Supplement of

The sensitivity of pCO_2 reconstructions in the Southern Ocean to sampling scales: a semi-idealized model sampling and reconstruction approach

Laique M. Djeutchouang et al

Correspondence to: Laique M. Djeutchouang (merlindjeutchouang@gmail.com)

Supplementary Materials

Acronym	Description
FNN	Feed-forward Neural Network
GBM	Gradient Boosting Machines
SAZ	Sub-Antarctic Zone
PFZ	Polar Frontal Zone
NEMO	Nucleus for European Modelling Ocean
PISCES	Pelagic Interactions Scheme for Carbon and Ecosystem Studies
CSIR	Council of Scientific and Industrial Research
SOCCO	Southern Ocean Carbon and Climate Observatory
SOCCOM	Southern Ocean Carbon and Climate Observation Modelling

SOCAT	Surface Ocean CO_2 Atlas
SOSCEx	Southern Ocean Seasonal Cycle Experiment
WG	Waveglider
nUSV	new unmanned surface vehicle
ML	Machine Learning

This supporting information document provides ancillary methodological details and results pertaining to (1) descriptions of the study domain and mode data variables including the motive of the selection of the experimental domain, the characteristics of the NEMO-PISCES model (BIOPERIANT12) data variables of interest and processing, the experimental setting and steps used in the pCO_2 reconstruction; (2) descriptions of the ML regression methods; and (3) additional components on the results and discussion including the model training errors or in-sample uncertainties and biases, and the overall results of the SHIP experiment. Accompanying this supporting information text are four supplementary figures and four supplementary tables.

S1 Descriptions of the study domain and mode data variables

S1.1 Selection of the study domain

Many studies, the seasonal cycle is known as the strongest mode of natural variability of carbon dioxide (CO_2) and also the one that most strongly links climate and ocean ecosystems. The seasonal cycle characteristics are largely shaped by higher frequency intra-seasonal modes defining the response modes in physics and biogeochemistry components (Mongwe et al., 2016, 2018). Therefore, the SOSCEx - an initiative of the SOCCO, a research programme led by the CSIR- was launched in 2013. The SOSCEx aimed to explore the nature and links in dynamics and scale sensitivities of atmospheric forcing, CO2 fluxes, and primary production, with a particular focus on the seasonal cycle mode as a test for the climate sensitivity of earth systems models in respect of the evolution of both atmospheric CO_2 and ocean ecosystems in the 21st century (Swart et al., 2012; Monteiro et al., 2010, 2015). The novel aspect of the third phase (SOSCEx III, 2015-2018) of the project was the integration of a multi-platform approach that consisted of combining gliders, ships, floats, satellites and prognostic models in order to explore new questions about climate sensitivity of CO_2 and ocean ecosystem dynamics and how these processes are parameterized in forced ocean models such as the high-resolution (±10km) forced NEMO-PISCES ocean model BIOPERIANT12.

