Comment on esd-2021-16
Anonymous Referee #1

Referee comment on "Exploring the ocean and atmosphere coupled system with a data science approach applied to observations from the Antarctic Circumnavigation Expedition" by Sebastian Landwehr et al., Earth Syst. Dynam. Discuss., https://doi.org/10.5194/esd-2021-16-RC1, 2021

Summary and overall impression

This manuscript makes use of a large interdisciplinary dataset from the Antarctic Circumnavigation Expedition, a 90-day cruise from December 2016 to March 2017, in combination with the sparse PCA (sPCA) method to extract process understanding from this comprehensive dataset. The study has a very broad scope, aiming to obtain a holistic understanding of the process biogeochemical and physical processes in the Southern Ocean and atmosphere. The method (sPCA), goes beyond standard PCAs, which are commonly used in oceanography and meteorology. sPCAs aim to increase interpretability when dealing with many variables and processes. In addition, the authors apply a bootstrapping approach in order to quantify the uncertainty of their sPCA results.

I find this a very exciting study and it has the potential to be relevant and valuable to the community. I see three main strengths of the manuscript. First, it presents a method (sPCA) that is relatively new in Earth System Science and may be useful for further studies analyzing ship data. Second, the method allows the authors to conduct an extremely multidisciplinary analysis including a broad range of observed variables and are able to extract an understanding of the dominant processes in the study region. Third, the study is based on a new comprehensive observational dataset from a historically under-sampled region (the Southern Ocean), and includes measurements in the ocean, atmosphere, and cryosphere, covering all sectors, different interfrontal zones, both open ocean and near islands and continents, and covering a broad range of physical and biogeochemical variables.

At the same time, I have several major comments that I believe need to be addressed before publication. My main concern is the description of the method. I have to admit that I am not too familiar with standard PCAs, and sPCAs are completely new to me. Assuming that this may be the same for many readers, I believe the manuscript can gain considerable clarity by improving the description of the methods (see general comments for more specific details on this and other major comments).

General comments
**sPCA method:** I suggest expanding Section 3.1. I would appreciate a discussion on why setting some weights to zero is ok and why this does not lose crucial information. In a standard PCA, we say e.g., 80% of the variability is linked to OV1, 5% each to OV2 and OV3. We then know that there is a remaining 10% of variability due to other processes. With sPCA (the way I understand it from the manuscript) we reduce the complexity, ignoring some variables, to explain all of the remaining variability. Here, we get to 100%, but we actually know that we weighted many variables with 0 in order to do so. Isn't the standard approach more complete in its interpretation? What are the pros and cons of each? It should also be mentioned if the user chooses which weights are set to zero, or if the algorithm does that. (My apologies if I have misunderstood the sPCA method. If that is the case, I suggest you clarify it).

**LVs:** I find the current explanation of what an LV is quite confusing (L85-87), which led to further confusion later in the document. I recommend making it really clear here what an LV is in an sPCA and how it is different to the OVs. I recommend explicitly stating that the LVs are the processes we want to understand (i.e., the output from the sPCA) with the help of OVs (i.e., the input to the sPCA). (It becomes clearer later in the document, but is needed early on). Please add a section that summarizes what happens during the sPCA to add clarity on the method for people unfamiliar with it. The way I understand how the sPCA works from your manuscript, the user chooses a set number of processes they want to know about (here: 14), feeds all OVs (here: 111) into the algorithm. Some of the OVs are set weighted 0 to reduce the number of OVs for each LV. (☐ This should be discussed and mentioned if this happens randomly.) The algorithm then identifies 14 different sets of OVs. The users then see which OVs have non-zero weights in each LV to determine which process each LV represents. i.e. the user has to make a choice: if sea surface temperature, salinity, and MLD are OVs in an LV, then the LV might represent a process linked to ocean circulation. (☐ For each LV, it would be good to know which OVs are in it so that the reader can understand how the label for each LV was chosen). We can then also see the percentage of the variability that process has on the variability in all of the 111 variables.

☐ Is this correct? If yes, it might give you hints about which pieces of information the reader might want to hear about. If not, my understood explanation might give you hints about which parts were confusing.

Unimportant variables for an LV “are forced to be zero”: could we accidentally lose information here? Is this a subjective choice by the authors or done by the algorithm? This should be discussed further.

**Research Question(s):** Another concern is linked to the research question(s) the article wants to answer. It is such a broad study that scratches on so many topics that it becomes a bit blurry in the introduction where this is all going. The way it is currently presented, it appears as a data mining approach of plugging in all the data and seeing what happens. Were there some hypothesis before that you wanted to test? I would find it helpful to add a (couple of) specific research question(s) and build on that in the introduction why we want to know about that. E.g., Is it about the processes? Is it about showing that sPCAs are a good tool? (or both). Are there some processes we are unsure about, which the sPCA might shine a light on?

Linked to my previous comment: it is not clear to me which findings are confirmations of processes we already knew, and which findings are new insights. This should be clarified.

**Eddies:** One process that doesn’t seem to be covered in this study, but is a known driver behind variability in the Southern Ocean are mesoscale eddies. This should be discussed.
**Seasonality:** Please add a discussion on the fact that the cruise is only 90 days long (i.e., during one season) and that the ship is moving during that time, making it difficult (or impossible?) to conduct a seasonal analysis. The discussion should include why it is possible (or not possible?) to robustly conclude on any seasonal signals with this data.

**Specific and minor comments to the text:**

L. 131: In this section, I would have liked to also find out a bit more about the measurements, e.g., if the ocean measurement are at the sea surface only (same for atmosphere) and I recommend adding a sentence or two stating the nature of the measurements (sensors, air/water/ice samples... were some data collected by platforms other than the ship, such as satellites/planes...?).

L. 145: It should also be mentioned here (and possibly in the abstract/introduction) that this is an unsupervised machine learning approach (as stated in the Conclusion).