**Authors' response to reviewer #1**

Please note that in this document, colors' codes are as the referee's comments in black, the authors' responses in blue. Authors' changes in the manuscripts are shown in blue bold. The tables and figures are referred based on their numbers in the revised manuscript with markup (RM).

The authors present a thorough investigation of the variability in surface ozone of two related CAMS products compared to a comprehensive set of ozone observations distributed over Iran. To account for the fact that the global model simulations are not optimized for these conditions the authors have developed a downscaling approach based on a so-called LSTM neural network method, with, apart from modeled ozone, also assimilated meteorological quantities, as well as lagged $O_3$ observations. They show the benefit of the LSTM method compared to using the raw CAMS products for providing $O_3$. Also particularly the importance of the lagged $O_3$ observations was quantified.

The authors highly acknowledge the referee for spending time and providing valuable and punctilious comments.

I consider this manuscript well suited for publication in GMD, considering the comprehensive analyses presented here, including the development of the LSTM method, and the analysis of the different model versions to represent the ozone variability at various temporal scales and different stations and regions over Iran, which is also of wider interest. My only main hesitation concerns the difficulty to follow exactly what approach the authors have taken in their methodology. Any further revisions that help to clarify (and improve readability) their methods is still welcome. Also some further discussion on the implications of your study, e.g., for the use of coarse-scale global model data such as CAMS for policy applications (?) or possibly the scope of the methods developed here for wider application (?) should be better highlighted, to better place this work in a wider context. Part of the scope of this work is indeed mentioned on line 56-60 of the introduction, but there is no explicit answer to these interesting resarch questions in the abstract or the conclusions - only indirectly.

- In the RM, the methodology was improved from two perspectives, i.e., grammatically and context (adding more details), as follows:

L174-176: **This section has been divided into three sections. Sect. 3.1 details the theory of decompositions and the method used in this study. Sect. 3.2 describes the procedure for neural network modeling and the pre-processing of its input. Sect. 3.3 defines the metrics (indicators) that have been used to assess the CAMS performance and error sources.**

L199-204: **As expected from Eq. (1), KZ (5, 5) filters all periods less than 11.2 time steps. This corresponds to 33.54 hours or 1.4 days, as the data are recorded in an interval of 3-hours. The same holds for KZ (35, 5), which filters all periods less than 9.8 days. Hence, the S refers to the short scale fluctuations which is done in less than 1.4 days. Similarly, M refers to the synoptic scales events with a time scales ranging from 1.4 to 9.8 days. The variations with the time scales of more than 9.8 days are represented in L term.**

L217-237: **A range of control values for several hyperparameters (Table A2) were tested by multiple trial-and-errors. The most effective hyperparameters (Table A5) were selected using the Random Search optimization method. To prepare the LSTM inputs, several meteorological**

**variables (Table A3) were obtained from the CAMSRA and CAMSFC datasets. To prevent overfitting of the model, a cross-validation Lasso regression was performed to identify the potential predictors at each station. The lagged $O_3$ (from OBS) was also considered as the model inputs, since the concentration of $O_3$ is not only affected by meteorological factors but also by the influence of the $O_3$ levels in the past.**

L259-264: **That could arise from overfitting associated with complex chemical processes in the model or imbalance among coupled components. The E3 represents an unexplained error, reflecting the lack of observed variability in the modelled data. That refers to the variabilities which are not captured by the models, even though those variabilities exist in the observations. The E3 can arise from random and non-representative errors caused by sub-scales and non-resolvable processes in the observations, or from a deficiency of the model in capturing meso-scale phenomena.**

- Our methodology can be used for the validation of the other chemical species (simulated by global models). Besides, it can be used for the predictions of chemical species. Accordingly, the following text was added to the end of abstract as:

L20-22: **This study demonstrates that coarse-scale global model data such as CAMS needs to be downscaled for regulatory purposes or policy applications at local scales. Our method can be useful not only for the evaluation but also for the prediction of other chemical species, such as aerosols.**

- Yes, that was a missing point. An explicit answer to the questions were added in the abstract:

L15-17: **Results show the benefit of the LSTM method compared to using the raw CAMS products for providing $O_3$ over Iran. It is found that lagged $O_3$ observation has a larger contribution than other predictors in improving the LSTM.**

- The implication and an explicit answer to the questions were added in the conclusions as:

L616-620: **To date, most of the studies of ozone and other pollutants in Iran rely on reanalysis products, without using decompositions or downscaling procedures. Our findings show that the CAMSRA and CAMSFC datasets have some deficiencies in simulating ozone, in particular over the cities with high emissions of ozone precursors. Downscaling improves these products and makes them suitable for the study of ozone in major metropolitan areas. The method used in this study is not only applicable for the evaluation of the global models but also for prediction purposes.**

**More specific comments:**

- L16: “correspondence precision” - not clear what this is - suggest to use another wording here.

