Review of Gregor et al: Empirical methods for the estimation of Southern Ocean CO2: Support Vector and Random Forest Regression. Submitted to Biogeosciences

Summary:

Gregor and colleagues present two novel data interpolation methods used to reconstruct the sea surface partial pressure of CO2 and further the air-sea exchange of CO2 from 1998-2014. In a first step, the authors compare their newly derived estimates to an existing estimate (SOM-FFN by Landschützer et al.) for 3 regions as defined by Fay and McKinley, whereas in the second step the authors combine their methods with output from a process model to estimate errors for both the domain where observations do and do not exist. In their work the authors discuss differences between their new estimates and the SOM-FFN method and they further suggest that the uncertainty of current estimates is likely underestimated.

Strengths:

The authors provide two complementary new estimates shedding light on the Southern Ocean as a carbon sink, as well as current limitations in estimating this sink. The study critically reflects on past research and discusses differences to their new estimates. While this paper is largely methodological, it also briefly discusses seasonal cycle and trends over the 1998-2014 period. Another strength of the paper is that the authors combine data-based and process–based modelling approaches to test whether the current observational network is sufficient to estimate the Southern Ocean carbon sink within a certain uncertainty level. The topic nicely fits the scope of the journal and I believe this paper is certainly of interest to the wider Biogeosciences readership.

Weaknesses:

During my review, I have encountered a few things that need clarification from the authors. They are listed from the most to least concerning. A full list of comments with line-numbers can be found at the end of this document:

- Method section: 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.

- 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.

- 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 a 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.

Recommendation:

I have found this study to be very interesting and to be of value to the readership. The results are interesting, as the Southern Ocean carbon uptake is a hot topic at the moments and the question arises to what extend we can estimate the sink using the sparse observations we have. The study shows to what extend methodological differences add to the uncertainty in estimates and it tries to show that ship-based estimates underestimate the overall uncertainty. While I do believe the study should be published I don't think it is publication-ready at the moment. The 3 points raised above are of major concern and at least points 1 and 2 need to be added and point 3 at least discussed before publication. I therefor **recommend major revisions of the manuscript**.

Specific and minor comments to the text:

Page 2 line 9: "were" – I suppose "where"

Page 2 line 10: "interannual pCO2 trends" – interannual trends? I suppose you mean interannual variability, otherwise please clarify

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.

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"

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.

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).

Page 7 lines 14-15: why Nightingale? there are newer transfer velocity estimates from Wanninkhof et al. (2013, 2014) using CCMP?

Page 8 Figure 3: It is confusing that the SOM-FFN method is called "SOM" here – please don't change abbreviations throughout the 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.

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.

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".