scientific reports

Check for updates

OPEN

The importance of adding unbiased Argo observations to the ocean carbon observing system

Thea H. Heimdal ✉ & Galen A. McKinley

The current coverage of direct, high-quality ship-based observations of surface ocean pCO_2 includes large gaps in time and space, and has been declining since 2017. These ocean observations provide the basis for the data products that reconstruct surface ocean pCO_2 and estimate ocean carbon uptake. Improved data coverage is needed to advance our understanding of the ocean carbon sink and air–sea CO_2 exchange. Targeted sampling from autonomous platforms, such as biogeochemical floats, combined with traditional shipboard measurements represents a promising path forward to improve surface ocean pCO_2 reconstructions. However, floats provide indirect pCO_2 estimates derived from pH, and thus have higher uncertainty and are biased compared to direct shipboard measurements. Here, we use a Large Ensemble Testbed (LET) of Earth System Models and the pCO_2-Residual method to reconstruct surface ocean pCO_2 globally to test the impact of additional float observations, both with and without measurement uncertainties. Through comparison to the 'model truth', the LET allows for robust evaluation of the reconstructions. With only shipboard sampling, surface ocean pCO_2 is overestimated, and the 2000–2016 global ocean carbon sink is underestimated by 0.1 Pg C year^{-1}. Additional float observations significantly reduce this underestimation, and deviate from the 'model truth' by as little as 0.01 Pg C year^{-1}, even when floats have random uncertainties of ± 11 µatm. However, systematic bias in the float observations significantly degrades the accuracy of pCO_2 reconstructions, leading to an even stronger underestimation of the global ocean carbon sink of up to 0.32 Pg C year^{-1}. We conclude that adding float-based observations to the global observing system can significantly improve reconstructions of global surface ocean pCO_2, but only if these data are unbiased.

The Surface Ocean CO_2 ATlas database (SOCAT[1]) provides the basis of observation-based data products that are used to reconstruct surface ocean pCO_2 globally in space and time. These products are used to constrain air–sea CO_2 fluxes, some of which contribute to the Global Carbon Budget (GCB)[2]. From 1850 to 2023, the oceans have removed a total of 180 ± 35 Pg of carbon[2]. Air–sea flux estimates from the data products show a large spread, and deviate from those of global ocean biogeochemistry models (GOBMs), leading to a large uncertainty of the global ocean carbon sink (0.4 Pg C year^{-1}; Ref.[2]). In order to fully understand the climate impacts from rising emissions, it is essential to reduce uncertainties and accurately quantify the ocean carbon sink in space and time.

SOCAT is the largest global database of high-quality surface ocean CO_2 observations, which have traditionally been gathered by ships since the 1950s[1]. The main synthesis and gridded products (flags A–D) contain direct measurements of fCO_2 (fugacity of CO_2) with an uncertainty of < 5 µatm[3]. However, the SOCAT database is highly spatially biased towards the northern hemisphere, and covers only about 2% of the global ocean (at monthly 1° × 1° spatial resolution over the period of 1982–2022), and the number of observations collected has slowly decreased since 2017 (Ref.[3]). Reasons for the scarce and declining SOCAT coverage include limited resources for ocean observing, limited number of ships/routes and inaccessible/unsafe ocean regions. Therefore, estimates of the ocean sink and air–sea CO_2 flux in space and time are uncertain, especially on interannual to decadal timescales[4,5]. Improved data coverage, especially from undersampled regions, such as the Southern Ocean, is needed to reduce these uncertainties[6,7].

Only direct pCO_2 measurements are currently included in the SOCAT database, and these are generally collected from ships. There are also some contributions from autonomous platforms, such as moorings and Uncrewed Surface Vehicles (USVs[7]). These platforms can obtain high-quality direct pCO_2 observations with uncertainties equivalent to the highest-quality shipboard measurements contained in SOCAT (flag A and B; ± 2

Columbia University and Lamont-Doherty Earth Observatory, Palisades, NY, USA. ✉email: theimdal@ldeo.columbia.edu

µatm[3,7,8]). Indirect pCO_2 estimates obtained from biogeochemical floats are however not included in SOCAT. The reason for this is that indirect pCO_2 estimates from floats have potentially high uncertainties (± 11.4 µatm) and may be positively biased by as much as ~ 4 µatm[9–14]. The large uncertainties arise as pCO_2 is not measured directly, but is rather estimated using measurements of pH combined with a regression-derived alkalinity estimate[9]. The global mean air–sea disequilibrium is only in the order of 5–8 µatm[4], so the biases and uncertainties of the magnitudes associated with the float estimates could potentially have significant impacts on reconstructed surface ocean pCO_2 and air–sea CO_2 flux estimates.

Biogeochemical floats of the Argo array have collected ocean data since 2000, and projects such as the Southern Ocean Carbon and Climate Observations and Modeling (SOCCOM) and the Global Ocean Biogeochemistry Array (GO-BGC) have been implemented more recently and will continue into the future. Combining these autonomous observations with those from SOCAT should significantly increase the global coverage of surface ocean pCO_2, especially in regions inaccessible by ships, such as the Southern Ocean. The Southern Ocean is a critical region for carbon removal from the atmosphere, being responsible for ~ 40% of the global ocean uptake of anthropogenic CO_2 (Ref.[15]). However, its remoteness and harsh conditions, especially during winter months, have led to large data gaps. Floats can however sample in these conditions, and these additional observations have the potential to substantially improve global and regional pCO_2 reconstructions[5–7,11,12]. However, before float-derived pCO_2 can be confidently used together with direct pCO_2 from SOCAT in reconstructions, impacts of uncertainty and bias must be quantified and appropriately addressed.

Here, we use a Large Ensemble Testbed (LET)[5] of Earth System Models and the pCO_2-Residual reconstruction method[16] to assess how bias and uncertainty in float observations impact global reconstructions of surface ocean pCO_2 and the air–sea CO_2 flux. Instead of using real-world observations, we sample the target variable (i.e., surface ocean pCO_2) and driver variables (i.e., atmospheric CO_2 mole fraction (xCO_2), SST, SSS, MLD and Chl-a) from the LET, based on SOCAT coverage, and historical or potential Argo float coverage. By using the LET, surface ocean pCO_2 is known at all times and model 1° × 1° points. Therefore, the reconstructed pCO_2 can be robustly evaluated in space and time against the 'model truth'. We present two experiments. First, to account for observational bias, 4 µatm is systematically added to each pCO_2 value sampled from the LET that represent float sampling. In a second experiment, a random value between − 11 µatm and + 11 µatm is added to each float pCO_2 value from the LET to account for measurement uncertainty. Two different float sampling schemes are compared ('historical' and potential 'optimized' sampling).

