for this and maintain a realistic sampling scenario, sampling occurs only to a maximum latitude of $55°$ S in these experiments. This alternative sampling pattern represents USVs sailing west to east in a north–south zigzag pattern covering 40 and $55°$ S for
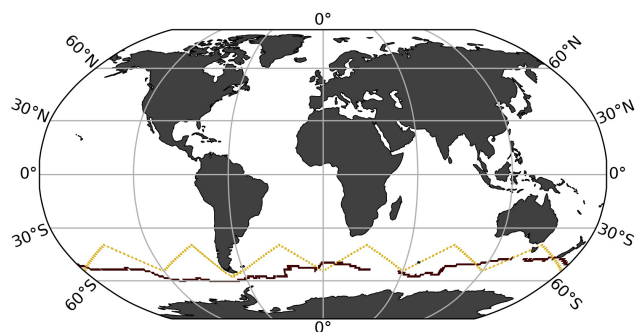

Figure 2. Saildrone uncrewed surface vehicle (USV) tracks representing the first circumnavigation around Antarctica from 2019 in maroon (one-latitude track; Sutton et al., 2021) and an alternative virtual route with meridional coverage (zigzag track).

every $30°$ of longitude (Fig. 2). We created two scenarios. For the first scenario, every $30°$ of longitude from 40 and $55°$ S is visited every 3 months within a single year, as suggested by Lenton et al. (2006). Assuming an average Saildrone USV speed, this scenario represents four platforms equally spaced around the Southern Ocean. This sampling pattern was repeated for 10 years, with year-round coverage (Zx4_10Y_YR; 7600 observations), and for Southern Hemisphere winter months only (Zx4_10Y_W; 2500 observations). The second scenario represents a high-sampling strategy, where every $30°$ of longitude from 40 and $55°$ S is visited approximately monthly. This can be achieved by deploying 10 platforms equally spaced around the Southern Ocean, running at an average Saildrone USV speed. This sampling pattern is repeated for 5 years, sampling year-round (Z_x10_5Y_YR; 11 400 observations) and during Southern Hemisphere winter months only (Z_x10_5Y_W; 3800 observations).

2.5 Air–sea $\mathrm{CO_2}$ flux

To assess the global ocean carbon sink associated with our $p\mathrm{CO_2}$ reconstructions, air–sea $\mathrm{CO_2}$ exchange was calculated for 1985 onward. Here, we computed air–sea $\mathrm{CO_2}$ fluxes using the bulk formulation with the Python package SeaFlux 1.3.1 (https://github.com/lukegre/SeaFlux, last access: August 2023; Gregor and Fay, 2021; Fay et al., 2021). We calculated global and Southern Ocean flux in the same manner for (1) the test bed model truth, (2) the SOCAT baseline and (3) the 10 experimental USV runs.

The net sea–air $\mathrm{CO_2}$ flux was estimated using

$$\mathrm{flux} = k_w \cdot \mathrm{sol} \cdot (p\mathrm{CO_2^{ocn}} - p\mathrm{CO_2^{atm}}) \cdot (1 - \mathrm{ice}),$$

where k_w is the gas transfer velocity; sol is the solubility of $\mathrm{CO_2}$ in seawater (in units of $\mathrm{mol\,m^{-3}\,\mu atm^{-1}}$); $p\mathrm{CO_2^{ocn}}$ is the partial pressure of surface ocean carbon (in µatm), either from the model truth or from the reconstructions; and $p\mathrm{CO_2^{atm}}$ (in µatm) is the partial pressure of atmospheric $\mathrm{CO_2}$

Table 1. Overview of the different sampling experiments tested in this study, mean bias and RMSE (in µatm) for various time periods, and latitude bands for all runs. Bold values represent the best score for each category. One-lat: one-latitude track, which incorporates the Saildrone USV route from Sutton et al. (2021). Zigzag: potential meridional sampling. Additional observations: number of 1° × 1° monthly Saildrone USV observations in addition to SOCAT. J–A: January–August. YR: year-round. W: Southern Hemisphere winter. ×4, ×5, ×10 and ×13 = 4, 5, 10 and 13 USVs. SO winter: Southern Ocean winter months, i.e., June, July and August, and also including September, not applicable – n/a.

Run name	SOCAT baseline	5Y_J-A	10Y_J-A	10Y_YR	x13_10Y_J-A	x13_10Y_YR	x13_10Y_W	x5_5Y_W	Z_x4_10Y_YR	Z_x4_10Y_W	Z_x10_5Y_YR	Z_x10_5Y_W
Saildrone track	n/a	One-lat	One-lat	One-lat	One-lat	One-lat	One-lat	One-lat	Zigzag	Zigzag	Zigzag	Zigzag
Years of sampling	n/a	5	10	10	10	10	10	5	10	10	5	5
Duration of sampling	n/a	Jan–Aug	Jan–Aug	Year-round	Jan–Aug	SO winter	SO winter	SO winter	Year-round	SO winter	Year-round	SO winter
Additional observations	n/a	2075	4150	4150	44250	25395	5022	7600	2500	11400	3800	
Global coverage increase (%)	n/a	0.01	0.02	0.02	0.3	0.1	0.03	0.04	0.01	0.07	0.02	
Mean bias (µatm)												
Test bed period (1982–2016)												
Globally	0.63	0.59	0.59	0.52	**0.39**	0.53	0.57	0.51	0.45	0.51	0.44	
North (35–90°N)	**0.11**	0.24	0.20	0.25	0.20	0.17	0.16	0.12	0.16	0.20		
Mid (35°S–35°N)	0.23	0.21	0.22	0.14	0.15	0.20	0.23	0.20	**0.13**	0.18	0.18	
South (90–35°S)	1.4	1.3	1.2	1.1	**0.80**	1.1	1.2	1.1	1.0	1.1	0.87	
SO winter months (JJA)	1.3	1.2	1.2	1.1	**0.90**	1.1	1.2	0.93	0.94	1.0	0.95	
SO summer months (DJF)	0.070	0.11	0.15	0.10	0.15	0.15	0.11	0.25	0.16	0.073	0.066	
2006/2012–2016												
Globally	0.51*	0.27	0.34	0.28	**0.03**	0.19	0.21	0.23	0.17	0.24	0.07	
South (90–35°S)	1.6*	0.93	1.1	1.0	**0.37**	0.72	0.73	0.89	0.92	0.67	0.55	
South (90–35°S) (JJA)	4.2*	2.6	2.7	2.8	**0.019**	2.2	2.5	1.8	**1.2**	2.4	2.0	
Mean RMSE (µatm)												
Test bed period (1982–2016)												
Globally	11.8	11.7	11.8	11.7	11.7	11.7	11.7	**11.5**	11.5	11.6	11.5	
North (35–90°N)	13.0	13.0	13.0	13.0	13.0	13.0	13.1	13.0	13.0	13.0	13.0	
Mid (35°S–35°N)	11.7	11.7	11.7	11.7	11.7	11.7	11.7	11.7	11.7	11.7	11.7	
South (90–35°S)	11.5	11.3	11.4	11.2	11.0	11.3	10.7	10.6	**10.6**	11.0	11.0	
2006/2012–2016												
Globally	11.6*	11.6	11.4	11.3	**11.0**	11.2	11.2	11.1	11.1	11.2	11.4	
South (90–35°S)	11.4*	11.1	11.0	10.7	10.6	10.4	10.0	**9.7**	10.6	10.6		
South (90–35°S) (JJA)	12.0*	11.3	11.2	10.9	10.5	10.3	10.3	**9.6**	10.6	10.3		

* Average value of the mean for 2006–2016 and 2012–2016. The global coverage increase was calculated based on the total number of available 1982–2016 monthly 1° × 1° observations from SOCAT (262 204 observations) and the large ensemble test bed (17 290 470 observations).

in the marine boundary layer. For GFDL, we used the direct model output of $p\mathrm{CO}_2^{atm}$, while for CESM and CanESM2, $p\mathrm{CO}_2^{atm}$ was calculated individually as the product of surface $x\mathrm{CO}_2$ and sea level pressure (the contribution of water vapor pressure was corrected for in CESM). Finally, to account for the seasonal ice cover in high latitudes, the fluxes were weighted by 1 minus the ice fraction (ice), i.e., the open-ocean fraction.

Winds have the largest impact on flux calculations (Fay et al., 2021), and temporally high-resolution output is not available for the LET. Monthly output is available, but this is not sufficient for the flux calculation due to the square dependency of wind speed (Wanninkhof, 2014). Given the necessity to use observed winds, for consistency, we use observations for all necessary variables for the flux calculation. Inputs to the calculation include EN.4.2.2 salinity (Good et al., 2013), SST and ice fraction from NOAA Optimum Interpolation Sea Surface Temperature V2 (OI SST v2) (Reynolds et al., 2002), and surface winds and associated wind scaling factor from the European Centre for Medium-Range Weather Forecasts (ECMWF ERA5) sea level pressure (Hersbach et al., 2020). Results presented show the global and Southern Ocean ($< 35°$ S) fluxes in units of $\mathrm{Pg\,C\,yr}^{-1}$.

