Review of "Earth System Model Aerosol-Cloud Diagnostics Package (ESMAC Diags) Version 2: Assessments of Aerosols, Clouds and Aerosol-Cloud Interactions Through Field Campaign and Long-Term Observations"

by Tang et al.,
submitted to Geosci. Model Dev. Discuss.

In this paper, the authors introduce version 2 of the Earth System Model Aerosol-Cloud Diagnostics Packages (ESMAC Diags), which is a tool that enables the comparison of observational datasets with the E3SM (Energy Exascale Earth System Model) model. While version 1 of ESMAC Diags mainly focused on the comparison of aerosol properties between the model and various observations, further capabilities have been added to version 2 to also enable an evaluation of cloud properties and aerosol-cloud interactions (ACI). Besides the technical aspects, which have already been described in its companion paper (Tang et al., 2022), the authors introduce the new capabilities of the revised ESMAC Diags by comparing E3SM to various observational datasets, ranging from in-situ to satellite observations, mainly focusing on clouds and ACI.
The manuscript is well-written and logically structured and fits well into the scope of GMD. The authors nicely outline the new capabilities of ESMAC Diags version 2. The source code of their tool is available to the general public, is accessible and well-documented. Within the analysis package, the authors also provide certain examples of how to preprocess input data and how to run this evaluation tool, enabling new users to quickly familiarize themselves with it, which I highly appreciate. In general, this manuscript merits publication provided that the following comments have been addressed.

General comments:

- The major issue I have regarding this manuscript is one that the authors are already aware of. To ensure a fair comparison between models and observations, such a comparison must be both, scale- and definition-aware. The authors state multiple times that using a satellite simulator would largely improve any comparison between the model and the satellite observations. They also show the example of the overestimated high-level cloud fraction (Fig. 9), which is larger than retrieved by the satellite observations. The authors attribute this bias to different ways cloud fraction is diagnosed in the model and retrieved in the satellite observations. I am wondering how valid any comparison of further cloud properties can be if a comparison already fails at something relatively simple as cloud fraction and how valid any findings with regards to ACI are that are presented in the manuscript. Here, the authors even state that retrieval biases between different observational datasets can strongly affect the analysis of ACI (P21, L483), and I was wondering why they then so confidently compare the model output to the observations if already different observational datasets are not comparable. To this end, the authors need to show that the differences between the native model output and the satellite-simulated model output are at least to some extent comparable to ensure that valid conclusions can be drawn from ESMACS Diags.

- P6, L140-141: Here the authors mention that they downscaled the model data to be in accordance with the aircraft observation. I wonder why they are not going the in opposite direction, namely upscaling the aircraft observation to agree with the model grid size and output frequency. Wouldn't this allow not only for a comparison of mean/median values but would also give information on whether the model values on grid-scale are within the variability of the aircraft observations looking at properties like confidence intervals or similar metrics?
- Following up on the last comment, I was wondering how missing subgrid-scale variability in cloud properties in the model output could potentially affect the comparison between the model and the observations, in particular when the scale of the observations and the model do not match.
- P11, L266-268: The distributions in the histograms look quite different, especially for the HI-SCALE comparison. For that reason, I wonder whether the means actually agree between the respective distribution (I could not find any values in the manuscript)? Anyway, using the mean as a metric for comparison might be not a good idea in my opinion. It is dependent on the underlying distributions (which don't look identical at first glance) and are only comparable when distributions are similar. Either verify that the underlying distributions are similar or replace the mean with a non-parametric estimator like the median, which is not dependent on the underlying distribution. Would the medians still be comparable?
- P14, Fig. 6: It is quite hard to make out differences between E3SM and the observations, in (a) and (b), as CN is almost an order of magnitude higher for HI-SCALE. One could revise this plot by using a log axis or by having a different axis limit for HI-SCALE

<u>Specific comments:</u>

- P2, L58-59: In a recent study (Choudhury and Tesche, 2022), aerosol information from CALIPSO have been used to create a satellite-derived, near-global dataset of CCN. While including it could be a nice addition to a future version of ESMAC Diags, I would at least refer to it in this manuscript to highlight that progress has been made for satellite-derived CCN.
- P3 Fig. 1: Use proper spelling of campaigns (HI-SCALE, ACE-ENA) in the inserts. I would also appreciate if you somewhere state in the plot what the contours are, as not everyone reads the figure caption first.
- P6, L133-134: I assume you rescaled to the output frequency of E3SM. If so, I would write it as such.
- P9, L220-222: The mergedSD dataset combines cloud probes with different size ranges. These size ranges could be different from the size ranges used to derive bulk N_d and R_eff in E3SM. Even though I don't expect major differences here, I would appreciate it if the authors could comment on that.
- P9, L237: Are you sure that Fig. 3 shows a PDF? I think it is rather a probability mass function (PMF) as you are using discrete intervals as depicted in the histogram in Fig 3. A PDF is for a continuous distribution, and you only get to probability once you integrate a PDF over a certain size range, which will then give you a PMF.

- P10 Fig. 2/3: Can you add what variables you show in the histograms, i.e. but N_d/R_eff on the x-axis lable.
- P23, L513-514: Here I think you refer to mean TOA albedo, averaged over all albedo bins. You should clarify that in the manuscript.

Technical comments:

- P1, L17: Change *…they are lack of the process-level…* to *…they are lacking process-level…*
- P1, L30: Change *…aerosol-cloud interactions…* to *…ACI…* as it has already been abbreviated.
- P2, L66 and P2, L71: DOE, spell out first before abbreviating.
- P11, L273: Change *profile* to *profiles*.
- P19, L443: Change over to overall.
- P19, L452: … sensitive Nd to CCN relationship …
- P24, L527: … the Earth System Model …

References:

- Choudhury, G. and Tesche, M.: Estimating cloud condensation nuclei concentrations from CALIPSO lidar measurements, Atmos. Meas. Tech., 15, 639–654, https://doi.org/10.5194/amt-15-639-2022, 2022.