By using a model testbed, it is not our intent to predict real-world surface ocean pCO_2 and air–sea CO_2 fluxes. Instead, our goal is to assess the accuracy with which a machine learning algorithm reconstructs the 'model truth' given inputs consistent with SOCAT and float data coverage. By comparing the different experimental runs, the goal is to assess how float measurement bias and uncertainty may impact the global surface ocean pCO_2 reconstruction and estimated air–sea flux.

Methods

Surface ocean variables (SST, SSS, xCO_2, MLD, Chl-a, pCO_2) were sampled from the Large Ensemble Testbed (LET[5]) based on SOCAT and two different Argo sampling schemes (historical vs. potential optimized float coverage; see Sect. Overview of sampling scenarios and experimental runs). The pCO_2-Residual method[16] was used to reconstruct surface ocean pCO_2 in space and time. A brief description is provided below, but for further details see Ref.[6].

The pCO_2-residual approach using the Large Ensemble Testbed (LET)

The LET includes 25 randomly selected members from three independent initial-condition ensemble of Earth System Models (ESMs). These models are CESM-LENS[17], GFDL-ESM2M[18] and CanESM2[19]. This 75-member testbed includes model output from 1982–2016 (Ref.[5]). For each ensemble member, surface ocean pCO_2 and co-located driver variables (i.e., SST, SSS, Chl-a, MLD, xCO_2) were sampled monthly at a 1° × 1° resolution, at times and locations equivalent to SOCAT observations and additional floats (see Sect. Overview of sampling scenarios and experimental runs).

Prior to algorithm processing, the direct effect of temperature on pCO_2 was removed[16]. This temperature-driven component (pCO_2-T) was calculated using the equation of Refs.[20,21]:

$$pCO_2 - T = pCO_2^{mean} * \exp\left[0.0423 * \left(SST - SST^{mean}\right)\right]$$

where pCO_2^{mean} and SST^{mean} is the long-term mean of surface ocean pCO_2 and temperature, respectively, using all 1° × 1° grid cells from the testbed (i.e., not only where SOCAT coverage exists). pCO_2-Residual is the difference between pCO_2 and the calculated pCO_2-T.

The eXtreme Gradient Boosting method (XGB[22]) was then used to develop an algorithm that allows the driver variables (SST, SSS, Chl-a, MLD, xCO_2) to predict the target variable (pCO_2-Residual). The XGB algorithm for this study used a learning rate of 0.3, 4,000 decision trees with a maximum depth of 6 levels, and this was fixed for all experiments[6]. For the final reconstruction of surface ocean pCO_2 across all space and time points, the previously calculated pCO_2-T values were added back to the reconstructed pCO_2-Residual values.

The full XGB process was repeated individually for each of the 75 LET members, providing a total of 75 reconstruction vs. 'model truth' pairs, which was statistically compared. Bias was calculated as 'mean prediction – mean truth', and the root-mean-squared error (RMSE) as:

$$\sqrt{[(prediction - truth)^2]^{mean}}$$

where, unless otherwise specified, the 'mean' represents all 1° × 1° grid cells globally and all months over the period of 2000–2016. Statistical comparisons between the test set and the reconstructions are equivalent to what would be derived using real-world data. Since we are using a testbed, we calculate error statistics by comparing the pCO_2 reconstruction to the 'full' LET model pCO_2 field, and not only the test set (i.e., all 1° × 1° grid cells, but excluding those used for training).

Air–sea CO_2 flux

Air–sea CO_2 exchange was calculated as in Ref.[6], using the bulk formulation with Python package Seaflux.1.3.1 (https://github.com/lukegre/SeaFlux; Refs.[23,24]). The air–sea flux was calculated in the same manner for both the ML reconstructions and the 'model truth', to allow for flux comparisons that reveal the influence of bias and uncertainty on the pCO_2 reconstruction. Since we are using a model testbed, the flux estimates presented here are only to quantify how bias and uncertainty in float measurements propagate through the pCO_2 reconstruction to impact fluxes; however, they do not represent real-world fluxes. Here, the sign convention used is positive fluxes to the atmosphere and negative fluxes to the ocean.

Overview of sampling scenarios and experimental runs

Sampling scenarios

We sampled target and driver variables from the LET based on (1) SOCAT sampling distributions, (2) SOCAT + 500 'Optimized' potential floats[25], and (3) SOCAT + 500 randomly selected 'Historical' Argo floats. The number of 500 floats was selected as it represents a realistic number for a sampling array; the active and currently funded GO-BGC sampling project aims to deploy 500 floats. The 'Historical' float scenario includes random sampling distributions of floats deployed in the years between 2004 and 2020 (https://fleetmonitoring.euro-argo.eu/dashboard) (Fig. 1a). The available LET output ends in year 2016 (Ref.[5]). To match the 17 years of 'Historical' Argo coverage (2004–2020), float observations were sampled from the LET starting in year 2000 until 2016, i.e., the final year of the testbed. The 'Optimized' float scenario includes potential float locations following Ref.[25], with each float sampling every month in the selected location (Fig. 1b). The 'Optimized' float observations were sampled from the LET covering the years 2000 through 2016 to match the 'Historical' scenario. The 'Historical' and 'Optimized' float coverage includes a total of 21,659 and 102,000 monthly 1°x1° observations, respectively (Fig. 1c). These float scenarios represent an increase in global surface ocean pCO_2 coverage by 0.1% and 0.6%, respectively, compared to using SOCAT alone that has about 1.5% coverage (considering all 1° × 1° grid points in the LET for 1982–2016).

