Biogeosciences Discuss., https://doi.org/10.5194/bg-2017-215
Manuscript under review for journal Biogeosciences
Discussion started: 12 June 2017

Empirical methods for the estimation of Southern Ocean CO_2: Support Vector and Random Forest Regression

Luke Gregor[1,2], Schalk Kok[3], and Pedro M. S. Monteiro[1]

[1]Southern Ocean Carbon-Climate Observatory (SOCCO), CSIR, Cape Town, South Africa
[2]University of Cape Town, Department of Oceanography, Cape Town, South Africa
[3]University of Pretoria, Department of Mechanical and Aeronautical Engineering, Pretoria, South Africa

Correspondence to: Luke Gregor (lukegre@gmail.com)

Abstract. The Southern Ocean accounts for 40% of oceanic CO_2 uptake, but the estimates are bound by large uncertainties due to a paucity in observations. Gap filling empirical methods have been used to good effect to approximate $p\mathrm{CO}_2$ from satellite observable variables in other parts of the ocean, but many of these methods are not in agreement in the Southern Ocean. In this study we propose two additional methods that perform well in the Southern Ocean: Support Vector Regression (SVR) and Random Forest Regression (RFR). The methods are used to estimate $\Delta p\mathrm{CO}_2$ in the Southern Ocean, achieving similar results to the SOM-FFN method by Landschützer et al. (2014). The RFR as able to achieve better RMSE (12.26 µatm) compared the SVR (16.04 µatm) and SOM-FFN (12.97 µatm). To assess the efficacy of the methods and the limits of the training dataset (SOCAT v3), SVR and RFR are applied in a modelled environment. Again the RFR method outperformed the SVR by a substantial margin. However, both methods achieved higher out-of-sample than in-sample errors, indicating that the SOCAT v3 dataset is not yet fully representative of the Southern Ocean. The SVR was able to generalise better to the training dataset than the RFR with lower ratio between the out-of-sample and in-sample errors, but not enough to compensate for its poorer performance. The ensemble of the estimates show that interannual variability of the Southern Ocean CO_2 sink is dominated by the Polar Frontal Zone, while the Sub-Antarctic Zone is the dominant sink.

1 Introduction

The global oceans have played an important role in mitigating the effects of climate change by taking up 25% of anthropogenic CO_2 emissions annually (Khatiwala et al., 2013; Le Quéré et al., 2016). The Southern Ocean has played a disproportionate role in this uptake, accounting for 40% of the oceanic anthropogenic CO_2 uptake (Khatiwala et al., 2013; Frolicher et al., 2015). Yet, despite the region's importance, first order CO_2 flux estimates are bound by large uncertainties due to sparse observations in the Southern Ocean (Lenton et al., 2006; Monteiro et al., 2010; Lenton et al., 2012; Takahashi et al., 2012; Bakker et al., 2016). These uncertainties limit our capacity to resolve variability and trends of CO_2.

Viable alternative methods to estimate net CO_2 flux are atmospheric CO_2 inversions, ocean biogeochemical process models and empirical models (Rödenbeck et al., 2015). As shown by Le Quéré et al. (2007), atmospheric CO_2 inversions are useful tools to estimate the net CO_2 fluxes, but fail to offer further understanding with spatially integrated air-sea flux estimates (Fay and McKinley, 2014). Conversely, ocean biogeochemical process models are good tools for mechanistic understanding, but fail to represent seasonality of CO_2 fluxes in the Southern Ocean (Lenton et al., 2013; Mongwe et al., 2016). Empirical modelling offers an opportunity to bridge the gap between sparse data in the Southern Ocean and correct parameterisation of future earth systems models.

Empirical models maximise the utility of existing surface ocean CO_2 observations ($p$$CO_2$) by interpolating these with satellite proxy data. Access to in-situ $p$$CO_2$ data, via platforms such as SOCAT (Surface Ocean CO_2 Atlas), has been crucial to the success of empirical methods (Rödenbeck et al., 2015; Bakker et al., 2016). This, in conjunction with the increasing use of machine learning, has seen a proliferation in the number and diversity of methods in the literature. Rödenbeck et al. (2015) compared a suite of fourteen methods using a regional framework provided by Fay and McKinley (2014). The authors found that methods agreed in regions were data coverage was adequate, but for data sparse regions, such as the Southern Ocean, interannual CO_2 trends of various empirical methods were not coherent.

The primary reason for the varied results in Rödenbeck et al. (2015) is thought to be the way in which the algorithms deal with sparse data in the Southern Ocean. These methods were typically variants of multiple linear regression (MLR) or artificial neural networks (ANN), with regression being applied in regional windows or clusters based on climatologies of satellite measurable variables. The SOM-FFN approach by Landschützer et al. (2014) exemplifies the combination of non-linear clustering coupled with regression. In a later work, Landschützer et al. (2015) used the SOM-FFN approach along with several other methods to show that Southern Ocean CO_2 uptake strengthened after 2000. However, the lack of measurements in the Southern Ocean meant that these methods could not be effectively tested with an independent dataset (Rödenbeck et al., 2015).

In the early 2000s, the North Atlantic experienced similar data paucity. Friedrich and Oschlies (2009) approached this problem by using process model output to evaluate the efficacy of an artificial neural network as well as finding the optimal proxy variables for estimating $p$$CO_2$. This idealised environment was also used to estimate the effect of including/excluding certain proxy variables where it was found that filling remote sensing gaps in temperature and chlorophyll-a with climatology improved the estimates. In the intercomparison study by Rödenbeck et al. (2015) proxies typically include, but are not limited to: sea surface temperature (SST), chlorophyll-a (Chl-a), mixed layer depth (MLD) and sea surface salinity (SSS).

In this study, we introduce and compare two empirical methods new to this ocean CO_2 application: Support Vector Regression (SVR) and Random Forest Regression (RFR). SVR is a method based on the theory of statistical learning, making the method robust to over-fitting by statistically determining the complexity of a problem rather than a heuristic approach as required in setting up an ANNs hidden layer structure (Vapnik, 1999; Smola et al., 2004). RFR uses an ensemble of decision trees to create robust estimates, often without requiring data pre-processing making it an effective "off the shelf" method (Louppe, 2014).

We use SVR and RFR to estimate CO_2 fluxes in the Southern Ocean to try to better resolve the seasonal cycle from 1998 to 2014. These methods are trained with SOCAT v3 data collocated with satellite proxies. We compare these results with those of Landschützer et al. (2014). In the next part we aim to better understand the limitations of these methods within the framework of the SOCAT v3 data. SVR and RFR are implemented in a simulated environment with a realistic sampling strategy to assess

