Summary of Revisions

RC = Reviewer comment

We thank all three reviewers for their positive and constructive feedback. In order to provide a quick overview of the changes to the to-be-revised manuscript, we give a summary here:

- The title has been changed to: “Exploring the ocean and atmosphere coupled system with a data science approach applied to observations from the Antarctic Circumnavigation Expedition” (following RC3.3).
- We have added research questions in the introduction for a framework that better structures the manuscript as a whole (following RC1.6).
- The methods description has been revised substantially to make the language more accessible to non-data scientists (following the general and several targeted comments of Reviewer #1).
- Section 5 (description of individual LVs) will be moved to a new appendix A to substantially shorten the manuscript. We now summarize the outcome of all LVs briefly in a revised section 4.1, and highlight the novel aspects we found there as well. We give give one condensed description of LV9 as example in a revised section 4.2. (following RC1.7, 3.1, 3.4)

RC = Reviewer Comment, AC = Author Comment, new suggested text in blue

Answers to Reviewer 2

Anonymous Referee #2, 01 Jun 2021

Comments on "Biogeochemistry and Physics of the Southern Ocean-Atmosphere System Explored With Data Science" by Landwehr et al.

RC2.1: This manuscript presents a detailed exploration of a very large ensemble of measurements of in-situ variables from the Southern Ocean and from the Southern Atmosphere. It emphasizes the technique of sparse principal component analysis which indicates possible causal relationships and tries to identify underlying processes explaining how the variations of the observed variables. As it is now the manuscript is well written but it could benefit from incorporating the minor remarks I have below. I also propose to more clearly delineate the advantages of the sPCA method to guide the reader about the choice of analysis made here.

AC2.1: We thank the reviewer very much for their positive and constructive remarks. We address all comments below in detail.

Major comments:

RC2.2: The following sentence at the end of the discussion (Page 60, lines 1102-1103) would need to be better backed up by the authors: “In summary, we find that the sPCA is not only capable of resolving many of the complex connections between the OVs (Observed Variables) but also to provide estimates of their relative importance for the observed variability of each OV.”

AC2.2: The reviewer makes a good point. Rereading the section and the first summary sentence, we find that this particular sentence is misplaced here and partly non-sensical. This is because the complex connections are discussed in detail in the individual LV descriptions, where we highlight a number of processes involving several OVs, but this is not the topic of this section 6.3. The second half of the sentence makes grammatically no sense, because “their” refers to “OVs” resulting in “... but also to provide estimates of OVs’ relative importance for the observed variability of each OV.” What we meant to say is that the occurrence of OVs in several LVs provides insight into where the OV variability might stem from. In order not to repeat information from the individual OV discussions and the section 6.3, we removed this sentence and start the summary as follows:

“In summary, we find that state variables of the environment such as the air-sea temperature difference,...”

RC2.3: I would welcome a paragraph stating, with possible examples from the results and the discussion, the strengths of the sPCA method. The weaknesses are well described but the reader would also like to have the view of the authors on what guided them to select this method for an analysis.

AC2.3: This is an important point, which we apparently did not communicate very clearly. To make this clearer, we have added the following short paragraph in section 3.5 for a general description of the advantages, and a second paragraph in the introduction to Section “6 Discussion” in l. 982. The discussion section actually highlights some of the aspects which we consider to be key advantages, that is the identification of “hotspots” and of “key OVs”. The attribution of a number of processes that explain the variability of each OV is discussed further upfront in the manuscript and is shown in Fig. 5. And last, but not least, the exploratory character of the sPCA allows researchers to conduct an untargeted analysis and potentially find relationships or (spatial / temporal) patterns which would have been left undiscovered in a targeted analysis because one did not think of the possibility.

