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 3

Data clustering has the potential to improve the representativity of data assimilation results. This is shown by Hermanns and co-authors in their paper. This is an interesting idea and would be something to incorporate in for instance the European CAMS ensemble analysis and reanalyses. However, the paper also raised multiple questions and I am not yet convinced that the potential of the method has been fully exploited. To my opinion a major revision is needed in response to my major and detailed comments.

We thank the reviewer for his/her careful revision of our manuscript. We believe that incorporating the reviewers comments and suggestions led to an improved manuscript.

Major comments:

The results for PM2.5/10 (and ozone) should also be shown and should be discussed in more detail. Why is the result so different? Can this be understood? In the abstract, last sentence, improvements are reported for NO2, O3, CO: Please report the PM results as well.

We have 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 and explains the difference in the results. The abstract has been changed to report the relative AV-difference, including for SO_2, PM_{10} and $PM_{2.5}$.
We have changed the last sentence of the Abstract as follows: "[...] the largest improvement in the relative representativity measure is evaluated for CO with 16 %, for NO_2 with 4 %, and for O_3 with 1 %. A reduction in the relative representativity measure is observed for SO_2 with -5 %, for PM_{10} with -2 % and for PM2.5 with -5%, although these differences do not lead to significant deviations in absolute values given the overall error and the improvement for CO outweighing the changes in the other species."

The motivation for - and introduction to - the clustering approach can be improved. In particular I was wondering why the diurnal cycle is used as property to distinguish stations for all species? The diurnal cycle of ozone is large during summertime pollution photochemical smog events when it builds up during the day. For other species(like NO2, PM) the diurnal cycle may have a very different interpretation, e.g. rush hour emission peaks or development of the PBL. The effectiveness of this choice may be quite different in summer and winter. Apart from the diurnal cycle, are there other properties that may be used for the clustering? Please add a discussion to answer these questions, and also add the seasonal results.

We have added the corresponding Fig. A1 to the appendix and referenced it in:
"Furthermore, while the evaluation shows fluctuations for each season, the general result holds

true for each season individually, see Fig. A1 in the Appendix." The choice of parameters regarding the averaged diurnal cycle of air pollutants (mean and standard deviation) was made to find a compromise between the number of parameters used and the accuracy of the used parameters. As was shown by Beyer et al., 1999, the distance calculated within the KSC method approaches its maximum value as the dimension of the problem increases. Therefore, we wanted to limit the number of parameters to a minimum. Further, the parameters of our choice are closely related to the parameters chosen by, e.g., Gaubert et al. 2014, and Joly and Peuch, 2012, which made us confident that the mean and standard deviation of the averaged diurnal cycle leads to good results, as was confirmed by our discussion of the clustering. We have added the following information to the manuscript: "While other features are also suitable for clustering as was shown e.g., in Joly and Peuch (2012), restricting the number of features is important to apply meaningful clustering. In high dimensions, the "nearest neighbour" problem cannot be solved in all cases (Beyer et al., 1999)."

The new Fig. A1:

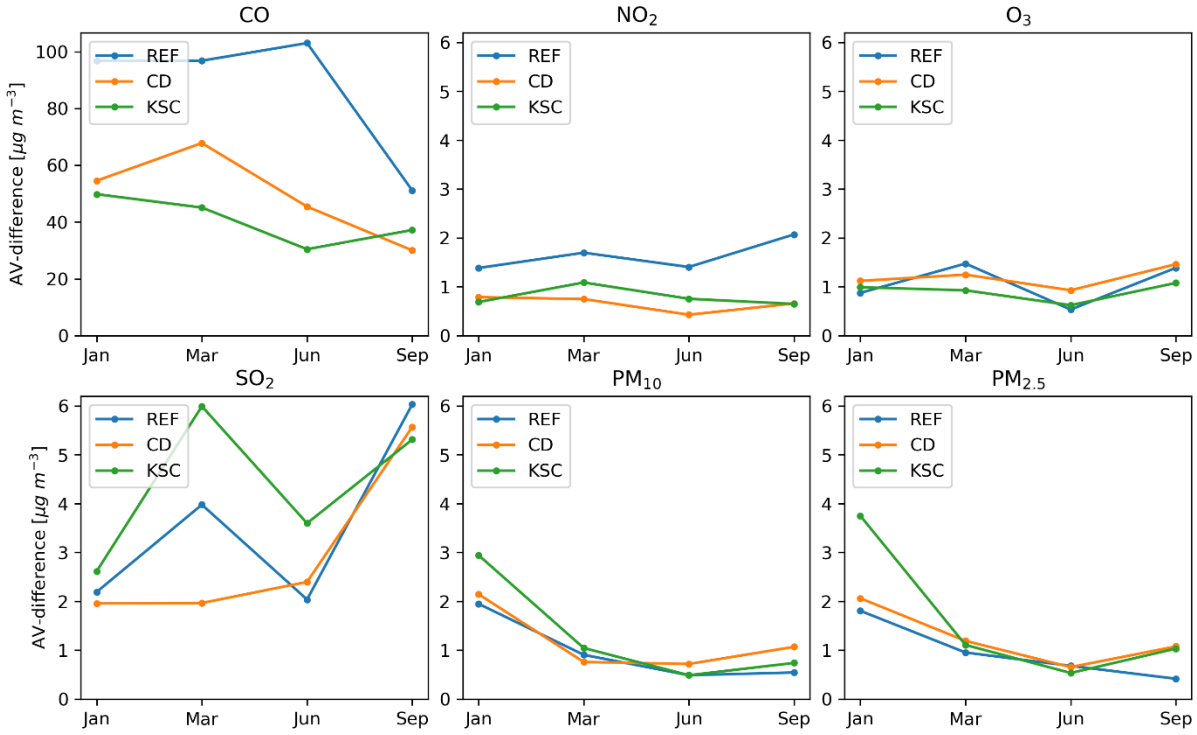

Figure A1. The evolution of the AV-difference for the REF, CD and KSC configurations for each species is shown over the four evaluated months of 2016.

Is the comparison with CAMS a fair comparison? Are the same stations used in both cases? The CAMS REF experiment is not well described in the paper and I have the impression that there are much less stations used by CAMS. Showing the distributions of assimilation and validation stations in all experiments could be a useful extra plot.

Thank you for pointing out that our description of the experimental setup was not sufficient. All presented experiments use the same observation stations as the CAMS REF experiment. The

configurations only differ in the assignment of individual observations to the assimilation or validation data set; thus, the same number of stations is used in all experiments, including CAMS REF. We have referenced the EEA data base for the general user who does not have access to the pre-selected data set from the CAMS REF experiment. We have updated the description of the observational data accordingly (see below). Anyway, we have decided to not include a plot for the individual observation splits as the large number of stations prohibit a detailed assessment of the assignment for single stations. The paragraph now reads: "The observational data used in this study are ground-based observations used within CAMS regional analysis. The data is pre-filter to the classes 1-7 according to Joly and Peuch 2012 from data accessible at the European Environmental Agency (EEA) via *https://eeadmz1-downloads-webapp.azurewebsites.net/* (last access: 09 June 2025). They consist of measurements of hourly concentrations of carbon monoxide (CO; 382 stations), nitrogen dioxide (NO2; 1743 stations), ozone (O3; 1712 stations), sulfur dioxide (SO2; 1005 stations), PM10 (1029 stations), and PM2.5 (509 stations) from 40 EEA member and cooperating countries."

The European stations also come with a site classification (rural/urban, background/traffic/industry). This by itself can be seen as a clustering. In CAMS the Joly-Peuch site classification is used (https://doi.org/10.1016/j.atmosenv.2011.11.025). Again, the categories 1-10 of Joly-Peuch are also a form of clustering. Please add this reference and discuss the relation to the present study.