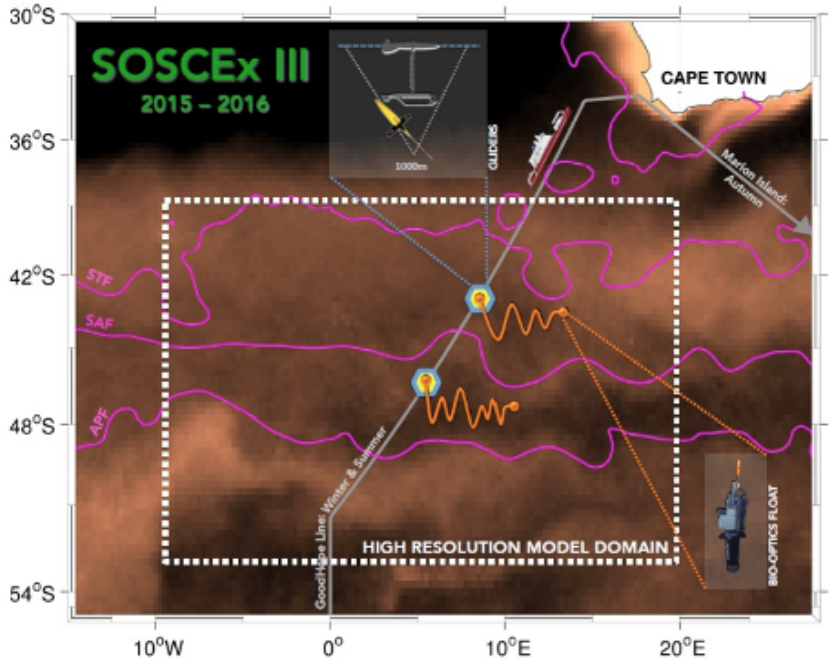

Figure S1: Schematic view of the observing strategy for the beginning of the third phase of SOSCEx project illustrating the use of multiple SO observing platforms, ships, gliders, floats, and numerical models. The hexagonal patterns (blue-yellow) depict the twined glider deployments; the orange curve shows the Lagrangian float sampling trajectories, while the high-resolution modelling domain is represented with the white dashed line. In magenta lines are the average locations of the oceanic fronts shown as derived from satellite altimetry data, whereas the underlying shading depicts the mean summer chlorophyll-a concentration in the region with lighter shading=high Chl-a areas. (Source: https://socco.org.za/news/plans-underway-for-soscex-iii/).

S1.2 Data variable characteristics and processing

Here we summarize in Table 1 the data variables of the forced high-resolution (±10km) NEMO-PISCES coupled ocean model BIOPERIANT12, and the processing techniques these variables have to undergo.

Table S1: Summary of the BIOPERIANT12 model variables of interest, and data processing steps applied on feature and target variables.

Variables	Abbreviations	Processing	Date range	Resolutions	
				Space	Time
Air-sea pCO_2 gradient	ΔpCO_2	Model simulations	1 year	1/12°	daily
Atmospheric (atm) pCO_2	pCO_2^{atm}	*In-situ*			
Surface ocean pCO_2	pCO_2 or pCO_2^{ocean}	$pCO_2^{atm} - \Delta pCO_2$			
Sea surface temperature	SST	Model simulations			

Sea surface salinity	SSS	Model simulations			
Mixed layer depth	MLD	Model simulations			
		$\log_{10}$ transformation			
Nano chlorophyll concentration	NChl	Model simulations			
Diatom chlorophyl concentration	DChl	Model simulations			
Chlorophyll-a	Chl-a	NChl + DChl			
Day of the year	J	$\left(\cos\left(j \times \frac{2\pi}{365}\right), \sin\left(j \times \frac{2\pi}{365}\right)\right)$		-	

55

S1.3 Summary of the experiment

Table S2: Summary of all the 8 semi-idealized ocean system simulations experiments (OSSE-8) that we conducted in this study. The simulated ocean observing platforms (SHIP, FLOAT, WG, and nUSV Saildrone) correspond to their real-world counterparts (ship, carbon-float, Waveglider, and Saildrone) used in the SOCAT project, SOCCOM initiative, SOCCO programme and by Saildrone Inc., respectively. The sampling regimes represent the periods in which the data sampling phase of different experiments occurred according to the temporal scales of the underlying platforms. Note that the observing platforms Wave glider and float have two scenarios each based on the fact that they are deployed either in the north (SAZ) or south (PFZ) of the 10º x 20º experimental domain. Experiment abbreviations together with their subsequent scenarios (defined by the sampling regimes/strategies) are used in figures and throughout the text.

Ocean Observing Platforms	Sets	Sampling Regimes	Experiments
Ships (SOCAT-like)	SHIP	Summer (smr)	SHIP(smr)
		Summer + Winter (smr+wtr)	SHIP(smr+wtr)
		Autumn + Spring (aut+spr)	SHIP(aut+spr)
Floats (SOCCOM-like)	SHIP + FLOAT	Summer (smr) + One year round	SHIP(smr) + FLOAT(SAZ)
			SHIP(smr) + FLOAT(PFZ)
			SHIP(smr) + FLOAT(SAZ+PFZ)
Wavegliders (SOCCO-like)	SHIP + WG		SHIP(smr) + WG(SAZ)
			SHIP(smr) + WG(PFZ)
Saildrones	SHIP + nUSV		SHIP(smr) + nUSV

