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 #1

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
As currently written, it is unclear what advantages k-means clustering provide over other methods. It seems the main accomplishment from using this method was to split the North Atlantic into 3 zones, which the authors refer to as the tropical, subtropical, and subpolar regions, and to identify winter from summer, with transition periods in between. This is hardly news.

The aim of the paper is to present and appraise the k-means clustering method as an alternative to univariate analyses, such as principal component analysis, for pattern recognition in the ocean carbon cycle data and models. We call this a proof-of-concept test study of the method, because we are using a small climatological dataset with known variability. We expect to recover what we already know and maybe show that there is merit in using this method for larger more complex datasets. So in a sense, the fact that we obtain "old news" can be reassuring. This is now better explained in the abstract lines 28-31, and in the text, page 9, lines 20-23, and in the conclusions page 17 lines 19-25.

In the revised manuscript, we include specific statements to explain and highlight the advantages of the k-means clustering over other methods. The new statements are included in the Introduction (page 2, lines 16-20), and in the Conclusions (page 18, lines 16-20)). Namely, we note that the first advantage the k-means clustering brings is that, being a two-variable method, describes co-variability patterns of two fields, sea surface temperature and partial CO_2 pressure. Typically, other methods (eg the widely used principal component analysis or EOF is using one variable at a time).

We have also added new language in the text to help describe the temporal and the geographical regions obtained from k-means clustering. This is all now better explained in

page 9-10 under Temporal and Spatial attribution sections in the North Atlantic and in pages 14 and 15 for the Southern Ocean.

The arguments presented for choosing the number of clusters, i.e k=3 in this case, are also very arm-wavy. The number of clusters seems like it could vary as a function of the bin size used in the histograms, a binning which is very coarse, and is a function of subjective decisions by the authors.

Yes, we agree with the reviewer that there is an element of subjectivity when we choose *apriori* the optimum number of clusters. This is a known obstacle, which however researchers have used to their advantage to obtain better understanding of datasets and processes. This is not unusual though in statistical analysis. It is also the case with EOFs, for example, where it is rather subjective which modes to analyze. Usually people look where the variance is larger, e.g. modes 1 and 2 or maybe 3. But someone might choose to look at mode 4, relating it to a certain process, and thus obtaining additional information about the system. This is precisely one of the reasons that we chose to do a proof-of-concept study, to investigate whether we can obtain known regimes in the two basins. Nevertheless, we have offered different ways to evaluate which is the best choice of k. First, visual inspection is mandatory, as others have noted before, but also we attempt to create a less subjective criterion, the "sensitivity test" based on scores.

We have tested the sensitivity of our results using different binning choices for the two parameter space (see Figure 1 below). Fig. 1 top row shows the regimes as in the paper (4x4 bins), the temporal and the spatial attribution. Fig. 1 bottom row shows the regimes using much finer binning detail (21x27) as well as the temporal and the spatial attribution. We find that the temporal attribution in both cases is almost the same, the first regimes spans Jan-Apr, the second regime June-September and the third regime May, November and December, with October as the only shifted month. The spatial attribution differs (the colors are different) because we have different frequencies in different bins, but the relative importance of each area on the map (tropical/subtropical etc) is the same in both cases. The subpolar North Atlantic is more important in the first regime relatively to the tropical North Atlantic, etc. We therefore conclude that our method does not depend on the choice of bins.

In hindsight, this is to be expected because since the k-means clustering technique groups the 2D histograms and not the raw data, the binnings are not important. We must, of course, keep the same binnings when we compare the model with observations, but the initial choice of bins is not restrictive.
This is also mentioned in the text, page 7, lines 14-15 and page 9 lines 4-6.