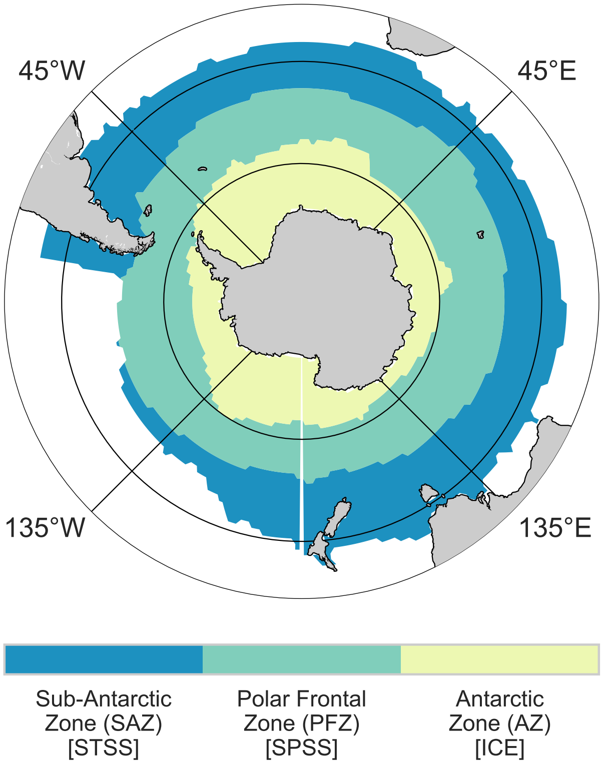
45°W
45°E
135°W
135°E
Sub-Antarctic Zone (SAZ) [STSS]
Polar Frontal Zone (PFZ) [SPSS]
Antarctic Zone (AZ) [ICE]

Figure 1. The three Southern Ocean biomes as defined by Fay and McKinley (2014). The common names for the biomes are shown in the key, with the abbreviations shown in the round brackets. The abbreviation in the square brackets show the abbreviations as given by Fay and McKinley (2014).

if there are biases to this sparse data. This approach allows us to test the impact of including various proxy variables as done by Friedrich and Oschlies (2009). Thereafter the methods are applied to observational data for actual estimates of $p\mathrm{CO}_2$.

2 Data and Methods

This study is presented in two parts. The first applies SVR and RFR to the SOCAT v3 dataset and compares these outputs with those of the SOM-FFN by Landschützer et al. (2014). These estimates will be referred to as the observational estimates. Here the domain is limited to the three Southern Ocean (SO) domains of Fay and McKinley (2014) that are shown in Figure 1. These biomes are used to assess the performance of each of the methods, as done in Rödenbeck et al. (2015). Fay and McKinley (2014) use a different nomenclature, which roughly corresponds to frontal zones. We rename the Sub-Tropical Seasonally Stratified biome (STSS) as the Sub-Antarctic Zone (SAZ); the Sub-Polar Seasonally Stratified biome (SPSS) becomes the Polar Frontal Zone (PFZ) and the ice biome (ICE) is the Antarctic Zone (AZ) (Mongwe et al., 2016).

Table 1. Information on data products used in this study. The temporal and spatial resolutions are for the raw data (before gridding). Dashes show that times are either not applicable or that the dataset is continually updated. Note that the start and end year show full years only. Links to download the data are given in the additional materials. The asterisk (*) indicates that variables are the output of a data assimilative model.

Group / Product	Variables	Date Range		Resolution		Reference
		Start	End	Time	Space	
SOCAT v3	fCO2sea	1970	2014	1 mon	1°	(Bakker et al., 2016)
CDIAC	xCO2atm	1970	2014	–	–	(CDIAC, 2016)
Globcolour	Chlorophyll	1998	–	1 day	0.25°	(Maritorena and Siegel, 2005)
GHRSST	Sea Surface Temperature	1981	–	1 day	0.25°	(Reynolds et al., 2007)
ECCO2 (cube92)	*Mixed Layer Depth *Salinity	1992	2015	1 day	0.25°	(Menemenlis et al., 2008)

The second part aims to better understand the limitations of these methods with the given dataset by implementing the methods to ocean biogeochemical model output. This will be referred to as the simulation experiment. Here the domain of the study is south of 34°S – the biomes Fay and McKinley (2014) are defined by oceanographic and biological parameters and would thus be different in the model.

2.1 Gridded Data

The data sources are shown in 1. These gridded data refer primarily to remotely sensed data, with the exception of MLD and SSS. These latter variables are output from ECCO$_2$, an assimilative model specific to the Southern Ocean. For the sake of brevity, these variables will be included under the description of "gridded observations".

All data are gridded to monthly x 1° using *iris* and *xarray* packages in Python (Hoyer et al., 2016; Met Office). Gridded $p\mathrm{CO}_2$ (SOCAT v3) is used to train the algorithms (Bakker et al., 2016). Surface station measurements (flask and tower) of atmospheric xCO$_2$ are interpolated to a regular grid using support vector regression (Masarie et al., 2014). Mean sea level pressure (NCEP2) is used in the conversion from xCO$_2$ to $p\mathrm{CO}_2$ (Kanamitsu et al., 2002).

Cloud coverage and low light at high latitudes during winter result in missing Chl-a data. Cloud gaps are filled with the climatology of Chl-a (from 1998 to 2014) and missing low light data are filled with a value of 0.1 ± 0.03 mg m^{-3} (uniformly distributed random noise).

2.2 Model Data

The prognostic coupled physics – biogeochemical model used in this study is a regional NEMO-PISCES configuration, BIOPERIANT05-GAA95b. This model is an updated version of PERIANT05 used by Dufour et al. (2012), where BIOPERIANT05-GAA95b includes biogeochemistry with PISCES-v2. The model has a peri-Antarctic domain with an open northern boundary at 30°S. The horizontal resolution of the configuration is 0.5° cos(latitude) with 46 vertical levels. The northern boundary is forced by a global 0.5° model, ORCA05 as presented in Biastoch et al. (2008). Output was saved as five-day averages. The simulation was run from 1992 to 2009. The data is resampled to 1.0° spatial resolution and monthly temporal resolution data to match observations.

2.3 Data transformation and derived variables

There are several transformations that are applied to data for both model output and gridded observations. The $\log_{10}$ transformations of MLD and filled chlorophyll (Chl-a_{clim}) are taken to return a normal distribution.

Several of the studies in Rödenbeck et al. (2015) included latitude, longitude and/or time as proxies of $\Delta p\mathrm{CO}_2$. However, many of the methods that are regional or cluster the data before regression did not include coordinates. In this study, we use a single large domain with no clustering or regional subsets. This then raises the question of whether including coordinates would improve estimates or not. Including the coordinates may create a model where the training location is too narrow.

Seasonality of the data is preserved by transforming the day of the year (j) and is included in both SVR and RFR analyses:

$$t = \begin{pmatrix} \cos\left(j \cdot \frac{2\pi}{365}\right) \\ \sin\left(j \cdot \frac{2\pi}{365}\right) \end{pmatrix} \tag{1}$$

Transformed coordinate vectors were passed to only SVR using n-vector transformations of latitude (λ) and longitude (μ) (Gade, 2010; Sasse et al., 2013), with n containing:

$$A, B, C = \begin{pmatrix} \sin(\lambda) \\ \sin(\mu) \cdot \cos(\lambda) \\ -\cos(\mu) \cdot \cos(\lambda) \end{pmatrix} \tag{2}$$

Co-located fCO_2 (y) and proxy data (X) were used to create training arrays (x). The final input for SVR were the following proxies (with 12 columns): $\log_1 0$(Chl-a_{clim}), SST, $f\mathrm{CO}_{2(atm)}$, ADT, $\log_{10}$(MLD), ICE, SSS, cos(j), sin(j) and n-vectors [A, B, C]. SVR requires each column of the proxies to be z-scored; *i.e.* normalized to the mean (μ) and standard deviation (σ) of each column ($\frac{x-\mu}{\sigma}$).

2.4 Empirical methods and implementation

Data is split randomly into a training and independent test dataset with a ratio $0.7:0.3$. The independent dataset is used to give a test error of the trained algorithm. The statistical learning package, *Scikit-Learn*, in Python is used for all regression and cross-validation methods (Pedregosa et al., 2011). The details on each cross-validation method are outlined in the subsections below.

