Response to referee comments
Paper #: essd-2017-113
Title: The "Ocean Carbon States" Database: a proof-of-concept application of cluster analysis in the ocean carbon cycle
Journal: Earth System Science Data

Reviewer #2

We wish to thank the reviewer for their thoughtful comments and helpful suggestions. We address below their comments (in bold), our responses shown in regular font. We have attempted to clarify any dubious statements and offer additional discussion of many points in the paper in a major rewrite of the manuscript.

General Comments

1. **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.**

We emphasize here (and in the new revised manuscript, page 6, lines 24-30, Fig. 1) that we are clustering 2d histograms, not raw scatterplots of pCO_2 and SST. The Euclidian distances are computed between the 12 monthly 2d histograms (Fig 2 in the text) and the centroid of each cluster. Therefore there is not really dependence on the range of each variable. Fig. 1, the schematic diagram of the methodology, now makes this point very clear.

Furthermore, we would like to point out that normalization of raw scatterplots before clustering them using k-means is important when one computes regression weights (as in Lefevre et al 2005). Here we are only interested in the pattern identification and characterization, hence it is not critical to normalize the data.

Lastly, we note the discussion in *Finding Groups in Data: An Introduction to Cluster Analysis* by Leonard Kaufman, Peter J. Rousseeuw. Wiley 2005.
(page 11):
"From a philosophical point of view, standardization does not really solve the problem. Indeed, the choice of measurement units gives rise to relative weights of the variables. Expressing a variable in smaller units will lead to a larger range for that variable, which will

then have a large effect on the resulting structure. On the other hand, by standardizing one attempts to give all variables an equal weight, in the hope of achieving objectivity. As such, it may be used by a practitioner who possesses no prior knowledge. However, it may well be that some variables are intrinsically more important than others in a particular application, and then the assignment of weights should be based on subject-matter knowledge (see, e.g., Abrahamowicz, 1985). On the other hand, there have been attempts to devise clustering techniques that are independent of the scale of the variables (Friedman and Rubin, 1967). The proposal of Hardy and Rasson (1982) is to search for a partition that minimizes the total volume of the convex hulls of the clusters. In principle such a method is invariant with respect to linear transformations of the data, but unfortunately no algorithm exists for its implementation (except for an approximation that is restricted to two dimensions). **Therefore, the dilemma of standardization appears unavoidable at present and the programs described in this book leave the choice up to the user.''** [emphasis our own]"

Additionally, in *Cluster Analysis for Applications* by Michael Anderberg. Academic Press 1973. (page 13):
"Some investigators recommend reducing all variables to standard form (zero mean and unit variance) at the outset. Such suggestions simplify the mechanics of analysis but constitute a complete abdication of the analyst's responsibilities and prerogatives to a mindless procedure. There are ways of dealing with heterogeneous data without surrendering control of the analysis."

As we feel the issue is important and will come up with the reader, all points above are now discussed in the paper, in a dedicated section Data Normalization (page 8).

Nevertheless, we have standardized the data and performed k-means analysis in the North Atlantic. We again find a preference for 3 regimes, as is visible in the 12 monthly 2D histograms (Figure 1 below), which is sensible since we are using an annual mean climatology. In Figure 2 (also below), we see the regimes as determined from the raw and the normalized data. They look somewhat similar, but we think that the normalized data do not provide some clear physical interpretation of the value ranges and are hard to map back again on the regional map of the North Atlantic. We do however recover the seasonal cycle. We feel that for the purposes of this study where we seek to identify how the histograms of the pairs pCO_2-SST change in time and space, it makes more sense to use non-normalized data.

