Reviewers initial comments are in dark blue
Responses to the comments are in green

Response to both reviewers

We would like to thank the reviewers for the comprehensive and constructive feedback on the manuscript. We feel that the comments that they made have contributed to a much better manuscript.

Some of the recommended changes were fairly large and thus the manuscript, primarily the results and discussion, have changed significantly. As recommended by both reviewers, we have deepened both the analyses and the discussion. In summary:

- A deeper analyses of the observational estimates assesses the performance based on the RMSE scores (and other metrics) and then assesses the difference between the estimates.
- The synthetic data experiments have been formalised. There are now two primary experiments: 1) what is the impact of including coordinates as proxies on the estimates; 2) what is the effect of the sampling biases in SOCAT v3. We also find that the ensemble mean of ΔpCO2 scores better than the individual methods.
- The discussion is structured after the results (two points above), but now goes much more in depth.
- The final figure focussing on the trends of the fluxes has been removed. This is due to the fact that the manuscript is now much longer and the trend analysis would distract the reader from the primary goal of the study; which is to introduce methods and the synthetic data experiments.
- These results will be published in a future publication.

The remaining comments have been addressed to each specific reviewer.

We hope that these changes make the manuscript suitable for publication.

Response to R1

Weaknesses

- Methods: The weakest bit of this paper certainly is the methods section at the moment. Particularly the 2 approaches are explained to briefly. It is very difficult to follow with many new terms being introduced but not explained, e.g: " A few slack variables ( ) are allowed, within the limits of a slack parameter" – what are "slack variables" and "slack parameter"? " versatile by mapping X onto a higher dimensional feature space using an interchangeable kernel" – feature space? interchangeable kernel? " decision trees" – to the average BG reader a tree has leaves or needles ... " bagging" – the meaning is not clear "K-fold cross-validation" – again, please explain what this means Without knowing all these terms the reader is lost and understanding a method means trusting a method.
  We have addressed this weakness by including more detail about each of the methods. This includes the basic formulation for SVR and RFR. The terms are also now explained more explicitly.
- Validation, comparison: It is disappointing that the authors only provide the RMSE MAE and r2 in the manuscript for the entire period, i.e only one number. Many statements in the text do require a more thorough analysis. E.G section 4.1: " One of the most marked differences is the weaker sink estimated by the SOM-FFN method in the SAZ (Figure 4)." – Figure 4 shows that the difference between the estimates is e.g. larger in the earlier analysis years – a error/RMSE/r2 analysis per year would be interesting and make a stronger case. Furthermore, it would be very interesting how the error/RMSE/r2 varies with data density, both in time and in space.
  We have included three regional time series of RMSE for each of the biomes. These include the data density. Note that the RMSE values have also increased as these were previously reported for only the SAZ and FPZ combined. There is also now an analysis of the RMSE in the SAZ included as additional material. This shows that the increase in error in the SAZ is primarily due to increased number of coastal measurements.
- The usage of space and time coordinates: Firstly, I am not surprised that additional data result in a smaller error, as they add additional degrees of freedom. Secondly, after reading the methods section I was puzzled why they were included? In the end, on page 12 line 21 I found the statement: " This implies that the available proxy variables are not able to capture the variability of pCO2 ." pCO2 is not affected by time and space, but by the environmental conditions reflected in proxies such as SST or biology. Space and time are in this case only placeholders for unknown proxies. This needs to be better discussed up-front.
  This is addressed a little better in section 2.3: Data transformation and derived variables. The paragraph now reads:
  To gain a better understanding of these methods' strengths and weaknesses we implement SVR and RFR in a synthetic data environment. A similar approach was taken by {Friedrich2009} in the North Atlantic, which experienced a similar data paucity to the Southern Ocean in the early 2000's. This idealised environment was also used to estimate the effect of including/excluding certain proxy variables as well as the optimal coverage of cruise tracks to constrain the North Atlantic ΔpCO_2 adequately. Similarly, we assess the efficacy of including coordinate variables as proxies of ΔpCO_2 in the empirical methods. In the intercomparison study by 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); however several methods in the study also include latitude and longitude. While coordinates do not mechanistically impact ΔpCO_2, they do help to constrain estimates where the available remote sensing proxies cannot adequately do so. The synthetic data is also used to test the ability of the SVR and RFR to approximate ΔpCO_2 in the seasonally sparse Southern Ocean.

Specific and minor comments

- Page 2 line 9: "were" – I suppose "where"
  changed to where
- Page 2 line 10: "interannual pCO2 trends" – interannual trends? I suppose you mean interannual variability, otherwise please clarify
  changed to interannual variability
- Page 2 lines 16-18: This statement is right but wrong: Rödenbeck et al indeed did argue that there is a lack of independent ship-based observations in the SO which prohibit an independent comparison – hence right. However, e.g. Landschützer et al 2015 used for their trend analysis also an atmospheric inverse estimate which is based on independent, namely atmospheric, observations – hence wrong. So, in combination with the text above this statement is misleading.
  Changed the paragraph completely. The intro should now read better.
- Page 4 line 8: "gridded observations" – I don't think – not even for the sake of brevity – you can call data from an assimilation model (ECCO2) "gridded observations"
  Corrected
- Page 5 line 5-6: You claim that log10 normalisation of CHL and MLD leads to normal distribution, but I doubt that – I suspect it rather comprised a fairly normal distribution in the center with long tails.
  Have changed this to "a distribution that closer resembles a normal distribution"
- Page 5 lines 9-10 and following: see major comment above. A bit more discussion is needed what these coordinates represent in terms of CO2 predictors. Page 5 line 25 and following: The methods are hard to follow. Too many unknown and specific wording is used (see major comment above).
  Details are now "fleshed out" on pages 6 and 7
- Page 7 lines 14-15: why Nightingale? there are newer transfer velocity estimates from Wanninkhof et al. (2013, 2014) using CCMP?
  Now using Wanninkhof et al., (2014)
- Page 8 Figure 3: It is confusing that the SOM-FFN method is called "SOM" here – please don't change abbreviations throughout the manuscript.
  SOM-FFN is continued throughout manuscript
- Page 9 Figure 4: In all the following text the difference between the lines is discussed, but not that they are based on different datasets, i.e. SOCATv2 and SOCATv3. It is certainly plausible that the availability of data in SOCAT also affects the difference? I suggest to discuss this also in the main text.
  We address this issue in two ways: estimates are compared with SOCAT v2 and v3; models trained with SOCAT v2 and v3 are compared – this is presented only in the appendix.
- Page 11 lines 5-12: This is very vague. Firstly, the authors have not properly calculated uncertainties for each region and timestep. Secondly – as mentioned above, the discussion is missing the difference between SOCATv2 and SOCATv3. How

many new data are included in SOCATv3 and where? Could this add to the difference? Thirdly, the statement about the influence of the tropics is vague.
Table 2 has been changed to a figure showing the spatial and temporal variability of RMSE for each of the methods. Moreover, there is also a table detailing the average regional RMSE, MAE, bias, r2 and n.
There are also two new figures in the additional materials that address the issue of SOCAT v2 vs v3. We show that the relative majority of points gained in the SAZ in SOCAT v3 are in the Argentine sea – a region of high complexity. The tropics point has been changed to a discussion around "remote knowledge transfer" and this should now be much clearer.

- Page 12 lines 15-16: I suppose discontinuity at a cluster, or biome border is a sign of bad model quality as well. In 2 adjacent biomes, that are well sampled, I would expect no hard border, whereas in more poorly reconstructed biomes this border effect is more prominent. However, continuity is no sign of quality, but rather comprises a "prettier picture".
Removed the statement about the discontinuity of clusters as Reviewer 2 pointed that this is a trivial issue to solve.

Response to R2

Evaluation

- Deepen analysis: While the manuscript is relatively thorough in the description of the two methods (with exceptions - see minor comments below), I find that the evaluation part has quite some room for an extension and some deepening. In particular, I am missing a thorough analysis of the residuals in time and space.
  The analyses have been extended significantly. The analysis around the RMSE estimates have been extended and the differences between the methods are now investigated in full.
- Extend scientific discussion of method: The paper would benefit substantially from an extension of the scientific discussion of the pros and cons of the method. Many issues are currently mentioned and investigated, but few of them are really discussed to the necessary level of detail and finality. Examples include the inclusion of the spatial variables in the regression, which is tested, but then only partially further investigated. Another good example is the more limited predictability of the RFR relative to the SVR. Finally, with this new method needing to compete with a range of already existing methods, the authors needs to demonstrate more clearly why it is better. I understand that these are difficult issues to discuss, and that it is likely not possible to give a definite answer. But it would behove the authors well to push the manuscript as far as possible in this direction.
  The results and discussion have been extended significantly. The synthetic data experiments have been formalised and are now discussed fully.
- Deepen scientific analyses and discussion of results: As it stands, the paper focuses nearly entirely on the method, and leaves only very little room for the scientific findings. This is a shame, in my opinion. I think that there is enough room in the manuscript to add a few more scientific analyses to the paper and to discuss them thoroughly. One example is the seasonal cycle, which differs quite substantially between the different estimates and is hugely important for determining the annual CO2 sink.
  As stated above, the scientific analyses have been deepened and we feel that the manuscript is now more complete.
- Language/Grammar: There are several places where the writing can be improved and be made more concise and precise. Further, the manuscript contains a number of grammatical/typographic errors that should be eliminated before the resubmission.
  Changed as recommended in the specific comments below

Specific and minor comments

- Abstract, p1, line 5: I suggest to add the source of the data already here, i.e., to write "The methods are used to estimate DpCO2 in the Southern Ocean based on SOCAT V3... ".
  SOCAT added to the abstract
- Abstract, p1, line 6: Typo. Change "The RFR as able" to "The RFR is able"
  corrected as to is
- Abstract, p1, lines 6-7 and elsewhere: I don't think that there is a statistically significant difference in the RMSE between 12.26 and 12.97 µatm. Please rephrase.

phrase removed – also note that these estimates have changed. The previous estimates were for the SAZ and PFZ biomes only.

- Abstract, p1, line 8: "modelled environment". The commonly used expression here is "synthetic data".
  synthetic data now used throughout the manuscript
- Abstract, p1, line 9: "achieved". Not sure that this is the best expression, since one commonly tries to achieve something that is desirable. I am not sure that having a higher error is a desired outcome. Perhaps simply write "have".
  this has been changed throughout the manuscript
- Abstract, p1, line 11: Add "a" to ratio, i.e., to read "with a lower ratio".
  added "a"
- Abstract, p1, general: Following up on my major comments (ii) and (iii), I think that also the abstract could benefit from a reshuffling with a bit more text devoted to the discussion of the methods and how they compare to others, and a bit more text about the results.
  A large portion of the abstract has been rewritten to accommodate the reviewer's suggestions
- Introduction, p1-3: general: The introduction reads well and contains the most important pieces, but I would love to see a bit more material with regard to the particular strengths and weaknesses of the existing methods. As it stands, it does not become clear to the average reader why we need yet another set of methods to interpolate the sparse data. This also helps to set up the later discussion on how this new set of methods stacks up against the existing ones.
  The introduction has been reformatted to include a motivation for each of the methods as well as the description of the different methods and why these were chosen.
  Data and methods: p4, line 1-4: It would be much cleaner if you used the same biomes for the synthetic data as for the real data. Of course are the model boundaries some- what different if one used the same criteria as used by Fay and McKinley, but this really should not matter much. Much more relevant is that you use the same approach when using the synthetic and the real data, so that you can really draw conclusions from one approach to the other. I strongly suggest to reconsider this choice.
  The northern boundary of the synthetic data has been changed from 30°S to the boundaries defined by Fay and McKinley (2014).
- Model data: p5, line 2: "resampled to ... monthly averaged resolution" This likely adds quite some smoothing to the data, something that does not really exist in the observations. Although the latter have been binned to 1x1 dg and month of the year, but many grid cells contain only a few observations, and therefore do not really represent a monthly average. Why not spot sampling the model following the sampling scheme of the observational programs?
  Change has been implemented to the data, and the text now reads: *The synthetic observations are sampled at the model resolution (5-day x 0.5°) to resemble the SOCAT dataset. Hereafter all data is resampled to 1.0° spatial resolution and monthly temporal resolution data to match observations.*
- Data transformation: p5, line 5 (and elsewhere). "There are several transformations that are applied..." This is awkward and can be simplified (and improved) to "The

