Biogeosciences Discuss.,
https://doi.org/10.5194/bg-2017-215-AC1, 2017

Biogeosciences
Discussions
Open Access
EGU

Interactive comment on "Empirical methods for the estimation of Southern Ocean CO_2: Support Vector and Random Forest Regression" *by* Luke Gregor et al.

Luke Gregor et al.

luke.gregor@uct.ac.za

Received and published: 13 September 2017

NOTE: Manuscript is added as supplimentary document

Contents

1 General comments to 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.

2 Response to R1

2.1 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 $\Delta pCO2$ adequately. Similarly, we assess the efficacy of including coordinate variables as proxies of $\Delta pCO2$ 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 $\Delta pCO2$, 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 $\Delta pCO2$ in the seasonally*

sparse Southern Ocean.

2.2 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.

  This discussion has changed significantly. We feel that this is now addresssed.

- 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.

3 Response to R2

3.1 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

3.2 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. CO2 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

Please also note the supplement to this comment:
https://www.biogeosciences-discuss.net/bg-2017-215/bg-2017-215-AC1-supplement.pdf

Interactive comment on Biogeosciences Discuss., https://doi.org/10.5194/bg-2017-215, 2017.