We would like to thank the reviewers for their thorough work and helpful feedback, which led to a significant improvement of the manuscript. In the following, all remarks by the reviewers are listed in black text and our corresponding replies are given in blue.

Reviewer 2

Paper summary

Thank you for providing these remarks and comments. We are confident that our answer led to a compelling and well-improved manuscript.

This manuscript discusses a method for sub-sampling observational data in the context of air quality data assimilation, which requires to prepare the observational data into two datasets, respectively assimilation and validation. The authors propose to use clustering algorithms to improve the representativity of the observations during such a sub-sampling. Their methodology has two practical advantages: on the one hand, it is independent from the assimilation model, and on the other hand, it only requires observational data as inputs.

To evaluate the benefits of their clustering approach, the authors introduce an AV-difference (assimilation/validation) metric, which is the difference between the RMSE (Root-Mean-Square Error) of the model w.r.t. the assimilation dataset and the RMSE of the model w.r.t. the validation datasets. As such, a AV-difference of zero is synonymous of perfect representativity, while a high AV-difference suggests overfitting by the model.

Using an operational CAMS assimilation/validation configuration for year 2016 as a reference, the authors apply their approach on observations over Europe for four months of year 2016 (January, March, June and September, picked for their seasonal representativity), and demonstrate a significant decrease in the AV-difference for several pollutant species, and most notably carbon monoxide (53%), nitrogen dioxide (50%) and ozone (18%). The improvement is particularily interesting in the case of carbon monoxide, due to the scarcity of the observations compared to other pollutant species (such as ozone).

General comments

The core content of the manuscript is interesting and provides some convincing results, and the authors did a nice job in presenting the K-means clustering algorithm and its soft constraint variant, and how they adapted their problem to both. This said, this manuscript could be improved in terms of presentation and could elaborate on a few points to ensure the final paper is compelling to all.

Possible presentation improvements

I found that the AV-difference metric was very important to understand the paper and its contributions, yet it's defined quite late in the manuscript (L193). The Introduction does make a review of the state-of-the-art in this regard, but only states that the present study will improve representativity through clustering without hinting at how it will measure it. A few sentences (if not a single one) in the Introduction to give the big picture may be enough.

We agree that highlighting the big picture of our analysis in the introduction will lead to an enhanced understanding of our results. We have added a description of the AV-difference as the representativity measure used in the paper to the introduction. This addition reads:

“Here, the relative representativity of two datasets is determined by quantifying and subsequently comparing the difference in the RMSEs between the model analysis and the observations from the assimilation and validation data set. This measure is hereafter called AV-difference and is described in detail in Sec. 3.4.”

The manuscript could also benefit from a few more figures to support its content. A possible addition could be a flow-chart in the Introduction, summarizing the proposed methodology (e.g., observations going through the clustering to be split into the two datasets, fed to a data assimilation model like EURAD-IM). Such a flow-chart would not only summarize the overall methodology to the reader in a single figure, but could also be used to picture its advantages in terms of input/output.

Thank you very much for this valuable input. We have added a flowchart (now Fig. 2) to give an overview over the proposed method. It details the process from the input data to the clustering up to the evaluation of the representativity that we employed in this manuscript. Further, we have added Fig. A1 to showcase the seasonality of the representativity enhancement to the manuscript as requested by reviewer #3.

The new Fig. 2:

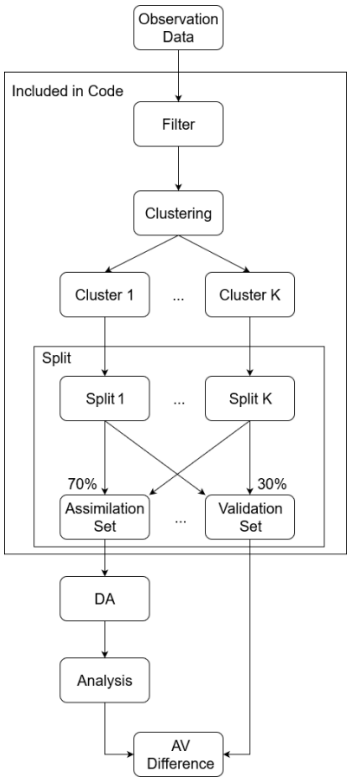

Figure 2. Flowchart of the proposed methodology. In the presented case, 'Observation Data' is the observation data used in the CAMS-project. The elements in the box titled 'included in code' are included in the provided software code (Hermanns, 2025), with the KSC algorithm as the 'Clustering'. The box 'split' is the extraction of the assimilation and validation set from a clustering result. 'DA' is short for data assimilation. The 'AV-difference' is the representativity measure and detailed in Sec. 3.4.