S2 Descriptions of the supervised ML regression methods

S2.1 Feed-forward Neural Network

The Feed-forward Neural Network (FNN) is a class of the neuronal network algorithms that is most commonly used as non-linear approach in the surface ocean pCO_2 community (Rödenbeck et al., 2015; Landschützer et al., 2016; Denvil-Sommer et al., 2019; Gregor et al., 2019; Gloege et al., 2021; Bushinsky et al., 2019; Gregor and Gruber, 2021). Solving pCO2 problem is within the capability of a single hidden layer of the neural network (Landschützer et al., 2013; Gregor and Gruber, 2021). Thus, we use the Multi-layer Perceptron regressor whose implementation is in the Scikit-learn Python package. The principle is summarized in Fig. 2 where a network with random weights is generated similarly to coefficients in the linear regression. Data are passed forward through the network in order to estimate the target values (pCO_2). The difference between estimates and true values is backpropagated through the weights until the targets are met with sufficient accuracy. In our study, we tuned the following primary hyper-parameters: the number of hidden layer and weights per layer (these the architecture of the network, e.g., Fig. 2), and the learning rate (α). We tuned these hyper-parameters with a Bayes-search cross-validation (BayesSearchCV) approach by making use of the Scitkit-optimize Python package.

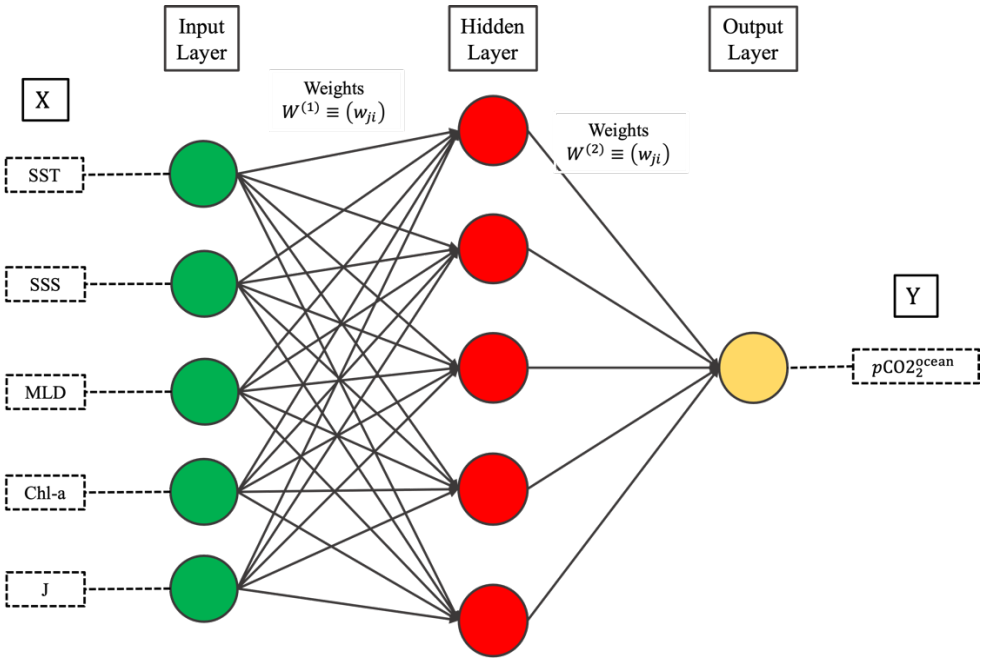

Figure S2: Depiction of a typical example of the architecture or graph of a single hidden layer Multi-layer Perceptron network. X is the array of features data [SST, SSS, MLD, Chl-a, J] whereas Y is the array of the target variable [pCO_2^{ocean}] as described in Table 1.

S2.2 Gradient Boosting Machines

Gradient Boosting Machines (GBM) is a widely used machine learning (ML) algorithm due its efficiency, accuracy, and interpretability (Chen and Guestrin, 2016; Ke et al., 2017; Gregor et al., 2019; Gregor and Gruber, 2021). It is a variant of the Gradient Boosting Decision Tree (GBDT) learning frameworks. GBM produce a prediction model in the form of an ensemble of weak prediction models typically called decision tree learners that increase the efficiency of the model and reduce memory usage during the training. It builds these multiple weak learners in a stage-wise or sequential fashion and generalize them by allowing optimization of an arbitrary differentiable loss function (Friedman, 2001; Ke et al., 2017). This can be known as aggregative learning, where in each stage algorithm improves what it is learnt. Although GBM have been proven to deal well with imbalanced or sparse datasets (Ke et al., 2017)), it is more likely to overfit the training data because of the model potential for high complexity (Frery et al., 2017). Thus, tuning GBM hyper-parameters to prevent overfitting is very important. In our study, the following hyper-parameters were tuned: number of trees or leaves, depth of the trees, learning rate, number of estimators, and boosting type. We use the LightGBM and Scikit-optimize Python packages for our implementation of GBM and optimization/tuning of hyper-parameters through BayesSearchCV module, respectively.