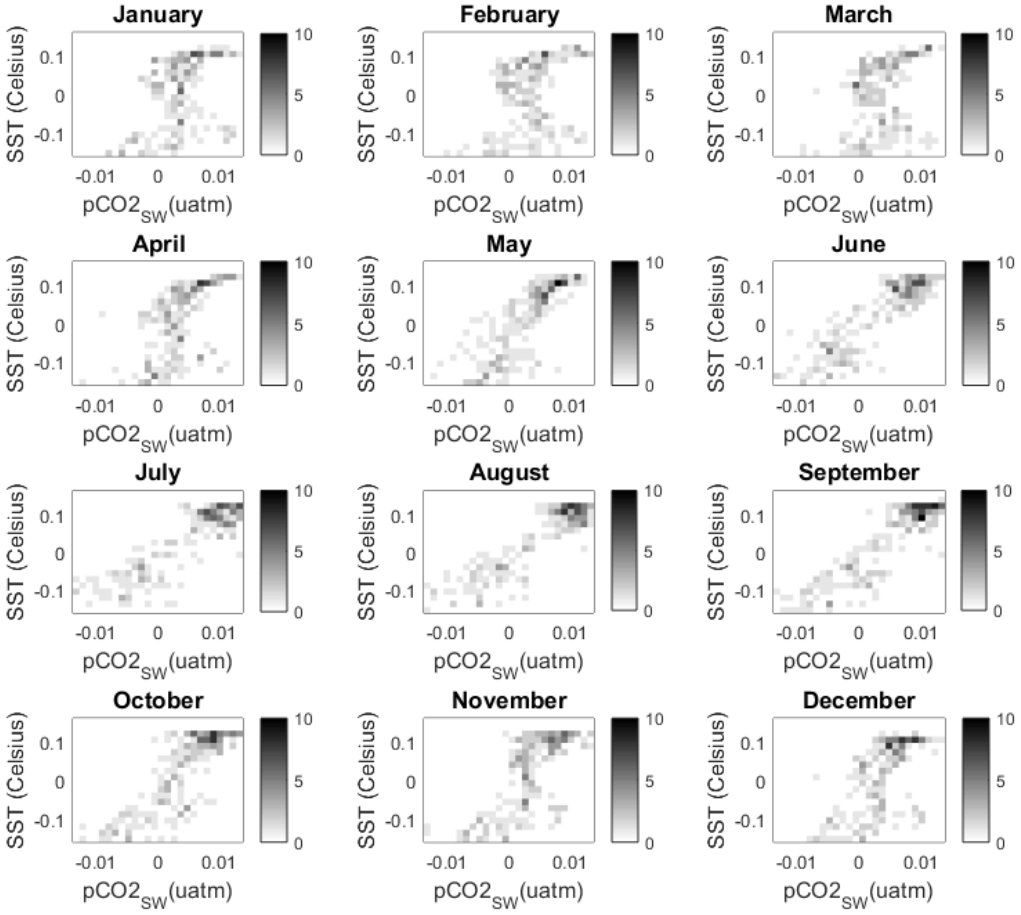

Figure 1: Monthly 2D histograms of partial pressure of CO2 and SST in the North Atlantic from the Takahashi dataset (normalized). The horizontal axis is pCO2 (uatm) and the vertical axis is SST (Celsius). The colorbar describes the actual frequency of occurrence of each bin.

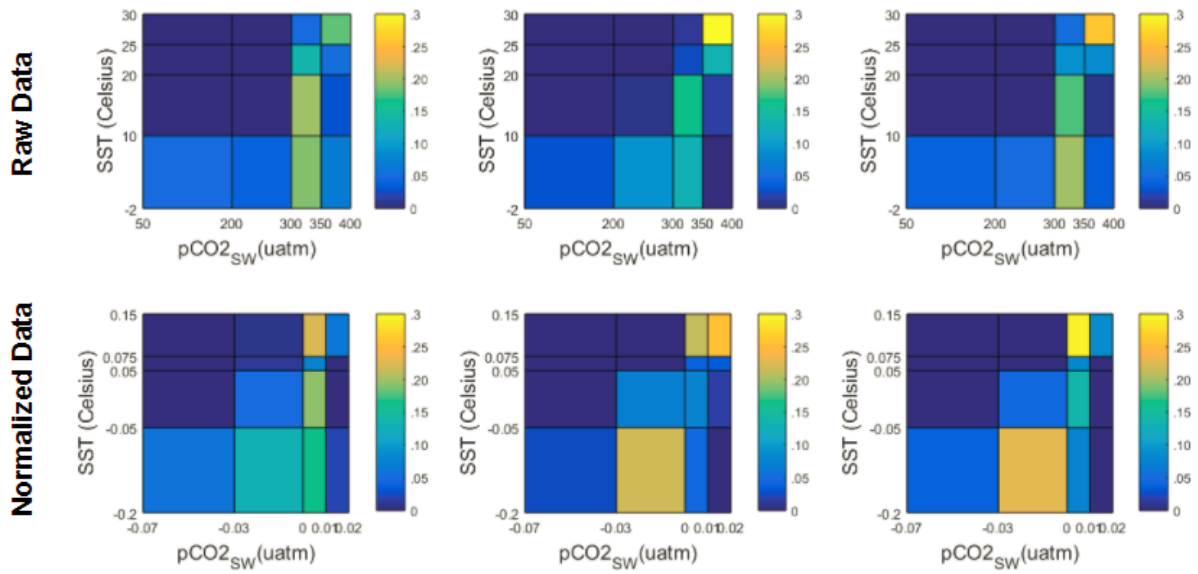

Figure 2: Top panel shows the cluster results from the paper (using a different color scheme but the same colorbar maximum). The normalized data panel shows the cluster regimes from normalized data. Note that the cluster regimes between raw and normalized data cannot be compared in a 1 to 1 framework.