- L277, new paragraph: "The main advantage of the sPCA approach over its standard counterpart is the automatic selection of OVs by assigning non-zero weights for a given LV. The automatic optimisation of the weights associated with the OVs is done sequentially for each LV, starting from the one corresponding to the largest mode of variance. This ensures that, although not exactly, all the LVs are as uncorrelated as possible. The use of sPCA has also the advantage of being less susceptible to noise and unimportant data variations. This advantage can be understood when contrasting the sPCA results with the large number of principal components with very low explained variance of the standard PCA. Although by considering these components the standard PCA is able to fully explain the data variance, such variance directions are of little practical use in our case, as it would be difficult to link them to natural processes. Compared to the standard PCA, sPCA is less likely to return components with very small explained variance, which are usually corresponding to noise. This advantage is further strengthened by our novel use of the bootstrap analysis, which promotes robustness to noise, meaning that OVs which contribute mainly through noise are identified as such. Data is resampled randomly, and the influence of noise can be observed in large fluctuations of the solution. Therefore, analyses relying on aggregated bootstrapped solutions are more robust to the influence of noise than the traditional PCA or even a single run sPCA. Moreover, using sPCA over the standard PCA has also the benefit of not being susceptible to rank-deficient covariance matrices, in particular when the number of data points is smaller compared to the number of OVs. And last, but not least, the exploratory character of the sPCA allows researchers to conduct an untargeted analysis and potentially find relationships or (spatial / temporal) patterns which would have been left undiscovered in a targeted analysis because one did not think of the possibility."

- L.982, new paragraph: “The key strengths of the method are: (a) Sparse PCA has an untargeted exploratory character, i.e. the possibility of relating many different OVs with each other and identifying correlations, which one might not intuitively address in a targeted analysis. (b) Because sPCA can easily relate geographical information with all OVs, it is possible to explore spatial patterns and obtain a geographic overview. This also allows us to identify geographical hotspots, as discussed in Section 6.1. (c) Sparse PCA can help to identify original variables which are key to many processes, as discussed in section 6.3. Due to the possibility of exploring a large number of OVs at the same time, it becomes straightforward to isolate those OVs that stand out. And finally (d), we can explore which processes (LVs) contribute to explaining the variability of the OVs, as is shown in Fig. 5.”

RC2.4: The distance to the continent (Latent Variable 5, LV5) is not the best indicator of land influences as the authors seem to suggest. A much better indicator would be a ^{222}Rn concentration measurement. Radon-222 is a radiogenic gas which emission flux is 100 times more important over land than over ocean. As such, you can use the concentration of ^{222}Rn to trace how long ago an air parcel was over a continent. Several authors have used this property as a measure of the continental influence of an air parcel travelling over the ocean (Heimann et al. (1990) Balkanski and Jacob (1990)).

AC2.4: We agree with the reviewer that 222Rn would be much better suited as a terrestrial tracer. Unfortunately, there were no such measurements undertaken. Hence, we resorted to a simple metric such as distance to land.

Minor comments:

RC2.5: Caption of Figure 1: do you really mean “microbial gases” or is it rather “biogenic gases”. If you use the terms ‘microbial gases’ you imply that these gases are exclusively emitted by microbial organisms.

AC2.5: This is correct, it should read “biogenic gases”. We have corrected the caption.

RC2.6: Was there any attempt made to tag the air masses or use back-trajectories to know how long ago this air mass was above continents? It could (for example) explain why certain air masses have a higher O_3 content as discussed in lines 457-458 page 26.

AC2.6: The ozone mixing ratio is relatively invariant across most of the expedition with two exceptions, that is during the passage of the Balleny Islands for a few days (see Fig. 8 negative activation of the LV East of 180°E) and from South Georgia to Cape Town. Particularly the latter, long period is reflective of the air mass transported between 60 °S and 50 °S as shown with the 48 hrs back trajectories in Fig. 8. This is not necessarily evidence of continental influence. However, using CO as a semi-conservative tracer for continental influence (combustion) beyond 48 hrs, we find a relative concentration increase between South Georgia and Cape Town, which might be an indication of continental influence. The first instance of higher ozone concentrations near Balleny is clearly characterized by Antarctic air mass outflow, where the higher elevation of the continent, from which air masses descended, might have played a role. In light of these observations, we have added the following in l. 459 after “(see Figure 8c)”:

“This might indicate enhanced vertical mixing in the marine boundary layer during cold air advection, which might lead to the entrainment of free tropospheric air masses with higher O3 concentration into the marine boundary layer (see Figure 8c). Such entrainment is particularly likely for the high ozone concentrations observed during a cold air outbreak from Antarctica, where air masses descended from further aloft. For elevated ozone concentrations between South Georgia and Cape Town, continental pollution outflow from South America cannot be ruled out, because CO concentrations are also slightly elevated.”

RC2.7: Lines 480-482: did you check whether the values of RH for these warm air masses. Could the values of RH be an indicator for prior precipitation?