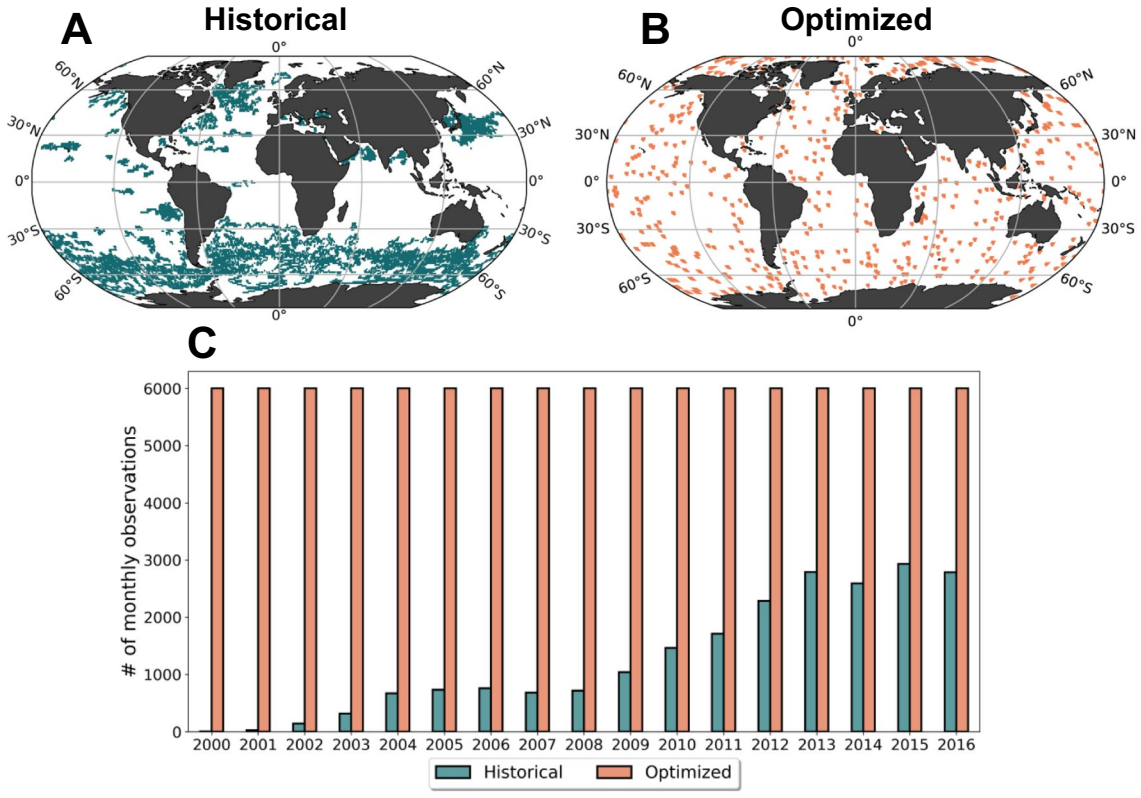

Fig. 1. Map showing the spatial extent of the 'Historical' (**A**) and 'Optimized' (**B**) floats, and the number of 1° × 1° monthly observations additional to SOCAT (**C**) for each float sampling scheme.

Experimental runs

To account for potential bias, 4 µatm was added to each pCO_2 value (float locations only, not SOCAT) sampled from the testbed ('biased' experiment). The value of 4 µatm is based on previous studies comparing offsets between float-based pCO_2 estimates and direct ship-based measurements[9,12]. In a second experiment, to account for measurement uncertainty, a random value between − 11 µatm and + 11 µatm was added to each pCO_2 value ('error' experiment; float locations only, not SOCAT). A unique random value was generated for each individual pCO_2 value sampled from the testbed using the NumPy package (NumPy.random.uniform). The value of ± 11 µatm was selected based on results from an uncertainty analysis of biogeochemical Argo float measurements, incorporating various uncertainty contributions, such as the pH sensor, alkalinity estimate and carbonate system equilibrium constants[9]. In addition, we present 'baseline' runs that include floats without any bias or random error. The 'SOCAT' scenario includes only SOCAT sampling locations and none from floats. In sum, there are seven experiments: 'SOCAT', 'SOCAT + FLOAT_hist', 'SOCAT + FLOAT_opt', 'SOCAT + FLOAT_hist_biased' 'SOCAT + FLOAT_opt_biased', 'SOCAT + FLOAT_hist_error' and 'SOCAT + FLOAT_opt_error'.

Results

We present results as mean 2000–2016 bias or RMSE for the 75-members of the LET with the interquartile range (IQR; Q3-Q1) in parentheses.

Performance metrics

Root-mean-squared error (RMSE)

The three different 'Historical' and 'Optimized' float experiments show similar global mean RMSE within its respective group (Fig. 2a). Both sampling schemes have consistently lower RMSEs compared to the 'SOCAT' run through the whole duration of the testbed period (1982–2016), even though float observations do not begin until 2000 (Fig. 2b). This demonstrates that, even though the data have substantial uncertainty, their addition provides a valuable constraint that improves the ability of the ML model to generalize, also prior to sample addition.

The 'baseline' (green) for each of the sampling schemes demonstrate slightly lower global mean RMSEs compared to the 'biased' (blue) and 'error' (pink) runs (Fig. 2a,b). The 'Optimized' float experiments consistently demonstrate lower RMSE compared to the 'Historical' ones (Fig. 2a,b). The global mean RMSE for the period of float addition (i.e., 2000–2016) for the 'SOCAT' run is 11.6 µatm (IQR = 2.1 µatm), which decreases to 10.5–10.7 µatm (2.2–2.3 µatm) when adding the 'Historical' floats, and to 9.6–9.8 (2.4–2.5 µatm) for the 'Optimized' floats (Fig. 2a; Table 1). While the 'Optimized' float experiments show improvement in RMSE on a global scale, the 'Historical' experiments show improvement mainly in the Southern Ocean (Fig. S1). This is not surprising considering the greater concentration of floats in the Southern Ocean for the 'Historical' scenario (Fig. 1).

There is significant spread in RMSE across the 75 testbed ensemble members for all experiments, which occurs because the CanESM2 experiments lead to consistently higher RMSE than in the experiments with CESM and GFDL (Fig. 2a). When comparing the experiments across ensemble members of each individual Earth System Model in the LET, the spread is reduced significantly (Fig. 2a). The IQR decreases from > 2 µatm (full testbed) to 0.1–0.4 µatm for individual models (Table S1).

Bias

The 'SOCAT' run and all 'Historical' float experiments show positive mean bias (i.e., overestimation of pCO_2 compared to the 'model truth') in the period of float addition (2000–2016), but there is significant discrepancy between the float experiments (Fig. 3a,b). Compared to the 'SOCAT' run with a mean bias of 0.6 µatm (0.5 µatm), bias improves (i.e., moves closer to zero) to 0.08 µatm (0.4 µatm) for the 'SOCAT + FLOAT_hist' and to 0.1 µatm (0.3 µatm) for the 'SOCAT + FLOAT_hist_error' runs (Fig. 3a; Table 1). However, when the float observations are biased high by 4 µatm, the global mean (2000–2016) bias increases dramatically in both the 'Historical' and 'Optimized' experiment, to 1.1 µatm (0.3 µatm) and 1.5 µatm (0.2 µatm), respectively (Fig. 3a; Table 1). Note also that the 'SOCAT + FLOAT_hist_biased' experiment starts to deviate from the 'SOCAT' run already at the initiation of sampling and bias increases with time (Fig. 3b). For both the 'SOCAT + FLOAT_hist_biased' and 'SOCAT + FLOAT_opt_biased' runs, overestimation of pCO_2 (positive bias) mainly occurs in the southern hemisphere (Fig. S2).