input data are transformed..."
Changed as recommended

- Data transformation: p5, lines 9-10: "This then raises the question..." I found this some- what confusing. I suggest that you simply describe what you did in the method section, i.e., that your standard model includes the spatial coordinates, but that you also tested a version without them, and then have a more thorough discussion in the discussion section.
  This has been introduced briefly in the methods – only the methodology is presented
- Data transformation: p5, lines 16-20: I suggest to add here somewhere the time period that these data cover.
  This was added at the end of the first paragraph in section 2.1 Gridded Data
- Empirical methods: p5, line 21, Data are plural. Thus "The data are split..." SVR:
  changed as recommended
- p5, line 26: "The formulation of the SVR is such..." Awkward writing. I suggest to simplify this to "The cost function of the SVR minimizes ..."
  This section has changed – more detail for each method added at the request of Reviewer 1
- SVR: p6, lines 1-7: I suggest to add a bit more text here to better explain the SVR, and in particular, to better explain the particular parameter choices.
  More detail has been added. The cost function has been included.
- RFR: p7, lines 1-6: As above, I also suggest here to better explain the method and the parameter choices.
  More detail has been added about the RFR, specifically, the theoretical model for a decision tree.
- RFR: p7, line 9 "The out-of-bag error is used to select the hyper-parameters..." This is extraordinary cryptic. Please explain better.
  This should be clearer with the additional information provided.
- CO2 fluxes: p7, lines 14-16. "calculated". This expression is used three times in a row in a very repetitive manner. This makes it boring and hard to read. Please reformulate.
  Restructured as suggested
- Results, p8, lines 3-6: This connects to my first major comment. In my opinion, this section needs to be substantially extended and strengthened. A comparison of correlation coefficients and RMSE is insufficient in my opinion. I would like to see an analysis of the pattern and structure of the residuals in time and space. I also would like to see the biases and perhaps a few other metrics.
  The results and discussion have been updated with a much more in depth look at the RMSE values for the observational estimates
- Results, p8, line 5: "slightly better...". I don't think that this statement holds up to further scrutiny. With a measurement error of about 1 µatm and data that are distributed in time and space anything but random, I don't think that this difference is significant. To me, all one can say is that the two results are comparable in performance.
  This has been changed
- Results, p8, Figure 3: I would love to see also the annual mean figure and its discussion added to the results section.
  The image has been changed and now includes the mean state.

- Results, p9, line 9: "Estimates are higher..." but also elsewhere This is a result that is picked up here, but it is not really discussed later on. This is just one example of a few such mismatches between results and the later discussion section.
  These issues have hopefully been ironed out. The results and discussion have been rewritten to a large extent.
- Results, p10, lines 15-16 "Out-of-bag error" and "Out-of-sample error". These terms are uncommon and thus need to be carefully defined and later repeated in order for the average reader to be able to follow the arguments.
  We define the in and out of sample errors adequately and are now used frequently enough for the reader to keep track. The out of bag errors are only referred to briefly
- Results, p11, "These results suggest that estimates would benefit from the inclusion of coordinates". This statement is problematic for various reasons. First, such a conclusion should not really be part of the results section. Second, I don't really buy the argument, since almost by definition, the inclusion of additional independent variables tends to improve the fit, i.e., it increases the degrees of freedom of the problem at hand. This does not imply an increase in predictability or a true increase in "knowledge", as tested, for example through an analysis of the Akaike information criterion (AIC). Personally, I also oppose the inclusion of such variables, as they do not include any process information, and, in fact, suppress the establishment of knowledge transfer between regions of similar dynamics, but distant in time/space. I suggest to reconsider this choice and conclusion.
  The reviewer makes a valid point. However, the whole point of the synthetic data experiment is to test this. We feel that the new synthetic data experiments should better show the pros and cons of coordinates as proxies. We still find that, in the case of RFR and SVR as implemented in this study, should be included as the current available proxies are likely not fully capturing the variability of ΔpCO2.
- Discussion, p11, line 6: "weaker sink". This is not really obvious from Figure 4. I suggest to add a figure showing the annual mean DpCO2 including the differences between the different estimates. With such a figure, the whole paragraph becomes much easier to follow.
  This region has now been highlighted with a red oval. This is primarily to avoid too many figures in the manuscript. The differences of summer ΔpCO2 have been added to the additional materials
- Discussion, p11, line 19: "sparse winter data". This is certainly a valid hypothesis, but couldn't the authors use the synthetic data to test this hypothesis?
  The manuscript now follows a format of two primary synthetic data experiments, where the first asks what the impact of coordinates as proxies is and the second addresses the issue of sampling bias in the SOCAT dataset
- Discussion, p12, line 1: "Ensemble estimate". This is not an unreasonable assumption, but it is again one that could be easily tested with the synthetic data.
  We now show, with the synthetic data that the ensemble estimate of RFR and SVR is in fact a better fit to the out-of-sample estimate than the standalone methods.

- Discussion, p12, line 15: "additional complexity of dealing with DpCO2 discontinuities" It turns out that this is a very small issue. You can test this by comparing the smoothed with the raw version in the pCO2 data sets provided by Landschützer et al. See http://cdiac.ornl.gov/oceans/SPCO2_1982_2011_ETH_SOM_FFN.html.
  This has been removed from the discussion
- Discussion, p12, lines 17-30: The conclusion stated on page 11 about the inclusion of a spatial variable should come, at the earliest here.
  This topic has been moved to the discussion
- Discussion, p12, in general: There are many other things that need to be discussed here (see also my second major comment above).
  The discussion should now be more comprehensive
- Discussion, p13, line 2, "Tuning the algorithm..." This sentence needs to be embedded better in order for it to make sense to the average reader.
  The discussion has changed – this sentence no longer exists.
- Discussion, p13, section 4.4. "Trends of ensemble estimates". This section and related ones needs to be substantially strengthened. As it stands, this small section is not much more than a teaser. This should not be.
  We removed the section on the trends as it may in fact distract the reader from the already dense material. This will be published in the near future.
- Conclusion, p13, line 32, "from satellite proxies..." This is not quite correct, since SSS, MLD, and atm. CO_2 stem from other sources. Please reformulate.
  Sentence now reads: The SOCAT v3 dataset was co-located with assimilative model output and satellite measurable proxy variables to create a training dataset.
- Conclusions, p14, lines 4-10: Some of these conclusions are not really that evident from the results provided earlier. This has a lot to do with the results section not having made the point well enough.
  This has been changed substantially and should no longer contain any surprise results.
- Data availability, p14: I think it would be much better if the data were hosted by an international database such as CDIAC (in the future NCEI) or Pangaea.
  This will be hosted by FigShare which has DOI

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 (luke.gregor@uct.ac.za)

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 pCO_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 ΔpCO_2 in the Southern Ocean based on SOCAT v3, achieving similar ~~results~~ trends 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~~ Results show that the SOM-FFN ~~(12.97~~ approach outperforms the RFR and SVR methods with respective RMSE scores of 14.84, 16.45 and 24.40 µatm~~). To assess the efficacy of the methods and the limits of the training dataset (SOCAT~~ . However, this is, in part, due to an increase in coastal observations from SOCAT v2 to v3~~), SVR and RFR are applied in a modelled environment. Again~~ . The success of the SOM-FFN and RFR both depend on the ability to adapt to different modes of variability. The SOM-FFN achieves this by having independent regression models for each cluster, while this flexibility is intrinsic to 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~~ . Analyses of the estimates shows that the SVR and RFR's respective sensitivity and robustness to outliers define the outcome significantly. Further analyses on the methods were performed by using a synthetic dataset to assess: which method (RFR or SVR) has the best performance?; what the effect of using time, latitude and longitude as proxy variables is on ΔpCO_2?; and what is the impact of the sampling bias in 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.~~ on the estimates? We find that while RFR is indeed better than SVR, the ensemble of the two methods outperforms either one, due to complementary strengths and weaknesses of the methods. Results also show that for the RFR and SVR implementations, it is better two include coordinates as proxy variables as RMSE scores are lowered and the phasing of the seasonal cycle is more accurate. Lastly we show that there is only a weak bias due to undersampling. The synthetic data provides a useful framework to test methods in regions of sparse data coverage and showing potential as a useful tool to evaluate methods in future studies.

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; Frölicher 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, 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 (pCO_2) by interpolating these with satellite proxy data. Access to in-situ pCO_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 majority of these methods are 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 authors found that methods agreed in regions ~~were~~ where data coverage was adequate, but for data sparse regions, such as the Southern Ocean, interannual CO_2 ~~trends~~ variability 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~~ Only two of the methods in Rödenbeck et al. (2015) were able to adequately represent interannual variability of ΔpCO_2, namely: the SOM-FFN (self-organizing map – feed forward neural network) from Landschützer et al. (2014), and the mixed layer scheme (MLS) from Rödenbeck et al. (2014). These two methods were used by Landschützer et al. (2015) to show that Southern Ocean CO_2 uptake strengthened after 2000. However, these methods often showed large interannual differences in flux estimates despite agreeing on the overall decadal trend. This shows that there is lack of coherence even amongst the methods that perform well, meaning that different methods may lead to different interpretation of 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 pCO_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)The primary reason for the varied results is thought to be the way in which the algorithms deal with sparse data in the Southern Ocean (Rödenbeck et al., 2015). This alludes to the importance of testing multiple approaches, as different methods may be able to better represent the CO_2 estimates in the data sparse Southern Ocean.

In this study , we introduce and compare two empirical we introduce two methods new to this ocean CO_2 applicationapplication, namely: 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; ?). (Vapnik, 1999; Smola et al., 2004). In a review on the use of Support Vector Machines (the broad category for regression and classification variants) in remote sensing, (Mountrakis et al., 2011) found that the method had the "ability to generalize well even with limited training samples". This makes SVR an appealing consideration for the sparsely sampled Southern Ocean. 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).

As with SVM, Random Forests (both classification and regression variants) have also been used in remote sensing applications, though it does not seem to be as widely used in earth systems sciences despite proving to be a powerful, yet easy to implement, learning algorithm (Caruana and Niculescu-Mizil, 2006; Hastie et al., 2009). 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 However, the lack of data in the Southern Ocean, particularly in winter, makes it difficult to understand the limitations of these methods within the framework of the SOCAT v3 data. context of SOCAT data.

To gain a better understanding of these methods' strengths and weaknesses we implement SVR and RFR are implemented in a simulated environmentwith a realistic sampling strategy to assess if there are biases to this sparse data . This approach allows us to test the impact of includingvarious in a synthetic data environment. A similar approach was taken by Friedrich and Oschlies (2009) in the North Atlantic, which experienced a similar data paucity to the Southern Ocean in the early 2000's. This idealised environment was also used to estimate the effect of including/excluding certain proxy variables as done by Friedrich and Oschlies (2009). Thereafter the methods are applied to observational data for actual estimates of pCO_2 well as the optimal coverage of cruise tracks to constrain the North Atlantic ΔpCO_2 adequately. Similarly, we assess the efficacy of including coordinate variables as proxies of ΔpCO_2 in the empirical methods. In the intercomparison study by Rödenbeck et al. (2015) proxies typically include, but are

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~~ fCO_2^{sea}	1970	2014	1 mon	1°	(Bakker et al., 2016)
CDIAC	~~xCO2atm~~ $xCO2_2^{atm}$	1970	2014	–	–	~~(?)~~ (Masarie et al., 2014)
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)

not limited to sea surface temperature (SST), chlorophyll-a (Chl-a), mixed layer depth (MLD) and sea surface salinity (SSS); however several methods in the study also include latitude and longitude. While coordinates do not mechanistically impact ΔpCO_2, they do help to constrain estimates where the available remote sensing proxies cannot adequately do so. The synthetic data is also used to test the ability of the SVR and RFR to approximate ΔpCO_2 in the seasonally sparse Southern Ocean.

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).

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)~~ The domain of this synthetic data experiments is defined by the three southern biomes of Fay and McKinley (2014). These are defined by observed oceanographic and biological parameters~~and would thus be different in~~ , but are used for the sake of consistency despite potential differences between observations and the model.