AC2.7: As RH in LV3 is a measure of the strength and direction of air-sea moisture fluxes and a tracer of large-scale moisture advection, it is not positively correlated with in situ measurements of rainfall. This is in contrast with LV4, which represents changes in RH due to precipitation events. RH and the amount of precipitation in the five days prior to arrival of the trajectories in the marine boundary layer are indirectly related as can be seen in a very weak, but significant correlation of the two variables in our dataset (Pearson correlation of 0.11 with a p-value of 0.005). We interpret this as a signal of precipitation occurring in the advected warm air mass, that is characterised by high RH. Due to several processes (meridional advection of moist air over a cold ocean surface, precipitation, and long-distance moisture advection), which affect RH in an air mass, and due to high variability of RH during a time period of five days, RH cannot directly be used as a tracer of precipitation during transport. As we can see in this study, the sPCA analysis succeeded in identifying these different time periods, which were affected by the aforementioned moist processes (see LV3, LV4 and LV9). Please, also be aware, that the simulated rainfall along the backward trajectories is only poorly constrained in the study area due to a lack of observations. Therefore, our results regarding precipitation during large-scale transport need to be interpreted carefully and further research is needed to understand the role of precipitation on the cycling of water vapour during warm air advection.

RC2.8: Page 31, lines 531-534, the following sentence comes a bit out of nowhere:

"There is no apparent explanation for the inclusion of carbon monoxide (CO), the mass concentration of sulfate in nonrefractory particulate matter (SO_4^{2-}), and the atmospheric isoprene concentration ($\text{Isoprene}_{\text{air}}$), and further analysis is beyond the scope of this work."

You might be missing something important here relative to isoprene. It would be worth investigating or asking other groups to think about this positive correlation between extratropical cyclone activity and isoprene in air. Isn't it simply that isoprene sources are abundant in the subtropical regions and the cyclones channel rapidly air from lower latitudes to the latitudes at which you are making these measurements?

AC2.8: It is true that this sentence might appear to come out of nowhere. In fact, it is there, because we consider OV contributions, if their median value of the contributing weight is larger than their single standard deviation from the bootstrap runs. We state this in l. 324f, but we cannot expect the reader to remember this. Hence we added after the sentence:

"We mention them here, because their contributing weight to the LV is larger than their single standard deviation from the bootstrap runs."

Thank you for the hint on the potential transport of isoprene from lower latitudes. This is an interesting point. However, if that were the case one would expect a similar behaviour for CO (more sources in the subtropics, and it is longer-lived than isoprene). Instead we see CO anticorrelating with LV13 (i.e., low CO and high isoprene when LV13 is activated). In addition to that, previous measurements (albeit sparse) of marine isoprene in subtropical regions (as summarised by Hackenberg et al., GBC, 2017) do not show that isoprene mixing ratios are higher in these regions than at higher latitudes. There are of course higher terrestrial emissions of isoprene in the subtropics, but 1) the short lifetime of isoprene at these latitudes would limit how far it can be transported and 2) one would expect a similar behaviour from CO (see also the previous point).

Reference: Hackenberg, S. C., et al. (2017), Potential controls of isoprene in the surface ocean, *Global Biogeochem. Cycles*, 31, 644– 662, doi:10.1002/2016GB005531.

RC2.9: With regards to the results described for LV2: Drivers of the cloud condensation nuclei population. You do not mention that small particle in the nucleation mode will eventually end up in the accumulation mode upon growth and coagulation. Condensation nuclei (CN) that are not activated will join the accumulation mode aerosol.

A very noteworthy reference concerning CCN is the one from Lee et al (2013). The authors studied twenty right parameters that cover all important aerosol processes to understand the cause of uncertainty for CCN.

AC2.9: It is correct that we have not discussed the nucleation and Aitken modes in the LV 2 section. This is because we limit the discussion to the OVs which are displayed in the specific LV figures, those are the ones that contribute with their weight beyond one single standard deviation from the bootstrap runs. We understand that for an audience who is more focused on aerosol science it might be unsatisfactory that the discussion is short from an expert's perspective. Given that the manuscript is already lengthy and we have been asked to shorten the discussion by Reviewer 3, we only added the following sentence in l. 680:

"We refer the reader to Lee et al. (2013) for a comprehensive investigation on aerosol processes relevant to CCN number concentrations and their uncertainty."

RC2.10: Lines 685-687 why is your hypothesis limited to rainout and does not include washout? "To check our hypothesis concerning rainout, we investigated the precipitation rate along the backward trajectories for the previous three days (see Figure 14)"

AC2.10: We actually meant "washout" in general, not specifically removal by rain. We have hence replaced "rainout" by "washout".