2.4.1 Support vector regression

The formulation of SVR is such that the cost function minimizes the number of points on or outside the allowable error margins (ϵ) as shown in 2a. A few slack variables (ξ) are allowed, within the limits of a slack parameter (C), which is set by the user.

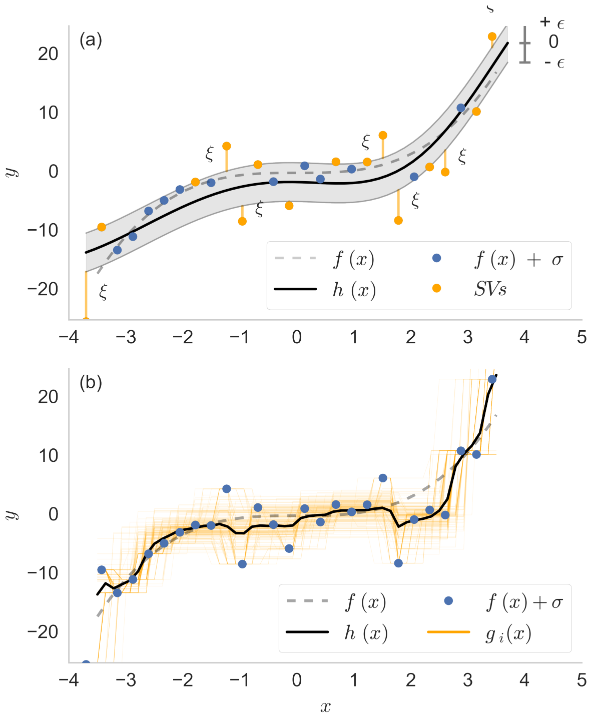

Figure 2. A simple example demonstrating the principle of (a) support vector regression and (b) random forest regression. The dashed grey line is the true function $f(x) = 0.4x^3$ with the blue dots representing a random sample taken from this function $f(x)+\sigma$, where σ is normally distributed noise. The black line in each figure, $h(x)$, show the estimate of the true function. The orange dots in (a) show the samples from the random subset chosen as support vectors from which $h(x)$ is estimated. The orange lines in (b) show 200 decision tree estimates, $g_i(x)$, which are averaged to create the ensemble, $h(x)$.

The points on or outside these margins are the support vectors and are used to construct the hypothesis function, $h(x)$. This elegant approach is made versatile by mapping $\mathbf{X}$ onto a higher dimensional feature space using an interchangeable kernel. In this study we used a Gaussian kernel (or radial basis function – RBF), which allows for potentially infinite complexity, determined by the number of support vectors (Vapnik, 1999). The assignment of the number of support vectors is analogous to defining the architecture of an ANN. The RBF kernel introduces an additional hyper-parameter (γ) that defines the width of the Gaussian. Selection of the SVR hyper-parameters (ϵ, C, γ) is done using a two-stage coarse–fine grid search approach using K-fold cross validation with $k = 8$.

2.4.2 Random Forest Regression

A random forest (RF) is an ensemble of decision trees, which means that the average estimate of n trees is taken (Breiman, 2001) (Figure 2b). Random forests reduce the high variance of decision trees by bagging (bootstrap aggregating) in which the training dataset is sampled with replacement resulting in a $\sim 63\%$ chance of being chosen at least once for a particular tree

Table 2. The scores for each of the empirical methods trained with SOCAT v3 data. The domain for these scores is the Southern Ocean as defined by Fay and McKinley (2014).

METHOD	RMSE	MAE	r2
SVR	16.04	10.55	0.6
RFR	12.26	7.43	0.77
SOM-FFN	12.97	8.56	0.7

(Louppe, 2014). A random forest typically performs better when number of trees (t) is large, but increasing the number of trees has diminishing returns in terms of performance vs. computation. Additional robustness is given to RFs by randomizing and/or limiting the number of variables (m) given to the nodes in each tree when splitting the data (hence random) (Louppe, 2014). The complexity of a RF can be adjusted by limiting the minimum number of leaves at a terminal branch (l), where a fully-grown tree would allow l to be one; tree depth can also be limited to reduce the complexity and has a similar effect to limiting l.

A useful feature of bagging is that it intrinsically provides a cross-validation dataset (a.k.a. out-of-bag samples) that is not part of the training procedure (for a specific set of trees). The advantage of this approach over K-fold cross-validation is that the full dataset can be used in the training procedure, as opposed to splitting the dataset for cross-validation. The out-of-bag error is used to select the hyper-parameters (t, m, l) for the RF.

2.5 CO_2 fluxes

Air-sea CO_2 fluxes are calculated from:

$$F\mathrm{CO_2} = K_0 \cdot k_w \cdot \Delta pCO_2 \cdot (1 - [\mathrm{ice}]) \tag{3}$$

The gas transfer velocity (k_w) is calculated using a quadratic dependency of wind speed with the coefficients of Nightingale et al. (2000). Wind speed is calculated from the u and v vectors of CCMP v2 (Atlas et al., 2011). Coefficients from Weiss (1974) are used to calculate K0 and $\Delta p\mathrm{CO_2}$ is estimated by the empirical models. The effect of sea-ice cover on CO_2 fluxes is treated linearly; the fraction of sea ice cover ([ice]) is converted to fraction of open water by subtracting one as shown in Equation (3).

These results are analyzed regionally with the three Southern Ocean biomes defined by Fay and McKinley (2014) (Figure 1). We compare our estimates of CO_2 fluxes with those of Landschützer et al. (2014) who used a two-step neural network method abbreviated to SOM-FFN (self-organizing map – feed forward neural network).

Biogeosciences Discuss., https://doi.org/10.5194/bg-2017-215
Manuscript under review for journal Biogeosciences
Discussion started: 12 June 2017

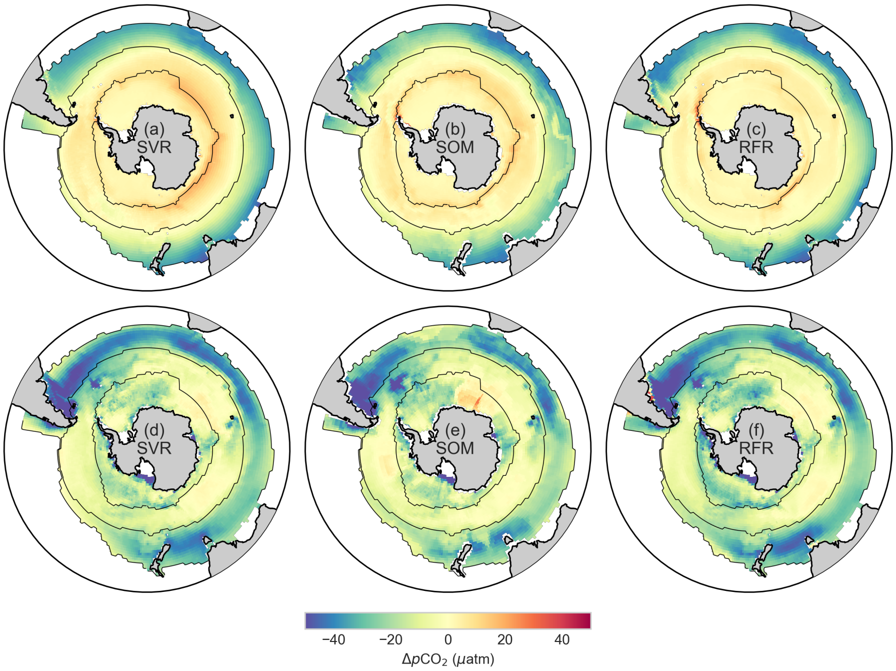

Figure 3. Seasonal averages for $\Delta p\mathrm{CO}_2$ from 1998 to 2014 for SVR, SOM and FRF. The mean winter (JJA) $\Delta p\mathrm{CO}_2$ is shown in the top row (a, b, c) and the mean summer (DJF) $\Delta p\mathrm{CO}_2$ is shown in the bottom row (d, e, f). The thin black lines denote the SAZ, PFZ and AZ from outside inward. Note that the $\Delta p\mathrm{CO}_2$ has been normalized to sea ice cover where $\Delta p\mathrm{CO}_2$ is multiplied by $(1 - [\mathrm{ice}])$.