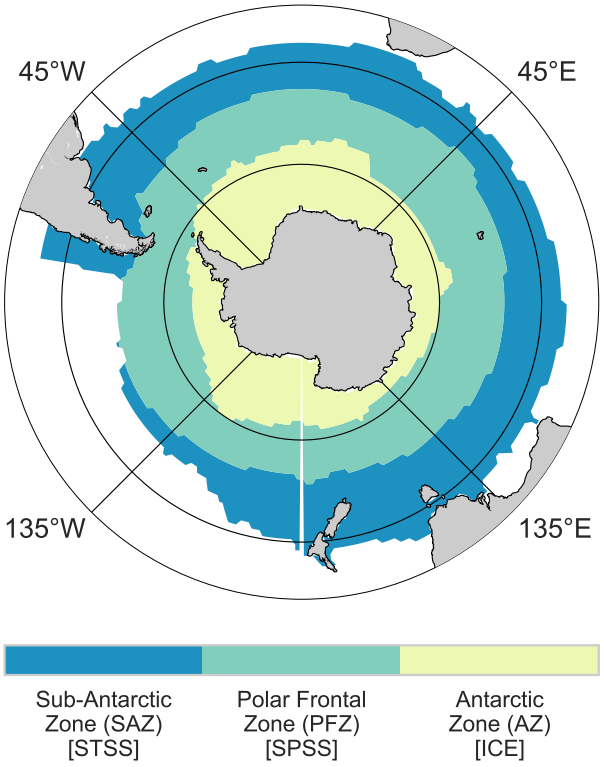

Figure 1. The three Southern Ocean biomes as defined by Fay and McKinley (2014). The common names for the biomes are shown in the key, with the abbreviations shown in the round brackets. The abbreviation in the square brackets show the abbreviations as given by Fay and McKinley (2014).

2.1 Gridded Data

The data sources are shown in Table 1. These gridded data refer primarily to remotely sensed data, with the exception of MLD and SSS. ~~These~~ The 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"~~. The temporal range of the data (1998 through 2014) is limited by the availability of Globcolour (Chl-a starting in 1998) and SOCAT v3 (fCO_2 ending in 2014).

All data are gridded to monthly x 1° using iris and xarray packages in Python ~~(?Met Office)~~(Hoyer and Hamman, 2017; Met Office). Gridded pCO_2 (SOCAT v3) is used to train the algorithms (Bakker et al., 2016). Surface station measurements (flask and tower) of atmospheric ~~xCO~~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~~xCO_2 to pCO_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~~ output from a regional NEMO-PISCES configuration ~~,~~ (BIOPERIANT05-GAA95b~~. This model~~) is used as the synthetic dataset. The configuration 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~~ is saved as five-day averages. The simulation was run from ~~1992~~ 1998 to 2009. The synthetic observations are sampled at the model resolution (5-day × 0.5°) to resemble the SOCAT dataset. Hereafter all data is resampled to 1.0° spatial resolution and monthly temporal resolution data to match observations. Finally, for the simulation experiment we define the Southern Ocean using the three southernmost biomes defined in Fay and McKinley (2014) as done for the observational estimates.

2.3 Data transformation and derived variables

~~There are several transformations that are applied to data for both model output and gridded observations~~Both gridded data and synthetic input data are transformed in preparation for the empirical algorithms. The $\log_{10}$ transformations of MLD and filled chlorophyll (Chl-a_{clim}) are taken to return a distribution that closer represents a normal distribution.

Several of the studies in Rödenbeck et al. (2015) included latitude, longitude and/or time as proxies of ΔpCO_2. ~~However, many of the methods that are regional or cluster the data before regression~~ It is important to note that coordinates do not drive mechanistic changes in ΔpCO_2. Rather, the inclusion of coordinates in the empirical methods account for unknown or regionally varying proxies that cannot be measured remotely. Many methods in the intercomparison by Rödenbeck et al. (2015) did not include coordinates, but account for unaccountable spatial variability by clustering or subsetting data regionally. 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.~~ Two scenarios for each method in the simulation experiment are run: no coordinate variables, and including coordinate variables (time, latitude and longitude).

The coordinates are transformed to preserve the continuity of the data as is shown below. Seasonality of the data is preserved by transforming the day of the year (j) and is included in both SVR and RFR analyses:

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

Transformed coordinate vectors were passed to ~~only SVR~~ both SVR and RFR 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~~ are used to create training arrays (x). The final input for SVR ~~were the following proxies~~ and RFR are (with 12 columns): $\log_10$(Chl-a_{clim}), SST, $fCO_{2(atm)}$, 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~~ Data are 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 basic formulation of SVR is ~~such that~~ similar to that of linear regression as described by Smola et al. (2004):

$$f(x) = \langle w, \mathbf{x} \rangle + b \qquad \text{with} \quad b \in \mathbb{R} \tag{3}$$

where b is an intercept, $\langle \cdot, \cdot \rangle$ denotes the dot product of 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. The~~ weights (w) and $\mathbf{x}$, the training data. The weights and intercept are found by solving the cost function:

$$\text{minimise} \quad \frac{1}{2}||w||^2 \quad \text{subject to} \quad \begin{cases} y_i - \langle w, x_i \rangle - b & \leq \epsilon \\ \langle w, x_i \rangle + b - y_i & \leq \epsilon \end{cases} \tag{4}$$

In this form, w is minimised according to the target values (y_i) to a precision of ϵ – i.e. there is no room for error greater than ϵ. However, with the majority of problems, meeting these constraints is not possible if data are noisy or ϵ is set small. The inclusion of slack variables (ξ_i, ξ_i^*) relaxes the constraints and the problem is now formulated as:

$$\text{minimise} \quad \frac{1}{2}||w||^2 + C\sum_{i=1}^{n}(\xi_i + \xi_i^*) \qquad \text{subject to} \quad \begin{cases} y_i - \langle w, x_i \rangle - b & \leq \epsilon + \xi_i \\ \langle w, x_i \rangle + b - y_i & \leq \epsilon + \xi_i^* \\ \xi_i, \xi_i^* & \geq 0 \end{cases} \tag{5}$$

Here C is a parameter that adjusts for the amount of error that the minimisation allows. The slack variable $|\xi|$ is only counted towards the cost if the point lies outside the margin ($|\xi| \geq \epsilon$). The points on or outside ~~these margins are the~~ the margins are called support vectors and are used to construct the hypothesis function, $h(x)$. This ~~elegant approach is made versatile by mapping **X**~~ is shown in Figure 2a where a linear SVR is fitted to noisy data produced from a cubic spline. The optimisation problem shown in Eq. 5 is solved in its dual formulation (see Hastie et al. 2009 for the full description). Importantly, solving the dual formulation allows for efficient kernelisation of SVR.

Kernelisation describes the process that maps the proxy variables ($\mathbf{x}$) 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~~ exhaustive grid search approach ~~using~~ with cross validation. We use K-fold cross validation~~with~~ , where the data is divided into eight equal "folds" ($k = 8$~~.~~). Seven of the folds are used to train the model, while the remaining fold is used for validation. This is done iteratively until each excluded fold has been used to test the results.

2.4.2 Random Forest Regression

~~A random forest~~ Decision trees form the basic building block of a Random Forest (RF)~~is an ensemble of decision trees, which means that the average estimate of n trees is taken~~ , with the average of n decision tress is taken as the ensemble estimate (Breiman, 2001) (Figure 2b). The basic idea of a decision tree is to iteratively partition data into boxes using simple rules that minimize the error at each split (referred to as a node) – these boxes would become hypercubes in higher dimensional problems. This is described by the basic formulation as described in Loh (2011):

1. Start at the root node
2. For each X, find the set S that minimizes the sum of the node impurities in the two child nodes and choose the split $X \in S$ that gives the minimum overall X and S.
3. If a stopping criterion is reached, exit. Otherwise, apply step 2 to each child node in turn.

Decision trees have high variance due to their discrete nature. Random forests reduce ~~the high variance of decision trees by bagging (bootstrap aggregating)in which the~~ this high variance by bootstrapping with aggregation (called bagging): a subset of the available training dataset is sampled with replacement ~~resulting in~~ for each decision tree in the RF. The sampling with replacement means that each training observation has a $\sim 63\%$ chance of being chosen at least once for a particular tree (Louppe, 2014). ~~A~~ This subsampling provides estimates that are robust to outliers as these have a chance of being omitted in training. This means that a random forest typically performs better when number of decision 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 proxy variables (m) given to the nodes in each tree when splitting the data

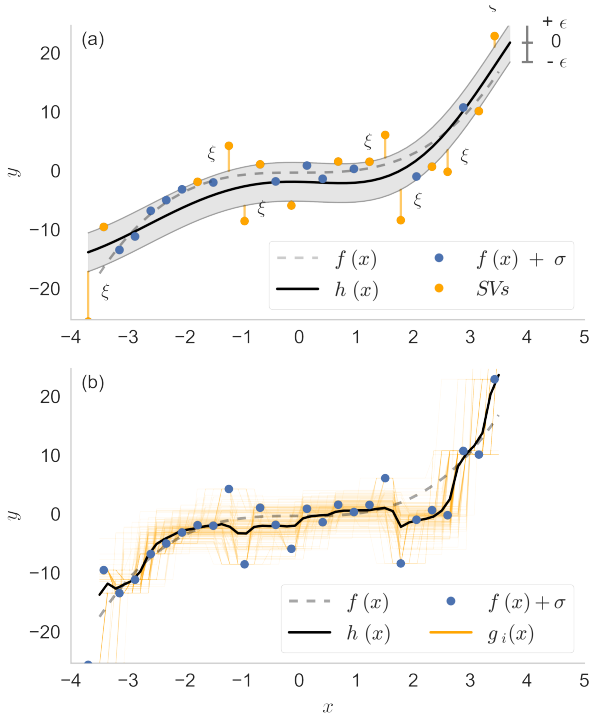

Figure 2. A simple example demonstrating the principle of (a) support vector regression and (b) random forest regression. The dashed grey line is the true function $f(x) = 0.4x^3$ with the blue dots representing a random sample taken from this function $f(x)+\sigma$, where σ is normally distributed noise. 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, $h(x)$.

(hence random) (Louppe, 2014). In this study, the maximum number of proxy variables ($m = 11$) was given to the RFR. 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 out-of-bag samples are those that are not selected during bagging. 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 cross-validate the model and select the hyper-parameters (t, m, l) for the RF.

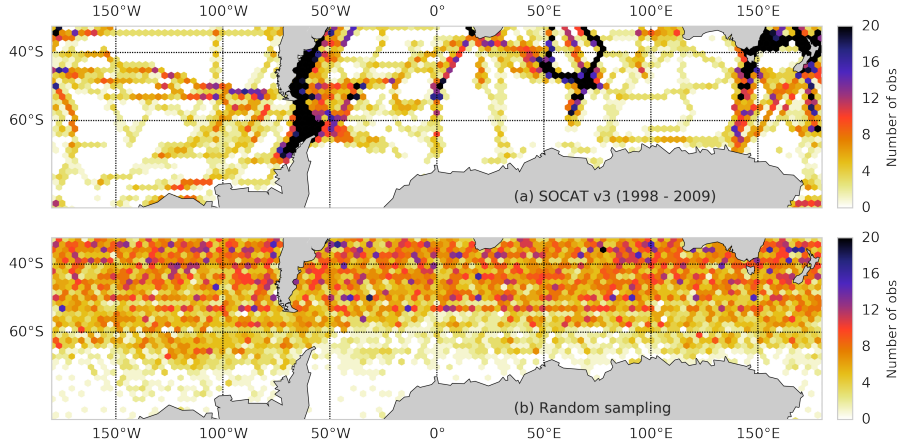

Figure 3. The spatial distribution of sampling locations in the synthetic dataset (BIOPERIANT05). The top panel (a) shows the sampling strategy using SOCAT v3 locations; and the bottom panel (b) shows the uniform random sampling distribution used in the second experiment.

2.5 CO_2 fluxes

Air-sea CO_2 fluxes are ~~calculated from~~quantified with:

$$FCO_2 = K_0 \cdot k_w \cdot \Delta pCO_2 \cdot (1 - [\text{ice}]) \tag{6}$$

The gas transfer velocity (k_w) is calculated using a quadratic dependency of wind speed with the coefficients of ~~?. Wind speed is calculated from the u and v~~ Wanninkhof (2014). The u and v vectors of CCMP v2 are used to compute the wind speed (Atlas et al., 2011). Coefficients from Weiss (1974) are used ~~to calculate K0~~ for K_0 and ΔpCO_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 (6).

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). Note that the SOM-FFN method was trained using SOCAT v2 compared to the methods in this study that used SOCAT v3.

2.6 Synthetic data experiments

Two experiments are run with the synthetic data. The first experiment aims to identify the efficacy of including or omitting coordinates as proxy variables on each method's ability to estimate ΔpCO_2 using SOCAT v3 locations. This is achieved by implementing the model with the transformed coordinate variables as proxies and then without. Note that the training procedure for the models remains the same as for the observational estimates of ΔpCO_2.

The second experiment assesses the impact that the seasonally sparse SOCAT v3 has on the ability of the methods to estimate ΔpCO_2. This is done by comparing the results of ΔpCO_2 estimates when trained according to: 1) SOCAT v3 locations trained with synthetic data (Figure 3a); 2) uniformly random sampling locations (random in space and time) with a sample size the same as SOCAT v3 (Figure 3b). Once again this the training procedure remains the same (as stated above).