RC2.11: Paragraph 5.5 why is LV12 not related to $N_{ccn,0.15}$, $N_{ccn,0.30}$ and $N_{ccn,1.0}$? Monahan et al. (1986) parametrization of sea salt emission predicts that these small seasalt aerosols would be abundantly produced at high wind speeds.

AC2.11: This is indeed an interesting point. The reviewer is correct that the Monahan et al., (1986) source function predicts the emission of small sea spray particles at high wind speeds. Furthermore, this is supported by more recent sea spay source functions that also predict substantial emissions of small sea spray particles (e.g. de Leeuw et al., 2011). These particles are composed of sea salt and organics, and are therefore hygroscopic and efficient CCN. However, recent aerosol-focused ship-based studies have found that on a number basis and excepting very high wind speed events, sea spray particles still only form minor fractions of the total marine aerosol (typically less than 20%), and consequently, only minor fractions of marine CCN populations (Modini et al., 2015; Quinn et al., 2017; Schmale et al., 2019). Instead, it appears that under typical conditions marine CCN populations are composed primarily of non-sea-salt sulfate aerosols.

The overall sPCA results are consistent with this picture. LV2 contains strong contributions from Nccn at all 3 supersaturations, as well as accumulation mode aerosol number concentrations and aerosol sulfate concentrations. This suggests high correlation between these variables and supports the recent ship-based studies mentioned in the paragraph above. On the contrary and as noted by the reviewer, the Nccn variables do not show up in LV12, which is the LV related to sea spray aerosol. We believe that this is because, on average, sea spray only contributes minor fractions to the Nccn populations, and thus, to a first order, the variability in Nccn is not driven by variability in the number of sea spray particles. This does not preclude the occurrence of very high wind speed events where sea spray completely dominates CCN populations (such extreme cases are discussed for the ACE cruise in Schmale et al., 2019), but it does suggest that these events do not occur frequently enough to be picked up the sPCA analysis.

This picture is also reflected in Fig. 5, which shows that the variability in the Nccn variables is dominated by LV2 and not LV12.

To answer the reviewers question we have added the following brief summary on line 700:

"Since SSA particles contain sea salt they are hygroscopic and efficient CCN. Therefore, it is interesting to note that all of the CCN OVs are absent in LV12. The absence can be explained by recent studies that suggest that, on average, SSA particles only form a minor fraction of the total marine CCN budget (Modini et al., 2015; Quinn et al., 2017; Schmale et al., 2019), which instead appears to be dominated by accumulation mode non-sea-salt sulfate aerosols (e.g. see discussion of LV2 in Appendix A)."

References: De Leeuw et al. (2011), doi: 10.1029/2010rg000349; Monahan et al. 1986, doi: https://doi.org/10.1007/978-94-009-4668-2_16, Modini et al. (2015), doi: 10.1002/2014JD022963, Quinn et al. (2017), doi: 10.1038/ngeo3003, Schmale et al. (2019), doi: 10.1175/bams-d-18-0187.1

RC2.12: Page 41, line 713: please be more specific than 'The relatively large size of airborne SSA droplets' since particles much larger than 2 or 3 um do not scatter as efficiently at visible wavelengths than particles between 0.2 and 2 um.

AC2.12: Thanks for pointing out this lack of clarity. We rephrased this sentence to indicate the specific size range of SSA particles that we were referring to in l. 713 :

"The size distributions of dried SSA particles peak at diameters of around 0.2 µm and therefore contain substantial contributions from particles with diameters in the range from ~0.1 to 1 µm (Prather et al., 2013). The strong contribution to this size range means that SSA particles are effective at scattering solar radiation and thereby reducing visibility through the atmosphere."

RC2.13: FVFM is defined line 1664**: ''FVFM** is the maximum photochemical efficiency of photosystem II' and used line 738 without definition.

AC2.13: Thank you for spotting this. We have added the definition in line 738 and removed it from l. 1664.

RC2.14: Lines 762-764: explain for the non-specialist what to look for in Figure 5: " Bacterial abundance has a relatively high negative contribution to LV11 (see Figure 5), as bacterial concentrations are linked to the availability of dissolved organic matter (a product of particulate organic matter including POC and PON) and nutrients (Church et al., 2000; Kirchman et al., 2009)."

AC2.14: We apologize, the reference should point to Figure 16. We have corrected it accordingly.

RC2.16: Page 54, line 989: You wrote "strong precipitation even", did you mean "strong precipitation event"?

AC2.16: Yes, this has been corrected to "event".