Note that reconstructions of $p\mathrm{CO}_2$ for the SOCAT baseline and the experimental USV runs are limited in their spatial extent to the open ocean (see Sect. 2.1; excluding coastal areas, the Arctic Ocean and marginal seas). The same mask was thus also applied when calculating the flux of the model truth, prior to comparison with the reconstructions.

3 Results

3.1 Performance metrics for the SOCAT baseline reconstruction

The mean bias for the entire test bed period (i.e., 1982–2016) is 0.63 µatm globally (Fig. 3a) and 1.4 µatm for the Southern Ocean ($< 35°$ S; Table 1). Bias is much closer to 0 for the mid-latitudes (between $35°$ S and $35°$ N; 0.23 µatm) and northern latitudes ($> 35°$ N; 0.11 µatm) (Fig. 3a). There is a significant difference in bias in the case of Southern Hemisphere winter months (June, July, August) versus summer months (December, January, February), with a global mean bias (for 1982–2016) of 1.3 µatm compared to 0.07 µatm, respectively (Table 1), due to the sparseness of SOCAT observations from the Southern Hemisphere during the harsh winter season (Fig. S5a). The mean RMSE for the entire test bed period (i.e., 1982–2016) is 11.8 µatm globally (Fig. 3b) and 11.5 µatm for the Southern Ocean (Table 1). RMSE is highest in the eastern tropical and southeastern Pacific Ocean and in the Southern Ocean, where the algorithm generally overestimates $p\mathrm{CO}_2$ (i.e., positive bias; Fig. 3a), with some exceptions in the Atlantic section. This is consistent with the areas significantly undersampled by SOCAT (Fig. S5b). Ex-
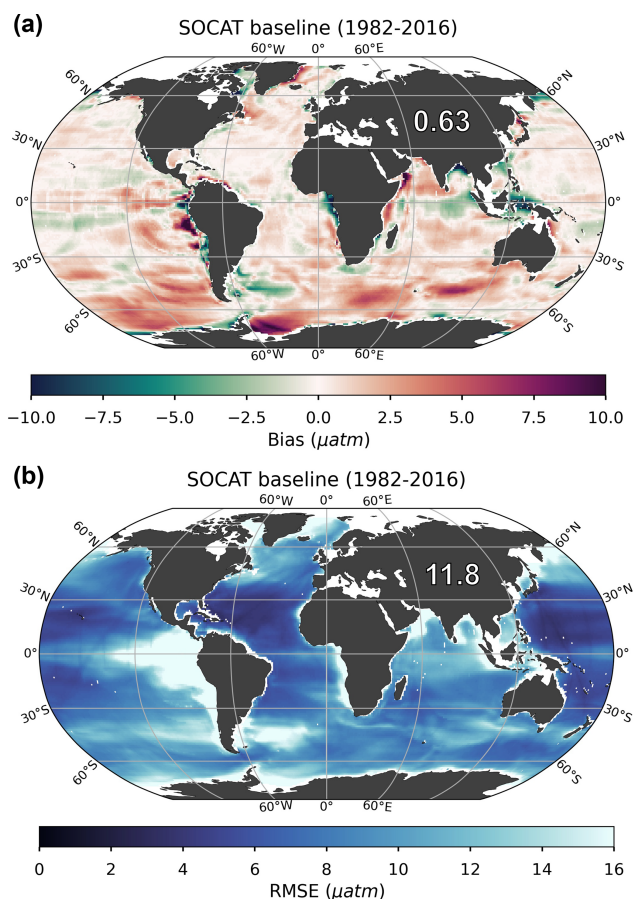

Figure 3. Bias **(a)** and root-mean-squared error (RMSE) **(b)** for the SOCAT baseline (i.e., no USV) over the period of 1982 through 2016. The global mean bias and RMSE are 0.63 and 11.8 µatm, respectively. Note that only the open ocean was considered in the reconstruction, so several areas were masked out prior to algorithm processing, such as the Arctic Ocean, coastal areas and marginal seas (no data; white areas in figures).

cept in these areas, RMSE and bias are generally low (close to 0) in the open ocean but show higher values along coastlines (Fig. 3b). The predicted $p\mathrm{CO}_2$ is thus more accurate in areas similar to and surrounding the SOCAT observations (i.e., monthly $1° \times 1°$ grid cells equivalent to SOCAT coverage but sampled from the LET). Figure 3 shows mean bias and RMSE for the full reconstruction (see Sect. 2.3), but note that there is a statistically significant difference between the training and testing set errors (Fig. S6). This indicates potential overfitting in our ML model (i.e., higher errors for the "unseen" reconstruction) and that further tuning of the hyperparameters could increase generalization skill (see Sect. S1).

3.2 Reconstruction improvements with Saildrone USV additions

Our presentation of global maps is limited to runs x5_5Y_W (5022 monthly $1° \times 1°$ observations) and Z_x4_10Y_YR
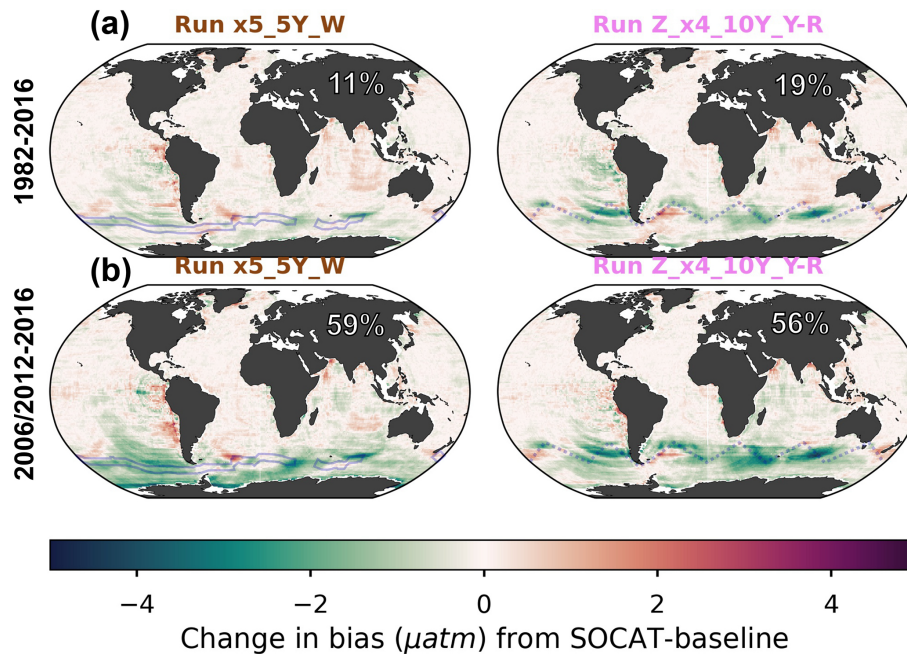
Figure 4. Change in bias when comparing runs x5_5Y_W and Z_x4_10Y_YR to the SOCAT baseline reconstruction averaged over the duration of the test bed period (**a**; 1982–2016) and the period of USV additions (**b**; 2006–2012 or 2012–2016). The percent global improvement in absolute bias is shown in each panel. The USV Saildrone tracks are shown in blue.

(7600 monthly $1° \times 1°$ observations). These runs were selected as they represent observational schemes that are realistic in the near-term future considering logistics and cost level, both non-meridional and meridional sampling, and different approaches to observing duration and seasonal coverage. For the remaining runs, equivalent maps can be found in the Supplement.

3.2.1 Bias

All Saildrone USV runs show a reduction in bias compared to the global mean 1982–2016 SOCAT baseline (Figs. 4a, S7). The improvement in bias is mainly due to lower reconstructed pCO$_2$ values at southern latitudes, where the SOCAT baseline reconstruction generally overestimates pCO$_2$ (Fig. 3a). The global mean bias for zigzag run Z_x4_10Y_YR is 0.51 µatm, a higher improvement (19 %) over the SOCAT baseline compared to the one-latitude run x5_5Y_W (11 % mean improvement; mean bias = 0.57 µatm) (Fig. 4a; Table 1). Generally, the zigzag runs show higher improvements from the SOCAT baseline (19 %–31 % improvement; resulting mean bias of 0.44–0.51 µatm) compared to the one-latitude runs (7 %–19 % improvement; resulting mean bias of 0.52–0.59 µatm) (Fig. S7; Table 1). However, the one-latitude run x13_10Y_W that samples Southern Hemisphere winter months only stands out with the lowest global mean (1982–2016) bias of 0.39 µatm, representing a 39 % mean improvement from the SOCAT baseline (Table 1; Fig. S7). This run, however, has 3 and 5 times more

observations (25 395) than Z_x4_10Y_YR and x5_5Y_W, respectively.

Compared to the entire test bed period, even larger improvements in global mean bias are shown for the period of Saildrone USV additions (2006–2016 and 2012–2016; Figs. 4a vs. 4b, S7 vs. S8). Compared to the SOCAT baseline, run x13_10Y_W results in a mean bias improvement of 95 %, while the remaining one-latitude runs and the zigzag runs show mean improvements up to 63 % and 86 %, respectively (Fig. S8). The spread in mean bias (2006/2012–2016) across the 75 test bed members for each experiment is shown in Fig. S9.