The 'SOCAT + FLOAT_opt' and 'SOCAT + FLOAT_opt_error' float experiments show near-zero global mean biases for the entire duration of sample additions (2000–2016; Fig. 3c), with negative global mean biases of − 0.04 µatm (0.1 µatm) and − 0.05 µatm (0.2 µatm), respectively (Fig. 3a; Table 1).

For all float experiments, reduced (improved) bias compared to the 'SOCAT' run occurs generally in the Southern Ocean, and extends back in time prior to the addition of the floats (2000–2016) (Fig. S3). In the high southern latitudes, this is also the case for the 'biased' experiments (Fig. S3).

As found with RMSE, there is spread in the bias across the 75 testbed ensemble members of the LET, but there is less difference across the ESMs (Fig. 3a). The 75-member ensemble spread is larger for the 'SOCAT' run (IQR = 0.5 µatm) and the 'Historical' experiments (IQR = 0.3–0.4 µatm) compared to the 'Optimized' experiments (IQR = 0.1–0.2 µatm) (Fig. 3; Table 1). The 'SOCAT' run and the two 'biased' float experiments always demonstrate a positive mean bias, regardless of ESM (Fig. 3a; Table S1). Mean bias for the 'SOCAT + FLOAT' and 'error' experiments vary in sign depending on the ESM and type of sampling scheme (Fig. 3a; Table S1). For CanESM2, the 'SOCAT + FLOAT' and 'error' experiments for both float sampling schemes have negative mean bias, as do the 'SOCAT + FLOAT_opt' and 'SOCAT + FLOAT_opt_error' experiments for CESM. The 'SOCAT + FLOAT_hist' and 'SOCAT + FLOAT_hist_error' experiments demonstrate positive bias for CESM and GFDL.

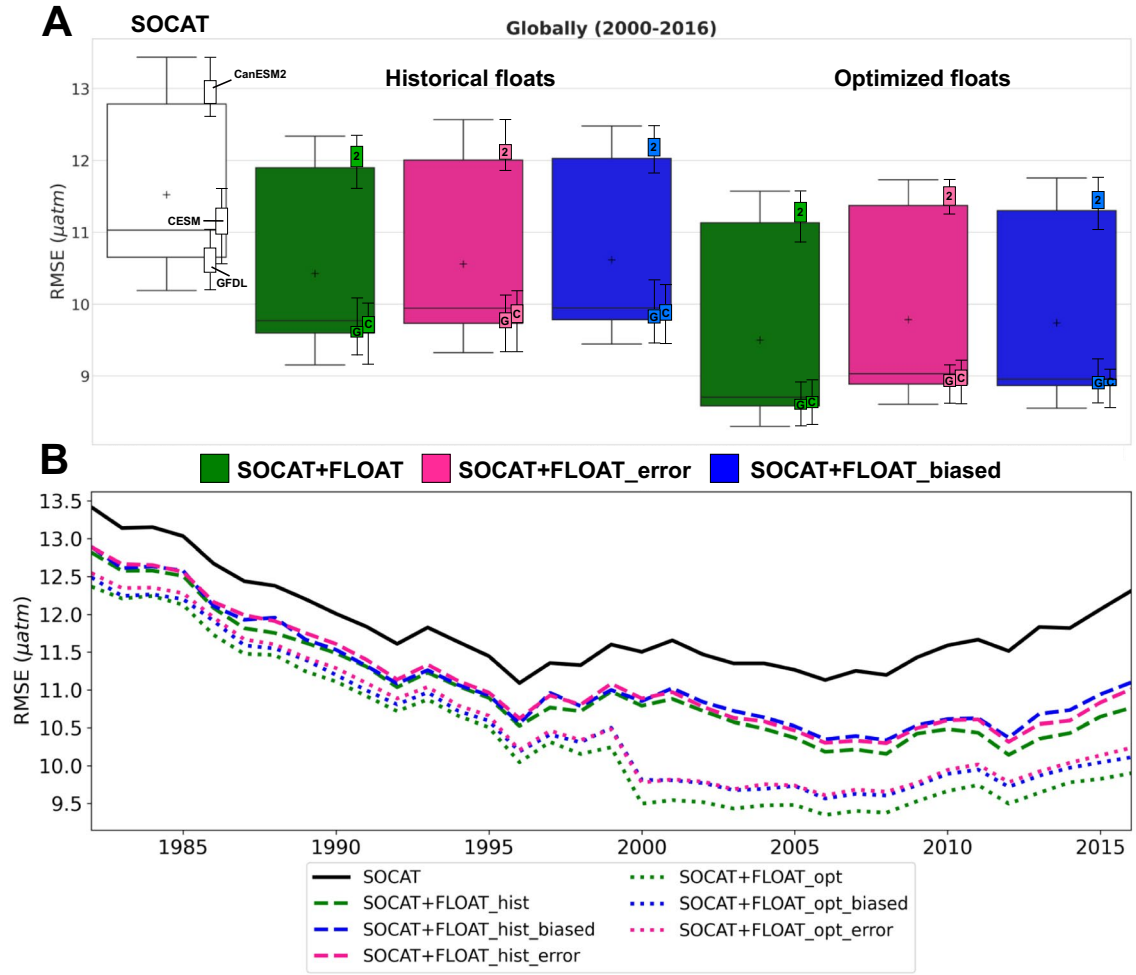

Fig. 2. Spread in RMSE globally for the duration of additional float sampling (2000–2016) for the full 75-member Large Ensemble Testbed (large boxes) and the three individual ESMs that each contributed 25 members (small boxes) (**A**). 2 = CanESM2. G = GFDL. C = CESM. The spread in RMSE for individual models includes outliers. Large colored boxes = interquartile range (IQR). Horizontal bars inside boxes = median. Horizontal bars outside boxes = minimum and maximum value. Crosses = mean. Annual global mean RMSE (for the 75 members) over the testbed period (1982–2016) for the six float experiments and the 'SOCAT' run (**B**).

2000–2016 global error metrics (in µatm)	SOCAT	Historical			Optimized		
		Baseline	Biased	Error	Baseline	Biased	Error
BIAS							
Testbed mean	0.6	0.08	1.1	0.1	−0.04	1.5	−0.05
1 IQR	0.5	0.4	0.3	0.3	0.1	0.2	0.2
Q1	0.9	0.3	1.3	0.3	0.05	1.6	0.03
Q3	0.4	−0.1	1.0	−0.1	−0.1	1.5	−0.1
RMSE							
Testbed mean	11.6	10.5	10.7	10.6	9.6	9.8	9.8
1 IQR	2.1	2.3	2.2	2.3	2.5	2.4	2.5

Table 1. Overview of global mean (2000–2016) bias and RMSE and the interquartile range (IQR) (in µatm) averaged over the full 75-member Large Ensemble Testbed.

