Overview:

The paper is an evaluation of biases, variability and multi-year trends of the CAMSRA and MERRA-2 reanalysis data surface sets for O3, NO2, CO, SO2, PM10 and PM2.5 for the 2003-2020 period over Europe. CAMSRA shows better agreement with surface observations than MERRA-2 with the exception of SO2.

General comment:

The evaluation procedure applied in the paper is sound and provides valuable insight for the user of the data sets.

My main concern of the paper is the selection of the stations for the evaluation. Urban stations, which are probably the largest fraction of the European AQ networks, were considered as “background” stations for the paper and combined with the rural stations. I think it is not common practice to include urban stations in the group of background stations. In any case, a stratification between “rural” and “urban” stations is strongly recommended as the scale represented by urban observations is often much smaller as the grid-point resolution of the evaluated models. I think the authors need to quantify the errors against rural stations and urban stations separately, and clarify if the presented results are dominated by the comparison with urban stations or not.

The paper can further be improved by adding more information about the speciation of the PM10 and 2.5 values. It would be interesting to know how large the sulphate (given the differences in SO2) contributions were, and to what extend dust and sea salt contributed to PM in both data sets.
Specific comments.

Please include more references in table 1 and discuss their findings. For example,


L 75: add references (links or DOI) of the observational data sets

L 93: Atmospheric composition fields are only available in grid-point space.

L 101: SOA is included in the two OM aerosols variables in CAMSRA.

L 120: Please mention here (and also for CAMSRA) the vertical extent of the model level at the surface, which may be important for the interpretation of the results.

L 125: Please discuss the formulae applied to derive PM10/2.5 for CAMS RA and compare it to the approach used for MERRA-2

L 152: To consider “urban” stations as “background” stations seems far-fetched. Excluding the urban regime, at least as option seems necessary for the study. (see my general
L 153 Provide reference for the classification

L 162 Provide information about the typical grid-box variability, i.e. the deviation (variance) of the observations from the mean. Please provide information how many of the “observation grid boxes” are actual composed of multi-station means. For example, a pdf of the number of stations per grid box would be informative.

L175 Please, clarify if PCC represents a spatial or a temporal correlation.

L 180 Please, clarify the accumulation index I, i.e. is it over time per station, or over stations per time instance)

Fig 4 Why is there no PCC for Tr of Merra-2

L 420 Please, add some information about speciation for PM10 & PM 2.5