Perhaps surprisingly, there is not a strong connection between the global or Southern Ocean mean bias and the number of added USV observations (Fig. 5). The one-latitude high-sampling run x13_10Y_J-A (44 250 observations) shows similar mean bias or is outperformed by all zigzag runs and the one-latitude runs that restrict sampling to Southern Hemisphere winter months (i.e., x5_5Y_W and x13_10Y_W).

Considering the change in bias from year to year, the SOCAT baseline shows positive bias at all latitudes in the beginning of the test bed period, before improvement occurs around 1990 (Fig. 6a). This is consistent with increasing SOCAT sampling with time for the period considered here (i.e., up to 2016; Fig. S5c). As SOCAT observations are biased towards the Northern Hemisphere (Fig. S5a, b), bias in the Southern Ocean ($< 35°$ S) increases significantly starting in the 2000s and remains high until the end of the test bed pe-
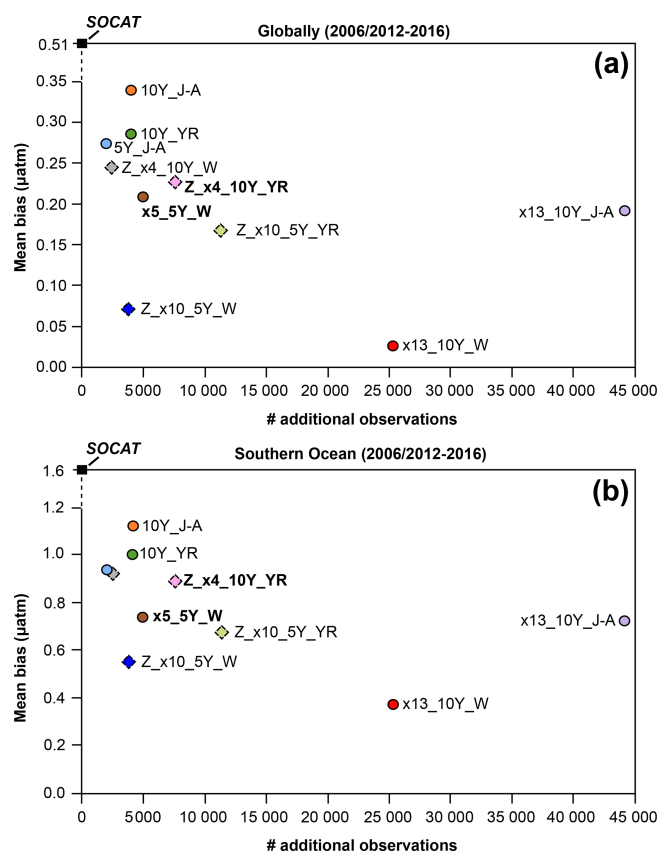
Figure 5. Mean bias globally **(a)** and for the Southern Ocean **(b)** for the duration of Saildrone USV sampling (2006–2016 or 2012–2016) for all runs presented in Table 1. Circles represent runs using the one-latitude track, while diamonds represent zigzag runs. Runs highlighted in bold correspond to the two selected runs mapped in Figs. 4, 6, 7 and 9. Global (0.51 µatm) and Southern Ocean (1.6 µatm) bias values shown for the SOCAT baseline (black squares) represent a mean of values for 2006–2016 (global mean of 0.52 µatm; Southern Ocean mean of 1.63 µatm) and 2012–2016 (global mean of 0.51 µatm; Southern Ocean mean of 1.56 µatm). The number of monthly $1° \times 1°$ USV observations in addition to SOCAT is denoted by # additional observations. Box plots illustrating the spread across the 75 ensemble members are shown in Fig. S9.

riod (Fig. 6a). By adding USV sampling, bias in the Southern Ocean improves over the SOCAT baseline around year 2000 (Figs. 6b–d; S10), up to 6–12 years before to the introduction of additional samples in either 2006 or 2012. This improvement is shown for the majority of the 75 ensemble members (Fig. S11). Run Z_x10_ 5Y_W, which has the lowest mean bias out of the zigzag runs (Fig. 5), shows improvement even further back in time, until the beginning of the test bed period (Fig. S10). While the annual mean bias of the zigzag runs varies rather consistently, there is a larger spread across the one-latitude runs (Fig. 6d).

3.2.2 Root-mean-squared error (RMSE)

Similar to bias, improvements in RMSE are most significant during the period of USV additions and within the Southern Ocean (Fig. 7a vs. 7b). For the duration of USV additions, the one-latitude runs show improvements in global mean RMSEs of 0.5 %–3 % (0.1 %–1 % for 1982–2016), while the zigzag runs show higher improvements between 2 %–5 % (1 %–3 % for 1982–2016) (Figs. 7, S12, S13). Mean RMSE is further reduced in the Southern Ocean by up to 16 % and during Southern Hemisphere winter months (JJA) by up to 21 % (run Z_x10_5Y_YR; mean RMSE of 9.6 µatm; Table 1). There is minimal change in RMSE (or bias) during Southern Hemisphere summer months (DJF; Fig. S14). The two zigzag runs sampling year-round (Z_x4_10Y_YR and Z_x10_5Y_YR) have the lowest RMSE values both globally and in the Southern Ocean (Fig. 8). The spread across the 75 test bed members for each experiment is shown in Fig. S15.

The zigzag runs, as well as the high-sampling one-latitude runs (i.e., x13_10Y_J-A and x13_10Y_W), show improvements compared to the SOCAT baseline from the initiation of sampling (Figs. 9, S16, S17). The year-round zigzag runs, however, show improvement in the Southern Ocean from the beginning of the test bed period (Figs. 9c, d, S16). RMSE improvements back in time are greater for all runs in the Southern Hemisphere winter months (Fig. S18).

3.3 Impact on the air–sea CO$_2$ flux with Saildrone USV additions

Air–sea flux was calculated in the same manner for both the ML reconstructions and the model truth, which allows for the isolation of the impact of different sampling strategies, as mediated by the pCO$_2$ reconstruction, on fluxes (see Sect. 2.5). These flux estimates are made to inform understanding of the errors that may exist in CO$_2$ flux estimates derived from pCO$_2$ reconstructions and how new sampling could address these errors. Flux estimates represent the average of the 75 members of the LET in each case and are not estimates of real-world fluxes.

Compared to the model truth, the SOCAT baseline reconstruction underestimates the global and Southern Ocean sink by 0.11–0.13 Pg C yr^{-1} over 1982–2016 (Fig. 10; Table S1 in the Supplement). Regardless of the sampling pattern, adding Saildrone USV observations increases both the global and the Southern Ocean mean sink compared to the SOCAT baseline (Figs. 10, S19). The one-latitude runs show an increase of 0.01–0.03 Pg C yr^{-1} (2 %–6 % strengthening) of the Southern Ocean sink (1982–2016), while the zigzag runs lead to an even stronger sink, increased by 0.04–0.06 Pg C yr^{-1} (7 %–11 % strengthening) (Table S2). When averaging over the years of Saildrone USV sampling addition (i.e., 2006–2012 and 2012–2016), the Southern Ocean sink increases up to 0.09 Pg C yr^{-1} (14 % strengthening) for the one-latitude runs and up to 0.1 Pg C yr^{-1} (15 % strength-
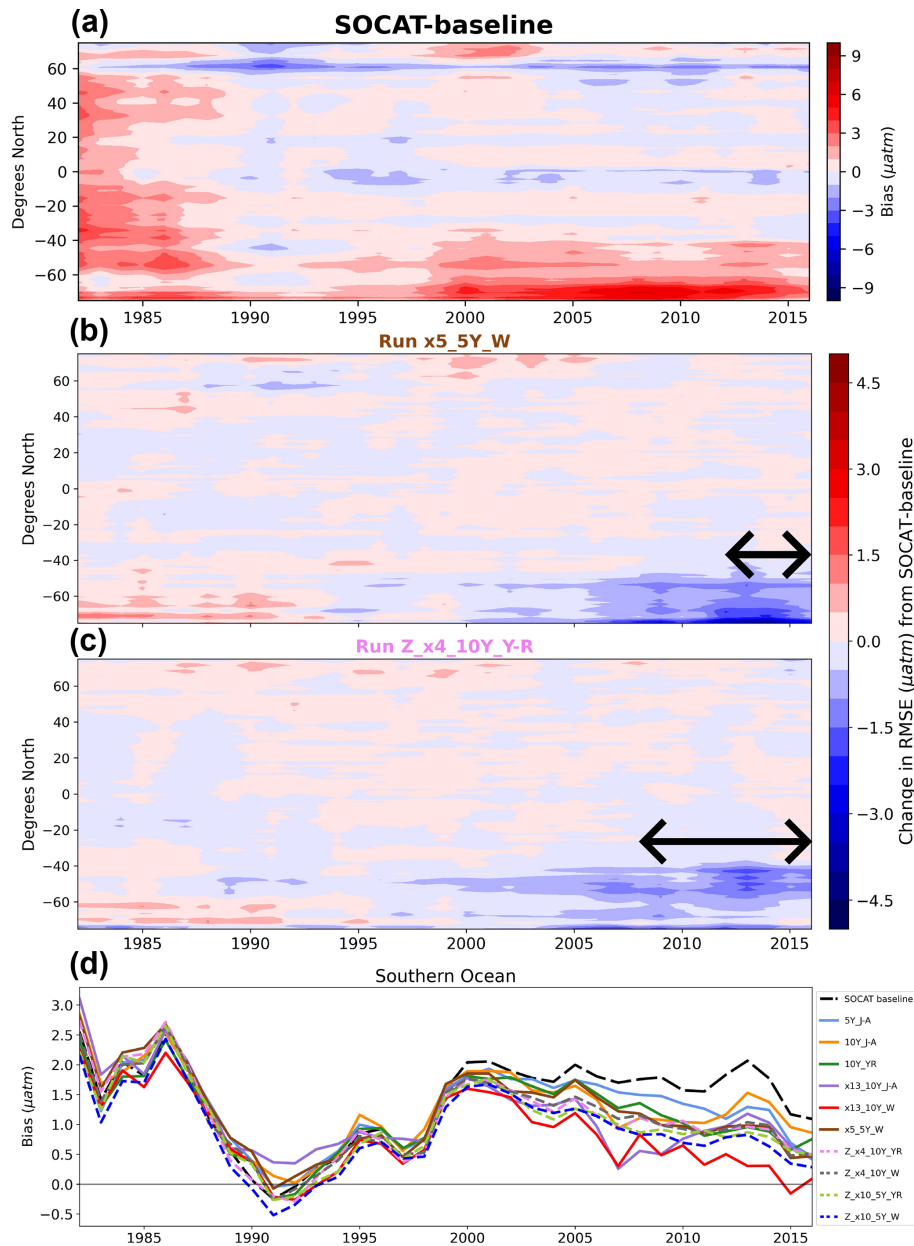
Figure 6. Zonal mean and annual mean Hovmöller of bias for the SOCAT baseline **(a)**. Change in bias for runs x5_5Y_W **(b)** and Z_x4_10Y_YR **(c)** compared to the SOCAT baseline shown in panel **(a)**. Improvement in bias in the Southern Ocean expands back in time well beyond the duration of USV additions for both runs (shown by arrows in each panel). Annual mean bias for the Southern Ocean (> 35° S) for all runs **(d)**.