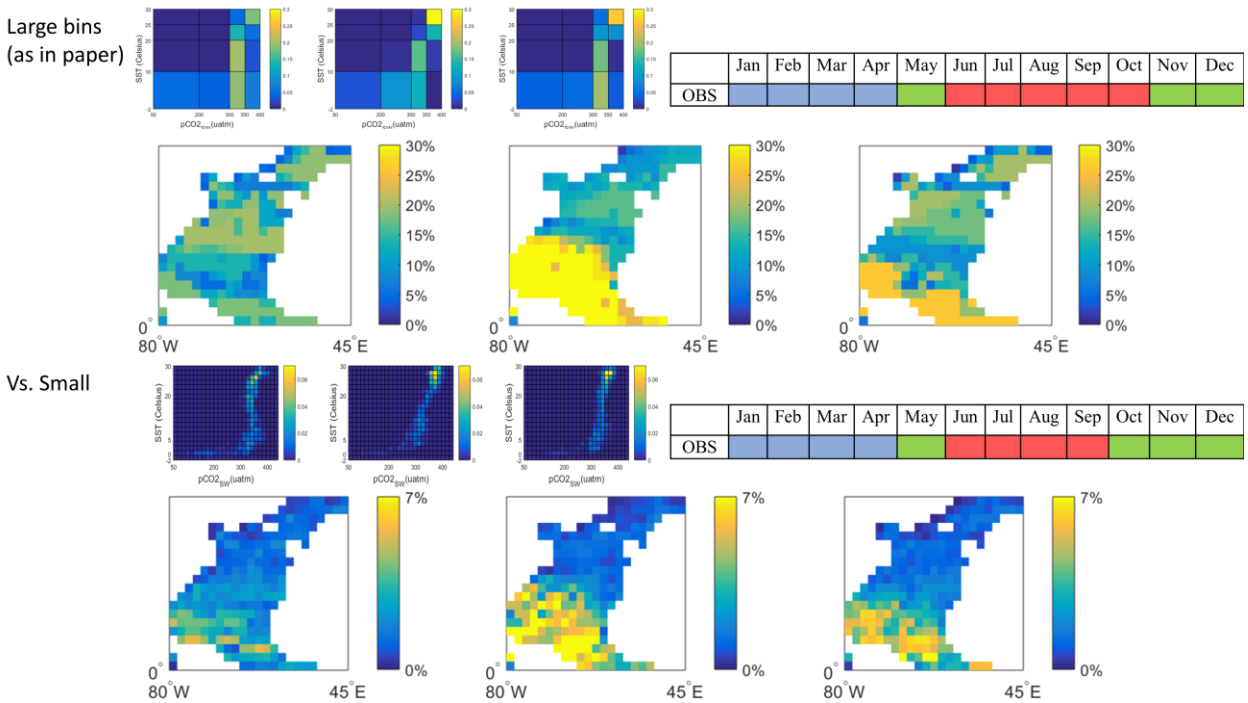

Figure 1: Effect of bin size change in the North Atlantic analysis. Top row: binning as in the manuscript: regimes, temporal and spatial attribution. Bottom row: larger number of bins (21x27), regimes, temporal and spatial attribution.

The paper contains also too many figures, 14 in the main text, and 15 in the supplement. Why some figures are put in the main text while others in the supplement is not totally clear from the presentation, though. It is also not clear at all what the main result is.

This being a proof-of-concept paper on this method, we want to provide as many figures as possible to help all readers (those from the carbon cycle science and oceanography as well as those of more data science background) assess the method. We did every effort to separate the really important ones in the main text. We have now grouped the main manuscript figures differently, to make more sense, brought some figures from the Supplemental Material (SM) to the main text, to help the reader and left out some figures out from the SM.

The main results are now better synopsized at the end of main sections of the paper. For example, abstract lines 27-31, page 10 lines 1-5, page 15 lines 21-26,and in the conclusions.

The paper is difficult to understand, the methodology not clear and I am left confused as to what the main point was and the merit of the method is. A comprehensive rethink of the manuscript is, in my view, necessary.

We have reorganized the paper and have given additional thought in the presentation of the method and the results. The revised manuscript is now largely re-cast to address these concerns, as it is also detailed in our previous responses to the reviewer.

<u>Specific comments</u>

P1, L9: pattern

Fixed.

P2, L22: column.

Fixed.

P2, L32: "Traditional methods of univariate analysis. . .". What are these methods exactly? Not sure what you are referring to here. K-means clustering can also be applied to a single variable.

The traditional methods we are referring to here are Empirical Orthogonal Function (EOF) Analyses/ Principle Component Analyses (PCA). These methods only look at the modes of variability of one variable. K-means cluster analysis for one variable would show how a single variable clusters in time and space but not in response to another variable. This is clarified in the revised manuscript in the Introduction page 2 lines 7-20.

P3, L6: I think kmeans clustering can also be applied to a single variable.

See above. We have also made note of this in the paper by referring to the method as a uni- or multivariate clustering method.

P3, L30: the actual address was https://data.giss.nasa.gov/oceans/carbonstates/

Fixed.

P6, L5: "in the North Atlantic basin"

Fixed.

