Review of "General circulation models simulate negative liquid water path–droplet number correlations, but anthropogenic aerosols still increase simulated liquid water path" by Johannes Mülmenstädt et al.

This paper presents an interesting "seeming paradox" in the GCM realm, that is the latest generation of GCMs reproduced the inverted-v Nd-LWP relationship as has been observed by satellite-based studies, whereas PD-PI causal experiments using the same set of models produce the opposite response, which is consistent with the parameterized precipitation-suppression mechanism. It's not really a "paradox" in the GCM realm, as we know only the latter is causal. To figure out the causes of the inverted-v Nd-LWP relationship, the authors provided a thorough investigation on sources of confounding/covariability that evidently contribute to the noncausal Nd-LWP relationship manifested in GCMs' internal variability.

By unraveling the cause of the seeming paradox, this study concludes several key points with profound impacts on the ACI community:
1) Using the GCM framework with the capability to test causality, the authors show the possibility that correlations observed in PD mean climate states are not necessarily causal, and my not even represent the correct sign of causal LWP adjustment to aerosol perturbations.
2) When interpreting LWP adjustments from multiple lines of evidence, cautions are needed, and the causal piece of information should be taken into account and carefully weighted when integrating these lines of evidence.
3) There is pressing need to address the representativeness and confounder questions in non-GCM lines of evidence.

For me personally, this manuscript is extremely intriguing and has stimulated a lot of fruitful thinking on my end! The text is constructed in a clear, easy-to-follow and attractive storytelling manner. I enjoyed reading it, and I believe there is no doubt that this manuscript is worthy of speedy publication to raise community awareness of these impactful conclusions.

Stimulated thoughts after reading (rather than comments):

- Regarding the disagreement between the causal experiment (PD-PI) and the internal variability, is there a possibility that the causal experiment is missing some feedback mechanisms (at longer timescales) that may be present in the internal variability, because of the fixed climatic boundary conditions? (I realize this may not contribute much to the disagreement, but just wondering…) About the climatic boundary conditions, you mentioned that SST is fixed, is the circulation (winds) also fixed? (I'm not very familiar with the setup of these experiments)
  For now, let's assume the internal variability (inverted v) captures the mean climate state where MET (large-scale conditions), Nd, and LWP are in balance (manifested in some climate scale correlations), perturbing Nd initially causes changes in LWP, which may later lead to circulation and/or SST changes (feedback from LWP to MET, and then possibly back to LWP). Is this potential feedback pathway artificially shutoff in these PI-versus-PD runs, based on the configuration?

- Regarding the “funny” “doubly surprising” thing happened in CMIP6 models. I’m just curious is there any clue on what causes the CMIP6 models to get this inverted v (I understand the case for ModelE)? Are there any speculations? Is this due to the fact that the newer version of models better capture the mean climate states, thereby closer to observationally derived correlations? A following question is that if you use AeroCom IND3 models to predict the PI-LWP, would you get agreement with the causal experiment?

Some notes:

- Line 77, check spelling “ObservaTon”

- Figure 7-12, perhaps it’s worth mentioning these are results from E3SM in the captions? (I know this is clearly indicated in the main text, so, feel free to ignore this).

- Figure 9, wind vectors are kind of hard to see, I suggest enlarging them (perhaps fewer of them will help too); is it better to indicate translated PBL depth in pressure or meter (more intuitive units)?

- Line 229 & Figure 11, regarding Nd-LWP correlation within each PBL depth bin (not shown), I wonder if it’s worth showing, as I am curious about whether they look similar to what have been shown in Figure 7, i.e., in classic Simpson fashion, or different?

- *Just want to say that I really enjoyed reading Section 3.4 and the conclusion part. Great discussions! and I think the ACI community should really think carefully along these lines (i.e., representativeness versus/and causality) before producing tons of papers on the topic while not sure about how much of the results are causal.*

Looking forward to reading part 2 of the series!

Regards,

Jianhao Zhang