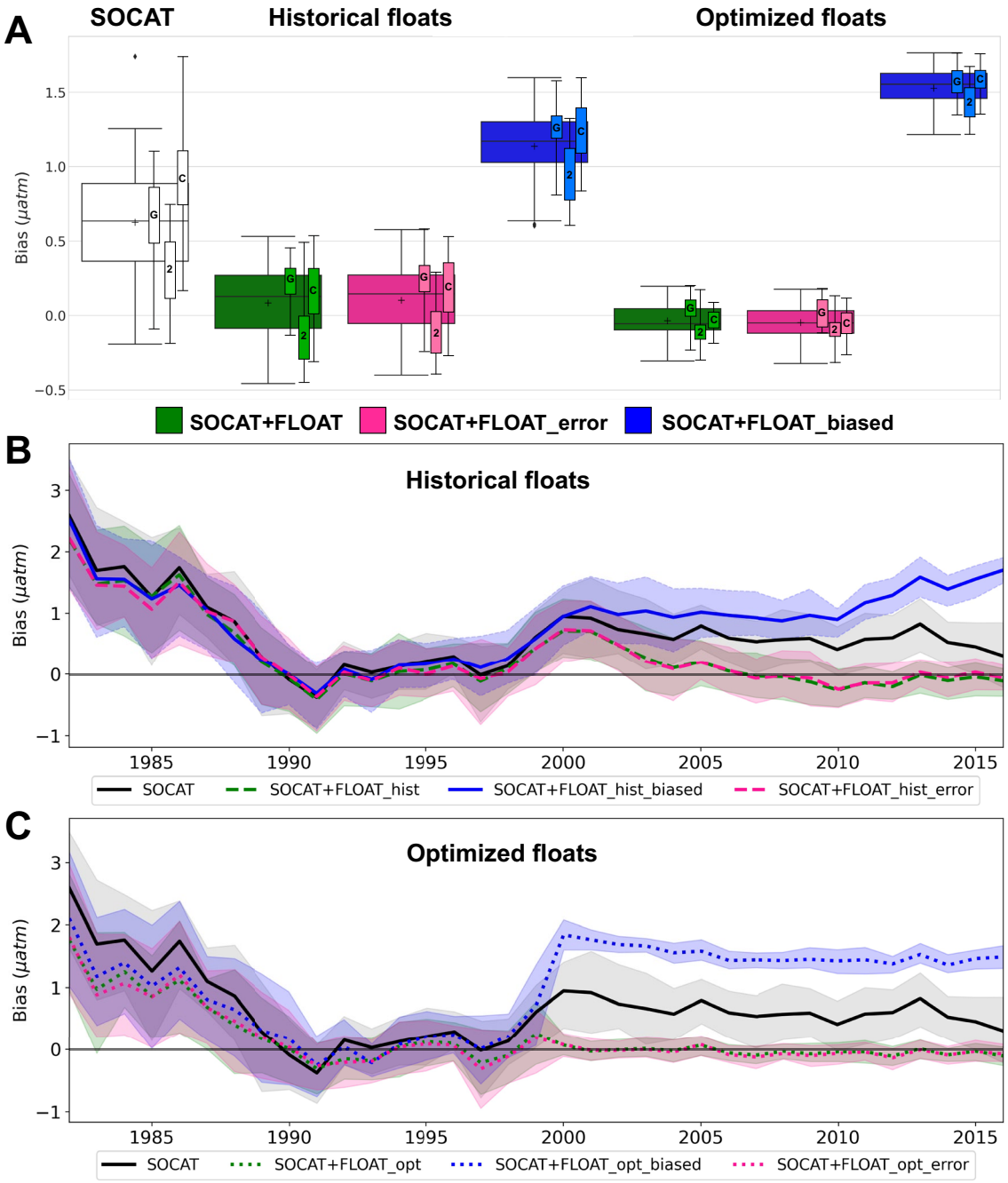

Fig. 3. Spread in bias globally for the duration of additional float sampling (2000–2016) for the full 75-member Large Ensemble Testbed (large boxes) and individual ESMs (small boxes) (**A**). 2 = CanESM2. G = GFDL. C = CESM. The spread in bias for individual models includes outliers. Large colored boxes = interquartile range (IQR). Horizontal bars inside boxes = median. Horizontal bars outside boxes = minimum and maximum value. Crosses = mean. Diamonds = outliers. Annual global mean bias (for the 75 members) over the testbed period (1982–2016) for the 'historical' (**B**) and 'optimized' (**C**) float experiments and the 'SOCAT' run, with shaded areas representing 1 IQR.

Air–sea CO_2 flux

The global air–sea flux was calculated in the same manner for the reconstructions and the 'model truth'. This allows for comparison of the differences in fluxes and attribution of flux differences solely to differences in the pCO_2 reconstructions due to biases and uncertainties in float observations. These are not estimates of real-world fluxes.

Compared to the 'model truth', the 'biased' experiments underestimate the mean annually averaged 2000–2016 global ocean sink by 0.26 Pg C $year^{-1}$ ('SOCAT + FLOAT_hist_biased') and 0.32 Pg C $year^{-1}$ ('SOCAT + FLOAT_opt_biased') (Fig. 4; Table S2). This is also reflected by the ensemble spread (Fig. S4). The 'baseline' and 'error'

