Biogeosciences Discuss.,
https://doi.org/10.5194/bg-2017-215-RC2, 2017

Biogeosciences
Discussions

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

Prof GRUBER (Referee)

nicolas.gruber@env.ethz.ch

Received and published: 29 July 2017

1 Summary

Gregor et al. introduce and test two new statistical methods to interpolate the relatively sparse surface ocean pCO_2 data in the Southern Ocean to produce time and space continuous distributions of this quantity, from which the air-sea CO_2 flux can be computed. Underlying both methods, i.e., Support Vector Regression (SVR) and Random Forest Regression (RFR), are machine-learning approaches, wherein non-linear relationships are established between a set of observed independent variables and the

target property surface ocean pCO_2. In addition to using standard evaluation metrics to determine the quality of the fits, the authors also use output from a model where the "truth" is known, permitting them also to assess the contribution of the sparse sampling. Gregor et al. find that both models perform overall well and comparable with some of the best performing models published. The RFR model tends to have overall a lower root-mean-square-error (RMSE) than SVR, although the SVR was performing better in a test where the predictability beyond the training data set was assessed. The ensemble of the two methods confirms recent findings that the Southern Ocean carbon sink appears to have recovered from a period of low uptake during the late 1990s by having strengthened substantially all the way into the 2010s.

2 Evaluation

This is a very valuable and important contribution to the long-standing quest of quantifying the ocean uptake of CO_2, particularly in a region that is undeniably key for determining the global uptake. While there exist already nearly two handful of approaches in the literature that pursue the same objective, the sophistication of the approach taken by Gregor et al., the depth of their analyses, and the independent checking of the result via synthetic data makes this a novel and highly relevant study. Therefore, I am very supportive of this manuscript and would love to see published as soon as possible. But before giving the final green light, there are a couple of major comments/concerns that I would like the authors to carefully consider.

- 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.

- 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 behoove the authors well to push the manuscript as far as possible in this direction.
- 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 CO_2 sink.
- 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.

3 Recommendation

I very much support the publication of this manuscript, but it requires a moderate revision before I am ready to fully endorse it. The revision needs to focus on the three

major issues identified above, i.e., more thorough analysis and discussion of method, and extended discussion of results.

4 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... ".

Abstract, p1, line 6: Typo. Change "The RFR as able" to "The RFR is able"

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.

Abstract, p1, line 8: "modeled environment". The commonly used expression here is "synthetic data".

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".

Abstract, p1, line 11: Add "a" to ratio, i.e., to read "with a lower ratio".

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.

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.

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 somewhat 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.

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?

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..."

Data transformation: p5, lines 9-10: "This then raises the question..." I found this somewhat 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.

Data transformation: p5, lines 16-20: I suggest to add here somewhere the time period that these data cover.

Empirical methods: p5, line 21, Data are plural. Thus "The data are split..."

SVR: 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 ..."

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.

RFR: p7, lines 1-6: As above, I also suggest here to better explain the method and the parameter choices.

RFR: p7, line 9 "The out-of-bag error is used to select the hyper-parameters..." This is extraordinary cryptic. Please explain better.

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.

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.

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.

Results, p8, Figure 3: I would love to see also the annual mean figure and its discussion added to the results section.

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.

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.

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.

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.

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?

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.

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.

Discussion, p12, lines 17-30: The conclusion stated on page 11 about the inclusion of a spatial variable should come, at the earliest here.

Discussion, p12, in general: There are many other things that need to be discussed here (see also my second major comment above).

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.

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.

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.

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.

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.

Nicolas Gruber July 2017

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