S2.3 Surface ocean pCO_2 reconstruction steps

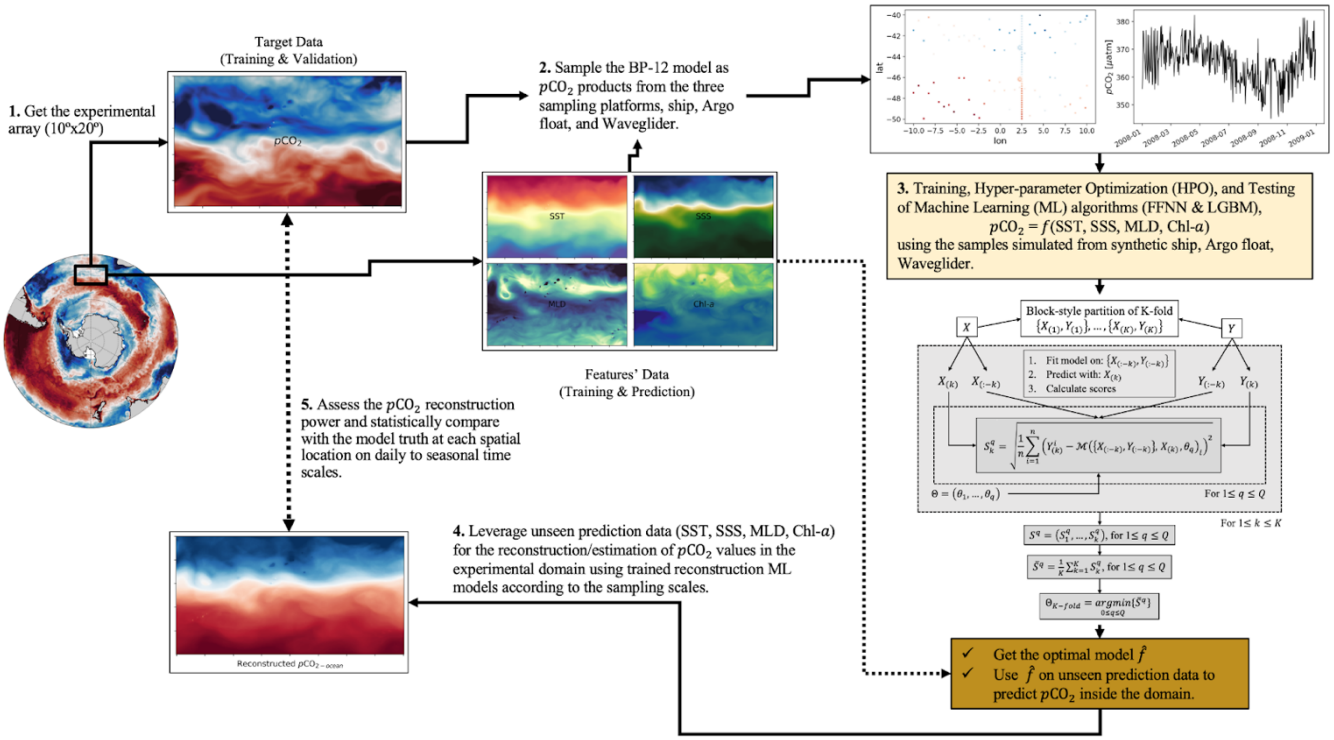

Figure S3: Schematic flow diagram showing the key steps required to reconstruct the surface ocean pCO_2 in the full experimental domain.

S2 Results and Discussion

S2.1 ML regression in-sample scores for individual methods

In-sample scores correspond to ML regression score calculated from all the training data points. This allowed to control the overfitting of the methods during the training. For instance, by focusing on the root mean square errors (RMSEs) and the mean bias errors (MBEs) or simply biases reported in Table 3, the nuanced differences between the two ML regression methods FNN and GBM show that GBM method was likely to overfit on training data compared to FNN method.