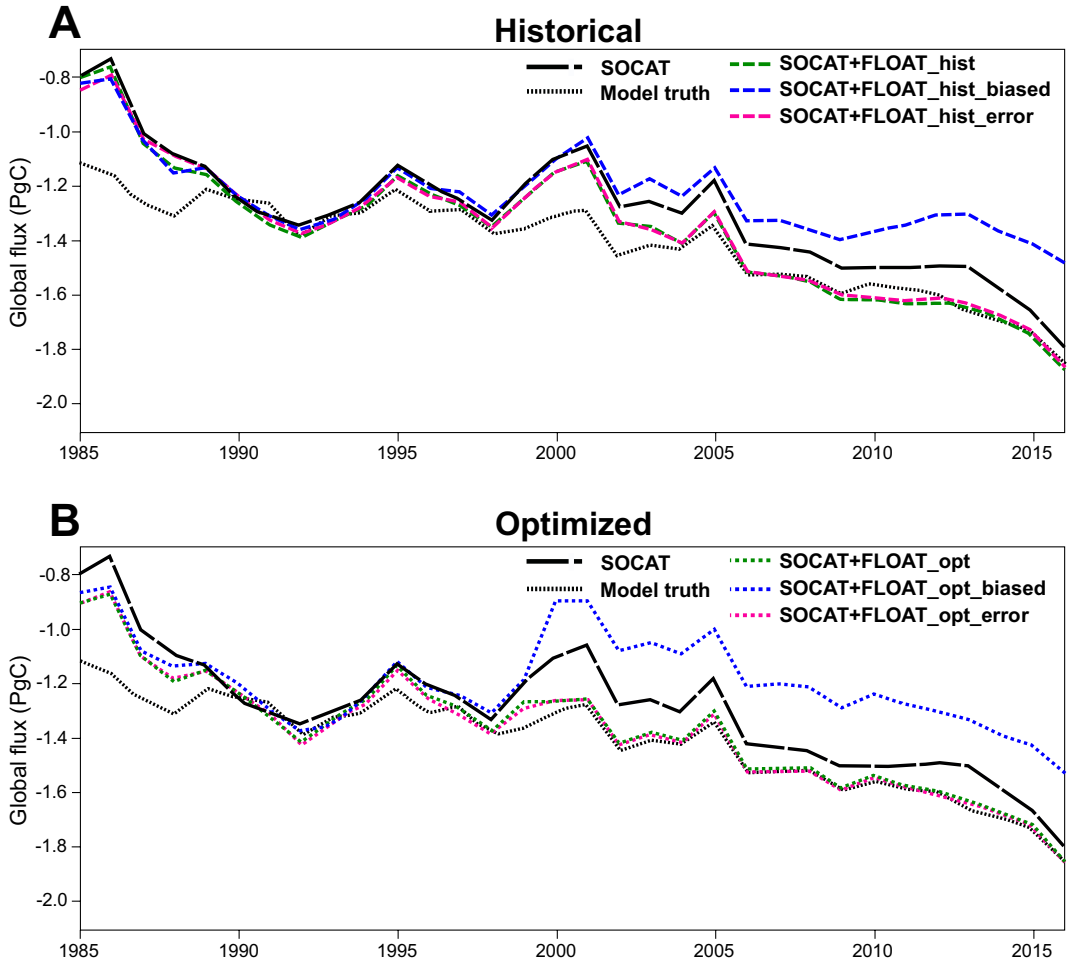

Fig. 4. Global annually averaged (all 75 members) air–sea CO_2 flux for the 'Historical' (**A**) and 'Optimized' (**B**) float experiments, compared to the 'SOCAT' run and the 'model truth'.

float addition experiments for both sampling schemes have a stronger global ocean sink, which is much closer to the 'model truth' (Fig. 4). These experiments deviate from the 'model truth' by as little as 0.02 Pg C $year^{-1}$ ('Historical') and 0.01 Pg C $year^{-1}$ ('Optimized') (Table S2), with a small spread across the ensembles (Fig. S4). The 'SOCAT + FLOAT_opt' and 'SOCAT + FLOAT_opt_error' experiments closely match the 'model truth' from the initiation of sampling (i.e., 2000) until the end of the testbed period (Fig. 4b). The majority of ensemble members underestimate the global ocean sink for the duration of float additions (2000–2016), or are indistinguishable from the 'model truth' (Table S2). However, some members do overestimate the sink for the 'baseline' and 'error' experiments, especially those of the CanESM2 model (Fig. S4). The CanESM2 model mean for the 'SOCAT + FLOAT_hist' and 'SOCAT + FLOAT_hist_error' experiments underestimates the global ocean sink by 0.03 Pg C $year^{-1}$ compared to the model truth (Table S2). All float experiments (except when floats are biased) show a negative mean bias of − 0.1 µatm for this model (Fig. 3; Table S1).

Discussion

We have used the pCO_2-Residual reconstruction method sampling from the Large Ensemble Testbed (LET[5]) to understand how bias and uncertainty in float-derived pCO_2 estimates may impact global reconstructions of surface ocean pCO_2 and the air–sea CO_2 flux. We find that a systematic bias in float observations significantly impacts the pCO_2 reconstruction globally (Fig. 3, Figs. S2, S3), leading to an underestimation of the mean 2000–2016 global ocean carbon sink of up to 0.32 Pg C $year^{-1}$ (Fig. 4; Table S2). The CO_2 flux between the ocean and atmosphere can be described as: $\Delta pCO_2 = pCO_2^{ocean} - pCO_2^{atm}$. If pCO_2^{ocean} is higher, ΔpCO_2 is positive, and this indicates outgassing as opposed to carbon uptake. The positive reconstruction bias shown by our 'biased' runs means that pCO_2^{ocean} is overestimated compared to the 'model truth'. Since the reconstructed pCO_2^{ocean} is higher than the 'truth', this leads to underestimation of the carbon uptake. Even if a small number of biased observations are introduced, pCO_2 is overestimated; the 'SOCAT + FLOAT_hist_biased' experiment starts to deviate from the 'SOCAT' run from the initiation of sampling (Fig. 3b), when float observations are limited (Fig. 1c). With an increasing number of biased sample additions, reconstruction bias increases (Fig. 3b). In contrast, when introducing stochastic uncertainty, the global mean bias and RMSE still improve compared to

the 'SOCAT' run (Figs. 2, 3). When accounting for measurement uncertainty of up to ± 11 µatm, the estimated 2000–2016 global mean air–sea flux deviates from the 'model truth' by as little as 0.01–0.02 Pg C year^{-1} (Table S2), which is comparable to the 'baseline' runs with no float bias or uncertainty (Fig. 4).

Despite the detrimental impacts to reconstruction bias, the 'biased' experiments show an improvement in RMSE compared to the 'SOCAT' run (Fig. 2). This improvement occurs mostly in the Southern Ocean (Figs. S1, S5a), which is the region with the sparsest coverage in SOCAT (Fig. S6). Regardless of sampling scheme, the float sampling significantly increases the total number of observations from the Southern Ocean (Fig. S6). Even if the float observations are biased, they still provide more information compared to the SOCAT database alone, resulting in the RMSE reduction. However, the bias in the float data strongly propagates into the reconstruction, resulting in a significant overestimation of pCO_2 (i.e., positive bias; Fig. 3, S2) and thus underestimation of the global and Southern Ocean sink (Fig. 4, Fig. S7). This suggests that improving reconstruction biases compared to RMSE is of greater importance in order to accurately estimate the air–sea flux.

Introducing biased samples from an already well covered region, such as the northern hemisphere, has less impact on the pCO_2 reconstruction in the same region. As shown by Fig. S5b, the 'biased' experiments show a significant reduction in bias over the northern hemisphere compared to the Southern Ocean and globally. The discrepancy between the 'model truth' and reconstructed fluxes shown globally, is mainly due to underestimation of the sink in the Southern Ocean (< 35° S; Fig. S7). Compared to the 'model truth', the 'biased' experiments underestimate the mean 2000–2016 northern hemisphere ocean sink by only 0.1 Pg C year^{-1} ('Optimized') and 0.03 Pg C year^{-1} ('Historical') (Fig. S7; Table S2). Particularly, the 'SOCAT + FLOAT_hist_biased' run shows lower bias over the northern hemisphere compared to the global and Southern Ocean, especially during the last years of the testbed period (Fig. S5b). This is likely due to the very small percentage of additional biased samples given the large number of SOCAT observations (Fig. S6).

