This paper investigates the use of Model Output Statistics (MOS) models for correcting the regional CAMS ensemble forecast at the monitoring stations available over Europe for different pollutants (O3, NO2, PM10, PM2.5). Several models (standard and regularized regression, random forest, gradient boosting machine) and approaches (global or station-specific local models, hourly or daily). The postprocessing step is crucial for improving the skills of the AQ forecasts at monitoring stations, and the literature on this topic remains limited. In this context, this study provides a sound and valuable contribution. I found the paper generally clear and well written, although some aspects of their methodology would need to be clarified. As a consequence, I am not entirely sure the authors always follow all the good practices in the field of machine learning, I raised some points that would deserve to be inspected and revised if necessary. I found the quality (resolution) of the figures generally too low, this should be improved. But overall, the study can be accepted for publication in ACP after addressing the comments below.

Introduction:

L48-50: I would not put the “compensation of systematic biases” as another application given that the spatial downscaling intrinsically include some correction of the bias. Also, “compensation” could be replaced by “correction”.

L59: I don’t think this “hierarchical” term fits with decision tree-based ensemble methods
L63-64: Some common traditional MOS methods like the aforementioned running-mean or Kalman filter do not require so large input dataset, so I would rephrase this sentence referring to “Some MOS methods…”

L66: “frequent upgrades”: please specify here the typical frequency of upgrade of CAMS

L66-67: To my opinion, the development of a “global model” including data at numerous monitoring stations will indeed benefit from a larger input dataset for training, but at the same time, it will have to make predictions over a larger range of conditions. Therefore, I don’t think there is an intrinsic guarantee of reaching better performance by doing so (although I acknowledge that the authors do get a slightly better performance). Another valuable aspect of the global approach is the ability to correct AQ forecasts at new stations, taking benefit from the other pre-existing ones and without the need to wait for accumulating a sufficiently large local dataset. Eventually, I would try to introduce this idea of the global approach in a more comprehensive way.

Section 2:

L84: More specifically, which new SIA were introduced in June 2019 (since I guess nitrates, sulfates and ammonium were already included before)? As some readers might think from this sentence that no SIA were included before that upgrade, I would be more specific here to avoid confusion.

L87-88: Given that this result is quite counter-intuitive (and somehow contradicts what is explained in the introduction to justify the interest of developing a global approach), I would eventually keep it for latter in the discussion and only mention here that this impact of this upgrade will be discussed later.

Section 3:

L94-96: This sentence could be better formulated. It corresponds to a supervised learning problem because you have at your disposal a labeled dataset composed of observations of the air pollutant concentrations you are trying to predict.
It would have been interesting to see to investigate the impact of the lead time on the performance of the different methods (although I understand this would have added more results to be discussed and thus less clear take-home messages).

Please provide the LASSO acronym when it first appears.

Please provide more information on the difference between ridge and LASSO regression. In contrast with the ridge regression, Tibshirani (1996) notably mentioned that the intrinsic nature of LASSO favors the production of coefficients that are exactly zero (in contrast with ridge regression that produces low non-zero coefficients), which comes with interesting benefits in terms of input features selection (especially when numerous input features) and interpretability.

"reached"

The text needs to be revised here. The tuning phase is not done prior to the training phase as it requires the ML model to be trained with different hyper-parameters configurations, in order to select the tuning providing the best performance on the validation dataset. However, when the optimal tuning is obtained, the ML model is trained again with both the training and validation dataset. In a last step, the performance of the obtained tuned ML model can be estimated on a test dataset entirely independent from the previously used training and validation data.

For the tuning of the ML models, do I correctly guess that the 5-fold cross-validation is performed exclusively on the training data, i.e. for the local approach only over 2017-2018, always excluding the 2019 test dataset? Please clarify this point in the text. (Obviously, this test dataset should not used in the training/tuning phases as the final evaluation of the performance of the ML models needs to be done on a independent test dataset).

I cannot understand what are the values tested here, please explain this point more clearly. Also, please explain the meaning of the different hyper-parameters tested. Finally, please indicate the default value kept for the other hyper-parameters.

Please indicate here that you refer to the MOS approaches that require relatively long input dataset, which is not the case of all MOS methods, as mentioned in the introduction.