3 Results

3.1 Observational CO_2 data results

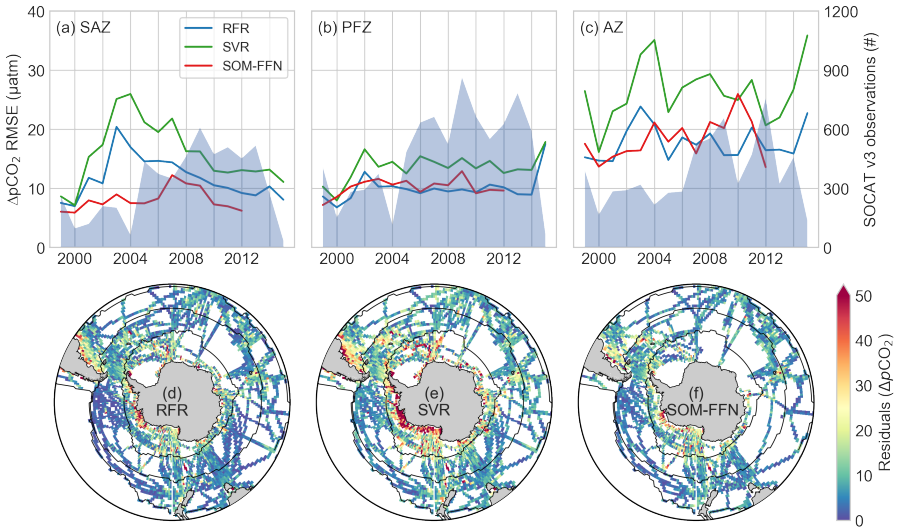

Figure 4. The ~~scores~~ RMSE (top row, a–c) for each of the ~~empirical methods trained with SOCAT v3 data~~three Southern Ocean biomes for RFR (blue), SVR (green) and SOM-FFN (red). The ~~domain~~ grey fill in the top row (a–c) shows the number of observations for ~~these scores is~~ each of the biomes for each year. The maps in the bottom row (d–f) show the spatial distribution of residuals in the Southern Ocean for SVR (d), SOM-FFN (e) and RFR (out-of-bag errors) (f). The thin black lines define the three Southern Ocean biomes as defined by Fay and McKinley (2014). Note that RFR and SVR are trained and tested with SOCAT v3 while SOM-FFN is trained and tested with SOCAT v2.

~~METHOD~~

We use the root mean squared error (RMSE) as the primary metric of the methods' performance as shown in Figure 4a–c. Note that the RFR RMSE is calculated from the out-of-bag error (effectively an independent error). SOM-FFN has the best RMSE score of 14.84 µatm (using SOCAT v2), which is better than the RMSE of RFR (16.45 µatm) and SVR (24.404 µatm), which are trained with SOCAT v3. The biases of the different methods are similar in magnitude for each of the biomes (-0.40, -0.03 and -0.75 µatm for the SOM-FFN, RFR and SVR respectively). The mean absolute errors (MAE) for the respective methods are 9.78, 9.85 and 15.27 µatm respectively.

Table 2. Various performance metrics for empirical estimates of ΔpCO_2 in the Subantarctic Zone (SAZ), Polar Frontal Zone (PFZ) and Antarctic Zone (AZ) (as defined by Fay and McKinley 2014). Results tested according to SOCAT v2 and SOCAT v3 are shown for the SVR and RFR methods.

	Method	RMSE	MAE	r~~2~~ ^{2}	Bias
~~SVR~~	SVR (v3)	~~16.04~~ 18.14	~~10.55~~ 11.28	0.48	0.61
	SVR (v2)	15.99	10.36	0.49	0.20
SAZ	RFR (v3)	13.67	8.16	0.70	-0.14
	RFR (v2)	12.66	7.65	0.68	-0.29
	SOM-FFN	10.07	7.04	0.76	-1.15
	SVR (v3)	14.45	10.06	0.48	0.31
	SVR (v2)	14.29	10.01	0.44	-0.01
PFZ	RFR (v3)	10.71	6.7	0.71	0.21
	RFR (v2)	10.56	6.77	0.69	-0.34
	SOM-FFN	11.01	7.68	0.6	0.26
~~RFR~~ height	~~12.26~~ SVR (v3)	~~7.43~~ 36.14	~~0.77~~ 25.19	0.56	-3.22
	SVR (v2)	35.69	25.01	0.59	-2.88
AZ	RFR (v3)	23.8	15.81	0.8	-0.27
	RFR (v2)	23.49	15.63	0.81	-0.62
	SOM-FFN	~~12.97~~ 21.32	~~8.56~~ 14.91	~~0.7~~ 0.82	-0.77

The difference between the mean absolute error (MAE) and the RMSE informs on the ability of methods to fit outliers or extreme points, as the RMSE scores larger errors much more severely than MAE. The SOM-FFN approach has the smallest difference between these two metrics (5.06, 6.60 and 9.13 µatm for the SOM-FFN, RFR and SVR respectively). This superior performance may be due to two factors. Firstly, the SOM-FFN method may be better at fitting the extreme points (those that are in the outer percentiles of the distribution). Second, it may allude to the fact that the SOCAT v2 dataset is less variable. Testing the SVR and RFR implementations against SOCAT v2 yields similar results, with the exception of in the SAZ, where both RMSE and MAE improve (results shown in Table 2).

The ~~RMSE, MAE and r^2 scores for each method applied to the data shown in Table 1 are shown in Table **??**. The RFR score is taken from the out-of-bag error, while the independent test set scores are used for SVR~~RMSEs and biases in the PFZ are least variable between methods. While there is a substantial increase in the number of observations from 2004 there is no appreciable change in the RMSE. The Antarctic Zone (AZ) is the primary contributor to these errors with much larger average RMSE values than for the SAZ and PFZ (36.14, 23.80 and 21.32 µatm for SVR, RFR and SOM-FFN respectively). This increase in the RMSE is likely driven by the larger variability of ΔpCO_2 observations in the AZ, where standard deviations of observations are 25.05, 20.01 and 54.65 µatm for the SAZ, PFZ and AZ respectively. ~~RFR achieves the best scores, with an~~

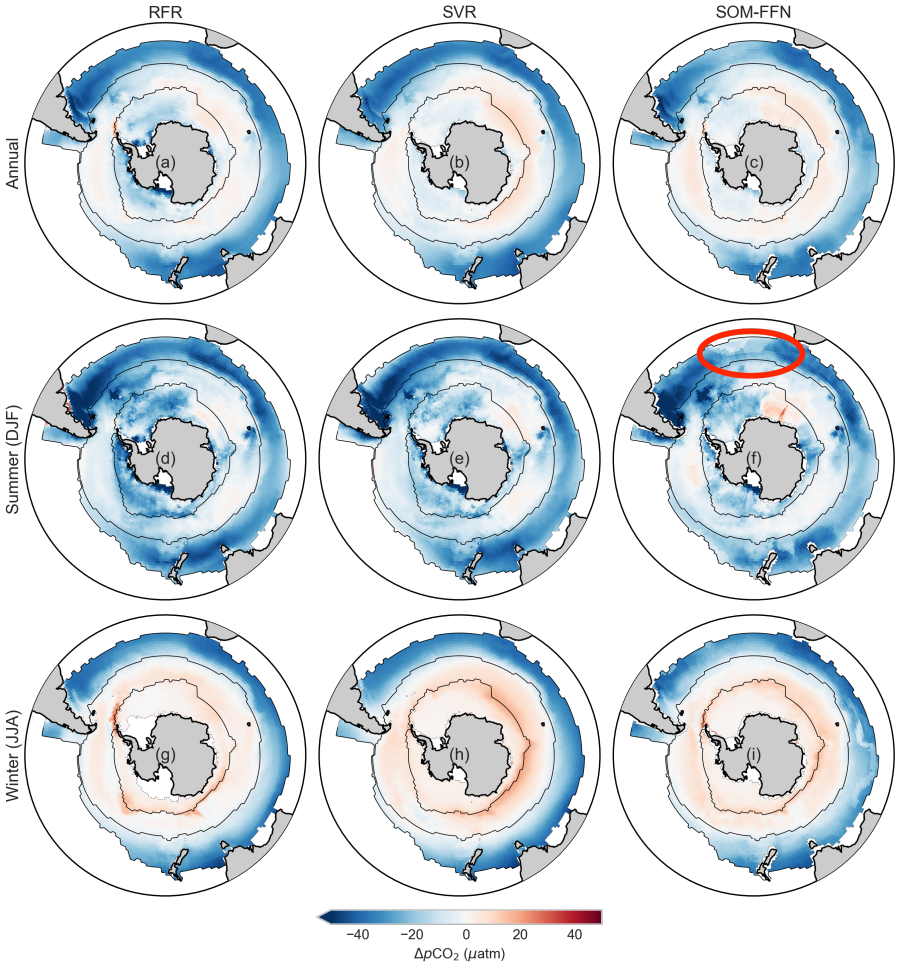

Figure 5. Seasonal averages for ΔpCO_2 from 1998 to 2014 for SVR, SOM-FFN and FRF. The annual mean is shown in the top row (a, b, c); the mean summer (DJF) ΔpCO_2 is shown in the middle row (d, e, f); and the mean winter (JJA) ΔpCO_2 is shown in the middle row (g, h, i). The thin black lines denote the SAZ, PFZ and AZ from outside inward. Note that the ΔpCO_2 has been normalized to sea ice cover where ΔpCO_2 is multiplied by $(1 - [\text{ice}])$. The red oval in (e) highlights the difference in SOM-FFN estimates of ΔpCO_2 during summer in the Atlantic compared to SVR and RFR.

~~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.~~ reflected in the highest r_2 scores in the AZ for the respective methods (Table 2).

~~The~~

The annual and seasonal averages (winter = JJA, summer = DJF) for ΔpCO_2 estimated by ~~SVR, SOM-FFN and RFR for the entire Southern Ocean region~~ RFR, SVR and SOM-FFN for the Southern Ocean are shown in Figure 5. ~~These show that there is, in general, good agreement in the spatial distribution between the methods. In winter (Figure 5a-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 ΔpCO_2 have thus Note that the estimates have been scaled to sea ice concentration ($\Delta pCO_2 \times (1 - [ice])$) as also done for fluxes in Equation 6 .

Seasonal averages for ΔpCO_2 from 1998 to 2014 for SVR, SOM and FRF. The mean winter (JJA) ΔpCO_2 is shown in the top row (a, b, c) and the mean summer (DJF) ΔpCO_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 ΔpCO_2 has been normalized to sea ice cover where ΔpCO_2 is multiplied by $(1 - [ice])$.

To the north of the Polar Front, in the SAZ, the ocean is a – this mutes winter estimates in the AZ. There is, in general, good agreement in the spatial distribution between the methods with the SAZ being a net sink of CO_2 (Figure 5) . The surface $\Delta p$and the region south of the Polar Front (PFZ and AZ) a source of CO_2 is more zonally symmetric in winter when compared to summer. The to the atmosphere as found by Metzl et al. (2006).

More specifically, there is stronger zonal asymmetry in summer compared to winter. This is driven, in part, by a strong reduction of $\Delta pp$$CO_2$ driven by biological production the Southern Ocean (Metzl et al., 2006; Lenton et al., 2012). There are three regions in the SAZ where ΔpCO_2 reduction is strongest and consistent between methods (Figure 5): east of South America (Malvinas Confluence), southeast south east of Africa (Agulhas retroflection) and between Australia and New Zealand (Tasman Sea). The reduction of ΔpCO_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 5d-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

However, there are more subtle 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 magnitudes and distributions of these patterns. The RFR underestimates winter outgassing south of the Polar Front (Figure 5g) compared to the SOM-FFN and RFR approaches other methods resulting in a weaker annual source. Conversely, the SVR has strong winter outgassing (Figure 5a-c) h) in the PFZ compared to other methods. In summer, the largest difference occurs in the eastern Atlantic sector of the SAZ where the SOM-FFN estimates higher ΔpCO_2 estimates (highlighted in Figure 5f) are larger compared to SVR and RFR(Figure 5d-f). . Other differences in the spatial output are more subtle.

Time-series of ΔpCO_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 ΔpCO_2 is not normalised to sea ice concentration in this figure.