Right, the word “association” could be a better choice. That was modified as

L17: **more associations**

- end of abstract (and end of conclusions): I expect a sentence that briefly describes the implications of your study. Same comment holds for the (end of) the conclusions.

  Right. Those were missing points which were added at the end of abstract and the end of conclusions as:

  L20-22: **This study demonstrates that coarse-scale global model data such as CAMS needs to be downscaled for regulatory purposes or policy applications at local scales. Our method can be useful not only for the evaluation but also for the prediction of other chemical species, such as aerosols.**

  L616-620: **To date, most of the studies of ozone and other pollutants in Iran rely on reanalysis products, without using decompositions or downscaling procedures. Our findings show that the CAMSRA and CAMSFC datasets have some deficiencies in simulating ozone, in particular over the cities with high emissions of ozone precursors. Downscaling improves these products and makes them suitable for the study of ozone in major metropolitan areas. The method used in this study is not only applicable for the evaluation of the global models but also for prediction purposes.**

- L41: suggest to change this sentence to: “In recent years, the Copernicus Atmosphere Monitoring Service (CAMS) has been mainly developed to assimilate observations of chemical composition to provide analyses of tropospheric ozone and aerosol concentrations,… ”

  That was modified as

  L61-62: **In recent years, the Copernicus Atmosphere Monitoring Service (CAMS) has been mainly developed to assimilate observations of chemical compositions to provide analyses of tropospheric ozone and aerosol concentrations, …**

- L43: “and a control run (no assimilation)” -> “and a control run (without assimilation of atmospheric composition)”

  That was modified as

  L64-65: **a control run (without assimilation of atmospheric composition).**

- L75: "…using a four-dimensional variational (4D-Var) scheme as…

  That was modified as

  L102: **using a four-dimensional variational (4D-Var) scheme as …**

- L80: "MERAA"-> "MERRA"

  A typo; that was corrected.

  L106: **MERRA**

- L84: "It is noteworthy that newer versions of data have been frequently adopted in CAMS." : it is unclear what the authors want to convey in this sentence. Is it that different satellite data have been used in the Reanalysis product, or that different CAMS reanalysis products exist, with CAMSRA the latest and most comprehensive, to date?

  The adoptions refers to the CAMS upgrades such as improving horizontal resolutions, vertical levels, newer version of the satellite retrievals. CAMS uses various satellite observations, covering different time periods (Table 2 in Innes et al., 2019). So the text was modified as:

  L111-112: **CAMS carries several upgrades, such as improving horizontal resolutions, vertical levels, and the newer version of the satellite retrievals. CAMS uses various satellite observations, covering different time periods.**

- L91: "Compared to CAMSRA, in CAMSFC only the initial conditions of each forecast are obtained from reanalysis datasets, i.e.,.." consider change to "Compared to CAMSRA, in CAMSFC the initial conditions of each forecast are obtained from analyses of atmospheric composition in near-real time, i.e.,…"

  That was modified as

  L122-123: **Compared to CAMSRA, in CAMSFC the initial conditions of each forecast are obtained from analysis of atmospheric composition in near-real time,**

- L98: "Biomass burning injects from GFAS" -> "biomass burning emissions are based on GFAS".

  That was modified as

L130: **Biomass burning emissions are based on GFAS.**

- L102: “from 9 July 2019 onwards,…”

  Right. That was modified.

  L133: **From 9 July 2019 onwards,**

- L149-150: “KZ(35,5)” - I understand that 35 here refers to ‘m’, the window size. But can the authors please explain why they choose the value of 35 here? (and a value of 5 in the definition of S in eqn. 2) Does this correspond to a filtering time scale of 35 x 3hr = approx. 105 hr, i.e. 4 days?

  Yes, the values are corresponds to the filtering time scales. Based on Eq. (1):

  $35 \times \sqrt{5} = 78.3$ (time steps)

  As the data are 3-hourly so:

  $78.3 \times 3 = 234.8$ (hours) $\approx 9.8$ (days)

  This point was clarified in RM as:

  L199-204: **As expected from Eq. (1), KZ (5, 5) filters all periods less than 11.2 time steps. This corresponds to 33.54 hours or 1.4 days, as the data are recorded in an interval of 3-hours. The same holds for KZ (35, 5), which filters all periods less than 9.8 days. Hence, the S refers to the short scale fluctuations which is done in less than 1.4 days. Similarly, M refers to the synoptic scales events with a time scales ranging from 1.4 to 9.8 days. The variations with the time scales of more than 9.8 days are represented in L term.**

- L189-191: As a modeler on initial reading I find this split in definition between ‘explained’ and ‘unexplained’ error a bit artificial. Different to what is suggested, I would also not have a direct understanding of the cause of ‘explained error’. After reading the manuscript, I think I better understand the arguments of calling errors either ‘explained’ or ‘unexplained’, but it might help to allude to that.