A recent study quantified the effect of introducing a ± 5 µatm measurement uncertainty or a 5 µatm bias in sailboat observations[26]. They reconstructed surface ocean pCO_2 globally by using the SOM-FFN[27] method. In agreement with our study, they found a negligible impact of random errors in the measurements, but demonstrate a significant global bias in the flux calculations when sailboat-based measurements are biased.

In the study presented here, and in the study by Ref.[6] in which USV Saildrone observations are added to SOCAT in the LET, we find a stronger global and Southern Ocean sink during the period of sampling addition (Fig. 4; Table S2). Previous testbed studies using the CarboScope/Jena-MLS[28] and/or SOM-FFN[27] reconstruction methods found that additional float observations lead to a decreased (weakened) Southern Ocean carbon sink[11,29]. In our study, only the 'biased' experiments predict a weaker sink compared to the 'SOCAT' run (Fig. 4).

The study by Ref.[29] used a single ensemble of a hindcast model as a testbed. They show negative reconstruction biases and find the global ocean carbon sink to be overestimated for 2009–2018 in most experiments with realistic or enhanced sampling. This difference from our findings may be due to the reconstruction approaches or the different enhanced sampling patterns, but the models used as a testbed also play a role. Our ensemble average indicates that with SOCAT sampling, the pCO_2-Residual method underestimates the sink (Fig. 4), but some individual members do overestimate the sink, especially those from CanESM2 (Fig. 3a). Given the clustering of skill metrics based on ESM (Figs. 2a, 3a), it is clear that model structure plays a non-negligible role in the detailed results. Coordinated studies using identical testbeds will be required to directly compare different reconstruction approaches, and to understand why the different reconstruction methods show a different direction (over- vs. underestimation) of the bias and the estimated ocean sink.

The 'SOCAT + FLOAT_opt' performs better globally compared to the equivalent 'SOCAT + FLOAT_hist' run, with 17% vs. 9% improvement in global mean (2000–2016) RMSE (Fig. S1), lower mean bias and less spread (− 0.04 µatm; 1 IQR = 0.1 µatm vs. 0.08 µatm; 1 IQR = 0.4 µatm, respectively; Fig. 3, Table 1), and less deviation from the 'model truth' global ocean sink (Fig. 4, Fig. S4). However, it is important to note that the 'Optimized' sampling scheme includes almost five times as many observations as the 'Historical' (Fig. 1c). The 'Optimized' floats also do not change their location over time, and samples in the same place every month for 16 years, which is not operationally realistic. Despite the notable differences in these float scenarios, we do find some convergence as their sampling become more similar: For the last four years of the testbed, when the number of sampling additions from the Southern Ocean is comparable (Fig. S6), RMSE values here are more or less identical (Fig. S5a), and the 'Historical' runs are able to reproduce the global ocean sink 'model truth' (Fig. 4). The addition of year-round samples from this poorly sampled region appears to be more important than the exact sampling pattern of the floats.

The greatly expanded spatiotemporal coverage by float-based estimates provides valuable data from regions and seasons that are severely undersampled by shipboard observations, particularly the Southern Ocean and especially during winter months[6,30,31]. Targeted sampling from autonomous platforms combined with ships, filling in the multi-dimensional state space of pCO_2 and its driver variables, represents a likely path forward to improve surface ocean pCO_2 reconstructions and air–sea CO_2 flux estimates[5–7,11,26,29–33]. However, although current studies agree that random measurement uncertainty has negligible impact on pCO_2 reconstructions, they also demonstrate the likely severe impact of bias in indirect float-based pCO_2 observations. Bias must be addressed before incorporating indirect pCO_2 estimates into global reconstructions, especially in areas with low coverage.

Data availability

The Large Ensemble Testbed is publicly available at https://figshare.com/collections/Large_ensemble_pCO2_testbed/4568555. Data analysis scripts and supporting files are publicly available in a GitHub repository at https://github.com/hatlenheimdalthea/Sampling_experiments_LET_Argo. The float+SOCAT sampling masks are publicly available at https://doi.org/10.5281/zenodo.13367537. Times and locations of floats of the 'Historical' sampling scenario are from https://fleetmonitoring.euro-argo.eu/dashboard.

Received: 8 February 2024; Accepted: 19 August 2024
Published online: 26 August 2024