The new Fig. A1:

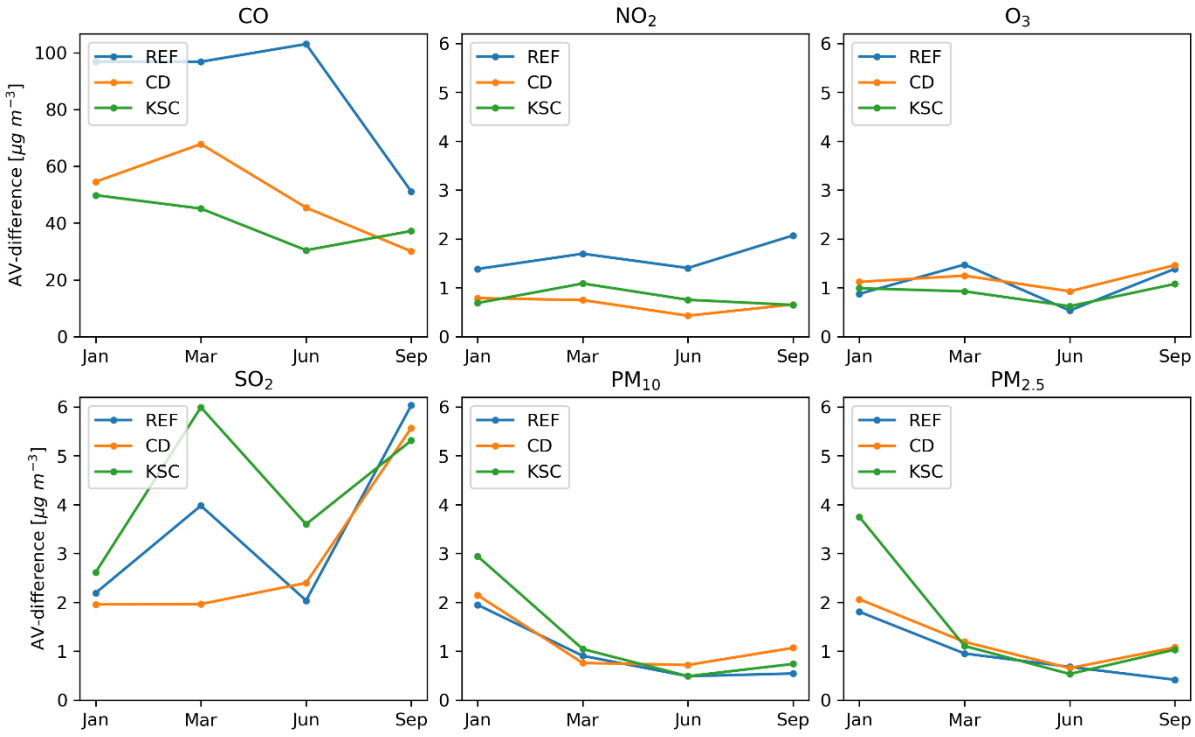

Figure A1. The evolution of the AV-difference for the REF, CD and KSC configurations for each species is shown over the four evaluated months of 2016.

I would also recommend moving Figure A1 from the Appendix back to the main body. Indeed, Figure A1 gives a very clear picture of how scarce CO observations are with respects to other pollutant species. Including it into the main body and making a few more references to it would strenghten the conclusion that the proposed clustering methodology significantly improves representativity of CO in the framework of air quality data assimilation.

Thank for the recommendation. We agree that the figure is better placed in the main body to strengthen the argumentation. It is moved to chapter 3. We have added this additional statement in the main text of chapter 3:
“The density of CO observations compared to the other measured species is substantially lower, even in highly populated areas. “
And this statement to the Conclusion:
“This is due to the relative scarcity of the CO observations (see Fig. 1) and a resulting strong influence on the clustering.”

Finally, on a side note, I would recommend using a gridded layout for most of the figures, especially line plots.