Table S3: Various in-sample errors (i.e., errors calculated from all the training points) for empirical estimates of the surface ocean pCO2 for different experiments we run. The configuration of these experiments is presented in Table 1 and clearly described in Section 2.3.2. The machine learning regression metrics we used to report this in-sample error are abbreviated as follows: RMSE is the root mean square error; MAE is the mean absolute error; MBE or Bias is the mean average error.

Sets	Sampling Regimes	Experiments	Algorithms	RMSE (µatm)	MAE (µatm)	MBE (µatm)
SHIP		SHIP(smr)	FNN	4.07	3.39	0.14

	Summer (smr)		GBM	0.84	0.65	0.01
	Summer + Winter (smr+wtr)	SHIP(smr+wtr)	FNN	5.19	4.09	-0.51
			GBM	1.41	1.07	0.02
	Autumn + Spring (aut+spr)	SHIP(aut+spr)	FNN	3.78	3.05	0.19
			GBM	2.25	1.75	0.08
SHIP + FLOAT	Summer (smr) + One year round	SHIP(smr) + FLOAT(SAZ)	FNN	6.21	5.06	0.21
			GBM	1.49	1.12	0.06
		SHIP(smr) + FLOAT(PFZ)	FNN	5.11	4.17	0.08
			GBM	0.85	0.64	0.02
		SHIP(smr) + FLOAT(SAZ+PFZ)	FNN	8.76	7.52	-2.01
			GBM	1.49	1.12	0.06
SHIP + WG		SHIP(smr) + WG(SAZ)	FNN	4.12	2.92	0.38
			GBM	0.54	0.35	0.01
		SHIP(smr) + WG(PFZ)	FNN	2.27	1.65	-0.12
			GBM	0.08	0.05	0.02
SHIP + nUSV		SHIP(smr) + nUSV	FNN	5.39	4.29	-0.11
			GBM	2.55	1.98	-0.03

