Review of Latto et al: The "Ocean Carbon States" Database: a proof-of-concept application of cluster analysis in the ocean carbon cycle. Submitted to ESSD

Summary:

Latto and Romanou present a cluster analysis in the North Atlantic and Southern Ocean using the K-means clustering approach. Using the analysis, the authors identify a subset of biogeochemical regimes, or as they call it "ocean carbon states". The authors combine the Takahashi et al sea surface pCO2 with NOAA OI SST to identify these carbon states in the North Atlantic and the Southern Ocean to then compare the clustering results with an ocean model. Using this comparison, the authors asses the model performance and identify model bias.

Strengths:

The authors combine both observation-based estimates with models. Using a statistical method, the authors provide a rather novel way to identify model biases.

Weaknesses:

Unfortunately, there are many things in the current manuscript that are misleading, need better discussion and would benefit from substantial revision. I will list them here from the most to the least concerning:

- K-means Method: The authors use k-means without data normalization. As the authors state, e.g. in the North Atlantic, the range of pCO2 is 50-450μatm whereas SST ranges from 2-30°C, i.e. the pCO2 range is an order of magnitude larger then SST, therefore, when distances are computed in "the Euclidian distance sense" the results will be biased towards the pCO2. The authors will need to provide some evidence that this is not problematic, discuss why it is favorable to bias towards pCO2 or to normalize the data first in order to give SST and pCO2 equal weight.

- Terminology: The authors ignore that the observation-based pCO2 and SST products are based on statistical interpolation methods as well. E.G the Takahashi climatology is created by interpolating observations using an advection-based interpolation algorithm, whereas the SST is interpolated using an optimal interpolation method. Therefore, the products are (a) NOT OBSERVATIONS as claimed in the text but OBSERVATION-BASED products and (b) they come with their own uncertainty. It is therefore questionable, given the data sparsity in the Southern Ocean e.g. to use the Takahashi preduct as "ground truth" (also given that in the Takahashi et al 2009 paper the authors themselves calculate a global flux uncertainty of 50%). This needs to be discussed instead of wrongly assuming observation-based product=observations.

- Discussion of method/parameter choice: Nowhere in the text it is properly discussed why pCO2 and SST are chosen, and why the reader should accept these proxies as representatives for processes in both ocean basins. Despite this, on page 2 line 29 the authors claim that pCO2 and SST are independent variables, which is just wrong. pCO2 is certainly not independent from SST. As Takahashi et

al 1993 and 2002 show, a change of 1°C in temperature results in a 4% pCO2 change due to the solubility effect.

- The authors present many figures, but provide too little explanation about their meaning. E.g. what is largely missing is a discussion on potential fields these clusters can be applied to.

- Introduction: The introduction is confusing rather than helping the reader build up the topic. The authors jump from paragraph to paragraph which to me seem to be very disconnected at times (e.g. paragraph 1 broadly discusses global warming, paragraph 2 jumps to gas exchange and paragraph 3 jumps to numerical simulations)

- Literature: I was disappointed that the authors missed to mention the already existing effort in using clustering techniques regarding the sea surface pCO2. Many studies (e.g. Lefevre et al 2005, Telszewski et al 2009, Sasse et al 2013, Landschützer et al 2013, 2014, Nakaoka et al 2013) use a self-organizing map (SOM) technique to build clusters in the surface ocean. Certainly, the aims of these other studies diverge from this one and certainly there are differences between AI methods (such as SOM) and K-means (despite the mathematical differences being actually very small), but nevertheless, the authors claim on page 3 that "To our knowledge, the ocean carbon cycle has not yet been evaluated using this technique" which might certainly hold true, but it behooves the authors well to at least discuss similar approaches to connect to the wider literature out there that indeed has applies similar methods for a similar purpose.

Recommendation:

While I value the effort and I certainly see the advantage of the technique and the resulting analysis, I believe the authors need to address the issues raised above before the manuscript can be considered for publication. I therefor **recommend at least major revisions of the manuscript**.

Specific and minor comments to the text:

Page 1 lines 21: "realistic, dynamical regimes" – I don't think the authors have shown anywhere that the regimes are "realistic"

Page 2 lines 24-31: More discussion is needed here. Furthermore, there is no citation backing the text.

Page 3 lines 1-4: I am confused here. I am familiar with the Fay and McKinley 2014 identification (not the Trochta et al 2015), and to the extent of my knowledge they do not "ignore the non-zonal, regional character of ocean biogeochemistry". Please explain. As it currently reads the statement is wrong.