P7, L6-p8-2: I find this explanation confusing and the conclusions of that section rather unconvincing and not objective at all. Based on what I understood from this section and fig 2b, I would select k=6. Also, the argument for picking k=3 because "Regime 4B and 4C appear to be almost equivalent" (L32) does not seem to hold as to me, regime 2A and 2B are also "almost equivalent". It is a mystery to me what sets of

rules the authors us to select their k. Also, the bin sizes of the histograms of Fig 2 are very large, spanning 5 degrees in temperature and 50uatm for the narrowest bins. What warrants this broad binning of the results in Fig 2? Presumably, the number of clusters k depends on the resolution of the data histogram (bin size). Finer bin sizes are able to pick out more patterns and so a higher k would be warranted. It is also not clear on what data the clustering algorithm is applied. Is it on each month independently or on all the data, or on some annual average?

We have addressed both the choice of k and the binning issues earlier in our responses to the reviewer. We have also re-written the text to be (hopefully) clearer to the reader as to what the procedure is. We have also added one additional figure (Figure 1) in the main paper to illustrate the methodology. We want to emphasize that the clustering (grouping) is done not on the raw data, but rather on the shape of the 2D histograms pCO_2-SST.

Essentially, the 12 2D histograms of the raw data are being clustered (using k-means method) into groups where the members of each group most closely resemble each other. We note that the sensitivity test, although not absolutely objective, it is more so than mere visual inspection. We do however always recommend visual inspection to ensure that the choice of clusters is reasonable. For example, simple visual inspection of Figure 2 shows that Jan-Feb-Mar-Apr are one group, June-July-Aug-Sept-Oct are another and finally May-Nov-Dec a third. The third cluster is somewhat not as clearly defined.

When looking at the actual cluster analysis output regimes, visual inspection is based not just on the distribution of the value-pairs pCO_2-SST but also on the frequency of the occurrence of those pairs. It is determined that 2A and 2B are not entirely equivalent based on this close inspection considering a higher frequency of pCO_2 and SST between 350-400 uatm and 25-30 Celsius in 2A than 2B (for example). Also, Regime 2B has higher frequency for SST between -2 and 20, and pCO2 between 300 and 350.

The bin sizes of the histograms are carefully chosen based on the 1D distributions of each variable, as seen in the supplementary figure 1 (Fig. S1). This method is subjective, but is fundamentally based on a statistically-minded approach where we attempt to assign even weight to each bin rather than have equal bins, for example, for low pCO_2 and low SST and high pCO_2 and high SST despite there being very few low values and much more high values that can be broken up so as to reveal more interesting patterns. Additionally, because we are estimating the frequency of the pairs, binning the 2d histograms differently will not alter the results of this analysis (see Figure 1 in this response).

P8, L15-17: "the seasons do not correspond to boreal seasons". This is a strange concept. How useful is it to pull together an entire region, define clusters, but then

ignore well-known events such as the Spring bloom, or spring restratification and sea ice melt. This also raises the question as to which region in the North Atlantic is the most variable and which region dominates the pCO2 and SST variability. What is gained in this analysis by removing the geographical aspect? What is the main conclusion of that section? Could it be that clusters identified in July have nothing to do with the clusters identified in February? Given that some features may dominate different regions at different times, does it even make time to link clusters in time? The only role of clusters is to minimized a statistical measure of misfit, how do you guarantee a logical link between clusters in time?

We only have a climatology, 12 monthly 2D histograms of pCO_2 and SST, therefore it is indeed probable that we obtain the seasonal cycle. We in fact identify what might be the "spring bloom" in the clusters (winter time regime in the North Atlantic) and discuss it (page 17 lines 20-22), although we acknowledge that it is not necessarily expected that we see the bloom given that we are not clustering with chlorophyll. However, inspection of Figure 5c shows that we do have a hint of that at high latitudes. This is now noted in the text. Other physical interpretations could be offered here. We wish that this kind of analysis will motivate others to also look at the regimes and help interpret them.

The purpose of the paper is not to identify all possible underlying processes, but at least point to the fact that the method can be used to analyze the data in different ways than other statistical methods.

For this proof of concept study, it is important to demonstrate that we can find a certain number of clusters using an unsupervised training methodology, and then be able to parameterize the clusters by well-known events because then we can show that we are producing realistic regimes. Although we do not find clusters that represent both July and February, if we had, this would show that similar modes and physical processes link pCO_2 and SST during these months. It should be clarified that we are not linking clusters with time, but rather characterize the clusters by their temporal occurrence time. This means that the cluster analysis automatically outputs which monthly 2D histograms are associated with which cluster, and this information can be used to explain the cluster outputs and which well-known events in those months could be driving pCO_2-SST relationships.