ening) for the zigzag runs (Table S2). These same features are found for the global ocean (Fig. S19; Table S2).

All of the zigzag runs quite closely match both the global and the Southern Ocean model truth air–sea CO$_2$ flux for the duration of sample additions (Figs. 10, S19). Except for the first couple of years of sample addition for the high-sampling run x13_10Y_J-A, none of the one-latitude runs can match the model truth air–sea CO$_2$ flux; instead, they all underestimate the flux (Figs. 10, S19). The zigzag runs have impact

on the air–sea flux from an earlier date, starting to pull the results away from the SOCAT baseline and toward the model truth already in the late 1990s, while the one-latitude runs do the same about a decade later (Figs. 10, S19).

4   Discussion

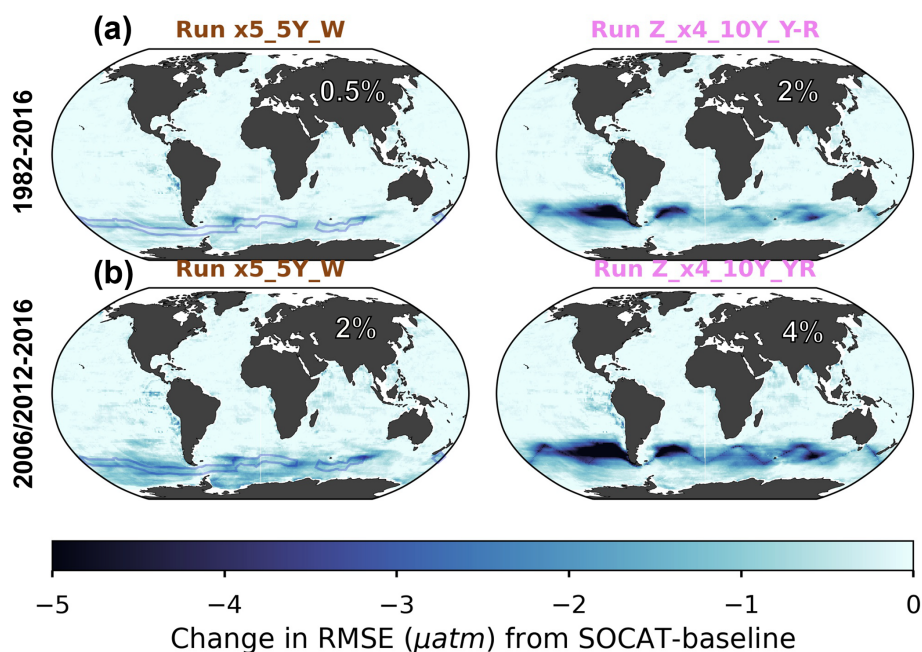We have tested the pCO$_2$-Residual reconstruction method with the large ensemble test bed (LET) to estimate its fi-

Figure 7. Change in RMSE when comparing runs x5_5Y_W and Z_x4_10Y_YR to the SOCAT baseline averaged over the duration of the test bed period (**a**; 1982–2016) and the period of Saildrone USV additions (**b**; 2006–2012 or 2012–2016). The percent global improvement is shown in each panel.

delity and understand how new samples could increase skill. We find that, regardless of the chosen Saildrone USV sampling pattern, the reduction in mean bias and mean RMSE compared to the SOCAT baseline is most prominent within the Southern Ocean, ($< 35°$ S) during the period of which Saildrone USV observations were added (Figs. 4, 6, 7, 9). However, it is important to mention that the additional Southern Ocean sampling also impacts (improves) the $p\mathrm{CO}_2$ reconstructions globally (Figs. 5a, 8a). Based on our experiments, a combination of factors improves global and Southern Ocean $p\mathrm{CO}_2$ reconstructions, including the type of sampling pattern and seasonality of sampling and, to some extent, the number of additional observations. Importantly, increasing the number of observations or duration of sampling (5 vs. 10 years) is not the sole determining factor for improving the reconstructions (Figs. 5, 8). This is best demonstrated by the high-sampling run x13_10Y_J-A (44 250 observations), which does not provide significantly better reconstructions or is even outperformed by runs with 2–18 times fewer observations. The runs that produce lower mean RMSE do include data throughout Southern Hemisphere winter (Fig. 8). Run x13_10Y_J-A does not include more than a few observations in the month of August, as it follows the temporal pattern of the real-world one-latitude Saildrone USV expedition (Figs. S3, S4; Sutton et al., 2021). The one-latitude runs 10Y_J-A and 10Y_YR are directly comparable in terms of sample duration, spatial extent and number of observations (Table 1), but the latter, which covers all months, always shows lower mean RMSE and bias (Figs. 5, 6d, 8,

9d). These examples attest to the importance of addressing the issue of significant undersampling in the Southern Ocean during the winter season (Fig. S5a).

Another important comparison is between the one-latitude run x5_5Y_ W (5022 observations) and zigzag run Z_x10_5Y_W (3800 observations) that both sample during Southern Hemisphere winter months over a 5-year period (Table 1), where the zigzag run consistently performs better even though it includes fewer observations (Figs. 5, 8). Most of the runs that perform similarly to, or outperform, the abovementioned high-sampling run x13_10Y_J-A (44 250 observations) sample in a zigzag pattern. Out of all 10 runs, the year-round zigzag runs (Z_x4_10Y_YR and Z_x10_5Y_YR) are most able to reduce the mean error, as shown by the lowest RMSE values (Figs. 8, 9d). A recent study performed similar sampling experiments, as shown here, by comparing sampling from different types of autonomous platforms to a SOCAT baseline (Djeutchouang et al., 2022). They emphasized the importance of capturing the significant differences in $p\mathrm{CO}_2$ that exist across meridional gradients during summer and winter months (up to 15 µatm; Djeutchouang et al., 2022). The meridional coverage provided by the zigzag runs could explain why these runs generally outperform the one-latitude runs in our study and show significant reduction in both RMSE and bias even though the global $p\mathrm{CO}_2$ data density is raised by as little as 0.01 %– 0.07 %.