Page 4 and following: Observation-based products, as the climatologies presented use observations and usually a statistical interpolation algorithm to fill data gaps in space

and time. Therefore, the final climatology cannot be called observation anymore, but rather observation-based!!!

Page 5 line 10: I suppose the authors mean wind speed at 10 meter height rather than surface wind speed. Most gas transfer estimates are based on the 10-meter wind speed (such as the used Wanninkhof 1992 formulation)

Page 5 line 13: The Wanninkhof 1992 formulation is outdated as also highlighted by the author in several following, more recent publications.

Page 5 line 16: The reference to Le Quéré et al 2015 should be replace with the original data reference (Dlugokencky and Tans 2014). The global carbon budget combines all measurements/estimates for the budget, but individual contributions, such as the atmospheric CO2 should be acknowledged when used (this is also noted on page 1 of the excel sheet provided by the Global Carbon Project).

Page 5 line 23: Please mention that the Takahashi grid is a simply 4x5 degree regular grid

Page 5 line 25: The Takahashi estimate excludes the arctic ocean north of 80N

Page 6: k-means clustering: Firstly, I think this would fit better in section 2. Secondly, the authors do not provide sufficient explanation: E.G. it is not clear to everyone how euklidian distances are calculated. Therefore, it is easy to miss that the authors actually bias towards pCO2 (see major comment). Other terms not explained include "centroid clusters", "gaining cluster" and "seeds". These are abstract terms that need to be understood by the readers. Understanding a method means trusting a method!

Conclusions: line 10-11: "accurately determine the optimal number of clusters for the cluster analysis" - I disagree given the methodological caveats raised above.

Conclusions lines 15-20: I cannot follow why the authors conclude that biases in salinity temperature and wind are responsible for the mismatch in the NA and nutrients as well as salinity is responsible for the SO mismatch. Firstly, this result is for this model only. E.G. Lenton et al. 2013 have shown that there is large disagreements in models even with regards to the seasonality in CO2 and the drivers of all sorts of variability. Secondly, given the uncertainty from the observation-based estimate I am not convinced this conclusion is solid.

References used in this review:

Lefèvre, N., Watson, A. J., and Watson, A. R.: A comparison of multiple regression and neural network techniques for mapping in situ pCO2 data, Tellus, 57B, 375–384, 2005

Telszewski, M., Chazottes, A., Schuster, U., Watson, A. J., Moulin, C., Bakker, D. C. E., González-Dávila, M., Johannessen, T., Körtzinger, A., Lüger, H., Olsen, A., Omar, A., Padin, X. A., Ríos, A. F., Steinhoff, T., Santana-Casiano, M., Wallace, D. W. R., and Wanninkhof, R.: Estimating the monthly pCO2 distribution in the North Atlantic using a self-organizing neural network, Biogeosciences, 6, 1405–1421, doi:10.5194/bg-6-1405- 2009, 2009

Sasse, T. P., McNeil, B. I., and Abramowitz, G.: A new constraint on global air-sea CO2 fluxes using bottle carbon data, Geophys. Res. Lett., 40, 1594–1599, doi:10.1002/grl.50342, 2013.

Landschützer, P., N. Gruber, D. C. E. Bakker, U. Schuster, S. Nakaoka, M. R. Payne, T. Sasse, and J. Zeng: A neural network-based estimate of the seasonal to inter-annual variability of the Atlantic Ocean carbon sink, Biogeosciences, 10, 7793–7815, doi:10.5194/bg-10-7793-2013, 2014

Landschützer, P., N. Gruber, D. C. E. Bakker, and U. Schuster: Recent variability of the global ocean carbon sink, Global Biogeochem. Cycles, 28, 927–949, doi:10.1002/2014GB004853, 2014

Nakaoka, S., Telszewski, M., Nojiri, Y., Yasunaka, S., Miyazaki, C., Mukai, H., and Usui, N.: Estimating temporal and spatial variation of ocean surface pCO2 in the North Pacific using a selforganizing map neural network technique, Biogeosciences, 10, 6093–6106, doi:10.5194/bg-10-6093-2013, 2013.

Takahashi, T., J. Olafson, J. Goddard, D. Chipman, and S. Sutherland:, Seasonal variations of CO2 and nutrients in the high-latitude surface oceans: A comparative study, Global Biogeochem. Cycles, 7(4), 843–878, 1993

Takahashi, T., et al.: Global sea-air CO2 flux based on climatological surface ocean pCO2, and seasonal biological and temperature effects, Deep-Sea Res. II, 49, 1601–1622, 2002