Thank you for pointing out that we missed to put the proper citation to the manuscript. We have included a discussion of the paper and described its relevance for CAMS. However, we also like to highlight that our approach does not aim to classify stations directly but to identify and derive parameters that can be used to find a representative split to assimilation and validation sites. The new paragraph reads: "The study by Joly and Peuch (2012) characterizes pollutant time series by 8 parameters derived from their daily-, monthly-, seasonally- and annually averaged diurnal cycles. With this, they were able to derive 10 classes of air pollution monitoring sites using linear discriminant analysis, which allow for the classification of new sites without the need to recalculate the classification. This classification is used by the Copernicus Atmosphere Monitoring Service (CAMS) to filter observational data in order to improve their regional products (Peuch et al., 2022)"

The EURAD assimilation, if I understand well, has been run at a 15 km resolution. Sites near busy roads or near industries will not be well represented at this resolution. A pre-filtering would be good, as is done in CAMS. On the contrary, from the paper I get the impression that all EEA sites are used by the authors. Please explain and motivate this choice.

Thank you for this comment. As stated above, the same preselected stations as in CAMS are used. This excludes sites near busy roads or industries. We clarified this in the text: "The REF experiment contains the same observation stations as the KSC and CD experiments. The experiments differ only in the assignment of the observation stations to the assimilation/validation set used in the data assimilation runs. Note that the CAMS REF configuration applied a bundling of NO_2 with O_3 and PM_{10}

with $PM_{2.5}$ observations (i.e., these species are always assigned to the same data set) before selecting semi-randomly, where validation stations are placed near assimilation stations. Here, spatially isolated observations are used for the assimilation."

The filtering that needs to be used to account for unrealistic results in EURAD (line 216-221) gives an uneasy feeling. Please add evidence that the EURAD system is working well overall, with reasonable increments and good reductions of the rmse differences with the stations in the analysis. How has EURAD been tested?

The filtering for unrealistic results is only relevant for the month of January as the other simulated months do not show this behaviour. We have investigated the critical areas and found errors in the wind field used to disperse the emitted pollutants. Due to these errors, the emission plume was slightly offset to the observation location. Thus, the modelled plume only hit the observation location in its edges. The assimilation system (as any other presently used assimilation system) is not aware of plume displacements and correctly tried to increase the emissions. However, this was an exceptional situation. We chose to implement a rigorous filter to exclude this situation. Further, the EURAD-IM is continuously evaluated within the CAMS Regional Air Quality Production System (https://doi.org/10.5194/egusphere-2024-3744). Further, the results of the four-dimensional variational data assimilation system have been used in peer reviewed publications, e.g., Franke et al., 2023, Erraji et al., 2024.

I did not find the 2017 results fully satisfactory. One would hope that the clusters are quite robust and do not change much from year to year.

The classification based the clustering algorithm is inherently sensitive to the addition or removal of elements. Furthermore, the characteristics of observation stations used in the evaluation are not static. We consider it an advantage to not define a fixed classification for the stations but rather a modular one where the output is dependent on the available data. This can ensure to derive a representative assimilation set from the available observations. We have addressed your concerns in the text: "The split of the available observation data derived for 2016 cannot be carried over directly to the year 2017. This is expected since clustering is inherently sensitive to the addition and removal of individual objects. Furthermore, the characteristics of observation stations are not static, as shown in Fig. B1 and Fig. B2 in the Appendix."

Introduction:

- Introduction: Please add more references on air quality data assimilation activities. For instance the CAMS ensemble activity is relevant for this paper. Here also a split in observation and validation sites is applied (l 27).

We have added further references.
“The Copernicus Atmosphere Monitoring Service (CAMS, https://atmosphere.copernicus.eu/regional-air-quality-production-systems (last access: 07 August 2025)) employs an ensemble of air quality models using data assimilation to provide daily air quality forecasts and reanalysis over Europe. An overview of the ensemble and the models used within can be found in Marécal et al. (2015)”

- Introduction (l 31). Classification of surface sites is a topic related to the current paper. The paper of Joly-Peuch is a basis for the CAMS work and is relevant to discuss. Also the methodology of EEA should be mentioned (with reference).