  The explained error refers to errors, which arise from the model, i.e., $\sigma_m - \mathrm{r}\, \sigma_o$, in Eq. (5). For instance, the model shows some variabilities, which are unseen in the observed data. On the other hand, the unexplained error refers to the variabilities which are not captured by the model, even though those variabilities exist in the observations $\sigma_o^2(1-r^2)$ in Eq. (5). This point was explained in the RM as:

L259-264: **That could arise from overfitting associated with complex chemical processes in the model or imbalance among coupled components. The E3 represents an unexplained error, reflecting the lack of observed variability in the modelled data. That refers to the variabilities which are not captured by the models, even though those variabilities exist in the observations. The E3 can arise from random and non-representative errors caused by sub-scales and non-resolvable processes in the observations, or from a deficiency of the model in capturing meso-scale phenomena.**

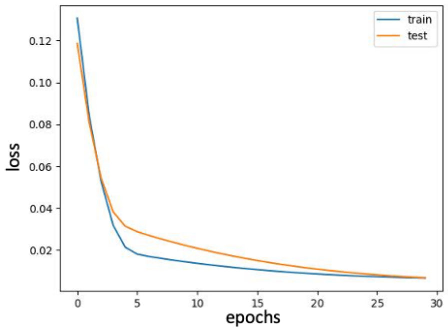
- L235: “for most of the station**s…**”

  Right. That was modified.

  L322: **for most of the stations**

- L246 and L248: The authors refer here to ‘opochs’. please provide an explanation what an ‘epoch’ is exactly, in this context. I missed that.

  LSTM model requires a specific configuration and tuning to work effectively with the datasets. A range of control values was tested by multiple trial-error evaluations using the Scikit function GridSearchCV. In our experiments, we tune one of the hyperparameters, i.e., epoch number from 1 to 30. This value is a hyperparameters for the learning algorithm, e.g., parameters for the learning process, not internal model parameters found by the learning process. The number of epoch is traditionally large, often hundreds or thousands, allowing the learning algorithm to run until the error from the model (loss function) has been sufficiently minimized. There are no given rules to set this parameter. One epoch leads to underfitting of the curve. As the number of epochs increases, more number of times the weight are updated in the neural network and the curve goes from underfitting to optimal curve.



  In the RM, we applied an “Early Stopping” option, which allows to specify an arbitrary large number of training epochs and stop training once the model performance stops improving on a hold out validation dataset.

  L340-348: **We tuned hyperparameters, which allow the learning algorithm to run until the error from the model, i.e., the loss function, has been sufficiently minimized. As there**

**are no given values to set these numbers, the optimum values were obtained by multiple trial-and-error tests (see Table A5).**

- L250-251: “That might reflect that the more predictors, the better the model would not be.” : as also reflected in the conclusions, I find this an important finding indeed. I’d suggest to stress this a bit better, also by re-formulating this sentence a little - now it reads a little clumsy. This finding may be worth a bit more statistical analysis, i.e. do the authors have any quantitative metric arising from the method which provides insight as to how much each of the individual parameters contributes to the quality of the end product? It might be a useful exercise to exclude some of the (physically) less obvious parameters from the list of fitting parameters, such as U, V, W, MSLP (?). Here an analysis of station 22 (Yazd), which performs relatively poor, while it uses an excessive list of input data, suggests indeed the limitations of this work. Can the authors comment?

That is indeed a very good point and suggestion. To decide on the importance of the variables we used LassoCV estimator. The variables with the highest absolute Lasso coefficient (importance weight) are considered the most important. For instance, Fig. 4 shows that the T2m is the most explanatory meteorological variable and NO, $NO_2$, and $O_3^{RA}$ are the main chemical variables for CAMSRA_S at most of the stations. The higher the weight value, the more the influence the variable has and hence more important. Table A6 lists the more influential variables on ozone variability at most of the stations.