Thank you for the suggestion. We have generated the plots using a gridded layout. However, this

adds a lot of lines to the figures, which we find distracting from the main results. Therefore, we decided to leave the figures as they were.

Questions regarding the content

1) Are there particular reasons for only simulating four months of 2016 ? I get the seasonality argument regarding the choice of the months, but why not simulating the entire year ?

We agree that it would have been interesting to simulate and evaluate the proposed method on the entire year. However, the employed four-dimensional variational data assimilation methodology requires an iterative optimization of the parameters to be optimized (here, initial values and emission factors). Therefore, simulating an entire year is computationally very demanding. The simulations conducted for this manuscript are based on the simulations by Lange et al., 2023, where we identified issues with the representativity of the utilized observational split provided within the CAMS project. The simulations conducted here to evaluate the model's performance given the split of our proposed method aim to illustrate its potential while keeping the computational burden feasible.

2) More broadly, it would be interesting to develop the seasonality of the results. At L276, there is this mention:

>Furthermore, while the evaluation shows fluctuations for each season, the general result holds true for each season individually.

but this is not enough to convince the reader about the seasonal trends of the results, especially considering point 1) (i.e., no full seasons, only sample months) and considering there is no figure or table detailing seasonal results. If these trends are indeed not significant, maybe a single table or figure would be enough to demonstrate that.

Thank you for the suggestion. We have added a figure detailing the seasonal evolution (as given by the four simulated months) in the appendix, Fig. A1. It shows the evolution of the AV-difference for each species for each season. The new Fig. A1 is shown above.

3) What about the slightly worse results for KSC in Tables C1 and D1 (Appendices C and D) ? Should we worry about them or are they small enough to be ignored ? While there is, indeed, an order of magnitude of difference between these results and those for carbon monoxide, additional details could show decisively whether or not the slightly higher AV differences are problematic, and at the very least, why the current manuscript does not elaborate further on them.

* For instance, the slighter higher AV-difference for ozone in D1 is probably not much of an issue given the thresholds for air quality. E.g., below 80 µg per cubic meter of ozone is considered to be good per CAMS, so 1.1 µg per cubic meter of AV-difference remains negligible. However, the reader does not necessarily know about such orders of magnitude depending on the species.

Thank you for this suggestion. We have included a reference to the RMSE target reference values in the CAMS quality control. "Furthermore, the RMSE target reference values in the quality control of the CAMS regional services are 16 µg m^{-3} for O_3, PM_{10} and $PM_{2.5}$ and 22 µg m^{-3} for NO_2 (Gauss et al., 2024), making AV-differences of ~ 1 µg m^{-3} negligible.". We also reworked chapter 4 to include the relative AV-difference. This is calculated by dividing the mean AV-difference by the mean of the assimilation and validation RMSE. The relative AV-difference better highlights the impact of the improvement depending on the magnitude of the RMSE.

* As far as I'm concerned, I would be also interested in learning if the given AV-differences are constant throughout each year (i.e. 2016 or 2017), or if they depend on the season, if not the day ? A plot of the AV-difference throughout the year for each species may be enough to address this concern.

Thank you for making this important statement. Indeed, we have analyzed the temporal evolution of the AV-difference and have decided to leave this discussion out of the manuscript to be more concise in our results. However, the reviewer comments suggest that the additional evaluation of the temporal evolution of the AV-difference increases the clarity of our results. Therefore, we have added a figure detailing the AV-differences during the seasons in the appendix (now Fig. A1). The AV-differences behave similarly for all months, although some variability on the daily AV-difference exist.

Specific comments

>L91: An overview of the geographic distribution of the available observation stations for each species is shown in the appendix in Fig. B2.

The reference seems to be wrong; the geographic distribution is shown in Fig. A1. Note that this overlaps with a previous comment on moving such figure back to the main text.

Done. We have put the Figure in the main text as requested by your previous comment.

>L128: "Is is termed to be violated [...]"

Done.

This looks like a typo. Shouldn't it be "It is termed to be violated..." ? Anyway, the full sentence is a bit unclear. What does "violated" mean precisely in this context ? Does it mean the sigma term only makes sense when the objects are assigned to distinct clusters ? Please clarify.

The term “violated” originates from Wagstaff, 2004. We agree that it is confusing in the presented context and does not serve the understanding. The term “violated” has been removed from the text and the descriptions altered accordingly.