The greatest reduction in mean bias out of all runs is shown by run x13_10Y_W (Figs. 5, 6d), which represents
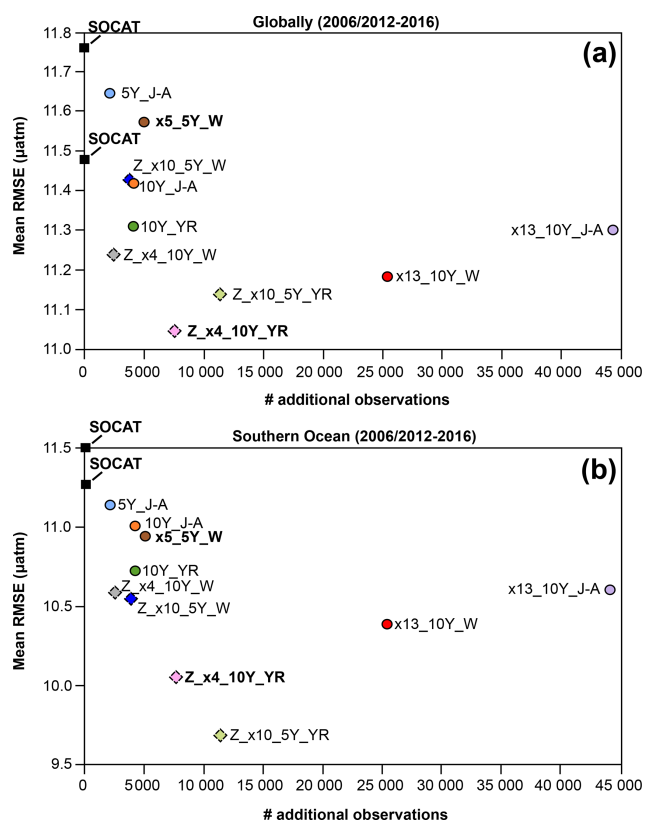
Figure 8. Mean RMSE globally (**a**) and for the Southern Ocean ($< 35°$ S; **b**) for the duration of Saildrone USV sampling (2006–2016 or 2012–2016) for all runs presented in Table 1. Circles represent runs using the one-latitude track, while diamonds represent zigzag runs. Runs highlighted in bold correspond to the two selected runs mapped in Figs. 4, 6, 7 and 9. RMSE values shown for the SOCAT baseline (black squares) represent a mean over 2006–2016 (global mean of 11.5 µatm; Southern Ocean mean of 11.3 µatm) and 2012–2016 (global mean of 11.8 µatm; Southern Ocean mean of 11.5 µatm). The number of monthly $1° \times 1°$ USV observations in addition to SOCAT is denoted by # additional observations. Box plots illustrating the spread across the 75 ensemble members are shown in Fig. S15.

one-latitude high-sampling (i.e., 25 395 observations) during Southern Hemisphere winter months only. This sampling strategy thus seems to have a higher ability to reduce the ML model's tendency to overestimate pCO$_2$ in the Southern Ocean compared to any of the meridional (zigzag) runs. However, it should be noted that run x13_10Y_W covers areas south of $55°$ S (Fig. S4), and its improvement in mean bias (and mean RMSE) is particularly prevalent at these high latitudes (e.g., Figs. S8, S10, S13, S16). Whether or not this run is, in fact, feasible with current or future technology is uncertain, as parts of the southernmost tracks potentially cover the Southern Ocean ice zone (Fig. S20), and solar radiation for solar-powered platforms and sensors becomes very limited during winter south of $55°$ S. Furthermore, this particular sampling strategy requires 13 USVs and

so would be the most costly of the observing scenarios. Although run x13_10Y_W demonstrates the highest reduction in mean bias out of all runs, the zigzag runs still reduce absolute mean bias (for 2006/2012–2016) in the Southern Ocean by 44 %–65 % (vs. 77 % for run x13_10Y_W).

Overall, the zigzag runs include significantly fewer observations, require fewer USVs, collect samples over the same duration (or even half the time, as run x13_10Y_W), cover areas north of $55°$ S and within the ice-free zone, and show major improvement in the reconstruction of pCO$_2$ (attested to by reductions in both bias and RMSE). The zigzag runs also closely match both the global and the Southern Ocean model truth air–sea CO$_2$ flux for the duration of sample additions (Figs. 10, S19). It also appears that the zigzag runs generally have a greater impact on both the pCO$_2$ reconstruction and the air–sea flux further back in time and start to deviate from the SOCAT baseline earlier compared to the one-latitude runs (Figs. 6, 9, 10, S10, S16, S18, S19). Even the zigzag scenarios with the least number of USVs (e.g., Z_x4_10Y_YR) reduce Southern Ocean reconstruction absolute mean (2006–2016) bias and RMSE by up to 46 % and 11 %, respectively, and could provide a basis for realistic future Southern Ocean pCO$_2$ sampling campaigns.

The main motivation for improving surface ocean pCO$_2$ reconstructions is so that we can more accurately estimate the current and future oceanic uptake of anthropogenic carbon. The Southern Ocean is a significant carbon sink, but estimates of the air–sea CO$_2$ flux diverge substantially in this region (Takahashi et al., 2009; Landschützer et al., 2014, 2015; Rödenbeck et al., 2015; Williams et al., 2017; Gray et al., 2018; Gruber et al., 2019; Bushinsky et al., 2019; Long et al., 2021; Fay and McKinley, 2021; Wu et al., 2022). Southern Ocean estimates incorporating observations from biogeochemical floats have shown a significantly weaker sink compared to those based only on observations from ships (Williams et al., 2017; Gray et al., 2018; Bushinsky et al., 2019). Bushinsky et al. (2019) and Hauck et al. (2023) performed similar sampling experiments, as presented here, by comparing ML surface ocean pCO$_2$ reconstructions based on SOCAT compared to additional SOCCOM or ideal virtual floats. These studies showed that SOCAT sampling alone overestimates the CO$_2$ uptake in the Southern Ocean and that additional floats reduce this overestimation, leading to a decreased (weakened) ocean carbon sink. In contrast, we find that the pCO$_2$-Residual method underestimates the CO$_2$ uptake with only SOCAT sampling and that adding USVs increased (strengthened) the Southern Ocean and global ocean sinks by up to 0.1 Pg C yr^{-1} (Figs. 10, S19; Table S2).

Going forward, additional studies are needed to better understand why these results suggest a different direction of the sink change with additional sampling. These differences could stem from the use of different reconstruction methods assessed. Hauck et al. (2023) used the MPI-SOMFFN and CarboScope/Jena-MLS reconstruction methods, while we use the pCO$_2$-Residual method. Another substantial dif-

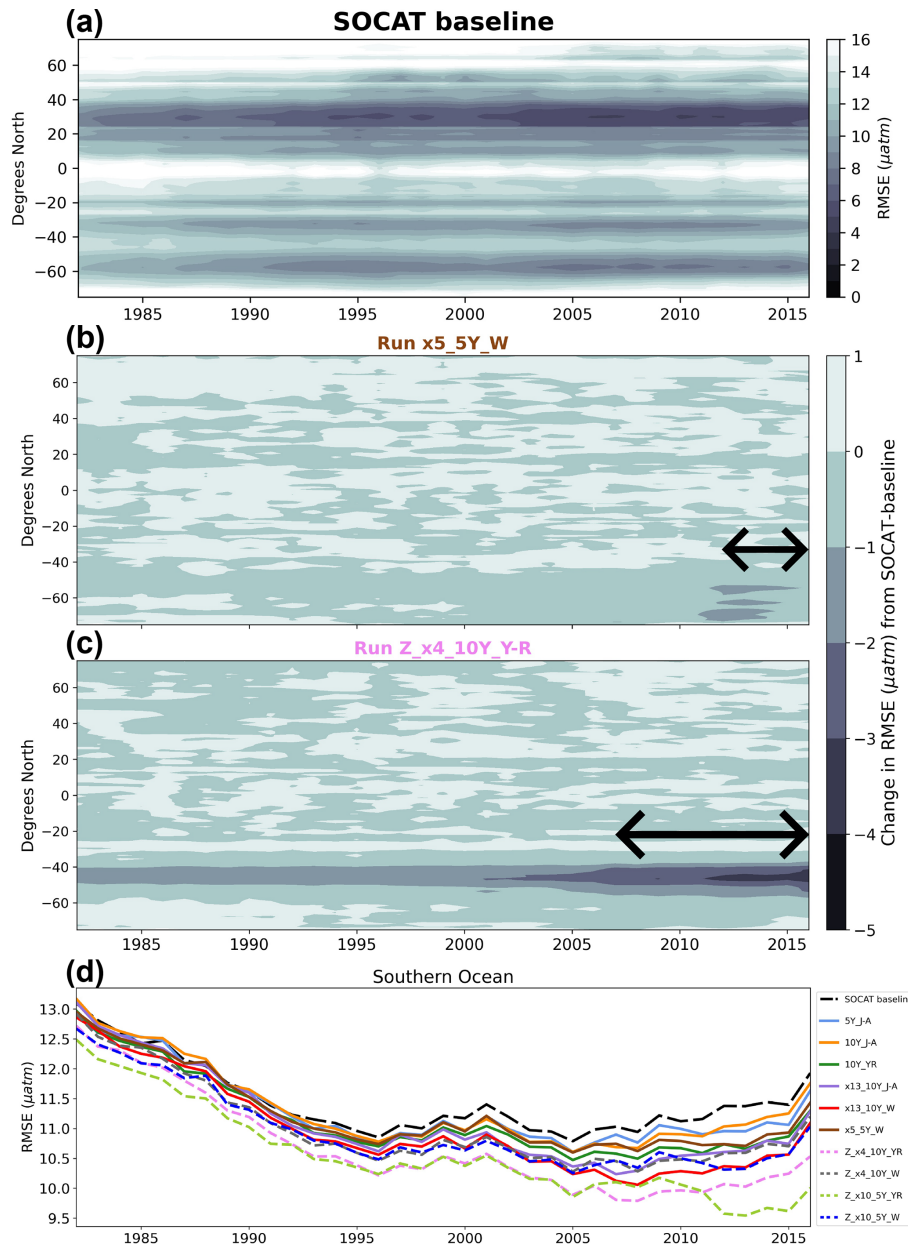**Figure 9.** Zonal mean and annual mean Hovmöller of RMSE for the SOCAT baseline **(a)**. Change in RMSE for run x5_5Y_W **(b)** and Z_x4_10Y_YR **(c)** compared to the SOCAT baseline. Run Z_x4_10Y_YR shows improvements in RMSE within the Southern Ocean, which expand well beyond the duration of Saildrone USV additions (shown by arrow in panel). Annual mean RMSE for the Southern Ocean ($> 35°$ S) for all runs **(d)**.

ference between the studies is the models and numbers of ensemble members used as the test bed. Hauck et al. (2023) use a single hindcast model, while we use 25 members each from three Earth system models. We find substantial spread across these 75 members (Figs. S9, S15), indicating that model structure and internal variability significantly impact results. Our study and Hauck et al. (2023) use different sampling masks and approaches for the calculation of fluxes, which could also be a factor. Targeted, coordinated studies using

multiple reconstruction approaches with consistent test bed structures, sampling masks and experimental approaches are clearly needed (Rödenbeck et al., 2015). Despite this need for this additional work, studies do agree that additional Southern Ocean observations could significantly improve reconstructions of air–sea CO$_2$ fluxes.