The time-series (1998 – 2014) for ΔpCO_2 for The agreements and differences between methods are also observed in the time series for each of the Southern Ocean biomes as defined by Fay and McKinley (2014) are shown in Figure 6. In general there is good coherence between the three methods with agreement in the biomes (Figure 6). Importantly there is coherence in the strengthening sink (2002 to 2012) and timing of the 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

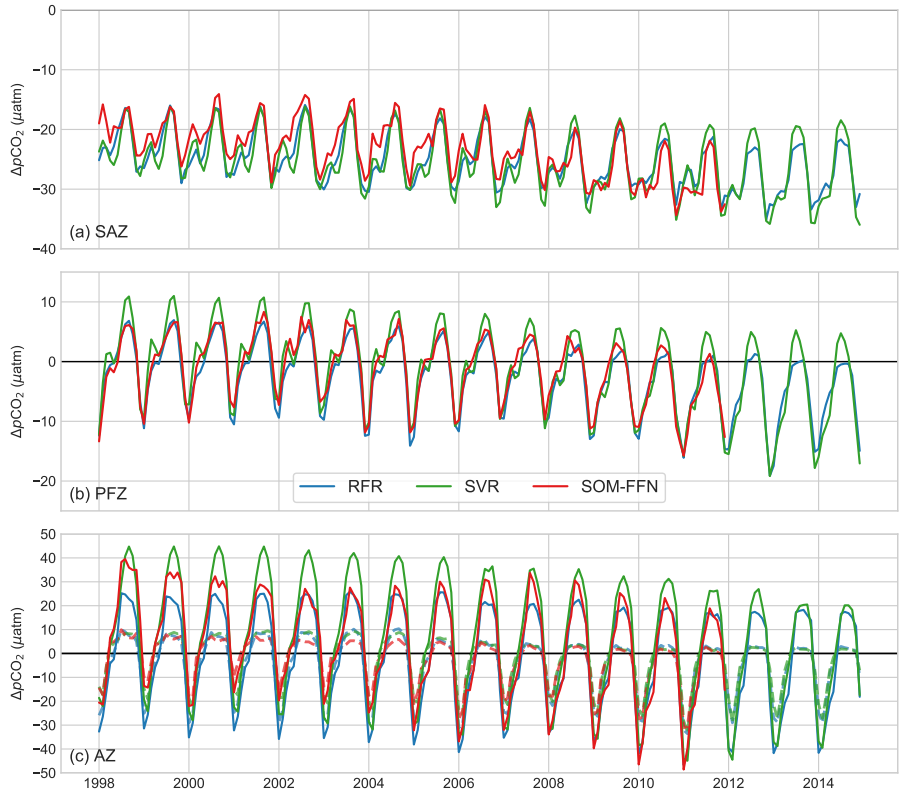

Figure 6. Time-series of ΔpCO_2 estimates for the three Southern Ocean biomes as defined by Fay and McKinley (2014): SAZ, PFZ and MIZ. The y-axis gridlines represent the same scale for figures (a) through (c). The SOM-FFN estimates are only available until 2011 as it is trained with SOCAT v2, while the SVR and RFR are trained with SOCAT v3. ΔpCO_2 normalised to sea ice cover is shown by dashed lines in the AZ.

~~the time-series representation~~between the three methods (Landschützer et al., 2015). The differences in the magnitude of the winter outgassing in the PFZ and AZ (Figure 6b,c) are also apparent, with the SVR overestimating ΔpCO_2 compared to other methods and the RFR with conservative outgassing estimates.

~~In the SAZ, the largest difference is~~ There is also a large difference between the SOM-FFN and the ~~other two methods. This is limited to the end of summerin the first half of the time-series. Comparatively, estimates~~ two other methods in summer, particularly from 1998 through 2006. Figure 5f shows that this could be driven by the difference in the eastern sector of the Atlantic (circled with red). Estimates of winter ΔpCO_2 ~~agree~~are in agreement, with the exception of the last four years when SVR winter estimates increase relative to RFR. The ~~overestimation of winter ΔpCO_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 ΔpCO_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~~

Table 3. ~~The performance metrics of~~ Root mean squared error (RMSE - µatm) for the three synthetic data experiments for RFR (left), SVR (middle) and ~~RFR in estimating ΔpCO_2 in a model simulation~~ the ensemble mean (~~BIOPERIANT05~~ENS) ~~using SOCATv3 cruise tracks as "sampling" locations~~of the two methods. Both ~~the~~ in- and out-of-sample errors are ~~shown~~reported (E[in] and E[out] respectively). ~~This~~ SOCAT experiments are those where the location of synthetic training data is ~~done~~ the same as SOCAT v3. This was run with (W coords) and without coordinates (W/O coords) namely time, latitude and longitude. A third experiment was run with random samples – coordinate variables are included as proxies. ~~The metrics are: RMSE = root mean squared error, MAE = mean absolute error, r-squared.~~

		~~MODEL~~ Experiment	~~INPUT~~ RFR	~~RMSE~~ SVR	~~MAE r²~~ ENS
~~ERROR~~	E[in]	SOCAT (W/O coords)	~~No coords~~6.65	~~8.2~~ 7.47	~~5.98 0.87~~ —
		SOCAT (W coords)	~~Coords~~ 5.12	~~6.26~~ 5.10	~~4.75 0.92~~ —
		Random Sampling	~~No coords~~7.23	~~6.27~~ 7.83	~~3.78 0.93~~ —
	~~Coords **4.7 2.72 0.95**~~ height	SOCAT (W/O coords)	~~No coords~~7.46	~~8.7~~ 7.46	~~6.51 0.67~~ 7.08
	E[out]	SOCAT (W coords)	~~Coords~~ 5.76	~~7.89~~ 6.19	~~5.99 0.72~~ 5.36
		Random Sampling	~~No coords~~4.88	~~7.87~~ 4.94	~~5.58 0.73~~ —
	~~Coords **6.33 4.5 0.82**~~ height				

~~The results also show that including time~~

The seasonal amplitude of ΔpCO_2 in the AZ ~~are~~ is far larger than ~~for both~~ in the SAZ and PFZ (Figure 6c) resulting in large methodological differences. However, this large differential ~~may not be realized as an outgassing~~ is not realized in calculated air-sea CO_2 ~~flux, particularly in winter,~~ fluxes, due to ice cover ~~.~~ as shown by the dashed lines (Figure 6c). Summer estimates are also influenced by sea ice cover, but not to the extent that winter fluxes would be reduced.

3.2 Simulation experiment results

~~The results from the simulation experiments are summarized in Table **??**. 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 **??**(a) and (b) respectively.~~

The advantage of using synthetic data is that both in- and ~~space coordinates~~ out-of-sample errors can be estimated, where the in-sample error is calculated from the training points and the out-of-sample error from the entire predicted domain. The latter gives a representation of the true error of the method. The results from these experiments are shown in Table 3. The detailed out-of-sample histograms are shown in Figure B1.

3.2.1 Coordinates as proxy variables

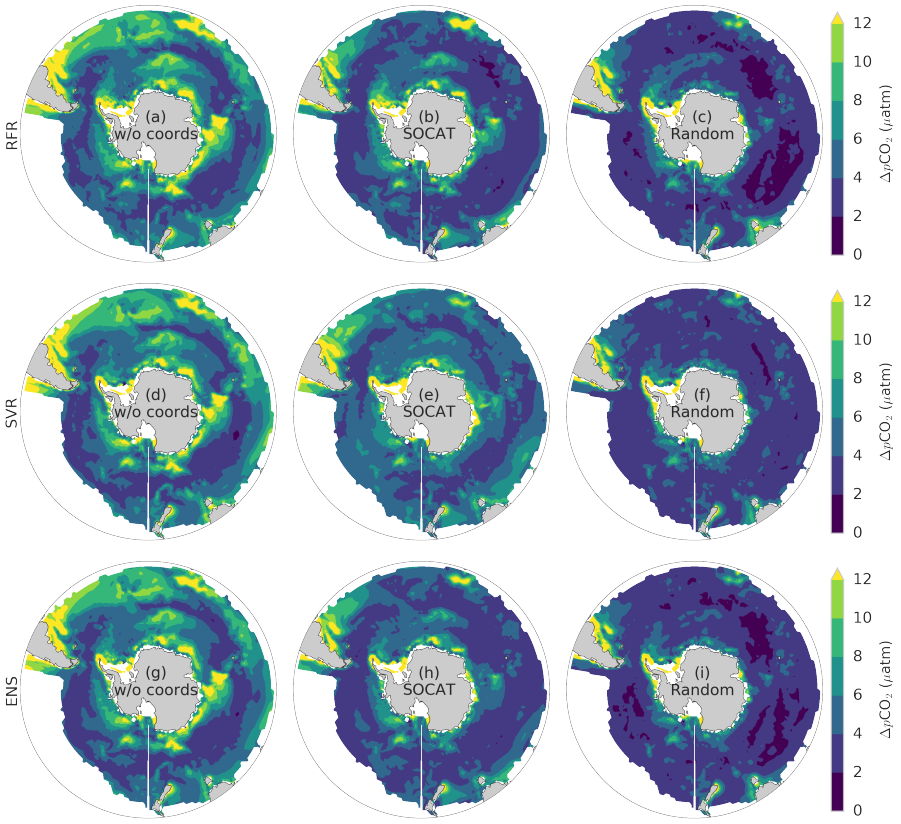

Figure 7. Distributions of root mean squared error (RMSE) for the three synthetic data experiments for RFR in the top row (a–c), SVR in the middle row (d–f) and the ensemble mean for RFR and SVR in the bottom row (g–i). The first column (a, d, g) shows the RMSE for synthetic SOCAT training locations without coordinates as proxies, while the second column (b, e, h) includes coordinates as proxies. The last column (c, f, i) shows the RMSE of randomly sampled training locations where coordinates are included as proxies.

This experiment used the synthetic dataset to test the influence of including or excluding transformed coordinates (time, latitude and longitude) as proxies of ΔpCO_2~~improves the estimates. This is shown in Figure 6 where the estimates~~ . There are four major results from the experiment results. Firstly, the RMSE estimates are smaller when coordinates are included as proxies for both in- and out-of-sample subsets (Table 3). Secondly, RFR achieves marginally better out-of-sample RMSE than SVR (5.76 and 6.19 µatm respectively) when 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~~ . Third, both RFR and SVR have comparable out-of-sample RMSE estimates (7.46 µatm) for ΔpCO_2 estimates trained with and without coordinate proxies. Lastly, the ensemble mean of SVR and RFR has lower out-of-sample RMSE estimates than the individual estimates for implementations with and without coordinate proxies, though these gains are marginal (Table 3).

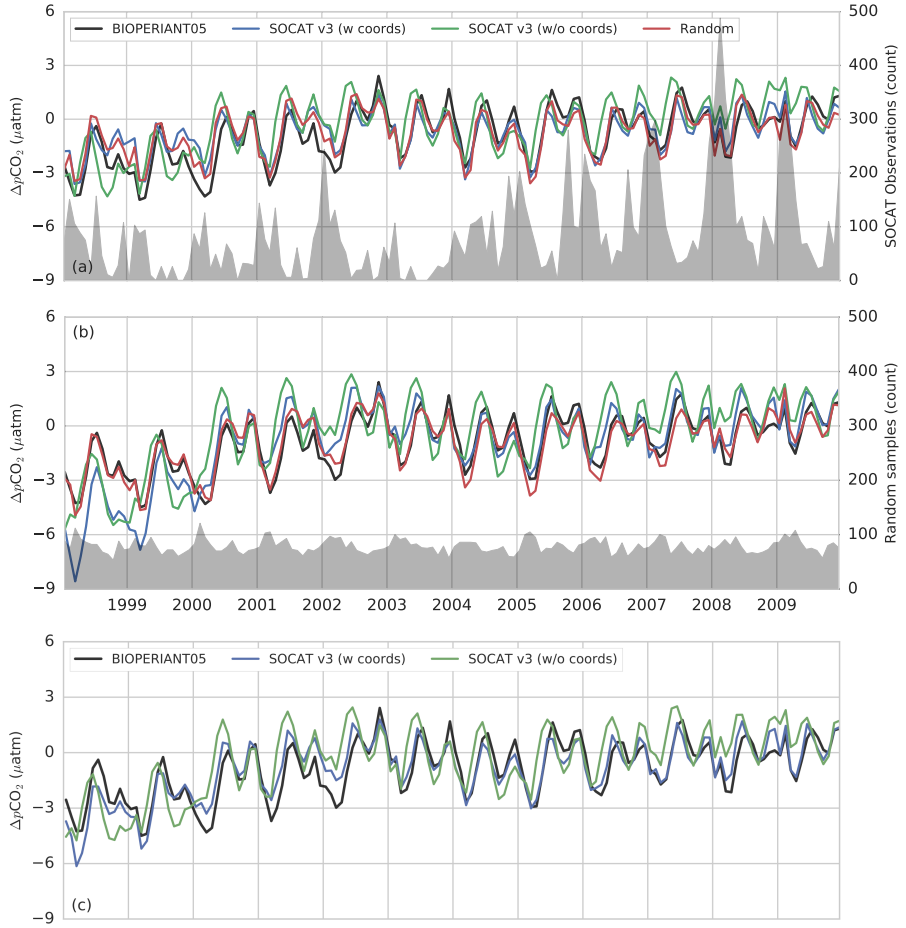

Figure 8. Time series of BIOPERIANT05 ΔpCO_2 (target) and empirical estimates of ΔpCO_2 for each of the experiments for RFR (a), SVR (b) and the ensemble mean estimates (c). The SOCAT v3 estimates are trained using the locations of SOCAT v3 data. The w coords variant includes coordinates as proxies of ΔpCO_2 while these are not included for w/o coords. The Random estimates are trained with uniformly distributed random sample locations. The number of samples per time step for SOCAT (a) and random sampling locations (b) are shown by the grey fill.