S2.2 Overall results from the SHIP experiment

S2.2.1 The spatial and seasonal cycle anomalies

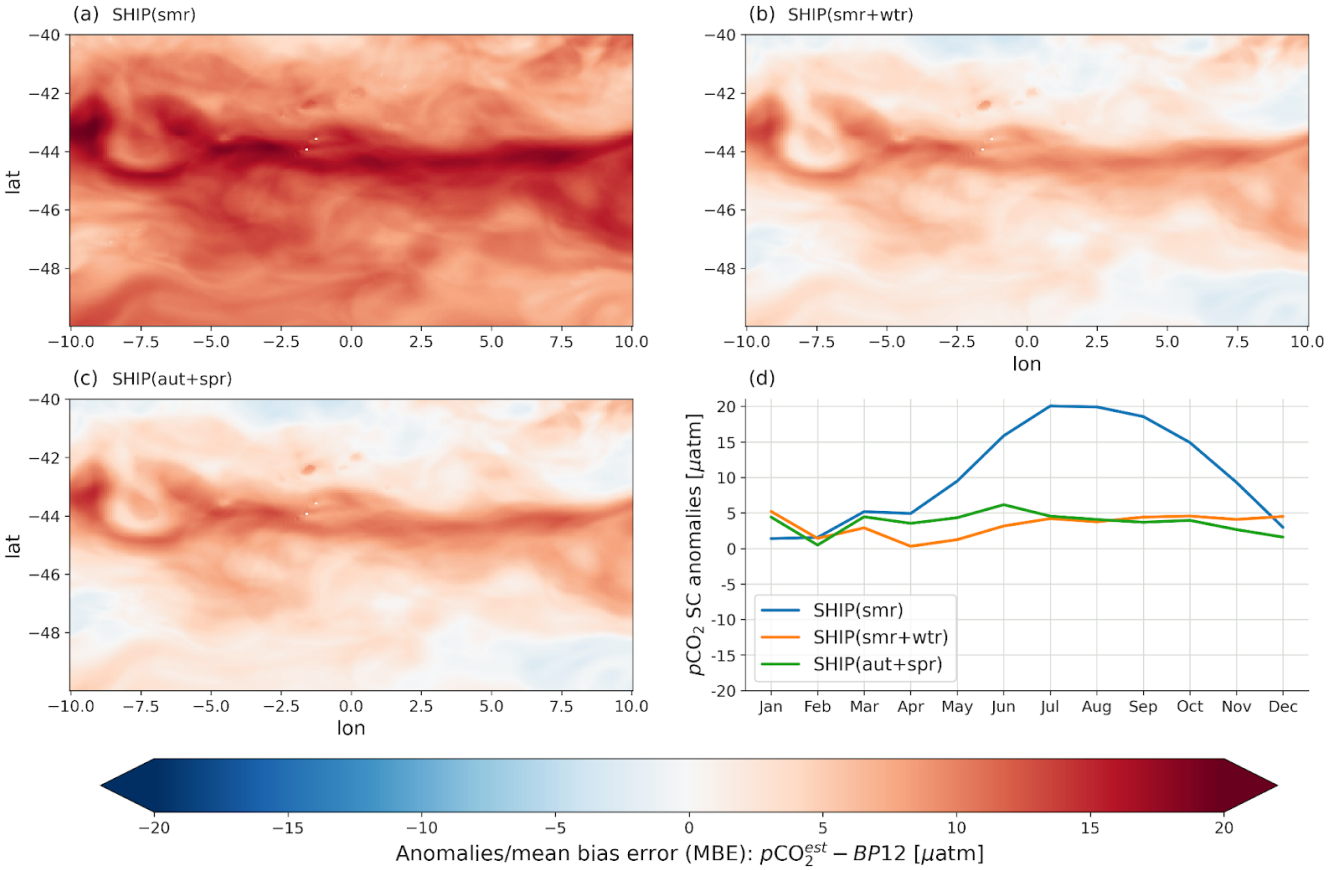

Figure S4: Reconstruction anomalies for the idealized SHIP experiment where the idealized ship sampled the domain according to the three sampling regimes/scenarios, summer (smr), summer + winter (smr+wtr), and autumn + spring (aut+spr). Panels (a), (b) and (c) show the maps of the reconstruction anomalies based these three sampling regimes, hence the three experiments SHIP(smr), SHIP(smr+wtr), and SHIP(aut+spr) respectively; panel (d) shows the anomalies of the mean seasonal cycle (SC) reconstruction based on these three sampling regimes; that is, SHIP(smr), SHIP(smr+wtr) , and SHIP(aut+spr).

S2.2.2 Reconstruction skills for the SHIP experiment

Table S4: ML regression modelling scores of the ensemble average (ML2) for the SHIP set of experiments: SHIP(smr) for summer sampling, SHIP(smr+wtr) for summer and winter sampling, and SHIP(aut+spr) for autumn and spring sampling. The configuration of this set of experiments is presented in Table S2. The first column of the table is the experimental set and the second one corresponds to the considered experiments. The statistical metrics used to assess ML2 for this set of experiments are abbreviated as follows: RMSE is the root mean square error; MAE is the mean absolute error; MBE or Bias is the mean average error; and r is the Pearson's correlation coefficient between the reconstructed and BP12 model truth pCO_2. Values in the table are significantly different from the mean for the corresponding column (with 95% confidence level or p-value < 0.05 for the two-tailed Z-test).

Sets	Experiments	RMSE	MAE	MBE	r

		(µatm)	(µatm)	(µatm)	
SHIP	SHIP(smr)	13.79	11.51	10.52	0.36
	SHIP(smr+wtr)	6.8	5.29	3.18	0.73
	SHIP(aut+spr)	7.07	5.5	3.57	0.72

140

References

Bushinsky, S. M., Landschützer, P., Rödenbeck, C., Gray, A. R., Baker, D., Mazloff, M. R., et al. (2019). Reassessing Southern
145 Ocean Air-Sea CO2 Flux Estimates With the Addition of Biogeochemical Float Observations. *Global Biogeochemical Cycles*, *33*(11), 1370–1388. https://doi.org/10.1029/2019GB006176

Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (Vol. 42, pp. 785–794). New York, NY, USA: ACM. https://doi.org/10.1145/2939672.2939785

150 Denvil-Sommer, A., Gehlen, M., Vrac, M., & Mejia, C. (2019). LSCE-FFNN-v1: A two-step neural network model for the reconstruction of surface ocean pCO2 over the global ocean. *Geoscientific Model Development*, *12*(5), 2091–2105. https://doi.org/10.5194/GMD-12-2091-2019

Frery, J., Habrard, A., Sebban, M., Caelen, O., & He-Guelton, L. (2017). Efficient Top Rank Optimization with Gradient Boosting for Supervised Anomaly Detection BT - Machine Learning and Knowledge Discovery in Databases. In M.
155 Ceci, J. Hollmén, L. Todorovski, C. Vens, & S. Džeroski (Eds.) (pp. 20–35). Cham: Springer International Publishing.