3 Results

3.1 Observational CO_2 data results

The RMSE, MAE and r^2 scores for each method applied to the data shown in Table 1 are shown in Table 2. The RFR score is taken from the out-of-bag error, while the independent test set scores are used for SVR and SOM-FFN. RFR achieves the best scores, with an RMSE of 12.26 atm. This is slightly better than the RMSE of the SOM-FFN (12.97 atm). The SVR performs poorly with an RMSE of 16.04 atm.

The seasonal averages (winter = JJA, summer = DJF) for $\Delta p\mathrm{CO}_2$ estimated by SVR, SOM-FFN and RFR for the entire Southern Ocean region are shown in Figure 3. These show that there is, in general, good agreement in the spatial distribution between the methods. In winter (Figure 3a-c), there is outgassing south of the Polar Front as previously found (Metzl et al., 2006). This is true also for the AZ, but sea ice cover suppresses the effect. The estimates of $\Delta p\mathrm{CO}_2$ have thus been scaled to sea ice concentration ($\Delta p\mathrm{CO}_2 \times (1 - [ice])$) as also done for fluxes in Equation 3.

To the north of the Polar Front, in the SAZ, the ocean is a sink of CO_2 (Figure 3). The surface $\Delta p\mathrm{CO}_2$ is more zonally symmetric in winter when compared to summer. The zonal asymmetry in summer is driven, in part, by a strong reduction of $\Delta p\mathrm{CO}_2$ driven by biological production the Southern Ocean (Metzl et al., 2006; Lenton et al., 2012). There are three regions in

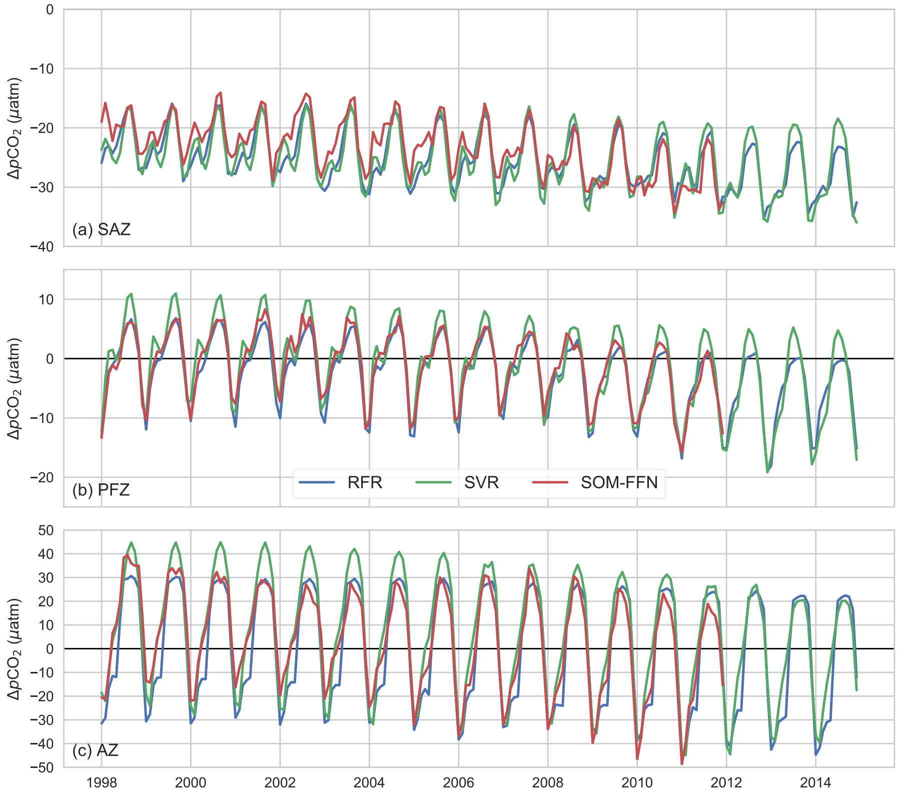

Figure 4. Time-series of $\Delta p\mathrm{CO}_2$ estimates for the three Southern Ocean biomes as defined by Fay and McKinley (2014): SAZ, PFZ and MIZ. The y-axis gridlines represent the same scale for figures (a) through (c). The SOM-FFN estimates are only available until 2011 as it is trained with SOCAT v2, while the SVR and RFR are trained with SOCAT v3. Note that $\Delta p\mathrm{CO}_2$ is not normalised to sea ice concentration in this figure.

the SAZ where $\Delta p\mathrm{CO}_2$ reduction is strongest and consistent between methods: east of South America (Malvinas Confluence), southeast of Africa (Agulhas retroflection) and between Australia and New Zealand (Tasman Sea). The reduction of $\Delta p\mathrm{CO}_2$ in the PFZ is strongest in the Atlantic sector downstream of the South Sandwich and South Georgia Islands and in the Indian sector downstream of the Kerguelen Plateau (Figure 3d-f). In both cases, SAZ and PFZ, these regions are consistent with regions of high biomass (Thomalla et al., 2011; Carranza and Gille, 2015).

There are clear differences in the spatial variability between methods. The most marked difference in winter is that the SVR estimates the PFZ as a stronger source of CO_2 to the atmosphere compared to the SOM-FFN and RFR approaches (Figure 3a-c). In summer, the largest difference occurs in the eastern Atlantic sector of the SAZ where the SOM-FFN estimates higher $\Delta p\mathrm{CO}_2$ compared to SVR and RFR (Figure 3d-f).

The time-series (1998 – 2014) for $\Delta p\mathrm{CO}_2$ for each of the Southern Ocean biomes as defined by Fay and McKinley (2014) are shown in Figure 4. In general there is good coherence between the three methods with agreement in the timing of the

Table 3. The performance metrics of SVR and RFR in estimating $\Delta p\mathrm{CO}_2$ in a model simulation (BIOPERIANT05) using SOCATv3 cruise tracks as "sampling" locations. Both the in- and out-of-sample errors are shown. This is done with and without coordinate proxies. The metrics are: RMSE = root mean squared error, MAE = mean absolute error, r-squared.

ERROR	MODEL	INPUT	RMSE	MAE	r^2
IN	SVR	No coords	8.2	5.98	0.87
		Coords	6.26	4.75	0.92
	RFR	No coords	6.27	3.78	0.93
		Coords	**4.7**	**2.72**	**0.95**
OUT	SVR	No coords	8.7	6.51	0.67
		Coords	7.89	5.99	0.72
	RFR	No coords	7.87	5.58	0.73
		Coords	**6.33**	**4.5**	**0.82**

seasonal cycle and the strengthening sink over the period 2002 – 2012 (Landschützer et al., 2015). However, the differences pointed out in the seasonally averaged maps are also present in the time-series representation.