These points can also be gleaned from RMSE maps (Figure 7). Both RFR and SVR errors are low; however the RFR outperforms the SVR marginally for the open ocean regions. Errors in coastal regions remain high for each of the experiments and methods (Figure 7a,b,d,e); such as in the Argentine Sea, the Agulhas retroflection, and the marginal ice zone. The ensemble mean of the estimates achieves a balance between the two methods with low and moderate RMSE scores in the open ocean and coastal regions. Lastly, the distributions of the errors for RFR and SVR without coordinate proxies (and thus the ensemble mean) are similar.

The time series (Figure 8) show that including coordinate variables plays an important role in achieving accurate phasing of the seasonal cycle. When coordinates are not included as proxies the phasing shifts earlier for both methods. There is also an improvement of estimates over time, where the first two years (1998 and 1999) have worse estimates for both SVR and RFR (Figure 8). This does not seem to be linked to the number of observations, but could be due to the distribution. The ensemble ΔpCO_2 in the 1998 to 1999 period is closer to BIOPERIANT05 output as the respectively over- and underestimates of RFR and SVR compensate for each other.

3.2.2 Random sampling regime

This experiment is performed to assess the inaccuracies that arise due to the spatial and temporal sampling biases in the SOCAT v3 dataset. A random sampling regime is compared to the Training locations are chosen at random and uniformly in time and space. This eliminates any summer/winter biases as well as clustering of cruise tracks in certain regions (such as the Argentine sea). Note that coordinates are included as proxies of ΔpCO_2 with the random sampling regime.

Firstly, the ~~RFR without coordinates is the same as the SVR~~ results show that the biases in SOCAT v3 do contribute to out-of-sample errors, as the random sampling regime achieved lower RMSE scores than any of the other experiments (4.88 and 4.94 µatm for RFR and SVR respectively as in Table 3). However, RFR is marginally less susceptible to sampling biases than SVR as the relative improvement for the latter is larger (with differences of 0.88 and 1.70 µatm respectively). The spatial distributions of RMSE for the random sampling implementations (Figure 7c,f) show that errors in coastal regions remain high (> 12 µatm) with uniform sampling. Lastly, there is an improvement in the estimates from 1998 to 2000 with the ~~inclusion of coordinates, again highlighting the superior accuracy of the RFR~~random sampling particularly for the SVR (Figure 8), suggesting that the method is more susceptible to the temporal bias than RFR (if coordinates are included as proxies). ~~These results suggest that estimates would benefit from the inclusion of coordinates .~~

~~(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.~~

4 Discussion

4.1 ~~Methodological differences in observational~~ Observational estimates

~~The differences observed~~ In this section we address the methods' ability to fit the training data, in other words an assessment of in-sample errors (Figure 4 and Table 2). Thereafter we investigate the differences in the estimates of ΔpCO_2 ~~are driven by differences in~~ (Figures 5 and 6).

4.1.1 Assessment of in-sample errors

Based on the results, the SOM-FFN method (by Landschützer et al. (2014)) proves to be an elegant implementation of neural network methods that is able to estimate SOCAT ΔpCO_2 (in-sample estimate) better than the RFR and SVR methods (with respective RMSE estimates of 14.84, 16.45 and 24.40 µatm). Here we assess these differences and try to identify the possible reasons for the differences.

One of the largest differences in the methods ability to fit the training data is in the SAZ where the RFR and SVR score poorly in comparison to SOM-FFN, particularly from 2000 to 2006 (Figure 4a). This is during a period where the number of observations are still relatively low in the SOCAT v3 database (Figure 4a). This may then be due to an increase in the complexity of ΔpCO_2 estimates in the SAZ from SOCAT v2 to v3 from 1998 through 2006, thus more challenging to fit accurately. This is exemplified in the maps of RMSE (Figure 4g-i), where coastal regions typically have larger error estimates. A comparison of SOCAT v2 and v3 for this period shows that the increase in the number of observations occurs primarily in the Argentine Sea, thus confirming this hypothesis (Figure A1a). The comparison of SOCAT v2 and v3 RMSE results for RFR and SVR confirm this (Table 2), where there is a marked improvement when using the older dataset. Importantly, this shows that increasing the number of measurements does not necessarily improve the in-sample error estimates, but may yield a more accurate out-of-sample estimate; however this is difficult to test with limited data.

Despite the improvement in performance when testing against SOCAT v2, SVR and RFR still have poorer performance than the SOM-FFN approach. We attribute this in part, to the SOM-FFN's ability to reduce the large RMSE contributions observed in the other two methods. This notion is supported by the smaller difference between RMSE and MAE, especially in the SAZ (Table 2). The SOM-FFN achieves this by increasing the flexibility of the algorithm by having multiple regression models that can each be optimised for data with a particular length scale of variability. This allows the SOM-FFN approach to adapt to short scales of variability in dynamic regions such as the ~~algorithms as well as the implementation of these methods.~~ Argentine Sea and the coastal Antarctic (Figure 4i).

In comparison, this implementation of SVR, which is theoretically similar to an artificial neural network, only has one length scale for the entire domain (Vapnik, 1999; Smola et al., 2004). This becomes apparent in the AZ, where many of the observations are in the more biogeochemically dynamic coastal Antarctic, where melting sea ice results in short decorrelation length scales (Bakker et al., 2008; Chierici et al., 2012; Jones et al., 2012). The SVR has much larger RMSE scores in the AZ than the RFR or SOM-FFN (35.69, 23.49 and 21.32 µatm respectively). This suggests that implementing the SVR approach without an initial clustering or regionalisation step, will not yield good results.

By comparison, the RFR approach is more adept at fitting various length scales of variability, accounting for both the higher and lower variability in the AZ and PFZ respectively (with SOCAT v3 standard deviations of 54.65 and 20.01 respectively). The high r^2 scores achieved by RFR in the AZ and PFZ (0.81 and 0.71 respectively) highlight the flexibility in the method (Table 2). This is due to the differences in the underlying mathematics of the methods. Decision trees, which are the building block of RFR, separate data at each decision node with a discrete boundary (Breiman, 2001). Conversely, ANNs and SVRs often use Gaussian functions in the cost function, resulting in smoother approximations (Vapnik, 1999). This makes decision trees prone to overfitting, but the ensemble implementation of Random Forests eliminates this to a large extent.

4.1.2 Differences in ΔpCO_2 estimates

One of the ~~most marked differences~~ largest differences in ΔpCO_2 is the weaker sink estimated by the SOM-FFN method in the SAZ (Figure 6). This difference can be traced to the eastern Atlantic SAZ~~(Figure 5e)~~, where the SOM-FFN has higher estimates of ΔpCO_2 ~~.~~ (Figure 5e shown by the red oval and the differences between the methods in Figure A3a,b). A comparison between the SVR and RFR trained with SOCAT v2 and v3 further eliminates the use of different training datasets as primary sources of difference, where methodology is a higher order driver of difference (Figure A2). The lack of this feature in the eastern Atlantic sector of the Southern Ocean in 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~~ The clustering process separates pCO_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~~ observations into clusters that are not restricted in time and space (Landschützer et al., 2014). This allows the SOM-FFN ~~is implemented in a global domain, meaning that the algorithm could be mapping the relationship between~~ to "transfer knowledge" from a remote location (even outside the Southern Hemisphere) if proxies are similar to the Southern Ocean. This knowledge transfer assumes that the relationship between pCO_2 and ~~its proxies from more tropical waters.~~ the measured proxies is globally consistent. Moreover, there is the assumption that all pCO_2 variability (within a cluster) can captured by the measured proxies. This assumption is not made when using coordinates or regional subsets as locations are isolated, but there is then the potential loss of knowledge from remote locations. This question will be addressed further in the discussion on the use of coordinate variables as proxies of ΔpCO_2.

Another difference between ΔpCO_2 estimates is the tendency for ~~the~~ SVR to overestimate ΔpCO_2 ~~compared~~ relative to the RFR and SOM-FFN approaches, particularly in the PFZ and AZ where winter data is sparse (Figure 6)b,c. We attribute this to the SVR's sensitivity to outliers, determined by the fact that the cost function penalises outliers heavily (Equation 5). In context of the SOCAT v3 dataset, the algorithm may treat the sparse winter data as outliers. This ~~means that the higher estimates~~ is due to the fact that sparse winter measurements of ΔpCO_2 ~~in winter could be extrapolated, leading to the relatively elevated winter estimates.~~ are positive, while the abundant summer measurements are negative Metzl et al. (2006); Lenton et al. (2013). This may then be a positive realisation of a methodological attribute that is typically considered a weakness.

Conversely~~the RFR~~, RFR winter estimates of ΔpCO_2 are often lower than the SOM-FFN and SVR estimates, again in the AZ and PFZ (Figures 5g–i and 6b,c). This may be due to the method~~'~~'s resilience against outliers~~. This is primarily due to the bagging approach, where~~ , which could be due to two attributes (Louppe, 2014). Firstly, outliers are less likely to dominate the feature space with the use of bootstrap aggregation as these points will be sampled less frequently. Secondly, 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~~ regress values by using the average of samples in a terminal node (or leaf), where the minimum number of samples per terminal node is set by the user. This second attribute means that estimates will never be outside the bounds of the minimum and maximum of the training dataset, thus leading to conservative estimates (as shown in Figure 2b).

The differences between the methods shown in Figure 6 could be a good case for an ensemble approach, where the strengths of one model compensate for the weakness of another. This is assessed in the synthetic data and will be discussed further.

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~~ These attributes may also be the reason for the differences in the magnitudes of the autumn peak in ΔpCO_2 in the SAZ and PFZ. Mechanistically this peak could be attributed to a sharp increase in cooling leading into winter, resulting in increased solubility of CO_2 and thus a sharp reduction of ΔpCO_2 (Metzl et al., 2006; Takahashi et al., 2002). Deeper mixing of the ~~noise~~ water column shortly thereafter would entrain CO_2 rich waters, thus increasing ΔpCO_2 (Lenton et al., 2013). However, the trend for this peak to shrink in the SAZ and PFZ for all methods suggests that this may an artefact that is 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.~~ SOCAT dataset.

~~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~~

4.2 Synthetic data experiments

In this section we discuss the outcomes of the two experiments performed on the synthetic dataset (BIOPERIANT05 model output). The first experiment addresses the efficacy of including coordinate variables as proxies of ΔpCO_2 ~~discontinuities of cluster boundaries.~~ . This is done by running two implementations RFR and SVR: without coordinates as proxies, and with coordinates as proxies. The second experiment addresses the impact that the SOCAT dataset, biased in both space and time has on ΔpCO_2 estimates.

~~The non-linearity of the RFR allows the implementation without coordinates to marginally outperform the SVR implemented with coordinates (Table ??). 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~~

4.2.1 Coordinate variables improve estimates

This topic has to some extent already been mentioned in the discussion of the observational data, where pointed out the case for and against the inclusion of coordinate variables as proxies for ΔpCO_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~~ . If coordinates are not included there is the benefit of potential information transfer from remote parts of the domain, but this assumes that the satellite observable proxies (and assimilative model output) constrain ΔpCO_2 in a globally consistent way. If coordinates are included the information transfer is lost and the assumption is made that the proxy variables are not able to ~~capture the variability of~~ constrain ΔpCO_2 in a globally consistent manner.

~~For example, there may be differences in the relationship between~~ The results of this experiment show that coordinates improve estimates of ΔpCO_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~~ with better RMSE scores for both SVR and RFR (Table 3). We are thus in favour of the second hypothesis that the available proxies cannot sufficiently constrain ΔpCO_2 ~~. This means that the RFR estimates are more conservative than SVR and~~ without coordinates. A two step clustering approach, such as 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~~ may be able to achieve comparable results without coordinates, but this would have to be tested with that specific method. However, this may also lead to trends in the data that may be artefacts of remote knowledge transfer, as potentially seen in the observational data (Figure 5f).