References

1. Bakker, D. C. E. *et al.* A multi-decade 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).
2. Friedlingstein, P. *et al.* Global carbon budget 2023. *Earth Syst. Sci. Data* **15**, 5301–5369. https://doi.org/10.5194/essd-15-5301-2023 (2023).
3. Bakker, D. C. E. *et al.* Surface Ocean CO_2 Atlas Database Version 2022 (SOCATv2022) (NCEI Accession 0253659), NOAA National Centers for Environmental Information. https://doi.org/10.25921/1h9f-nb73 (2022).
4. McKinley, G. A., Fay, A. R., Eddebbar, Y. A., Gloege, L. & Lovenduski, N. S. External forcing explains recent decadal variability of the ocean carbon sink. *AGU Adv.* **1**(2), e2019AV000149. https://doi.org/10.1029/2019AV000149 (2020).
5. Gloege, L. *et al.* Quantifying errors in observationally based estimates of ocean carbon sink variability. *Glob. Biogeochem. Cycles.* https://doi.org/10.1029/2020gb006788 (2021).
6. Heimdal, T. H., McKinley, G. A., Sutton, A. J., Fay, A. R. & Gloege, L. Assessing improvements in global ocean pCO_2 machine learning reconstructions with Southern Ocean autonomous sampling. *Biogeosciences* **21**, 2159–2176. https://doi.org/10.5194/bg-21-2159-2024 (2024).
7. Sutton, A. J., Williams, N. L. & Tilbrook, B. Constraining Southern Ocean CO_2 flux uncertainty using uncrewed surface vehicle observations. *Geophys. Res. Lett.* **48**(3), e2020GL091748. https://doi.org/10.1029/2020GL091748 (2021).
8. Sabine, C. *et al.* Evaluation of a new carbon dioxide system for autonomous surface vehicles. *J. Atmos. Oceaenic Technol.* **37**(8), 1305–1317. https://doi.org/10.1175/JTECH-D-20-0010.1 (2020).
9. Williams, N. L. *et al.* Calculating surface ocean pCO_2 from biogeochemical Argo floats equipped with pH: An uncertainty analysis. *Glob. Biogeochem. Cycles* **31**(3), 591–604. https://doi.org/10.1002/2016GB005541 (2017).
10. Fay, A. R. *et al.* Utilizing the Drake Passage Time-series to understand variability and change in subpolar Southern Ocean pCO_2. *Biogeosciences* **15**(12), 3841–3855. https://doi.org/10.5194/bg-15-3841-2018 (2018).
11. Bushinsky, S. M. *et al.* Reassessing Southern Ocean air–sea CO_2 flux estimates with the addition of biogeochemical float observations. *Glob. Biogeochem. Cycles* **33**(11), 1370–1388. https://doi.org/10.1029/2019GB006176 (2019).
12. Gray, A. R. *et al.* Autonomous biogeochemical floats detect significant carbon dioxide outgassing in the high-latitude Southern Ocean. *Geophys. Res. Lett.* **45**(17), 9049–9057. https://doi.org/10.1029/2018GL078013 (2018).
13. Mackay, N. & Watson, A. Winter air–sea CO_2 fluxes constructed from summer observations of the polar Southern Ocean suggest weak outgassing. *J. Geophys. Res. Oceans.* **126**(5), e2020JC016600. https://doi.org/10.1029/2020JC016600 (2021).
14. Wu, Y. *et al.* 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).
15. Khatiwala, S., Primeau, F. & Hall, T. Reconstruction of the history of anthropogenic CO_2 concentrations in the ocean. *Nature* **462**(7271), 346–349. https://doi.org/10.1038/nature08526 (2009).
16. Bennington, V., Galjanic, T. & McKinley, G. A. Explicit physical knowledge in machine learning for ocean carbon flux reconstruction: The pCO_2-residual method. *J. Adv. Modeling Earth Syst.* https://doi.org/10.1029/2021ms002960 (2022).
17. Kay, J. E. *et al.* The Community Earth System Model (CESM) large ensemble project: A community resource for studying climate change in the presence of internal climate variability. *Bull. Am. Meteor. Soc.* **96**(8), 1333–1349. https://doi.org/10.1175/BAMS-D-13-00255 (2015).
18. Rodgers, K. B., Lin, J. & Frölicher, T. L. Emergence of multiple ocean ecosystem drivers in a large ensemble suite with an Earth system model. *Biogeosciences* **12**(11), 3301–3320. https://doi.org/10.5194/bg-12-3301-2015 (2015).
19. Fyfe, J. C. *et al.* Large near-term projected snowpack loss over the western United States. *Nat. Commun.* **8**, 14996. https://doi.org/10.1038/ncomms14996 (2017).
20. Takahashi, T., Olafsson, J., Goddard, J. G., Chipman, D. W. & Sutherland, S. C. Seasonal variation of CO_2 and nutrients in the high-latitude surface oceans: A comparative study. *Glob. Biogeochem. Cycles* **7**(4), 843–878. https://doi.org/10.1029/93GB02263 (1993).
21. Takahashi, T. *et al.* Global sea-air CO_2 flux based on climatological surface ocean pCO_2, and seasonal biological and temperature effects. *Deep Sea Res. Part II* **49**(9–10), 1601–1622. https://doi.org/10.1016/S0967-0645(02)00003-6 (2002).
22. Chen, T. & Guestrin, C. Xgboost: A scalable tree boosting system. in *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 785–794) (2016). https://doi.org/10.1145/2939672.2939785.
23. Gregor, L. & Fay, A. R. Air–sea CO_2 fluxes for surface pCO_2 data products using a standardized approach, Zenodo [code. https://doi.org/10.5281/zenodo.5482547 (2021).
24. Fay, A. R. *et al.* SeaFlux: Harmonization of air–sea CO_2 fluxes from surface pCO_2 data products using a standardized approach. *Earth Syst. Sci. Data* **13**, 4693–4710. https://doi.org/10.5194/essd-13-4693-2021 (2021).
25. Chamberlain, P., Talley, L. D., Cornuelle, B., Mazloff, M. & Gille, S. T. Optimizing the biogeochemical Argo float distribution. *J. Atmos. Oceanic Tech.* **40**(11), 1355–1379. https://doi.org/10.1175/JTECH-D-22-0093.1 (2023).
26. Behncke, J., Landschützer, P. & Tanhua, T. A detectable change in the air–sea CO_2 flux estimate from sailboat measurements. *Sci. Rep.* **14**, 3345. https://doi.org/10.1038/s41598-024-53159-0 (2024).
27. Landschützer, P. *et al.* The reinvigoration of the Southern Ocean carbon sink. *Science* **349**(6253), 1221–1224. https://doi.org/10.1126/science.aab2620 (2015).
28. Rödenbeck, C. *et al.* Interannual sea–air CO_2 flux variability from an observation-driven ocean mixed-layer scheme. *Biogeosciences* **11**, 4599–4612. https://doi.org/10.5194/bg-11-4599-2014 (2014).
29. Hauck, J. *et al.* Sparse observations induce large biases in estimates of the global ocean CO_2 sink: An ocean model subsampling experiment. *Philos. Trans. R. Soc. A* **381**, 20220063. https://doi.org/10.1098/rsta.2022.0063 (2023).
30. Djeutchouang, L. M., Chang, N., Gregor, L., Vichi, M. & Monteiro, P. M. S. The sensitivity of pCO_2 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).
31. Mackay, N., Watson, A. J., Suntharalingam, P., Chen, Z. & Landschützer, P. Improved winter data coverage of the Southern Ocean CO2 sink from extrapolation of summertime observations. *Commun. Earth Environ.* **3**, 265. https://doi.org/10.1038/s43247-022-00592-6 (2022).
32. Gregor, L., Lebehot, A. D., Kok, S. & Monteiro, P. M. S. 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. *Geosci. Model Develop.* **12**, 5113–5136. https://doi.org/10.5194/gmd-12-5113-2019 (2019).
33. Landschützer, P., Tanhua, T., Behncke, J. & Keppler, L. Sailing through the Southern Ocean seas of air–sea CO2 flux uncertainty. *Philos. Trans. R. Soc. A*. https://doi.org/10.1098/rsta.2022.0064 (2023).

Acknowledgements
We acknowledge funding from NSF through the LEAP STC (Award #2019625). We thank Paul Chamberlain for discussions regarding the 'Optimized' float mask and for providing the code to generate the mask. We would like to acknowledge Val Bennington, Devan Samant, Julius Busecke, Amanday Fay and Abby P. Shaum for providing technical support.

Author contributions
THH and GAM designed the experiments. THH performed the simulations and calculated air–sea fluxes. THH and GAM wrote the manuscript.

Competing interests
The authors declare no competing interests.

Additional information
Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-024-70617-x.

Correspondence and requests for materials should be addressed to T.H.H.