We have added the reference to the Joly-Peuch paper including the relevance for the CAMS work (see also our comment above). Further, we have clarified the use of EEA data in the text, see also our answer to your comment above.

- l 58: Please provide a motivation why the diurnal cycle is used.

In Lyapina et al. (2016) the monthly averaged diurnal cycles were chosen as clustering parameters since they yield the most stable clustering results. We have chosen not to elaborate on their decision process since it would not benefit our paper.

- l 80: (IFS) please mention the European Centre of Medium-Range Weather Forecasts

Done

l 87: There is no pre-selection made for the stations? Would it be better to remove roadside traffic stations?

As mentioned above, we used the same data as in the CAMS assimilation system, thus the filtering according to the Joly-Peuch classification has been applied to the EEA data. We have added clarifications in the main text, see above.

- l 92: "An overview of the geographic distribution of the available observation stations for each species is shown in the appendix in Fig. B2" Do you mean A1?

We changed the reference in the text as we also moved the Figure to the main text as suggested by reviewer 2.

- Figure 1 is not a figure. A table could be an option, or a listing in the text would be possible as well.

We have moved the content of figure 1 into the text.

- l 104: Normalised: how? Also units should be removed for this formula to make sense (e.g. concentration).

We have added an explanation about the normalization of the feature vectors. The sentence now reads: “The elements of the feature vectors are normalized to a dimensionless value between [0,1] in order to ensure that each feature affects the distance value with the same weight.”

- l 152: What is the standard deviation of the mean and variance of the diurnal cycle? A formula would be good to be precise on what is computed.

We have given a more detailed explanation in the text. We feel that a formula would hinder readability due to the complexity of the necessary indexing. The description now reads: “The annually averaged diurnal cycle is calculated from the filtered observational data by computing the average for each hour of the day for the entire year. From this annually averaged diurnal cycle, the mean and the variance are calculated and form one set of the features used in the clustering algorithms.”

- l 184: "REF experiment" Please provide more details what this is and how it compares. Does CAMS include a similar number of stations? CAMS uses Joly-Peuch classification to pre-select the stations that are compared to the models.

In accordance with a previous remark, we have added more information about the REF experiment in the main text (see our comment above).

- l 193: The AV-difference: is this normalised? Does this have a unit (e.g. ug/m3)?

In line 193 the AV-difference has no unit, as it is derived in a general context and the unit depends on the evaluated quantities (RMSE in our case). Later, where the AV-difference is evaluated, the unit is given. It is not normalised.

- l 194: "split into two different observation configurations" This was unclear to me. Why is this done, and how are these two constructed?

Thank you for pointing this out. We added a more detailed explanation in the text: “A set of available observations (OBS) is split into an assimilation and validation set (A + V), hereafter observation configuration. To evaluate the split in terms of representativity, consider two distinct observation configurations with assimilation sets (A_1 and A_2) and validation sets (V_1 and V_2), [...]”

- l 203: Is the RMSE computed for individual model/measurement pairs (hourly observations)?

The RMSE is computed for each hourly model/measurement pair and for each day separately. We have added some clarifying remarks: “Amongst several possible options, the RMSE is chosen as an error metric. Here, the RMSE of the analysis is calculated for each simulation day for the assimilation and validation sets for individual hourly data pairs.”

- Fig. B2: please add diurnal cycles for all species.

Thank you for the suggestion. Our point is that the annually averaged diurnal cycle of a single species at a single station may be different from year to year. Therefore, we have chosen the CO time series to illustrate this. We like to emphasize that this behaviour is possible but not necessarily true for all species/stations. Therefore, we would like to keep the figure as is. Adding more species, which may not change strongly among the years, will hamper readability and distracts from the point we want to make with this illustration.