An important outcome of this experiment is that the inclusion of coordinates improves the seasonal phasing of the methods (Figure 8a,b). It is critical for the empirical methods to correctly estimate the phasing of ΔpCO_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~~as the seasonal cycle phasing may be a useful indicator of anthropogenic driven changes to the marine carbonate system.

4.3 ~~Limitations of SOCAT v3~~

~~A key finding of the simulation experiment is that out-of-sample RMSEs arelarger than in-sample RMSEs, implying that error estimates for observational ΔpCO_2 would also be underestimated. This is due to the paucity of measurements in the Southern Ocean, meaning that~~ One of the assumptions in these synthetic data experiments is that the models are, to some extent, representative of the variability in the observed ocean. However, the BIOPERIANT05 output does not achieve this, with a standard deviation of 19.80 µatm for synthetic 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. data compared to 38.20 µatm of the gridded SOCAT v3 observations (according to Southern Ocean as defined by Fay and McKinley 2014). This could be a cause for concern. However, we believe that this creates an even a stronger case for the use of coordinates as proxy variables. The increased variability in the observations could be due to processes that deterministic models can not yet constrain due to our lack of understanding of the marine carbonate system (Lenton et al., 2013; Mongwe et al., 2016).

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

4.2.1 SOCAT biases

The lack of winter pCO_2 data is a problem throughout the mid and high latitude oceans, but is particularly severe in the Southern Ocean (Bakker et al., 2016), but the impact of the lack of data in the Southern Ocean is not known. Moreover, the efficacy of various methods to fill this large temporal gap is unknown (Rödenbeck et al., 2015). Here we show that there is a considerable impact in this synthetic data environment, but the effect of the sampling bias is perhaps smaller than we would have anticipated. Both methods are able to estimate the spatial distribution and the seasonal cycle of ΔpCO_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. with relative accuracy (Figures 8 and 7 and Table 3). This could be due to two factors.

In summary, Firstly, winter data is less variable than summer data and requires less sampling. Mechanistically, this is a likely scenario. In summer ΔpCO_2 is spatio-temporally heterogeneous in the Southern Ocean due to the uptake of CO_2 by phytoplankton (Metzl et al., 2006; Bakker et al., 2008; Thomalla et al., 2011; Chierici et al., 2012; Lenton et al., 2013). The drivers of phytoplankton are complex due to the correct implementation of machine learning algorithms should minimize over fitting to the training datasetco-limitation of light and iron (as a micronutrient) in the Southern Ocean (Boyd and Ellwood, 2010; Thomalla et al., 201 This complexity would require more sampling, perhaps additional proxies or increased spatial resolution to capture the variability of ΔpCO_2. Conversely, processes driving winter ΔpCO_2, namely the interaction of mixing and buoyancy, act on larger scales, potentially leading to less spatio-temporal heterogeneity. However, in the case of the lack of observations means that we simply cannot know with certainty. This makes a strong case for autonomous sampling platforms to the Southern Oceansector of the SOCAT v3 dataset, the data is not yet representative of 's winter sampling gap. The SOCCOM float project may soon yield such measurements with pH derived estimates of pCO_2 for the entire domain. This means that there will be biases in estimates that generalization techniques (Russell et al., 2014; Johnson et al., 2017; Williams et al., 2017).

Secondly, the model used to generate the synthetic data may not be representative of the Southern Ocean. This has been discussed in the previous section, but here, rather than increasing our confidence, it diminishes our confidence in the result. Studies have shown that process models are not able to resolve for which more representative data is required.

4.3 Trends of ensemble estimates

While methodological differences exist, the trends of Δpaccurately represent the seasonal cycle of CO_2 and air-sea in the Southern Ocean (Lenton et al., 2013; Mongwe et al., 2016). Moreover, ΔpCO_2 flux (as shown in Figure 6) 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 is often driven by processes that are not representative of observations (Mongwe et al., 2016).

The most likely scenario is likely a combination of these two factors, where winter data is in fact less variable than summer data, but the error is larger than the experiment shows due to incomplete knowledge of the processes that describe pCO_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 SVRin the process models.

The trends of the ensemble of FCO_2 for the SAZ, PFZ and AZ are shown in Figure **??**. These are in agreement with the trends explained in Landschützer et al. (2015): a slight weakening of the sink from

4.2.1 The best method: the ensemble average

The synthetic data also allows us to compare the two methods relative to each other, in the context of the SOCAT v3 data. The data show that the RFR method performs better than the SVR (trained with coordinates as proxies) with respective out-of-sample RMSEs of 5.76 and 6.19 µatm (Table 3). However, it is the average of these two methods (ensemble mean) that achieves the lowest RMSE (5.36 µatm), albeit marginal. The time series in Figure 8c shows that the improvement may come from the period 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 to 2000, when RFR is plagued by underestimation of the sink strength, while SVR overestimates the sink strength. This supports the notion that the strengths and weaknesses of these two methods compliment each other. Moreover, it supports the merit of multiple approaches and further development of empirical methods for the estimation of ΔpCO_2observed 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 meanuptake 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.

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

5 Conclusions

In this study two empirical methods (SVR and RFR) are presented as alternative (and perhaps ~~complimentary) methods to estimating Δpcomplementary) $p\mathrm{CO_2}$ from satellite proxies by tuning the methodsto best predict ship-based measurements~~gap filling methods. These algorithms are 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~~ . We apply the methods to the Southern Ocean where the paucity of ship based measurements during winter is one of the major challenges. The SOCAT v3 dataset was co-located with assimilative model output and satellite measurable proxy variables to create a training dataset (Bakker et al., 2016). These estimates were compared with the SOM-FFN ~~method~~ approach 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 datasetwas 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~~

We found that the SOM-FFN method outperformed the new approaches with lower RMSE estimates than the RFR and SVR. The RFR performed comparably to the SOM-FFN approach when compared with the SOCAT v2 dataset, with which SOM-FFN was trained. The increase in the number of measurements in the highly variable coastal ocean between SOCAT v2 and v3 leads to increased RMSE values, particularly in the Subantarctic Zone (SAZ). Despite accounting for the increase in coastal data, the SOM-FFN still outperformed the SVR and RFR approaches in the SAZ. We attribute this to the methods ability to cluster the training data into regions of different modes of variability to which individual regressions are then applied. The SVR method performed poorly due to its inability to adapt to various modes of variability, while the RFR is intrinsically much more flexible, thus performed well in fitting the training data.

There was good agreement amongst the three methods with respect to the overall trend of $\Delta p\mathrm{CO_2}$~~compared to~~ , but there were also differences. The primary difference was in the the ~~other two approaches. Conversely, the RFR produced lower estimates of~~ Atlantic sector of the SAZ, where the SOM-FFN overestimated $\Delta p\mathrm{CO_2}$ ~~in winter. The ensemble fluxes showed that the SAZ region as responsible for the majority of~~ relative to the other methods. This is likely due to remote knowledge transfer within a data sparse cluster; however, we cannot identify this as right or wrong due to the lack of data in this region. Other differences were due to intrinsic attributes of the methods: SVR was sensitive to outliers resulting in relatively large winter $\Delta p\mathrm{CO_2}$ ~~uptake over the period (1998~~ estimates ~~– 2014), while the PFZ dominated interannual variability. Ice cover in the AZ muted the large seasonal amplitude of~~ potentially a desirable feature for sparse winter data; RFR underestimated $\Delta p\mathrm{CO_2}$ relative to the other methods due to its robustness to outliers.

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 $\mathrm{CO_2}$, that is the~~ applied to a synthetic dataset (process model output). Two major questions were asked: 1) what is the efficacy of including coordinate variables (time, latitude and longitude) as proxy variables? 2) What is the impact of sampling biases in the SOCAT v3 dataset~~. ?~~ The results showed that ~~RFR is better able to estimate~~

including coordinate variables improved the estimates of ΔpCO_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~~ for SVR and RFR. Moreover, the phasing of the seasonal cycle was also improved with the inclusion of coordinates. The second experiment showed that there is only a small bias in the estimates of ΔpCO_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 showsthat reported errors of empirical~~ ; however, the inability of process models to represent Southern Ocean ΔpCO_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 pCO_2 from pH sensors (Williams et al., 2017).~~ variability accurately places uncertainty on this result.

Lastly we show that while the RFR approach outperforms the SVR approach, the ensemble mean of the two methods scores better than either individual methods. This motivates for continued research on methods that complement each other in strengths and weaknesses.

Data availability. The data are available at (https://figshare.com/s/dd034ad593cfd8c5188a)

Appendix A: Comparison of SOCAT v2 and v3

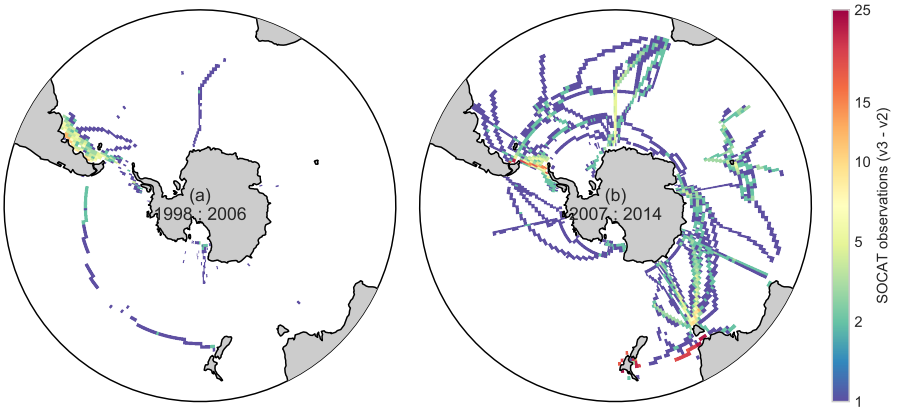

Figure A1. The increase in the number of observations between SOCAT v2 and SOCAT v3 for two periods: (a) 1998 through 2006, and 2007 through 2014.

One of the shortcomings of this study is that the SOM-FFN method used SOCAT v2 as a training dataset, while the SVR and RFR methods were trained with SOCAT v3. Figure A3 shows, there is a marked difference between the two datasets.

Importantly, the increase in the number of observations between 1998 and 2006 between SOCAT v2 and v3 are almost exclusively in the Argentine Sea.

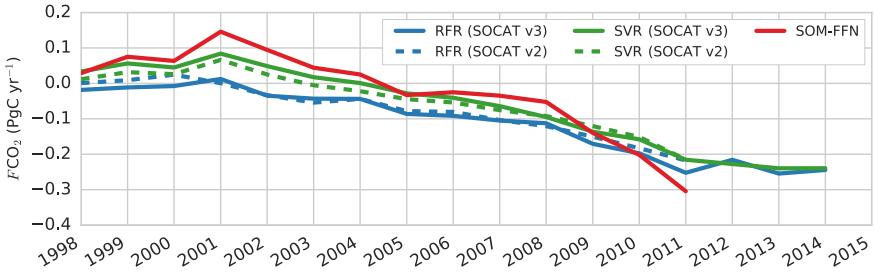

Figure A2. Comparison air-sea CO_2 flux RFR and SVR when trained with SOCAT v2 and v3. The SOM-FFN method, trained with SOCAT v2, is also shown. The figure demonstrates that methodology plays a larger role in determining the outcome of the estimate than the availability of data (for these two methods).

These differences may have an impact on the estimates of ΔpCO_2. To test this, the methods were implemented as explained in Section 2.4 with the exception that RFR and SVR methods were trained with both SOCAT v2 and v3. Figure A2 shows that, on average, there is a larger difference between the RFR and SVR methods than the different training datasets.

The differences between the different methods are shown in Figure A3. Figures (a) and (b) show that the SVR and RFR methods estimate a stronger sink in the Atlantic sector of the SAZ. Here (Figure A3b) the tendency of the SVR method to estimate strong outgassing south of the Polar Front relative to SOM-FFN and RFR is also seen. Conversely, the RFR, on average, underestimates ΔpCO_2 south of the Polar Front.

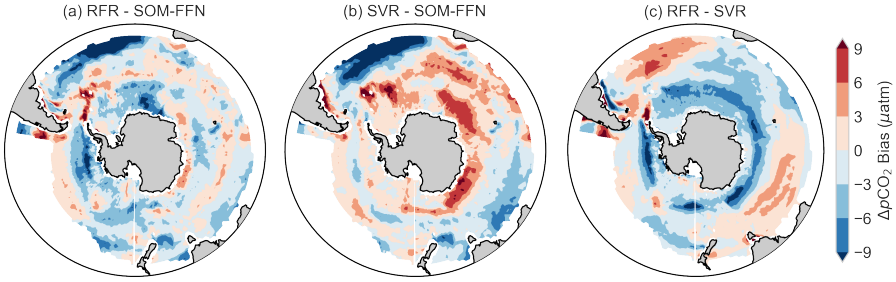

Figure A3. The differences between annual averages of each of the approaches for the period 1998 to 2006: (a) RFR – SOM-FFN; (b) SVR – SOM-FFN; (c) RFR – SVR.