In the SAZ, the largest difference is between the SOM-FFN and the other two methods. This is limited to the end of summer in the first half of the time-series. Comparatively, estimates of winter $\Delta p\mathrm{CO}_2$ agree, with the exception of the last four years when SVR winter estimates increase relative to RFR. The overestimation of winter $\Delta p\mathrm{CO}_2$ by the SVR is also observed in the PFZ, but for the majority of the time series. The SAZ and PFZ also show variability in the magnitude of a seasonal shoulder in late summer, where increasing $\Delta p\mathrm{CO}_2$ is briefly delayed by a short sharp decrease resulting in a saw-tooth pattern. This effect is the strongest for the SVR and weakest for the RFR. The seasonal amplitudes of $\Delta p\mathrm{CO}_2$ in the AZ are far larger than for both the SAZ and PFZ. However, this large differential may not be realized as an outgassing CO_2 flux, particularly in winter, due to ice cover.

3.2 Simulation experiment results

The results from the simulation experiments are summarized in Table 3. RFR consistently performs better than the SVR approach. This is consistent for both in- and out-of-sample errors, where in-sample errors represent only the SOCAT dataset and the out-of-sample errors represent the entire domain. The in-sample error is representative of the error that would be reported in the application of the data to observed data. Note that the in-sample error for the RFR methods is estimated using the out-of-bag errors. The out-of-sample error is considerably larger for each respective method, indicating that reported error estimates for the Southern Ocean could be underestimated. These in sample and out of sample errors are illustrated in Figure 5(a) and (b) respectively.

The results also show that including time and space coordinates as proxies of $\Delta p\mathrm{CO}_2$ improves the estimates. This is shown in Figure 4 where the estimates trained with coordinates (dashed-lines) achieve lower RMSE scores relative to the estimates trained without coordinates (solid lines). Importantly, this is true for both in- and out-of-sample errors. The RMSE of the RFR

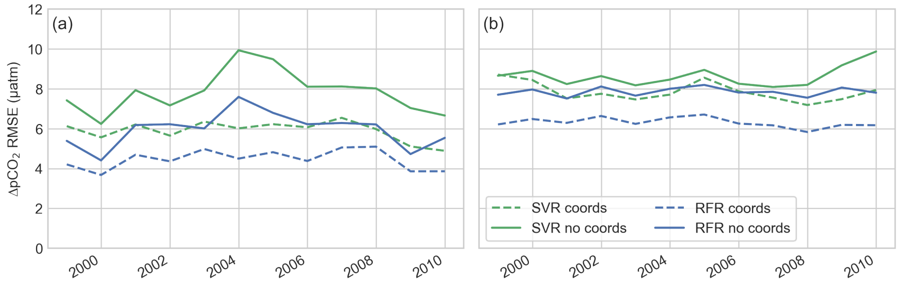

Figure 5. (a) In sample errors and (b) out of sample errors. Two SVR models are shown, one with the same variables as the SVR and another without space and time coordinates. The RFR outperforms the SVR, but the RFR without coordinates does not perform as well as the SVR. Clearly, adding the coordinates improves estimates.

without coordinates is the same as the SVR with the inclusion of coordinates, again highlighting the superior accuracy of the RFR. These results suggest that estimates would benefit from the inclusion of coordinates.

4 Discussion

4.1 Methodological differences in observational estimates

The differences observed in the estimates of $\Delta p\mathrm{CO}_2$ are driven by differences in the algorithms as well as the implementation of these methods. One of the most marked differences is the weaker sink estimated by the SOM-FFN method in the SAZ (Figure 4). This difference can be traced to the eastern Atlantic SAZ (Figure 3e), where the SOM-FFN has higher estimates of $\Delta p\mathrm{CO}_2$. The lack of this feature in the SVR and RFR estimates suggests that this is a function of the initial clustering step in the SOM-FFN. This clustering step separates the global $p\mathrm{CO}_2$ dataset into distinct clusters defined by oceanographic and biological properties rather than region (Landschützer et al., 2014). Thus a cluster in the subtropical South Atlantic could be grouped to the same cluster as the tropical South Atlantic. The SOM-FFN is implemented in a global domain, meaning that the algorithm could be mapping the relationship between CO_2 and its proxies from more tropical waters.

Another difference is the tendency for the SVR to overestimate $\Delta p\mathrm{CO}_2$ compared to the RFR and SOM-FFN approaches. We attribute this to the SVR's sensitivity to outliers. In context of the SOCAT v3 dataset, the algorithm may treat the sparse winter data as outliers. This means that the higher estimates of $\Delta p\mathrm{CO}_2$ in winter could be extrapolated, leading to the relatively elevated winter estimates.

Conversely the RFR estimates of $\Delta p\mathrm{CO}_2$ are often lower than the SOM-FFN and SVR estimates. This may be due to the method's resilience against outliers. This is primarily due to the bagging approach, where individual decision trees are trained with a subset of data that is sampled with replacement, thus the chance of sampling sparse winter data is lower. Moreover, the estimates will be more conservative due to the methods inability to estimate beyond the training data (as shown in Figure 2b).

The differences between the methods shown in Figure 4 could be a good case for an ensemble approach, where the strengths of one model compensate for the weakness of another.

4.2 Performance and caveats of methods in simulation experiment

In both the simulation and observations, the RFR achieved the lowest RMSE for in- and out-of-sample scores. We postulate that RFR is able to outperform both SVR and SOM-FFN due its ability to model data that contains a higher degree of non-linearity. The high degree of non-linearity stems from the discrete decision boundaries associated with decision trees, the building blocks of RFR. Such non-linearity increases the risk of over fitting to the noise specific to the training dataset. However, over fitting is minimized by using a large number of trees in a random forest, which, combined with bagging, results in good generalization (Louppe, 2014). However, if the training dataset is not representative of the entire domain, generalization techniques such as bagging will not be able to reduce the over fitting.

In contrast to RFR, the non-linearity of SVR is fixed by the selection of a constant width of the Gaussian kernel for the entire domain, thus applying the assumption of constant variability to the domain (both temporally and spatially). This can be overcome by clustering regions of similar variability, as was done in the two-step SOM-FFN approach by Landschützer et al. (2015). In fact the similarity between FFN and SVR (Vapnik, 1999), could lead to similar results if a clustering technique was applied to the latter. However, this introduces the additional complexity of dealing with $\Delta p\mathrm{CO}_2$ discontinuities of cluster boundaries.

The non-linearity of the RFR allows the implementation without coordinates to marginally outperform the SVR implemented with coordinates (Table 3). Though the inclusion of coordinates improves the RFR and, to a lesser extent, SVR error estimates. This indicates that SST, Chl-a, MLD and SSS are able to represent CO_2 relatively well, but the relationship between these variables changes by region and period. The inclusion of coordinates decomposes the problem to specific regions or periods as clustering approaches achieve. This implies that the available proxy variables are not able to capture the variability of $\Delta p\mathrm{CO}_2$.

For example, there may be differences in the relationship between CO_2 and SSS in the western Atlantic compared to the eastern Indian sector. A prior clustering step, or the addition of coordinate proxies would account for these differences.

