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.

Summary:

The paper addresses strengths and weakness with respect to causation and representativeness of correlations of different strains of evidence assessing the liquid water path (LWP) adjustment due to changes in anthropogenic aerosol and consequently droplet number concentration (N_d). In particular, methods used to establish correlations between LWP and N_d in present-day remote sensing climatology are applied to global climate model (GCM) simulations in four CMIP6 models of which three do indeed simulate a negative correlation between LWP and N_d in present-day (PD) aerosol perturbation experiments as observed in non-precipitating stratocumulus clouds.
However, the manuscript quickly shows that these negative correlations are not causal. The models still simulate a positive relationship between simulated LWP and N_d (likely through precipitation suppression). The remainder of the study provides further evidence that the GCM diagnosed relationship is indeed non causal and exemplifies how so-called confounding factors (in this case precipitation or PBL-depth), or simply statistical sampling and parameter-space limitations.
The authors argue, that such confounding factors may not only be limited to GCMs, but may also impact observation-based PD climatologically diagnosed relationships between LWP and N_d. And it concludes with a general need in the community to find joint avenues forward to disentangle correlation and causation between these to cloud properties in order to reduce uncertainty of the effective radiative forcing contribution from LWP adjustments.

The study raises some very important conundrums in understanding and quantifying aerosol-cloud interactions for the scientific community. It also very clearly shows that capturing the -ve slope between LWP and N_d in GCMs alone is insufficient to increase our confidence in the LWP adjustment and its correct mechanistic implementation in GCMs (which often miss the proposed causal mechanism altogether). The manuscript is very well written and follows a nice and clear story line. I thus recommend publication following minor revisions.

Comment on Methodology:

My most general comment is with respect to the statistics used in this study in Figs 7. and 11., which are discussed in sections 3.3.1 and 3.3.2. You introduce two distinct confounding variables here in these sections: surface precipitation rate or boundary layer depth. Both of these variables, as you state are not independent from your predictor variable N_d (and indeed your response variable LWP). The problem in binning in one variable, say PBL depth, and then looking at the slope in linear log space between averaged N_d and LWP is that you are already averaging out some of the co-variability that undoubtedly exists between predictor and response variable in each PBL depth bin. It thus skews your statistic (unless you got lucky) and the slope of the linear regression you obtain. It would be more accurate to assume that your response LWP variable co-varies with Nd and PBL and do a multi-variate fit. Or said differently: if you have an expression LWP= const. $H_{pbl}^a N_d^b$, then you can determine “a” and “b” using partial derivatives in log space. Note though that when integrated, these are only valid up to a constant! Therefore when determining a:=dln(LWP)/dln(N_d) at constant H_{pbl}, don’t average, but fit slope instead.

Minor Comments:

- Can you provide a solid argument for the 30% occurrence threshold. if not, how sensitive are your results to that parameter choice?

- L212: PBL depth only goverened by anticyclonic subsidence? What about the gradient in SST?

- Figs 8 and following: model level is not a meaningful quantity for people not directly involved in the study. Please provide more meaningful height intervals.

Edits:

- L53: I would remove brackets, its a stand-alone sentence
- L75: Please state explicitly that all other experiments use model diagnosed LWP and N_d.
- Figs 3 and following: Are these normalised PDFs around the edges? I don’t remember seeing this written anywhere.
- Fig4 caption: I would include info that its CMIP6 era experiments in caption
- Fig7: clarify that rain intervals intervals are given in brackets
- L184: sentence containing „N_d distribution is noticably lower“ is ambiguous to me. You mean the peak in the distribution is situated at lower N_d? All the distributions overlap, so how do you quantify „noticeably“?
- L232: Please rephrase „... equally accessible to clouds“. LWP and N_d are cloud properties, so how can they not be accessible to a real cloud? You mean accessible to an observed or a simulated cloud, upon which limiters are imposed? Or do you just want to point out that the LWP and N_d phase space is not populated uniformly at equal density? Please clarify.