What else can we learn using the model test bed? The SOCAT baseline demonstrates a weakening of the global and Southern Ocean carbon sinks starting in the 1990s, with a
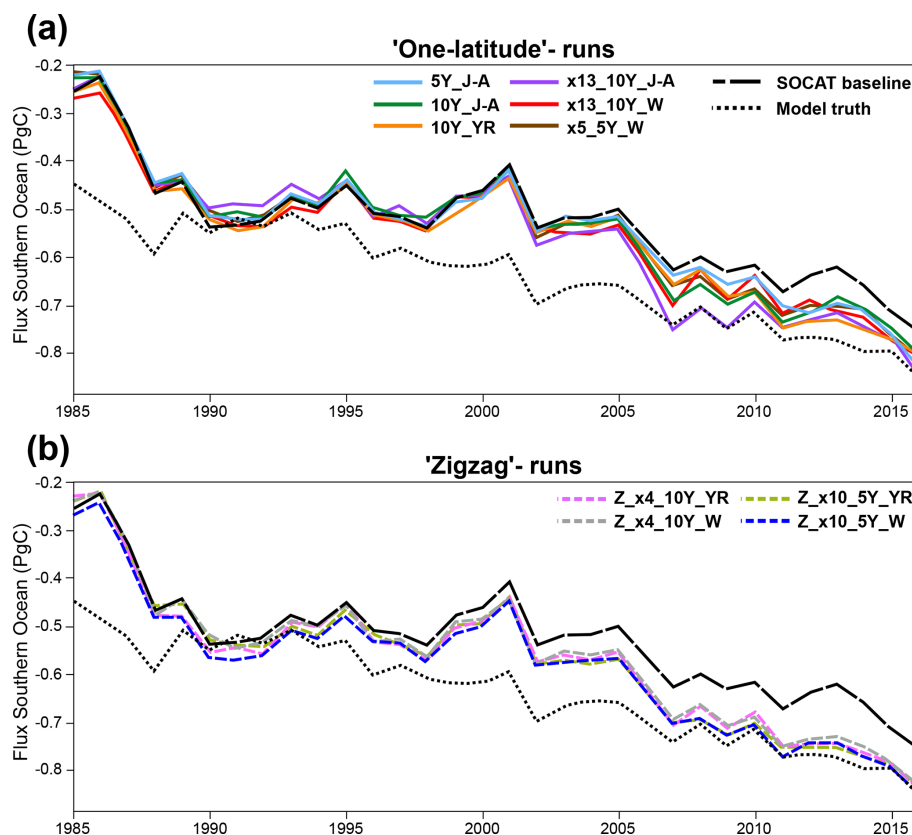
Figure 10. Southern Ocean ($< 35°$ S) annually averaged air–sea CO$_2$ flux for the SOCAT baseline (dashed black line) and model truth (dotted black line) for one-latitude runs (**a**; solid lines) and zigzag runs (**b**; dashed lines).

peak around year 2000 (Figs. 10, S19), which is in broad agreement with various data products using real-world SO-CAT data (e.g., Gruber et al., 2019; Landschützer et al., 2015; Bushinsky et al., 2019; Bennington et al., 2022a; Gloege et al., 2022). Peaks in bias and RMSE coincide in time with the weakening sink (Figs. 6d, 9d). As shown in Fig. 10, this low sink is significantly exaggerated compared to the model truth. To better understand this discrepancy, we performed an additional experiment based on run Z_x10_5Y_YR but assumed sampling every year for the entire test bed period (i.e., 1982–2016). There is now a significant reduction in the temporal variability of reconstruction bias; with the additional 35-year USV sampling, the reconstructed Southern Ocean air–sea CO$_2$ flux closely matches the model truth for the entire test bed duration (Fig. S21). This suggests that the large decadal variability of air–sea CO$_2$ fluxes since the 1980s and the weak anomaly in the Southern Ocean carbon sink in the early 2000s (Le Quéré et al., 2007; Landschützer et al., 2015; Gruber et al., 2019; Bennington et al., 2022a, b; Friedlingstein et al., 2023) may be at least partially attributable to undersampling of the Southern Ocean. This is in agreement with the float sampling experiments performed by Hauck et al. (2023), attributing the strong decadal variability to sparse and skewed SOCAT data distributions. We will fur-

ther explore this issue in future work. Still, this preliminary experiment suggests that interpretations of trends and variability of the global and Southern Ocean carbon sink should be considered with caution.

5 Conclusions

By using the large ensemble test bed (LET), we show that targeted meridional and winter sampling in the Southern Ocean can improve global and Southern Ocean ML surface ocean pCO$_2$ reconstructions. Significant improvements are possible by raising the global pCO$_2$ data density by as little as 0.01 %–0.07 %. Further, we find that this modest amount of additional Saildrone USV sampling increases the global and Southern Ocean air–sea CO$_2$ flux by up to 0.1 Pg C yr^{-1}, a quantity equivalent to 25 % of the uncertainty in the ocean carbon sink (0.4 Pg C yr^{-1}; Friedlingstein et al., 2023). Our findings are consistent with previous studies suggesting that additional observations during Southern Hemisphere winter months and covering meridional gradients can reduce uncertainties and biases in the reconstructions (Lenton et al., 2006; Monteiro et al., 2010; Djeutchouang et al., 2022; Mackay et al., 2022). As opposed to other autonomous platform approaches, Saildrone USVs obtain in situ pCO$_2$ observations,

with uncertainties equivalent to the highest-quality observations collected by research ships ($\pm 2\,\mu$atm; Sabine et al., 2020; Sutton et al., 2021), and can operate at a high speed so that the spatial extent and seasonal cycle of meridional gradients can be covered. The approach of combining high-accuracy Saildrone USV and SOCAT observations thus represents a promising solution to improve future surface ocean pCO$_2$ reconstructions and the accuracy of the ocean carbon sink. Lastly, we show that the large variability in bias and the weakening of the global and Southern Ocean carbon sink in the 2000s may partially be an artifact of Southern Ocean undersampling.

Author contributions. THH, GAM and AJS designed the experiments, and THH performed the simulations. THH, ARF and LG developed the code. THH and ARF calculated the air–sea fluxes. THH prepared the paper with contributions from all co-authors.

References

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 fCO$_2$ data in version 3 of the Surface Ocean CO$_2$ Atlas (SOCAT), Earth Syst. Sci. Data, 8, 383–413, https://doi.org/10.5194/essd-8-383-2016, 2016.

Bakker, D. C. E., Alin, S. R., Becker, M., Bittig, H. C., Castaño-Primo, R., Feely, R, A., Gkritzalis, T., Kadono, K., Kozyr, A., Lauvset, S, K., Metzl, N., Munro, D, R., Nakaoka, S., Nojiri, Y., O'Brien, K, M., Olsen, A., Pfeil, Benjamin, P., Denis, S., Tobias, S., Kevin F., Sutton, A. J., Sweeney, C., Tilbrook, B., Wada, C., Wanninkhof, R., Willstrand W. A., Akl, J., Apelthun, L. B., Bates, N., Beatty, C. M., Burger, E. F., Cai, W., Cosca, C. E., Corredor, J. E., Cronin, M., Cross, J. N., De Carlo, E. H., DeGrandpre, M. D., Emerson, S. R., Enright, M. P., Enyo, K., Evans, W., Frangoulis, C., Fransson, A., García-Ibáñez, M. I., Gehrung, M., Giannoudi, L., Glockzin, M., Hales, B., Howden, S. D., Hunt, C. W., Ibánhez, J. S. P., Jones, S. D., Kamb, L., Körtzinger, A., Landa, C. S., Landschützer, P., Lefèvre, N., Lo Monaco, C., Macovei, V. A., Maenner J. S., Meinig, C., Millero, F. J., Monacci, N. M., Mordy, C., Morell, J. M., Murata, A., Musielewicz, S., Neill, ., Newberger, T., Nomura, D., Ohman, M., Ono, T., Passmore, A., Petersen, W., Petihakis, G., Perivoliotis, L., Plueddemann, A. J., Rehder, G., Reynaud, T., Rodriguez, C., Ross, A. C., Rutgersson, A., Sabine, C. L., Salisbury, J. E., Schlitzer, R., Send, U., Skjelvan, I., Stamataki, N., Sutherland, S. C., Sweeney, C., Tadokoro, K., Tanhua, T., Telszewski, M., Trull, T., Vandemark, D., van Ooijen, E., Voynova, Y. G., Wang, H., Weller, R. A., Whitehead, C., and Wilson, D.: Surface Ocean CO$_2$ Atlas Database Version 2022 (SOCATv2022)

(NCEI Accession 0253659), NOAA National Centers for Environmental Information [data set], https://doi.org/10.25921/1h9f-nb73, 2022.

