Earth Syst. Sci. Data Discuss.,
https://doi.org/10.5194/essd-2017-113-RC1, 2017

Earth System Science Data Discussions

Interactive comment on "The "Ocean Carbon States" Database: a proof-of-concept application of cluster analysis in the ocean carbon cycle" *by* Rebecca Latto and Anastasia Romanou

Anonymous Referee #1

Received and published: 28 November 2017

In this study, the authors use a technique called k-means clustering to delineate oceanic regions based on SST and pCO2 data (and model). They then use these domains to look for the likely sources of model biases on pCO2 fluxes in these clusters. They conclude that wind speed is a major culprit in the North Atlantic, while a faulty biology (as diagnosed by nitrate) is likely the main source of error in the Southern Ocean.

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 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.

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.

Overall, I would not recommend publication of this manuscript in its current state. 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.

Specific comments

P1, L9: pattern

P2, L22: column.

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.

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

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

P6, L5: "in the North Atlantic basin"

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?

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?

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.

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.

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?

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?

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.

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

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?

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

Interactive comment on Earth Syst. Sci. Data Discuss., https://doi.org/10.5194/essd-2017-113, 2017.