While the RFR method achieved the lowest RMSE scores, it is not without limitations. The RFR method, unlike SVR, is not able to extrapolate estimates of CO_2 beyond the bounds of the observations (Louppe, 2014). This is due to the structure of decision trees, where estimates are based purely input and cannot extrapolate beyond the minimum and maximum observed $\Delta p\mathrm{CO}_2$. This means that the RFR estimates are more conservative than SVR and SOM-FFN, which are able to extrapolate. Moreover, the relative paucity of winter data combined with the bagging approach exacerbates the relative underestimates of winter $\Delta p\mathrm{CO}_2$. In bagging sampling with replacement would result in far more frequent selection of summer data than winter data. More winter data is needed to improve this imbalance.

4.3 Limitations of SOCAT v3

A key finding of the simulation experiment is that out-of-sample RMSEs are larger than in-sample RMSEs, implying that error estimates for observational $\Delta p\mathrm{CO}_2$ would also be underestimated. This is due to the paucity of measurements in the Southern

Ocean, meaning that SOCAT v3 is not yet representative of the full Southern Ocean domain, despite significant increases in the number of samples (Bakker et al., 2016). Tuning the algorithms to generalize to the dataset is crucial to avoid over fitting to the noise of the training subset. However, in this case, more strategic measurements are needed to make SOCAT more representative of the Southern Ocean.

The ratio of in-sample and out-of-sample errors for SVR and RFR can be used to gain insight about the ability of the respective methods to generalize to the training dataset. This ratio ($\frac{E_{out}}{E_{in}}$) is 1.26 for SVR and 1.35 for RFR, showing the SVR has the ability to generalize better to the training dataset, but this needs to be viewed in context of the methods' RMSE scores. These ratios can be applied to the in-sample errors in the observational estimates of $\Delta p\text{CO}_2$. This results in a theoretical out-of-sample RMSE of 20.21 µatm for SVR and 16.76 µatm for RFR for the estimates calculated from SOCAT v3. There may be variations of RFR, such as Extremely Randomized Trees (Geurts et al., 2006), that are perhaps better at generalizing to a sparse dataset, but investigating this requires additional work.

In summary, the correct implementation of machine learning algorithms should minimize over fitting to the training dataset. However, in the case of the Southern Ocean sector of the SOCAT v3 dataset, the data is not yet representative of CO_2 for the entire domain. This means that there will be biases in estimates that generalization techniques are not able to resolve for which more representative data is required.

4.4 Trends of ensemble estimates

While methodological differences exist, the trends of $\Delta p\text{CO}_2$ and air-sea CO_2 flux (as shown in Figure 4) are mostly in agreement. Moreover, the algorithmic differences that each method exhibit lend themselves to an ensemble approach. This approach allows for more robust estimates of $p\text{CO}_2$ and air-sea CO_2 fluxes (FCO_2). For instance, the conservative estimates of the RFR could be offset by the relative overestimation by the SVR.

The trends of the ensemble of FCO_2 for the SAZ, PFZ and AZ are shown in Figure 6. These are in agreement with the trends explained in Landschützer et al. (2015): a slight weakening of the sink from 1998 into the early 2000s (as also found by Le Quéré et al. 2007) followed by a reinvigoration of CO_2 uptake through to the end of the time series in 2014. The PFZ dominates this interannual variability of FCO_2 with a strong reduction in outgassing between 2002 and 2010. The relatively large seasonal amplitude of $\Delta p\text{CO}_2$ observed in the AZ is damped by weaker winds and winter ice cover resulting in relatively weak fluxes (compared the PFZ). Compared to the PFZ and AZ, the SAZ is a strong and consistent sink (with mean uptake of -0.042, -0.025 and -0.55 PgC yr^{-1} respectively) that strengthens slightly throughout the period, but the seasonal signal and amplitude are dominated by intra-seasonal modes as was found in observations (Monteiro et al., 2015). To understand the driving mechanisms behind these trends, an in depth study needs to be undertaken.

5 Conclusions

In this study two empirical methods (SVR and RFR) are presented as alternative (and perhaps complimentary) methods to estimating $\Delta p\text{CO}_2$ from satellite proxies by tuning the methods to best predict ship-based measurements. These algorithms are

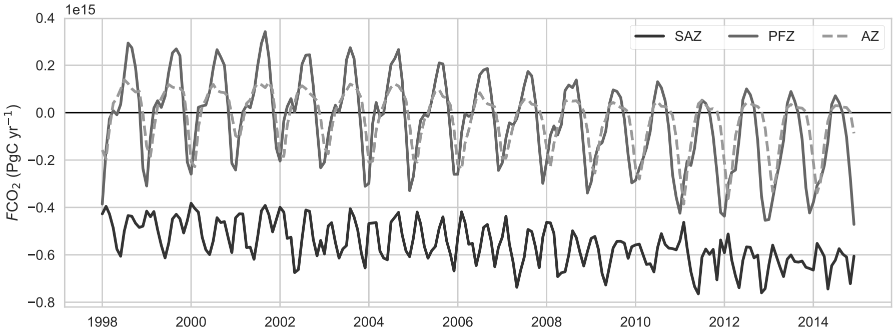

Figure 6. Ensemble air-sea CO_2 fluxes for each region as defined by Fay and McKinley (2014). Flux is calculated as shown in Equation 3. The SAZ = Sub-Antarctic Zone, PFZ = Polar Frontal Zone, and AZ = Antarctic Zone.

established in other fields, but have not been applied for the estimation of surface ocean $\Delta p\mathrm{CO}_2$ to overcome the limitations of the existing paucity of in situ observations, particularly in the Southern Ocean. The seasonal bias in observations is particularly evident during winter.

Both methods, with coordinate proxies, were applied to observational data and compared with the SOM-FFN method by Landschützer et al. (2014). There is good agreement between the trends of each of the methods, though an absolute assessment of the results to an independent dataset was not possible due to the paucity of data. Methodological differences were apparent over and above the dominant trend. The SVR is more likely to produce overestimates of winter $\Delta p\mathrm{CO}_2$ compared to the other two approaches. Conversely, the RFR produced lower estimates of CO_2 in winter. The ensemble fluxes showed that the SAZ region as responsible for the majority of CO_2 uptake over the period (1998 – 2014), while the PFZ dominated interannual variability. Ice cover in the AZ muted the large seasonal amplitude of $\Delta p\mathrm{CO}_2$.

To test the efficacy of these methods, they were first applied in an idealized model environment that simulates the distribution of the current ship based measurements of CO_2, that is the SOCAT v3 dataset. The results showed that RFR is better able to estimate $\Delta p\mathrm{CO}_2$ from the SOCAT v3 data. The experiment also confirmed that both SVR and RFR estimates are improved by including transformations of time and space coordinates as proxies of CO_2. It is shown that the SOCAT v3 dataset is not yet completely representative of the Southern Ocean. The in-sample error estimates were smaller than the out-of-sample estimates, but this varied according to each method's ability to generalize to the data. This shows that reported errors of empirical $\Delta p\mathrm{CO}_2$ estimates in the Southern Ocean are likely underestimated. More representative data will thus have to be collected to reduce the uncertainty of the mean annual flux to the < 10% threshold (Lenton et al., 2006). This may already be an achievable goal with biogeochemical Argo floats able to estimate $p\mathrm{CO}_2$ from pH sensors (Williams et al., 2017).

Data availability. Data will be hosted at ftp://anonymous@socco.chpc.ac.za/Gregor2017_JAMES

Acknowledgements. This work is part of a PhD funded by the ACCESS program. The authors would like to thank Marina Lévy for use of the BIOPERIANT05-GAA95b model data. This work was partly enabled by the Centre for High Performance Computing (CSIR).