Behncke, J., Landschützer, P., and Tanhua, T.: A detectable change in the air-sea CO₂ flux estimate from sailboat measurements, Sci. Rep.-UK, 14, 3345, https://doi.org/10.1038/s41598-024-53159-0, 2024.

Bennington, V., Galjanic, T., and McKinley, G. A.: Explicit Physical Knowledge in Machine Learning for Ocean Carbon Flux Reconstruction: The pCO2-Residual Method, J. Adv. Model. Earth Sy., 14, 3345, https://doi.org/10.1029/2021ms002960, 2022a.

Bennington, V., Gloege, L., and McKinley, G. A.: Variability in the global ocean carbon sink from 1959 to 2020 by correcting models with observations, Geophys. Res. Lett., 49, e2022GL098632, https://doi.org/10.1029/2022GL098632, 2022b.

Bushinsky, S. M., Landschützer, P., Rödenbeck, C., Gray, A. R., Baker, D., Mazloff, M. R., Resplandy, L., Johnson, K. S., and Sarmiento, J. L.: Reassessing Southern Ocean air-sea CO2 flux estimates with the addition of biogeochemical float observations, Global Biogeochemical Cycles, 33, 1370–1388, https://doi.org/10.1029/2019GB006176, 2019.

Chen, T. and Guestrin, C.: Xgboost: A scalable tree boosting system, in: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, San Francisco, California, USA 14–17 August 2016, 785–794, https://doi.org/10.1145/2939672.2939785, 2016.

Denvil-Sommer, A., Gehlen, M., and Vrac, M.: Observation system simulation experiments in the Atlantic Ocean for enhanced surface ocean pCO2 reconstructions, Ocean Sci., 17, 1011–1030, https://doi.org/10.5194/os-17-1011-2021, 2021.

Djeutchouang, L. M., Chang, N., Gregor, L., Vichi, M., and Monteiro, P. M. S.: The sensitivity of *p*CO₂ reconstructions to sampling scales across a Southern Ocean sub-domain: a semi-idealized ocean sampling simulation approach, Biogeosciences, 19, 4171–4195, https://doi.org/10.5194/bg-19-4171-2022, 2022.

Fay, A. R. and McKinley, G. A.: Observed regional fluxes to constrain modeled estimates of the ocean carbon sink, Geophys. Res. Lett., 48, e2021GL095325, https://doi.org/10.1029/2021GL095325, 2021.

Fay, A. R., Lovenduski, N. S., McKinley, G. A., Munro, D. R., Sweeney, C., Gray, A. R., Landschützer, P., Stephens, B. B., Takahashi, T., and Williams, N.: Utilizing the Drake Passage Time-series to understand variability and change in subpolar Southern Ocean *p*CO₂, Biogeosciences, 15, 3841–3855, https://doi.org/10.5194/bg-15-3841-2018, 2018.

Fay, A. R., Gregor, L., Landschützer, P., McKinley, G. A., Gruber, N., Gehlen, M., Iida, Y., Laruelle, G. G., Rödenbeck, C., Roobaert, A., and Zeng, J.: SeaFlux: harmonization of air–sea CO₂ fluxes from surface *p*CO₂ data products using a standardized approach, Earth Syst. Sci. Data, 13, 4693–4710, https://doi.org/10.5194/essd-13-4693-2021, 2021.

Friedlingstein, P., O'Sullivan, M., Jones, M. W., Andrew, R. M., Bakker, D. C. E., Hauck, J., Landschützer, P., Le Quéré, C., Luijkx, I. T., Peters, G. P., Peters, W., Pongratz, J., Schwingshackl, C., Sitch, S., Canadell, J. G., Ciais, P., Jackson, R. B., Alin, S. R., Anthoni, P., Barbero, L., Bates, N. R., Becker, M., Bellouin, N., Decharme, B., Bopp, L., Brasika, I. B. M., Cadule, P., Chamberlain, M. A., Chandra, N., Chau, T.-T.-T., Chevallier, F., Chini, L. P., Cronin, M., Dou, X., Enyo, K., Evans, W., Falk, S., Feely, R. A., Feng, L., Ford, D. J., Gasser, T., Ghattas, J., Gkritzalis, T., Grassi, G., Gregor, L., Gruber, N., Gürses, Ö., Harris, I., Hefner, M., Heinke, J., Houghton, R. A., Hurtt, G. C., Iida, Y., Ilyina, T., Jacobson, A. R., Jain, A., Jarníková, T., Jersild, A., Jiang, F., Jin, Z., Joos, F., Kato, E., Keeling, R. F., Kennedy, D., Klein Goldewijk, K., Knauer, J., Korsbakken, J. I., Körtzinger, A., Lan, X., Lefèvre, N., Li, H., Liu, J., Liu, Z., Ma, L., Marland, G., Mayot, N., McGuire, P. C., McKinley, G. A., Meyer, G., Morgan, E. J., Munro, D. R., Nakaoka, S.-I., Niwa, Y., O'Brien, K. M., Olsen, A., Omar, A. M., Ono, T., Paulsen, M., Pierrot, D., Pocock, K., Poulter, B., Powis, C. M., Rehder, G., Resplandy, L., Robertson, E., Rödenbeck, C., Rosan, T. M., Schwinger, J., Séférian, R., Smallman, T. L., Smith, S. M., Sospedra-Alfonso, R., Sun, Q., Sutton, A. J., Sweeney, C., Takao, S., Tans, P. P., Tian, H., Tilbrook, B., Tsujino, H., Tubiello, F., van der Werf, G. R., van Ooijen, E., Wanninkhof, R., Watanabe, M., Wimart-Rousseau, C., Yang, D., Yang, X., Yuan, W., Yue, X., Zaehle, S., Zeng, J., and Zheng, B.: Global Carbon Budget 2023, Earth Syst. Sci. Data, 15, 5301–5369, https://doi.org/10.5194/essd-15-5301-2023, 2023.

Fyfe, J. C., Derksen, C., Mudryk, L., Flato, G. M., Santer, B. D., Swart, N. C., Molotch, N. P., Zhang, X., Wan, H., Arora, V. K., Scinocca, J., and Jiao, Y.: Large near-term projected snowpack loss over the western United States, Nat. Commun., 8, 14996, https://doi.org/10.1038/ncomms14996, 2017.

Gloege, L.: Large ensemble pCO2 testbed, Figshare [data set], https://doi.org/10.6084/m9.figshare.c.4568555.v2, 2019.

Gloege, L., McKinley, G. A., Landschützer, P., Fay, A. R., Frolicher, T. L., and Fyfe, J. C.: Quantifying Errors in Observationally Based Estimates of Ocean Carbon Sink Variability, Global Biogeochem. Cy., 35, e2020GB006788, https://doi.org/10.1029/2020gb006788, 2021.

Gloege, L., Yan, M., Zheng, T. and McKinley, G. A.: Improved quantification of ocean carbon uptake by using machine learning to merge global models and *p*CO₂ data, Journal of Advances in Modeling Earth Systems, 14, e2021MS002620, https://doi.org/10.1029/2021MS002620, 2022.

Good, S. A., Martin, M., and Rayner, N. A.: EN4: Quality controlled ocean temperature and salinity profiles and monthly objective analyses with uncertainty estimates, J. Geophys. Res.-Oceans, 118, 6704–6717, 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., 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. and Fay, A. R.: Air-sea CO₂ fluxes for surface pCO2 data products using a standardized approach, Zenodo [code], https://doi.org/10.5281/zenodo.5482547, 2021.

Gregor, L., Lebehot, A. D., Kok, S., and Scheel Monteiro, P. M.: A comparative assessment of the uncertainties of global surface ocean CO₂ estimates using a machine-learning ensemble (CSIR-ML6 version 2019a) – have we hit the wall?, Geosci. Model Dev., 12, 5113–5136, https://doi.org/10.5194/gmd-12-5113-2019, 2019.

Gruber, N., Landschützer, P., and Lovenduski, N. S.: The variable Southern Ocean carbon sink, Annu. Rev. Mar. Sci., 11, 159–186, https://doi.org/10.1146/annurev-marine-121916-063407, 2019.

Hauck, J., Nissen, C., Landschützer, P., Rödenbeck, C., Bushinsky, S., and Olsen, A.: Sparse observations induce large biases in estimates of the global ocean CO$_2$ sink: and ocean model sub-sampling experiment, Philos. T. Roy. Soc. A, 381, 20220063, https://doi.org/10.1098/rsta.2022.0063, 2023.

Heimdal, T. H., McKinley, G., Sutton, A., Fay, A., Gloege, L., and Bennington, V.: Code for ML reconstruction of surface ocean pCO2 using the Large Ensemble Testbed (v1.0), Zenodo [code], https://doi.org/10.5281/zenodo.10966977, 2024a.