Appendix B: Synthetic data experiments

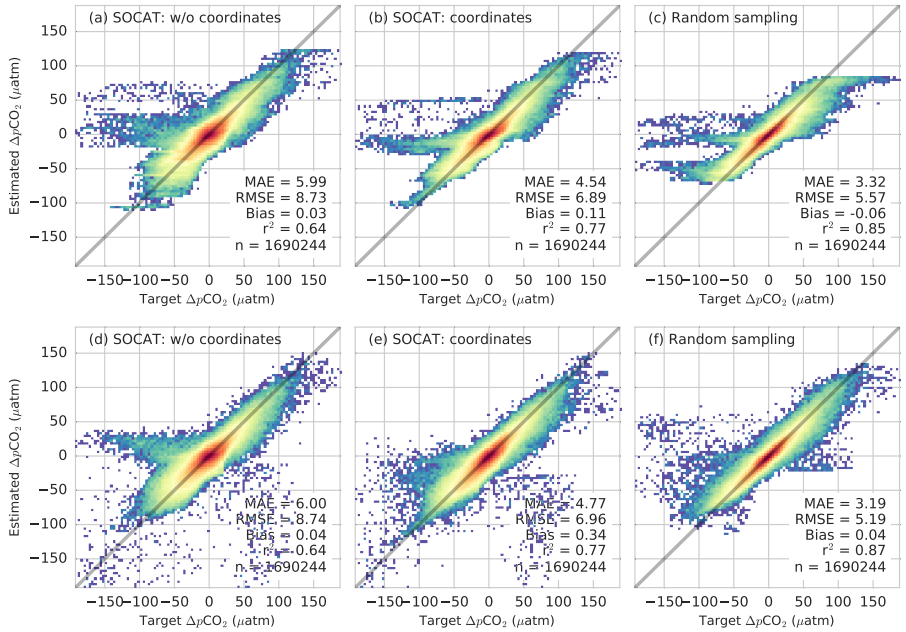

Figure B1. Two dimensional histograms for the distributions of out-of-sample estimates of ΔpCO_2 relative to target ΔpCO_2 (BIOPERIANT05). The top row (a–c) shows estimates made by RFR and the bottom row (d–f) shows estimates of SVR. The first column (a,c) shows those estimates trained SOCAT v3 locations without coordinate variables (time, latitude and longitude) as proxy variables and the second column (b,e) shows those with coordinate proxies. The last column (c,f) shows estimates trained with random locations (uniform in time and space) with coordinate proxies. The metrics are shown on each plot where MAE and RMSE are Mean Absolute Error and, Root Mean Squared Error respectively. n shows the number of observations in the estimate.

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., Hoppema, M., Schr, M., Geibert, W., and Baar, H. J. W. D.: A rapid transition from ice covered CO2-rich waters to a biologically mediated CO2 sink in the eastern Weddell Gyre, Biogeosciences, 5, 1373–1386, 2008.

Bakker, D. C., Pfeil, B., Landa, C. S., Metzl, N., O'Brien, K. M., Olsen, A., Smith, K., Cosca, C. E., Harasawa, S., Jones, S. D., Nakaoka, S.-i., Nojiri, Y., Schuster, U., Steinhoff, T., Sweeney, C., Takahashi, T. T., Tilbrook, B., Wada, C., Wanninkhof, R. H., 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. F., Kitidis, V., Körtzinger, A., Kozyr, A., Krasakopoulou, E., Kuwata, A., Landschützer, P., Lauvset, S. K., Lefèvre, N., Lo Monaco, C., Manke, A., Mathis, J. T., Merlivat, L., Millero, F. J., Monteiro, P. M. S., Munro, D. R., Murata, A., Newberger, T., Omar, A. M., Ono, T., Paterson, K., Pearce, D., Pierrot, D., Robbins, L. L., Saito, S., Salisbury, J., Schlitzer, R., Schneider, B., Schweitzer, R., Sieger, R., Skjelvan, I., Sullivan, K. F., Sutherland, S. C., Sutton, A. J., Tadokoro, K., Telszewski, M., Tuma, M., Van Heuven, S. M., 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, 8, 383–413, doi:10.5194/essd-8-383-2016, 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.

Boyd, P. W. and Ellwood, M. J.: The biogeochemical cycle of iron in the ocean, Nature Geoscience, 3, 675–682, doi:10.1038/ngeo964, http://www.nature.com/doifinder/10.1038/ngeo964http://dx.doi.org/10.1038/ngeo964, 2010.

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: Oceans, 120, 304–323, doi:10.1002/2014JC010203, http://doi.wiley.com/10.1002/2014JC010203, 2015.

Caruana, R. and Niculescu-Mizil, A.: An empirical comparison of supervised learning algorithms, Proceedings of the 23th International Conference on Machine Learning, pp. 161–168, doi:10.1145/1143844.1143865, http://doi.acm.org/10.1145/1143844.1143865, 2006.

Chierici, M., Signorini, S. R., Mattsdotter-Björk, M., Fransson, A., and Olsen, A.: Surface water fCO 2 algorithms for the high-latitude Pacific sector of the Southern Ocean, Remote Sensing of Environment, 119, 184–196, doi:10.1016/j.rse.2011.12.020, http://dx.doi.org/10.1016/j.rse.2011.12.020, 2012.

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:10.5194/essd-6-273-2014, 2014.

Friedrich, T. and Oschlies, A.: Neural network-based estimates of North Atlantic surface pCO2 from satellite data: A methodological study, Journal of Geophysical Research, 114, 1–12, doi:10.1029/2007JC004646, 2009.

Frölicher, 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, http://www.journals.cambridge.org/abstract{_}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.

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

Hoyer, S. and Hamman, J. J.: xarray: N-D labeled Arrays and Datasets in Python, Journal of Open Research Software, 5, 1–6, doi:10.5334/jors.148, http://openresearchsoftware.metajnl.com/articles/10.5334/jors.148/, 2017.

Johnson, K. S., Plant, J. N., Coletti, L. J., Jannasch, H. W., Sakamoto, C. M., Riser, S. C., Swift, D. D., Williams, N. L., Boss, E., Haëntjens, N., Talley, L. D., and Sarmiento, J. L.: Biogeochemical sensor performance in the SOCCOM profiling float array, Journal of Geophysical Research: Oceans, TBD, TBD, doi:10.1002/2017JC012838, http://doi.wiley.com/10.1002/2017JC012838, 2017.

Jones, S. D., Le Quéré, C., and Rdenbeck, C.: Autocorrelation characteristics of surface ocean pCO2 and air-sea CO2 fluxes, Global Biogeochemical Cycles, 26, 1–12, doi:10.1029/2010GB004017, 2012.

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., 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., 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., Van Heuven, S. M., Hoppema, M., Metzl, N., Sweeney, C., Takahashi, T. T., Tilbrook, B., and Wanninkhof, R. H.: The reinvigoration of the Southern Ocean carbon sink, Science, 349, 1221–1224, doi:10.1126/science.aab2620, http://science.sciencemag.org/content/349/6253/1221.abstracthttp://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 CO2 Sink Due to Recent Climate Change, Science, 316, 1735–1738, doi:10.1126/science.1137004, 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. R., 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. S., Munro, D. R., Nabel, J. E., Nakaoka, S.-i., O'Brien, K. M., 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. 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. T., Kuchinke, M., Matear, R. J., Roy, T., Sutherland, S. C., Sweeney, C., and Tilbrook, B.: The observed evolution of oceanic pCO2 and its drivers over the last two decades, Global Biogeochemical Cycles, 26, 1–14, doi:10.1029/2011GB004095, http://doi.wiley.com/10.1029/2011GB004095, 2012.

Lenton, A., Tilbrook, B., Law, R. M., Bakker, D. C., Doney, S. C., Gruber, N., Ishii, M., Hoppema, M., Lovenduski, N. S., Matear, R. J., McNeil, B. I., Metzl, N., Fletcher, S. E., Monteiro, P. M. S., Rödenbeck, C., Sweeney, C., and Takahashi, T. 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.

Loh, W.-y.: Classification and regression trees, Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 1, 14–23, doi:10.1002/widm.8, http://linkinghub.elsevier.com/retrieve/pii/0169743991801135http://doi.wiley.com/10.1002/widm.8, 2011.

Louppe, G.: Understanding random forests: from theory to practice, Ph.D. thesis, University of Liege, 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. N., Lee, T., Nguyen, A., Schodlok, M., Zhang, H., and J-M. Campin: ECCO2 : High Resolution Global Ocean and Sea Ice Data Synthesis, Mercator Ocean Quarterly Newsletter, 31, 13–21, http://www.mercator-ocean.fr/content/download/691/5904/version/1/file/lettre{_}31{_}en.pdf{#}page=13, 2008.

Met Office: Iris: A Python library for analysing and visualising meteorological and oceanographic data sets, Exeter, Devon, v1.2 edn., 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, 2006.

Mongwe, N. P., Chang, N., and Monteiro, P. M. S.: 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/ocemodhttp://dx.doi.org/10.1016/j.ocemod.2016.09.006http://www.elsevier.com/locate/ocemod, 2016.

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

Mountrakis, G., Im, J., and Ogole, C.: Support vector machines in remote sensing: A review, ISPRS Journal of Photogrammetry and Remote Sensing, 66, 247–259, doi:10.1016/j.isprsjprs.2010.11.001, http://dx.doi.org/10.1016/j.isprsjprs.2010.11.001, 2011.

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., Metzl, N., Olsen, A., Sabine, C. L., Cassar, N., Reum, F., Keeling, R. F., and Heimann, M.: Interannual sea-air CO2 flux variability from an observation-driven ocean mixed-layer scheme, Biogeosciences, 11, 4599–4613, doi:10.5194/bg-11-4599-2014, 2014.

Rödenbeck, C., Bakker, D. C., Gruber, N., Iida, Y., Jacobson, A. R., Jones, S. D., Landschützer, P., Metzl, N., Nakaoka, S., Olsen, A., Park, G. H., Peylin, P., Rodgers, K. B., Sasse, T. P., Schuster, U., Shutler, J. D., Valsala, V., Wanninkhof, R. H., and Zeng, J.: Data-based estimates of the ocean carbon sink variability - First results of the Surface Ocean pCO2 Mapping intercomparison (SOCOM), Biogeosciences, 12, 7251–7278, doi:10.5194/bg-12-7251-2015, 2015.

Russell, J. L., Sarmiento, J. L., Cullen, H., Hotinski, R., Johnson, K. S., Riser, S. C., and Talley, L. D.: The Southern Ocean Carbon and Climate Observations and Modeling Program (SOCCOM), Ocean Carbon and Biogeochemistry Article, 7, 1–28, http://rsta.royalsocietypublishing.org/content/372/2019/20130057.short, 2014.

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., Schölkopf, B., and Olkopf, B.: A Tutorial on Support Vector Regression, Statistics and Computing, 14, 199–222, doi:Doi 10.1023/B:Stco.0000035301.49549.88, 2004.

Tagliabue, A., Sallée, J.-B., Bowie, A. R., Lévy, M., Swart, S., and Boyd, P. W.: Surface-water iron supplies in the Southern Ocean sustained by deep winter mixing, Nature Geoscience, 7, 314–320, doi:10.1038/NGEO2101, 2014.

Takahashi, T. T., Sutherland, S. C., Sweeney, C., Poisson, A., Metzl, N., Tilbrook, B., Bates, N. R., Wanninkhof, R. H., Feely, R. A., Sabine, C. L., Olafsson, J., and Nojiri, Y.: Global sea-air CO2 flux based on climatological surface ocean pCO2, and seasonal biological and temperature effects, Deep-Sea Research Part II: Topical Studies in Oceanography, 49, 1601–1622, doi:10.1016/S0967-0645(02)00003-6, 2002.

Takahashi, T. T., Sweeney, C., Hales, B., Chipman, D. W., 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. S.: 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, 10, 988–999, doi:10.1109/72.788640, http://www.ncbi.nlm.nih.gov/pubmed/18252602, 1999.

Wanninkhof, R. H.: Relationship between wind speed and gas exchange over the ocean revisited, Limnology and Oceanography: Methods, 12, 351–362, doi:10.4319/lom.2014.12.351, http://doi.wiley.com/10.4319/lom.2014.12.351, 2014.

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. H., 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.