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 #3

We wish to thank the reviewer for their thoughtful comments and suggestions. We address each of their comments below (in bold) and our responses are in regular font. We have largely rewritten the text, focusing more on addressing some of the key concerns of the reviewer, which we outline below.

General comments
My main concern here is the apparent subjectivity is this determination, described in "Sensitivity to predefined number of clusters". The optimal number of clusters is determined *"…by identifying the K with the highest score and no significant change of the score thereafter."* (page 7, line 21). This is based on a visual inspection of Fig.2b, 10b, S3b, and S10b. Regarding the case for the North Atlantic, the authors mention that the optimal K number is 3 for both observations and model output. Based alone on the visual assessment of Fig. S3b, I am inclined to say that 5 is a more appropriate number of K for model outputs. How would this discrepancy affect your results? Furthermore, it seems convenient that in both cases, North Atlantic and Southern Ocean, the optimal number of K in model and observations is the same, how would you compare model and observations if they had a different number of K?

We agree with the reviewer that k-means clustering, as well as any other statistical grouping technique, involves a highly subjective element which is the *apriori* choice of number of clusters. This is a known issue, but the community finds that it offers an opportunity for further exploration of the data set as there is always the option of increasing k and trying to find meaning to additional clusters. Note, this is an issue with EOF analysis as well, which mode does one stop the analysis on? Often the third mode is not taken into account, although in other cases it is proved useful. In any case, we have toned down the language and refer to our Sensitivity test as a "less subjective" way of determining k than for example visual inspection. This is now better discussed in the text in a separate section (page 7-8). We have discussed in more detail in order to better justify the choices for k. It needs to be emphasized that our dataset (a 12 month climatology) is really small and it does not make much sense to have clusters with only one member. Presumably, in applications to higher order datasets, these additional clusters might emerge more clearly, as they will express interannual variability or longer term trends.

Specific comments:

1. **Abstract: This is a problematic section. There is a lot of emphasis on the utility of data- mining techniques and little or none mention of the biogeochemical findings regarding model biases in the North Atlantic or the Southern Ocean. It is not clear what are the main findings/conclusions of the study.**

The study aims to present k-means clustering as a data-mining technique to analyze observational data and model output. This is the first study of the sort, for ocean carbon cycle. Hence the emphasis on the data-mining techniques, given that we feel the main audience might be unfamiliar with these techniques. However, we agree with the reviewer that some description of the results of applying this method is needed, and have added it in the revised manuscript. We have reworked the abstract.

2. **Methods, Section 3.1.1: It took me several reading attempts to understand the way in which the K-means clustering works. It may be useful to include a diagram or include equations describing the iterative clustering process. As mentioned above, the method used to determine the final number of K clusters needs to be improved, or at least, better justified.**

Figure 1 now is a schematic diagram of the methodology. We definitely agree with the reviewer that it helps clarify this section. There is a section dedicated to the choice of k as well.

3. **Conclusions: *"A method is provided for an objective way to accurately determine the optimal number of clusters for the cluster analysis."* I highly disagree with this statement based on my Major scientific comments, above.**

Agreed. We have reworded the conclusion in order to suggest that there is some corroboration from the less subjective method we devised here. We also note that visual inspection is always recommended to ensure that the resulting regimes are distinct and meaningful.

4. **The discussion of the attribution of model errors is well written. However, the conclusions are compromised by the subjective determination of the number of K.**

We are glad that the reviewer appreciates this section, as we also think this is an interesting way to study the regimes in models and help improve model biases. However, we think that the choice of k is a necessary "evil", which many researchers find useful as

well, as by trying to interpret additional clusters for higher values of k might identify key processes we have overlooked. Again the analogy to EOFs is revealing, there is a subjective choice there to how many modes of variability one considers.

5. **Figures. The distribution of figures between the main manuscript and the supplementary material is confusing. I recommend merging the frequency histograms in figures S1 and S8 in one figure, and including it in the main manuscript. In general figures are duplicated for the North Atlantic and Southern Ocean. With the exception of the 2D histograms, I recommend merging some of the similar figures for different regions (e.g., Fig 2 and Fig 10.), and also include in the main manuscript the analysis of the Southern Ocean (Fig S9, S10, S11, and S13). Fig S2 is not clear. It seems that some of the lines in the label are not included in the figure (e.g., K=2 and K=9 are both dark blue, but there is only one black blue line in the figure, and they would be impossible to distinguish). Fig 2a, 10a, 3a, and 10a: For more clarity, it would help to label the month corresponding to each bar.**

We have cleaned up the figures, merged several ones and generally added more discussion to all of them. We would like to keep the two regions separate. The reason is that 1) the regimes are different, 2) having 4 panels of regimes (2 for obs and 2 for model results) can get confusing and 3) the GISS model has much more important biases in the Southern Ocean than in the North Atlantic, in addition to the fact that the biases have different causes in the two basins. We have corrected the convergence diagram in the Supplemental Material.

6. **Page 10, Line 21: *"As shown in Fig. S5, we identify a subpolar region..."* What is the basis for the determination of these sub-regions in the North Atlantic and Southern Ocean? If they are arbitrarily chosen, "*we define*", would be a more appropriate wording. The criteria used to define these regions is important in the final attribution of processes driving the errors between models and observations.**

Agreed, the language is a bit confusing, so we have reworded the text. However, the subregions are not arbitrary, they are based on the groups of bins in the regimes (Fig 5a, 7a, 14).

7. **Page 11, Line 25: "*Here, nitrate biases are probably due to misrepresentation of nitrogen fixation in the GISS climate model...*" In oligotrophic regions, relative errors in nitrate may be higher as the absolute concentration is overall lower, with much less seasonal amplitude when compared with higher latitudes. I agree**

that nitrogen fixation is likely a problem, but it might be useful to consider this aspect as well (see Arteaga et al, 2015, GRL, doi:10.1002/2014GL062937).

Indeed, this is an important point. We include it now in the text along with the citation.

8. **Page 12, Line 29: "*For almost all regimes and regions, biases in nitrate are large partly because of lack of a closed, state-of-the art nitrogen cycle representation in the climate model. However, observations are too scarce in the region, due to inclement weather and biases to specific seasons, so the model skill would be more adequately assessed as more in situ measurements are made (e.g. from the SOCCOM experiment; Johnson et al., 2017).*" This is also problematic regarding observations of CO2 flux, particularly in winter. How do you account for a possible bias towards summer fluxes due to a lower amount of observations during winter months?**

In this analysis, we consider that the observations are "perfect" but we do recognize that there is sparcity of measurements especially in the Southern Ocean. Takahashi et al 2009 discuss the bias of the flux and pCO2 in the Southern Ocean. The Error Attribution sections in our paper estimate the model bias based on whatever observations we have. It is beyond the scope of this paper to develop a method to assess model bias when the observations are sparse. It would be a very interesting work however and would provide a better estimate of our model bias.