Heimdal, T. H., McKinley, G., Sutton, A., Fay, A., and Gloege, L.: SOCAT+USV sampling masks for ML reconstruction of surface ocean pCO2 using the Large Ensemble Testbed, Zenodo [data set], https://doi.org/10.5281/zenodo.10811018, 2024b.

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., Nicolas, J., Peubey, C., Radu, R., Schepers, D., Simmons, A., Soci, C., Abdalla, S., Abellan, X., Balsamo, G., Bechtold, P., Biavati, G., Bidlot, J., Bonavita, M., De Chiara, G., Dahlgren, P., Dee, D., Diamantakis, M., Dragani, R., Flemming, J., Forbes, R., Fuentes, M., Geer, A., Haimberger, L., Healy, S., Hogan, R.J., Hólm, E., Janisková, M., Keeley, S., Laloyaux, P., Lopez, P., Lupu, C., Radnoti, G., de Rosnay, P., Rozum, I., Vamborg, F., Villaume, S., and Thépaut, J.: The ERA5 global reanalysis, Q. J. Roy. Meteor. Soc., 146, 1999–2049, https://doi.org/10.1002/qj.3803, 2020.

Kay, J. E., Deser, C., Phillips, A., Mai, A., Hannay, C., Strand, G., Arblaster, J. M., Bates, S. C., Danabasoglu, G., Edwards, J., Holland, M., Kuschner, P., Lamarque, J-F., Lawrence, D., Lindsay, K., Middelton, A., Munoz, E., Nealse, R., Oleson, K., Polvani, L., and Vertenstein, M.: The Community Earth System Model (CESM) large ensemble project: A community resource for studying climate change in the presence of internal climate variability, B. Am. Meteor. Soc., 96, 1333–1349, https://doi.org/10.1175/BAMS-D-13-00255.1, 2015.

Khatiwala, S., Primeau, F., and Hall, T.: Reconstruction of the history of anthropogenic CO$_2$ concentrations in the ocean, Nature, 462, 346–349, https://doi.org/10.1038/nature08526, 2009.

Landschützer, P., Gruber, N., Bakker, D. C. E., and Schuster, U.: Recent variability of the global ocean carbon sink, Global Biogeochem. Cy., 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., Hoppema, M., Metzl, N., Sweeney, C., Takahashi, T., Brook, B., and Wanninkhof, R.: The reinvigoration of the Southern Ocean carbon sink, Science, 349, 1221–1224, https://doi.org/10.1126/science.aab2620, 2015.

Landschützer, P., Tanhua, T., Behncke, J., and Keppler, L.: Sailing through the Southern Ocean seas of air-sea CO$_2$ flux uncertainty, Philos. T. Roy. Soc. A, 381, 20220064, https://doi.org/10.1098/rsta.2022.0064, 2023.

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.

Lenton, A. B., Matear, R. J., and Tilbrook, B.: Design of an observational strategy for quantifying the Southern Ocean uptake of CO$_2$, Global Biogeochem. Cy., 20, 1-11, https://doi.org/10.1029/2005GB002620, 2006.

Le Quéré, C., Rödenbeck, C., Buitenhuis, E. T., Conway, T. J., Lagenfelds, R., Gomez, A., Labuschagne C., Ramonet, M., Nakazawa, T., Metzl, N., Gillett, N., and Heimann, M.: Saturation of the Southern Ocean CO$_2$ sink due to recent climate change, Science, 316, 1735–1738, https://doi.org/10.1126/science.1136188, 2007.

Long, M. C., Stephens, B. B., McKain, K., Sweeney, C., Keeling, R. F., Kort, E. A., Morgan, E. J., Bent, J. D., Chandra, N., Chevallier, F., Commane, R., Daube, B. C., Krummel, P. B., Loh, Z., Luijkx, I. T., Munro, D., Patra, P., Peters, W., Ramonet, M., Rödenbeck, C., Stavert, A., Tans, P., and Wofsy, S. C.: Strong Southern Ocean carbon uptake evident in airborne observations, Science, 374, 1275–1280, https://doi.org/10.1126/science.abi4355, 2021.

Mackay, N. and Watson, A.: Winter air-sea CO$_2$ fluxes constructed from summer observations of the polar Southern Ocean suggest weak outgassing, J. Geophy. Res.-Oceans, 126, e2020JC016600, https://doi.org/10.1029/2020JC016600, 2021.

Mackay, N., Watson, A., Suntharalingam, P., Chen, Z., and Rödenbeck, C.: Improved winter data coverage of the Southern Ocean CO$_2$ sink from extrapolation of summertime observations, Commun. Earth Environ., 3, 265, https://doi.org/10.1038/s43247-022-00592-6, 2022.

McKinley, G. A., Fay, A. R., Eddebbar, Y. A., Gloege, L., and Lovenduski, N. S.: External forcing explains recent decadal variability of the ocean carbon sink, AGU Advances, 1, e2019AV000149, https://doi.org/10.1029/2019AV000149, 2020.

Mongwe, N. P., Vichi, M., and Monteiro, P. M. S.: The seasonal cycle of pCO$_2$ and CO$_2$ fluxes in the Southern Ocean: diagnosing anomalies in CMIP5 Earth system models, Biogeosciences, 15, 2851–2872, https://doi.org/10.5194/bg-15-2851-2018, 2018.

Monteiro, P. M. S., Schuster, U., Hood, M., Lenton, A., Metzl, N., Olsen, A., Rogers, K., Sabine, C., Takahashi, T., Tilbrook, B., Yoder, J., Wanninkhof, R., and Watson, A. J.: 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, 2, 702–714, https://doi.org/10.5270/OCEANOBS09.CWP.64, 2010.

Reynolds, R. W., Rayner, N. A., Smith, T. M., Stokes, D. C., and Wang, W.: An improved in situ and satellite SST analysis for climate, J. Climate, 15, 1609–1625, https://doi.org/10.1175/1520-0442(2002)015<1609:aiisas>2.0.co;2, 2002.

Rodgers, K. B., Lin, J., and Frölicher, T. L.: Emergence of multiple ocean ecosystem drivers in a large ensemble suite with an Earth system model, Biogeosciences, 12, 3301–3320, https://doi.org/10.5194/bg-12-3301-2015, 2015.

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 pCO$_2$ Mapping intercomparison (SOCOM), Biogeosciences, 12, 7251–7278, https://doi.org/10.5194/bg-12-7251-2015, 2015.

Sabine, C., Sutton, A., McCabe, K., Lawrence-Slavas, N., Alin, S, Feely, R., Jenkins, R., Maenner, S., Meinig, C., Thomas, J., van Ooijen, E., Passmore, A., and Tilbrook, B.: Evaluation of a new carbon dioxide system for autonomous surface vehicles, J. Atmos. Oceaen. Tech., 37, 1305–1317, https://doi.org/10.1175/JTECH-D-20-0010.1, 2020.

Stamell, J., Rustagi, R. R., Gloege, L., and McKinley, G. A.: Strengths and weaknesses of three Machine Learning methods for pCO$_2$ interpolation, Geosci. Model Dev. Discuss. [preprint], https://doi.org/10.5194/gmd-2020-311, 2020.

Sutton, A. J., Williams, N. L., and Tilbrook, B.: Constraining Southern Ocean CO$_2$ flux uncertainty using uncrewed surface vehicle observations, Geophys. Res. Lett., 48, e2020GL091748, https://doi.org/10.1029/2020GL091748, 2021.

Takahashi, 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., Sutherland, S. C., Wanninkhof, R., Sweeney, C., Feely, R. A., Chipman, D. W., Hales, B., Friederich, G., Chavez, F., Sabine, C., Watson, A., 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., Wong, C. S., Delille, B., Bates, N. R., and de Baar, H. J. W.: Climatological mean and decadal change in surface ocean pCO$_2$, and net sea-air CO$_2$ flux over the global oceans, Deep-Sea Res. Pt II, 56, 554–557, https://doi.org/10.1016/j.dsr2.2008.12.009, 2009.

Toms, B. A., Barnes, E. A., and Ebert-Uphoff, I.: Physically interpretable neural networks for the geosciences: Applications to earth system variability, J. Adv. Model. Earth Sy., 12, e2019MS002002, https://doi.org/10.1029/2019MS002002, 2020.

Wanninkhof, R.: Relationship between wind speed and gas exchange over the ocean revisited, Limnol. Oceanogr.-Methods, 12, 351–362, https://doi.org/10.4319/lom.2014.12.351, 2014.

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., Russell, J. L., Riser, S. C., and Takeshita, Y.: Calculating surface ocean pCO$_2$ from biogeochemical Argo floats equipped with pH: An uncertainty analysis, Global Biogeochem. Cy., 31, 591–604, https://doi.org/10.1002/2016GB005541, 2017.

Wu, Y., Bakker, D. C. E., Achterberg, E. P., Silva, A. N., Pickup D. P., Li, X., Hartman, S., Stappard, D., Qi, D., and Tyrrell, T.: Integrated analysis of carbon dioxide and oxygen concentrations as a quality control of ocean float data, Commun. Earth Environ., 3, 92, https://doi.org/10.1038/s43247-022-00421-w, 2022.