The Surface Ocean CO2 Atlas (SOCAT) is an international effort, endorsed by the International Ocean Carbon Coordination Project (IOCCP), the Surface Ocean Lower Atmosphere Study (SOLAS) and the Integrated Marine Biogeochemistry and Ecosystem Research program (IMBER), to deliver a uniformly quality-controlled surface ocean CO2 database. The many researchers and funding agencies responsible for the collection of data and quality control are thanked for their contributions to SOCAT.

References

Atlas, R., Hoffman, R. N., Ardizzone, J., Leidner, S. M., Jusem, J. C., Smith, D. K., and Gombos, D.: A Cross-calibrated, Multiplatform Ocean Surface Wind Velocity Product for Meteorological and Oceanographic Applications, Bulletin of the American Meteorological Society, 92, 157–174, doi:10.1175/2010BAMS2946.1, http://journals.ametsoc.org/doi/abs/10.1175/2010BAMS2946.1, 2011.

Bakker, D. C. E., Pfeil, B., Landa, C. S., Metzl, N., O'Brian, K. M., Olsen, A., Smith, K., Cosca, C., Harasawa, S., Jones, S. D., Nakaoka, S.-i., 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., Munro, D. R., Murata, A., Newberger, T., Omar, A. M., Ono, T., Paterson, K., Pearce, D., Pierrot, D., Robbins, L. L., Saito, S., Salisbury, J., Schlitzer, R., Schneider, B., Schweitzer, R., Sieger, R., Skjelvan, I., Sullivan, K. F., Sutherland, S. C., Sutton, A. J., Tadokoro, K., Telszewski, M., Tuma, M., Van Heuven, S. M. A. C., Vandemark, D., Ward, B., Watson, A. J., and Xu, S.: A multi-decade record of high-quality fCO2 data in version 3 of the Surface Ocean CO2 Atlas (SOCAT), Earth System Science Data Discussions, pp. 1–55, doi:10.5194/essd-2016-15, http://www.earth-syst-sci-data-discuss.net/essd-2016-15/, 2016.

Biastoch, A., Böning, C. W., Getzlaff, J., Molines, J.-M., and Madec, G.: Causes of Interannual–Decadal Variability in the Meridional Overturning Circulation of the Midlatitude North Atlantic Ocean, Journal of Climate, 21, 6599–6615, doi:10.1175/2008JCLI2404.1, http://journals.ametsoc.org/doi/abs/10.1175/2008JCLI2404.1, 2008.

Breiman, L.: Random forests, Machine Learning, 45, 5–32, doi:10.1023/A:1010933404324, 2001.

Carranza, M. M. and Gille, S. T.: Southern Ocean wind-driven entrainment enhances satellite chlorophyll-a through the summer, Journal of Geophysical Research C: Oceans, 120, 304–323, doi:10.1002/2014JC010203, 2015.

CDIAC: Multi-laboratory compilation of atmospheric carbon dioxide data for the period 1968-2015, doi:10.15138/G33W2G, https://doi.org/10.15138/G33W2G, 2016.

Dufour, C. O., Sommer, L. L., Zika, J. D., Gehlen, M., Orr, J. C., Mathiot, P., and Barnier, B.: Standing and transient eddies in the response of the Southern Ocean meridional overturning to the Southern annular mode, Journal of Climate, 25, 6958–6974, doi:10.1175/JCLI-D-11-00309.1, 2012.

Fay, A. R. and McKinley, G. A.: Global open-ocean biomes : mean and temporal variability, Earth System Science Data, 6, 273–284, doi:10.1594/PANGAEA.828650, 2014.

Friedrich, T. and Oschlies, A.: Neural network-based estimates of North Atlantic surface pCO 2 from satellite data: A methodological study, Journal of Geophysical Research, 114, C03 020, doi:10.1029/2007JC004646, http://doi.wiley.com/10.1029/2007JC004646, 2009.

Frolicher, T. L., Sarmiento, J. L., Paynter, D. J., Dunne, J. P., Krasting, J. P., and Winton, M.: Dominance of the Southern Ocean in anthropogenic carbon and heat uptake in CMIP5 models, Journal of Climate, 28, 862–886, doi:10.1175/JCLI-D-14-00117.1, 2015.

Gade, K.: A Non-singular Horizontal Position Representation, Journal of Navigation, 63, 395–417, doi:10.1017/S0373463309990415, 2010.

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

Hoyer, S., Fitzgerald, C., Hamman, J., and Others: xarray: v0.8.0, doi:10.5281/zenodo.59499, http://dx.doi.org/10.5281/zenodo.59499, 2016.

Kanamitsu, M., Ebisuzaki, W., Woollen, J., Yang, S. K., Hnilo, J. J., Fiorino, M., and Potter, G. L.: NCEP-DOE AMIP-II reanalysis (R-2), Bulletin of the American Meteorological Society, 83, 1631–1643+1559, doi:10.1175/BAMS-83-11-1631, 2002.

Khatiwala, S., Tanhua, T., Mikaloff Fletcher, S. E., Gerber, M., Doney, S. C., Graven, H. D., Gruber, N., McKinley, G. A., Murata, A., RÍos, A. F., and Sabine, C. L.: Global ocean storage of anthropogenic carbon, Biogeosciences, 10, 2169–2191, doi:10.5194/bg-10-2169-2013, 2013.

Landschützer, P., Gruber, N., Bakker, D. C. E., and Schuster, U.: Recent variability of the global ocean carbon sink, Global and Planetary Change, pp. 927–949, doi:10.1002/2014GB004853.Received, http://onlinelibrary.wiley.com/doi/10.1002/2014GB004853/full, 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., Tilbrook, B., and Wanninkhof, R.: The reinvigoration of the Southern Ocean carbon sink, Science, 349, 1221–1224, doi:10.1126/science.aab2620, http://www.sciencemag.org/cgi/doi/10.1126/science.aab2620, 2015.

Le Quéré, C., Rödenbeck, C., Buitenhuis, E. T., Conway, T. J., Langenfelds, R., Gomez, A., Labuschagne, C., Ramonet, M., Nakazawa, T., Metzl, N., Gillett, N. P., and Heimann, M.: Saturation of the Southern Ocean C02 Sink due to recent climate change., Science, 316, 1735–1738, doi:10.1126/science.1136188, http://www.ncbi.nlm.nih.gov/pubmed/17510327, 2007.

Le Quéré, C., Andrew, R. M., Canadell, J. G., Sitch, S., Ivar Korsbakken, J., Peters, G. P., Manning, A. C., Boden, T. A., Tans, P. P., Houghton, R. A., Keeling, R. F., Alin, S., Andrews, O. D., Anthoni, P., Barbero, L., Bopp, L., Chevallier, F., Chini, L. P., Ciais, P., Currie, K., Delire, C., Doney, S. C., Friedlingstein, P., Gkritzalis, T., Harris, I., Hauck, J., Haverd, V., Hoppema, M., Klein Goldewijk, K., Jain, A. K., Kato, E., Körtzinger, A., Landschützer, P., Lefèvre, N., Lenton, A., Lienert, S., Lombardozzi, D., Melton, J. R., Metzl, N., Millero, F. J., Monteiro, P. M., Munro, D. R., Nabel, J. E. M. S., Nakaoka, S.-i., O'Brien, K., Olsen, A., Omar, A. M., Ono, T., Pierrot, D., Poulter, B., Rödenbeck, C., Salisbury, J., Schuster, U., Schwinger, J., Séférian, R., Skjelvan, I., Stocker, B. D., Sutton, A. J., Takahashi, T., Tian, H., Tilbrook, B., Van Der Laan-Luijkx, I. T., Van Der Werf, G. R., Viovy, N., Walker, A. P., Wiltshire, A. J., and Zaehle, S.: Global Carbon Budget 2016, Earth System Science Data, 8, 605–649, doi:10.5194/essd-8-605-2016, 2016.