In regards to the geographical aspect, we do not remove it, but actually use it as another aid to help explain what phenomenon or process each regime represents. Figure 5c shows how we iterate through each cluster output to determine where regimes are occurring. The clusters are meant to capture any variability and to parse the region so that it can be

learned when and where specific regions of the North Atlantic basin dominate pCO_2-SST variability.

P8, L26: Contrary to what is stated here, it doesn't look like the clusters in the observations and in the model look very much alike. The magnitudes of the color bar is quite different between the two.

We discuss in the text in more detail how the model clusters compare with observations for both the North Atlantic and the Southern Ocean regions. With regards to the actual regimes (Figs. 5a and 7a) the shapes are similar and although the absolute frequencies differ in magnitude between the model and the observations, the relative importance of the frequencies is the same. This is all discussed in the text now, see page 10-11, lines 26-11. The model regimes are similar to the observed in terms of temporal occurrence and spatial attribution (e.g. the winter regime in the subtropical North Atlantic) (pg 11).

P8, L27-27: "the same bins of the most likely values are identified. . .". This seems to be the case in all the analyses so far. Not clear how this statement can be used to justify the previous statement that obs and model agree in that context.

The same pCO_2-SST bins with the highest frequency of occurrence in the observation and model regimes tend to correspond. An example is that in the summer regime, both model and observations show high frequency of occurrence for high pCO_2- high SST (350-400 uatm, 25-30 Celsius) and also mid-range value pairs (300-350 uatm, 10- 20 Celsius). Even though the actual values (magnitude) of the frequency of occurrence are different, these value pairs have the highest frequency of occurrence within their own cluster. This is better now explained in the text (page 10-11, lines 26-11)

P8, L28: Comparison between Fig 4 and Fig S4 shows the cluster variability is totally different between the two cases. Is the main conclusion here simply that there are seasons in the data and there are seasons in the model?

Old Figures 4 and S4 are now presented differently to be more clear (see Figs 5b and 7b). When we present the temporal attribution this way, it is more evident how the observed and model regimes correspond.

P9, L21-23: not clear what you mean by "composited". Also the rationale for doing this totally eludes me. What is gained/lost from doing this?

Composited here is used in the sense of "averaged over" each regime. So, the winter regime based on the observational data includes the months J-F-M-A, when we composite we

average over these months. We composite the model data on the observational winter regime, in order to better understand the model biases.

P9, whole section 3.1.3: What is gained from this analysis that is not achievable simply by looking at the observed and simulated CO2 fluxes? What information does k-means clustering contribute here? It would seem that a similar analysis done on each grid box would result in a much better and detailed analysis than if the domain is first split into various domains.

It is true, this type of model error analysis is possible without clustering analysis or when using a different analysis method. However, doing this type of error analysis in every grid point at every temporal scale would be rather cumbersome, especially for larger, more complex datasets. What clustering does is to group together similar states (the regimes) and then we can only apply error analysis not at every grid point and every time scale, but rather at each regime. For a small dataset, it is overkill to do clustering, but we are using it here as a proof-of-concept study.

P12: same issues here in the Southern ocean as above for the North Atlantic about the arguments for choosing k=3.

See above. These are addressed in the revised manuscript in part 4.2.

Figures:

Fig 1: poor choice of color scheme. Hard to distinguish between dark blue/light blue. Are all dark blue squares zero or non-zero?

We have reworked the figure now with different color scheme where non color squares correspond to zero. This is now Figure 2.

Fig 8-9: Since pCO2 is calculated from other variables, including SST, SSS, WSPD and NIT, what is the point of performing the analysis in two steps (i.e. fig 8)? It seems that Fig 9 simply shows that biases in SST, SSS and WSPD dominate the full bias, since pCO2 is just a function of these variables

We want to make the point that in a climate model, the air-sea flux of CO_2 is critical, because it modifies the atmospheric CO_2 and hence equilibrium and transient climate sensitivity. Fig. 9 shows that pCO_2 bias is largely responsible for flux biases in the model. Then Fig. 10 digs deeper to help us understand where the pCO_2 bias comes from. We agree

with the reviewer, pCO_2 depends on all these variables but in a non-linear way, hence the bias attribution needs to be done per regime and per region.