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, *29*(5), 1189–1232. https://doi.org/10.1214/aos/1013203451

Gloege, L., McKinley, G. A., Landschützer, P., Fay, A. R., Frölicher, T. L., Fyfe, J. C., et al. (2021). Quantifying Errors in Observationally Based Estimates of Ocean Carbon Sink Variability. *Global Biogeochemical Cycles*, *35*(4),
160 e2020GB006788. https://doi.org/10.1029/2020GB006788

Gregor, L., & Gruber, N. (2021). OceanSODA-ETHZ: A global gridded data set of the surface ocean carbonate system for seasonal to decadal studies of ocean acidification. *Earth System Science Data*, *13*(2), 777–808. https://doi.org/10.5194/ESSD-13-777-2021

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

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., et al. (2017). LightGBM: A highly efficient gradient boosting

decision tree. *Advances in Neural Information Processing Systems*, *2017-Decem*(Nips), 3147–3155.

Landschützer, P., Gruber, N., Bakker, D. C. E., Schuster, U., Nakaoka, S., Payne, M. R., et al. (2013). A neural network-based estimate of the seasonal to inter-annual variability of the Atlantic Ocean carbon sink. *Biogeosciences*, *10*(11), 7793–7815. https://doi.org/10.5194/bg-10-7793-2013

Landschützer, P., Gruber, N., & Bakker, D. C. E. E. (2016). Decadal variations and trends of the global ocean carbon sink. *Global Biogeochemical Cycles*, *30*(10), 1396–1417. https://doi.org/10.1002/2015GB005359

Mongwe, N. P., Chang, N., & Monteiro, P. M. S. (2016). The seasonal cycle as a mode to diagnose biases in modelled CO2 fluxes in the Southern Ocean. *Ocean Modelling*, *106*, 90–103. https://doi.org/10.1016/j.ocemod.2016.09.006

Mongwe, N. P., Vichi, M., & Monteiro, P. M. S. (2018). The seasonal cycle of <i>p</i>CO₂ and CO₂ fluxes in the Southern Ocean: diagnosing anomalies in CMIP5 Earth system models. *Biogeosciences*, *15*(9), 2851–2872. https://doi.org/10.5194/bg-15-2851-2018

Monteiro, P. M. S., Monteiro, P. M. S., Monteiro, P. M. S., Monteiro, P. M. S., Monteiro, P. M. S., Monteiro, P. M. S., et al. (2010). A Global Sea Surface Carbon Observing System: Assessment of Changing Sea Surface CO2 and Air-Sea CO2 Fluxes, 702–714. https://doi.org/10.5270/OCEANOBS09.CWP.64

Monteiro, P. M. S. S., Gregor, L., Lévy, M., Maenner, S., Sabine, C. L., & Swart, S. (2015). Intraseasonal variability linked to sampling alias in air-sea CO2 fluxes in the Southern Ocean. *Geophysical Research Letters*, *42*(20), 8507–8514. https://doi.org/10.1002/2015GL066009

Rödenbeck, C., Bakker, D. C. E., Gruber, N., Iida, Y., Jacobson, A. R., Jones, S., et al. (2015). Data-based estimates of the ocean carbon sink variability - First results of the Surface Ocean pCO2Mapping intercomparison (SOCOM). *Biogeosciences Discussions*, *12*(16), 14049–14104. https://doi.org/10.5194/bgd-12-14049-2015

Swart, S., Chang, N., Fauchereau, N., Joubert, W., Lucas, M., Mtshali, T., et al. (2012). Southern Ocean Seasonal Cycle Experiment 2012: Seasonal scale climate and carbon cycle links. *South African Journal of Science*, *108*(3–4), 3–5. https://doi.org/10.4102/sajs.v108i3/4.1089