Lenton, A., Matear, R. J., and Tilbrook, B.: Design of an observational strategy for quantifying the Southern Ocean uptake of CO2, Global Biogeochemical Cycles, 20, GB4010, doi:10.1029/2005GB002620, 2006.

Lenton, A., Metzl, N., Takahashi, T., Kuchinke, M., Matear, R. J., Roy, T., Sutherland, S. C., Sweeney, C., and Tilbrook, B.: The observed evolution of oceanic pCO 2 and its drivers over the last two decades, Global Biogeochemical Cycles, 26, n/a–n/a, doi:10.1029/2011GB004095, http://doi.wiley.com/10.1029/2011GB004095, 2012.

Lenton, A., Tilbrook, B., Law, R. M., Bakker, D. C. E., Doney, S. C., Gruber, N., Ishii, M., Hoppema, M., Lovenduski, N. S., Matear, R. J., McNeil, B. I., Metzl, N., Fletcher, S. E. M., Monteiro, P. M., Rödenbeck, C., Sweeney, C., and Takahashi, T.: Sea-air CO2 fluxes in the Southern Ocean for the period 1990-2009, Biogeosciences, 10, 4037–4054, doi:10.5194/bg-10-4037-2013, 2013.

Louppe, G.: Understanding Random Forests, Phd, University of Liege, doi:10.13140/2.1.1570.5928, 2014.

Maritorena, S. and Siegel, D. A.: Consistent merging of satellite ocean color data sets using a bio-optical model, Remote Sensing of Environment, 94, 429–440, doi:10.1016/j.rse.2004.08.014, 2005.

Masarie, K. A., Peters, W., Jacobson, A. R., and Tans, P. P.: ObsPack: A framework for the preparation, delivery, and attribution of atmospheric greenhouse gas measurements, Earth System Science Data, 6, 375–384, doi:10.5194/essd-6-375-2014, 2014.

Menemenlis, D., Campin, J.-m., Heimbach, P., Hill, C., Lee, T., Nguyen, A., Schodlok, M., and Zhang, H.: ECCO2 : High Resolution Global Ocean and Sea Ice Data Synthesis, Mercator Ocean Quarterly Newsletter, 31, 13–21, 2008.

Met Office: Iris: A Python library for analysing and visualising meteorological and oceanographic data sets, http://scitools.org.uk/.

Metzl, N., Brunet, C., Jabaud-Jan, A., Poisson, A., and Schauer, B.: Summer and winter air–sea CO2 fluxes in the Southern Ocean, Deep Sea Research Part I: Oceanographic Research Papers, 53, 1548–1563, doi:10.1016/j.dsr.2006.07.006, http://linkinghub.elsevier.com/retrieve/pii/S0967063706001944, 2006.

Mongwe, N. P., Chang, N., and Monteiro, P. M.: The seasonal cycle as a mode to diagnose biases in modelled CO 2 fluxes in the Southern Ocean, Ocean Modelling, 106, 90–103, doi:10.1016/j.ocemod.2016.09.006, www.elsevier.com/locate/ocemod, 2016.

Monteiro, P. M., Schuster, U., Hood, M., Lenton, A., Metzl, N., Olsen, A., Rodgers, K. B., Sabine, C. L., Takahashi, T., Tilbrook, B., Yoder, J., Wanninkhof, R., and Watson, A. J.: A Global Sea Surface Carbon Observing System: Assessment of Changing Sea Surface CO2 and Air-Sea CO2 Fluxes, Proceedings of OceanObs'09: Sustained Ocean Observations and Information for Society, pp. 702–714, doi:10.5270/OceanObs09.cwp.64, http://www.oceanobs09.net/proceedings/cwp/cwp64, 2010.

Monteiro, P. M., Gregor, L., Lévy, M., Maenner, S., Sabine, C. L., and Swart, S.: Intraseasonal variability linked to sampling alias in air-sea CO <sub>2</sub> fluxes in the Southern Ocean, Geophysical Research Letters, pp. n/a–n/a, doi:10.1002/2015GL066009, http://doi.wiley.com/10.1002/2015GL066009, 2015.

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

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, C., Thirion, B., Grisel, O., Blondel, M., Prettenhoffer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., and Cournapeau, D.: Scikit-learn: Machine learning in Python, Journal of Machine Learning Research, 12, 2825–2830, doi:10.1007/s13398-014-0173-7.2, http://dl.acm.org/citation.cfm?id=2078195, 2011.

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

Rödenbeck, C., Bakker, D. C. E., Gruber, N., Iida, Y., Jacobson, A. R., Jones, S. D., Landschützer, P., Metzl, N., Nakaoka, S.-i., 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 \emph{p}CO$_2$ Mapping intercomparison (SOCOM), Biogeosciences, 12, 7251–7278, doi:10.5194/bg-12-7251-2015, http://www.biogeosciences.net/12/7251/2015/, 2015.

Sasse, T. P., McNeil, B. I., and Abramowitz, G.: A novel method for diagnosing seasonal to inter-annual surface ocean carbon dynamics from bottle data using neural networks, Biogeosciences, 10, 4319–4340, doi:10.5194/bg-10-4319-2013, 2013.

Smola, A. J., Olkopf, B., and Schölkopf, B.: A tutorial on support vector regression*, Statistics and Computing, 14, 199–222, doi:Doi 10.1023/B:Stco.0000035301.49549.88, 2004.

Takahashi, T., Sweeney, C., Hales, B., Chipman, D., Newberger, T., Goddard, J. G., Iannuzzi, R., and Sutherland, S. C.: The Changing Carbon Cycle in the Southern Ocean, Oceanography, 25, 26–37, doi:10.5670/oceanog.2012.71, 2012.

Thomalla, S. J., Fauchereau, N., Swart, S., and Monteiro, P. M.: Regional scale characteristics of the seasonal cycle of chlorophyll in the Southern Ocean, Biogeosciences, 8, 2849–2866, doi:10.5194/bg-8-2849-2011, http://www.biogeosciences.net/8/2849/2011/, 2011.

Vapnik, V.: An overview of statistical learning theory., IEEE transactions on neural networks / a publication of the IEEE Neural Networks Council, 10, 988–99, doi:10.1109/72.788640, http://www.ncbi.nlm.nih.gov/pubmed/18252602, 1999.

Weiss, R.: Carbon dioxide in water and seawater: the solubility of a non-ideal gas, Marine Chemistry, 2, 203–215, doi:10.1016/0304-4203(74)90015-2, 1974.

Williams, N. L., Juranek, L. W., Feely, R. A., Johnson, K. S., Sarmiento, J. L., Talley, L. D., Dickson, A. G., Gray, A. R., Wanninkhof, R., Russell, J. L., Riser, S. C., and Takeshita, Y.: Calculating surface ocean pCO2 from biogeochemical Argo floats equipped with pH: An uncertainty analysis, Global Biogeochemical Cycles, 31, 591–604, doi:10.1002/2016GB005541, 2017.