"close"
L197-199: Over which period of time these correlations are calculated? Over the training period? Is this changing dynamically? Regarding the global approach, for instance with 3-days training phase, if there are 1000 stations available with full data availability, this means that the correlations are calculated based on 3000 points? Or this is done at each individual station, thus considering only 3 points? Please explain in more details this aspect of your methodology, for both the local and global approach.

Table 1: Please clarify if the ML model for O3 only includes O3 raw forecast or also the other pollutant raw forecast in its set of input features (I guess it is the first option). Also, for the hourly approach, please clarify if all the features at the different hours of the day are gathered together without providing any information to the ML model on this hour (or if a specific ML model is trained and used for each individual hour of the day). Please also explain what is done when the observation at D-1 is missing (I guess you simply don’t correct the raw forecast?). If we take the example of a global hourly model with 3-days training set and consider that there are 100 stations available, this means that the training dataset will include 3x100x24=7200 instances, am I understanding correctly?

Finally, the use of ridge and LASSO regression requires the input data to be standardized ((X-mean)/std), otherwise the differences of scales of the different predictors can impact in an undesired way the regularization term (lambda). Citing Tibshirani (1997) [The LASSO method for variable selection in the Cox model]: “The lasso method requires initial standardization of the regressors, so that the penalization scheme is fair to all regressors. For categorical regressors, one codes the regressor with dummy variables and then standardizes the dummy variables.” Have you applied such standardization? (For RF and GBM, this standardization is not required). If not, I am not sure to which extent this can impact the skills obtained here, but some additional tests would be required to clarify this point.

Just an additional comment but given that the errors of the raw ensemble differ substantially from one region to another (Fig. 3), it might be interesting to allow your global ML model to learn more directly some geographical aspects of the raw ensemble behavior (and errors); one possible approach would be to provide as additional features the climatological meteorological parameters. I am not entirely sure but this might eventually help the ML model to learn specific relationships in European regions of different climate and further improve the skills obtained.

Section 4:

L229: Apart from specific MOS approaches (e.g. persistence, running-mean) that are unbiased by construction, it is not correct to say that the MOS models are by construction
unbiased, as demonstrated by the results obtained here and shown in Fig. B1. The MOS models used in this study are indeed likely unbiased on the training dataset, but this does not guarantee the absence of bias on the testing dataset (although it is true that if atmospheric conditions are not that different between training and testing datasets and if the test period is sufficiently long, then the bias is expected to be low). Please correct the sentence.

L229-230: Could you please provide some information regarding the feature importance obtained with RF and GBM models?

Figure 1: The quality of the figures is generally not very good, please increase the resolution. Also, given the small differences and the scale used, please add an horizontal grid on these plots to make it easier to compare.

Figure 2: I think it should be “global” rather than “local” in the caption.

Section 5:

L269-270: For the global model, it is not clear why the authors finally selected the one based on only 3 days. Even with several thousands to stations, I would expect a still reasonable computation cost for this global model, no? Could the author provide some information on the computational cost (number of CPUs and hours)?

Table 2: Eventually, remove the digit of the relative reduction of RMSE, to make the table easier (lighter) to read (a figure in the Appendix would have been useful to facilitate even more the comparison).

L324: “represents”

L325: I would also include the persistence method in the previous section. This would provide a very useful reference for judging the results obtained with the different MOS approaches.

L328: “Figs.”
“Figures”

“allows”

Introduce the acronym before

Figures 5-8: Again the quality of these figures is very low, please increase the resolution, the size of the labels as well as the size of the legend (grey circle and square cannot be distinguished in this legend panel). Also, you could regroup them into two 4-panels figures, one for O3, the other for PM, to facilitate comparison between daily and hourly results.

Pages 14-15: Please organize your text in several paragraphs.

Please provide some numbers of the duration of the episodes in your domain and period of study. Also, in this section, please indicate the frequency of observation of these O3 and PM exceedances (base rate), this is an important piece of information that is currently missing.

To further demonstrate your point, could you provide the scores obtained over these 34% stations and over the remaining 66%? (same for PM)

I think there might be an error here. The global/hourly GBM approach gives better results on PM10 (Fig. 8) than O3 (Fig. 7). Please clarify.

Conclusion:

Based on all these results and acknowledging the persistent limitations that remain to be overcome, could the authors make any personal recommendation on which approach would be the best for an operational implementation in CAMS? Based on the experience acquired, could the authors provide some perspectives to be explored in order to further improve the skills obtained here?