The authors propose a treatment to an ensemble of CTMs to improve air quality forecasting. This has been done repeatedly in the past using various approaches. The novelty here is the systematic use of ML methods which seems to produce promising results, that is, outscoring the current ensemble method. As I am not a ML expert, I cannot judge the technical implementation of the various algorithms tested. Overall though, I would say the analysis presented makes sense and I trust the authors that the treatments they propose 'are doing the right thing for the right reason'.

My advice to the editor is to accept the manuscript for publication, pending some clarifications that I invite the authors to consider:

We would like to thank the referee for its positive comments and helpful suggestions. Our answers for each point are detailed below.

- Quality of figure needs improving

  The quality of all figures has been improved, with better resolution and increased size of legends and labels.

- Specify in plain words what is meant by raw ensemble – is that the unbiased ensemble mean? Possibly I have overlooked, but I cannot locate a definition in the text

  Thank you, this has been clarified line 218: For the global approach, tests have been performed using as a predictor either the raw Ensemble (i.e. the median of the 7 individual deterministic models) forecast or the unbiased Ensemble concentration of the target pollutant. The unbiased concentration is defined as the forecasted Ensemble concentration minus the bias observed at the station during the previous days (days of the chosen training period).

- Please comment on what would make your ML methodology better/preferred to other ensemble-improving methods (as for example: https://acp.copernicus.org/articles/14/11791/2014/, https://acp.copernicus.org/articles/13/7153/2013/)

  We would not say that our MOS methodology is better than the ensemble methods you mention but rather that they are complementary. The MOS could be applied to any type of ensemble model to downscale gridded concentration forecasts and further improve the performances at the locations of monitoring sites. Of course, the closer the ensemble is to the observations, the harder it will be to improve the performances with the MOS. Also, note that the MOS methodology we present was applied to the CAMS ensemble outputs but it could also be applied to a single deterministic model's outputs. This might be an advantage in situations where ensemble of several models is not available.

- I believe the authors could make stronger conclusions had they tested their methodology on high pollution episodes, which are notoriously more difficult to predict.

  We did test the ability of the methodology to detect high pollution episodes (exceedances of the European regulatory threshold values) for ozone and PM10

pollutants. This evaluation is made using the performance diagram which synthesize 4 detection scores in section 5

- On the same line of the comment above, would the use of the proposed ML method improve on the predicting of exceedances for regulated pollutant? Please consider adding a comment on these.

  Prediction of threshold exceedances has been evaluated (see response for the previous point) and results are mentioned in the conclusion section. Indeed, the use of a MOS method usually improves the ability to detect threshold exceedances.

- What do you think are the implication of your proposed methodology on gridded output?

  The MOS methodology we propose is designed to correct/improve concentration forecasts at the locations of monitoring sites. Additional post-processing needs to be applied to the MOS forecasts in order to spatialize the output concentrations at any grid cells of the modelling domain. Such post-processing could be based on land use regressions or kriging technics as in the French national forecasting system PREV'AIR (Honoré et al., 2008; Rouïl et al., 2009).

- Please avoid the use of acronyms in the conclusion*

  Thank you, CSI (Critical Success Index) acronym has been suppressed in the conclusion, referring in a more general way to detection performances. CAMS, MOS and GBM acronyms are now re-introduced in the conclusion section.