References:

Anderberg, M. R, 1973. Cluster analysis for applications. New York: Academic Press.

Kaufman, L. and P. Rousseauw, 2005. Finding Groups in Data: An Introduction to Cluster Analysis. Hoboken, New Jersey: John Wiley & Sons, Inc.

Lefèvre, N., A.J. Watson, and A.R. Watson, 2005. A comparison of multiple regression and neural network techniques for mapping in situ pCO_2 data. Tellus B, 57, 375–384, doi:10.1111/j.1600-0889.2005.00164.x

References used by Kaufman and Rousseau, 2005.:

Abrahamowicz, M., 1985. The use of non-numerical a priori information for measuring dissimilarities, paper presented at the Fourth European Meeting of the Psychometric Society and the Classification Societies, 2-5 July, Cambridge (UK). [1.2.1]

Friedman, H.P. and J. Rubin, 1967. On Some Invariant Criteria for Grouping Data, Journal of the American Statistical Association, 62:320, 1159-1178,

https://doi.org/10.1080/01621459.1967.10500923

Rasson, J.-P. and A. Hardy, 1982. A Global Algorithm and a Criterion for Grouping Data, COMPSTAT, 129-180.

2. **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.**

We agree with the reviewer. It was an oversight on our part to keep using the term "observations" for the Takahashi climatology without qualification. We now make this point clear in the text (Section 2) and substitute the word "Observations" with "Observation-based" as much as possible. However, and as it is stated in the text (page 6, lines 15-19) in the Figure legends and captions we keep using the term "observations" for simplicity. Note, that this being a proof-of-concept study, we aim to show how we can use Regimes to help assess model bias. Any observational uncertainty will have to be taken into account when we discuss the model bias (as done here, Section on Bias Attribution, page 11-13).

3. **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 choice of pCO_2 and SST is discussed now in its own section on page 3. There we explain that we choose pCO_2 (as a representative of biogeochemical and biological processes) and SST (as a representative of physical processes) since both types of processes affect the

oceanic CO_2 uptake which is critical for climate studies. Of course, pCO_2 and SST are also influenced by the underlying ocean circulation, mixing, air-sea fluxes and SST directly affects pCO_2, but there significant differences in the two of them to render them, we believe, good choices for our clustering analysis. Others might choose a different pair of variables for different reason and that would be an interesting comparison with our study.

4. **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.**

There is now better discussion on the potential uses of k-means clustering both in the Introduction section (page 2, lines 11-30) but also in the Conclusions (page 18, lines 16-20).

5. **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)**

The Introduction is largely rewritten. We aimed to cover the importance of ocean carbon cycle, the novelty of using clustering analysis in the context of pattern recognition and the model evaluation part. References were made to other relevant works.

6. **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.**

The reviewer is correct. We do know of these studies and greatly appreciate their place in the ocean carbon cycle literature. However, we did not mention them because a) are not employing the same technique (k-means) and b) they use clustering for a different purpose.

Here we aim to find the regimes of co-variability of pCO_2 and SST and use them to assess model biases. This is an entirely different vein of analysis, using the same tools, indeed, but for another purpose. However, we agree it makes sense and it is rather useful to show that this technique is familiar to the community, just in a different context. This is now part of the new Introduction section (page 2-3 lines 31-3).

<u>Specific and minor comments to the text:</u>

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

We have toned down the language and instead of realistic show what the regimes correspond to in the sections for temporal and spatial attribution (page 9-10)

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

We have expanded this discussion now in a separate section (page 3, lines 20-31).

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.

The description of the Fay and McKinley paper as well others in the same vein are provided under a new description. Please see Introduction, page 3, first paragraph).

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!!!

Corrected throughout the text and discussed in the Data section as noted above.

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)

In the model, we use the surface wind speed, and by surface we mean the first model layer which is about 10m height.

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

Correct, however, in our model representation (CMIP5) that is what we used. In newer model runs we have updated the formulation.

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).

Corrected.

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

Done

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

Done

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!

We have rewritten this section, and in doing so removed any ambiguous language. We have defined Euclidian distance and taken extra effort to explain the method. However, we still keep the discussion of k-means clustering inside the North Atlantic section. We find it is easier for us to explain and possibly the reader to follow the method if we have an example at hand, which here is the application of the method in the North Atlantic data.

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

We tone down the language. We now claim that the sensitivity method is less subjective than pure visual inspection. We do recommend however always to consult visual inspection as the results might vary.

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.

We agree with the reviewer and we apologize for the confusion. We never meant to generalize our conclusions to other models. The biases in the GISS model salinity, temperature and wind are responsible for the flux bias in the North Atlantic and conversely for the Southern Ocean. We have now clarified the text (pages 12-13, lines 26-21 and page 16 lines 14-25).