To assess the sensitivity of the model to the less obvious predicators, we designed two experiments. In first experiment ($MLR^{no\_lag(expr1)}$), the model was trained only using $O_3^{RA}$ and $O_3^{FC}$, while in second experiment ($MLR^{no\_lag(expr2)}$), the model was trained using the meteorological variables with high priority (as listed in Table A6). Both experiment were preform using $MLR^{no\_lag}$.

Our analysis shows a low value of $R^2$ for the S component at station Yazd (22), but that is relatively significant for the M term. Besides, the MSE at this station is very low (so good model performance). That could be associated with the station locations, which are less populated and less affected by local anthropogenic emissions sources (and easier to model). That is not related to the excessive list of input data. As this experiment shows by excluding most of the parameters, the MSE at this stations changes from 14.94 to 16.06 (does not change that much).

L568-575: **Two experiments were designed to assess the sensitivity of the model to less obvious predictors. In the first experiment, i.e., $MLR^{no\_lag(expr1)}$, the model was trained only using $O_3^{RA}$ and $O_3^{FC}$. In the second experiment, i.e., $MLR^{no\_lag(expr2)}$, the model was trained using the most influential meteorological variables (see Table A6). For the sake of simplicity (and being less expensive), both experiments were performed using the $MLR^{no\_lag}$ model. Table A7 lists the results of these experiments for station 22 (Yazd). As can be seen, the MSE of $MLR^{no\_lag(expr1)}$ and $MLR^{no\_lag(expr2)}$ are larger than that of $MLR^{no\_lag}$. That shows that part of the $O_3$ variability is explained by meteorology and**

**partly by the chemistry ($O_3^{RA}$ or $O_3^{FC}$). Separating these two factors causes a decline of *r* (see Fig. A9).**

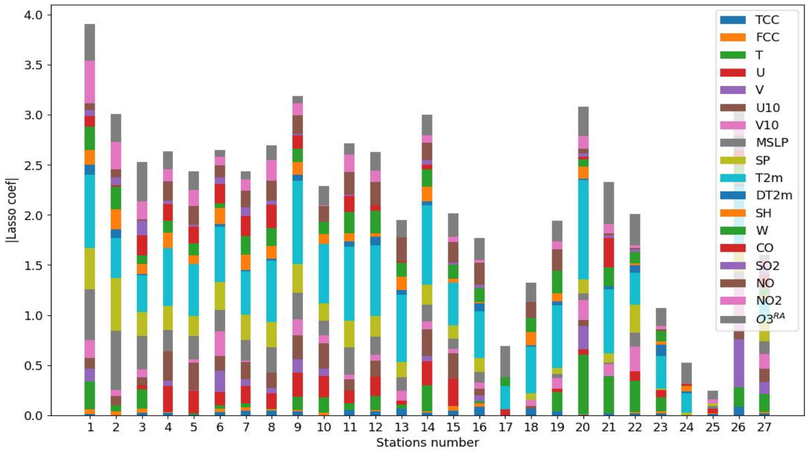


**Figure 4. Cross-validation Lasso regression to identify the potential predictors for ozone modeling. The higher absolute Lasso coefficient, the most important would be the variable.**

**Table A6. The most important explanatory variables of the models at most of the stations.**

| | **Chemical species** | **Meteorological variables** |
|---|---|---|
| **CAMSRA_S** | **NO, $NO_2$, $O_3^{RA}$** | **T2m** |
| **CAMSFC_S** | **$O_3^{FC}$** | **BLH, V10m** |
| **CAMSRA_M** | **$O_3^{RA}$** | **TCC, U** |
| **CAMSFC_M** | **$O_3^{FC}$** | **-** |

**Table A7. The results of the experiments (1) $MLR^{no\_lag(expr1)}$: the model was trained only using $O_3^{RA}$ and $O_3^{FC}$, (2) $MLR^{no\_lag(expr2)}$: the model was trained using the meteorological variables with high priority (listed in Table A6) at station 22 (Yazd). The *r* refers to the correlation coefficient between $O_3^{SD}$ and measured $O_3$.**

| | **$MLR^{no\_lag}$** | | **$MLR^{no\_lag(expr1)}$** | | **$MLR^{no\_lag(expr2)}$** | |
|---|---|---|---|---|---|---|
| | MSE | *r* | MSE | *r* | MSE | *r* |
| CAMSRA_S | 14.94 | 0.41 | 16.06 | 0.33 | 16.09 | 0.32 |
| CAMSFC_S | 14.69 | 0.43 | 16.30 | 0.31 | 16.01 | 0.33 |
| CAMSRA_M | 1.85 | 0.61 | 2.81 | 0.22 | 2.92 | 0.10 |
| CAMSFC_M | 1.77 | 0.63 | 2.90 | 0.12 | - | - |

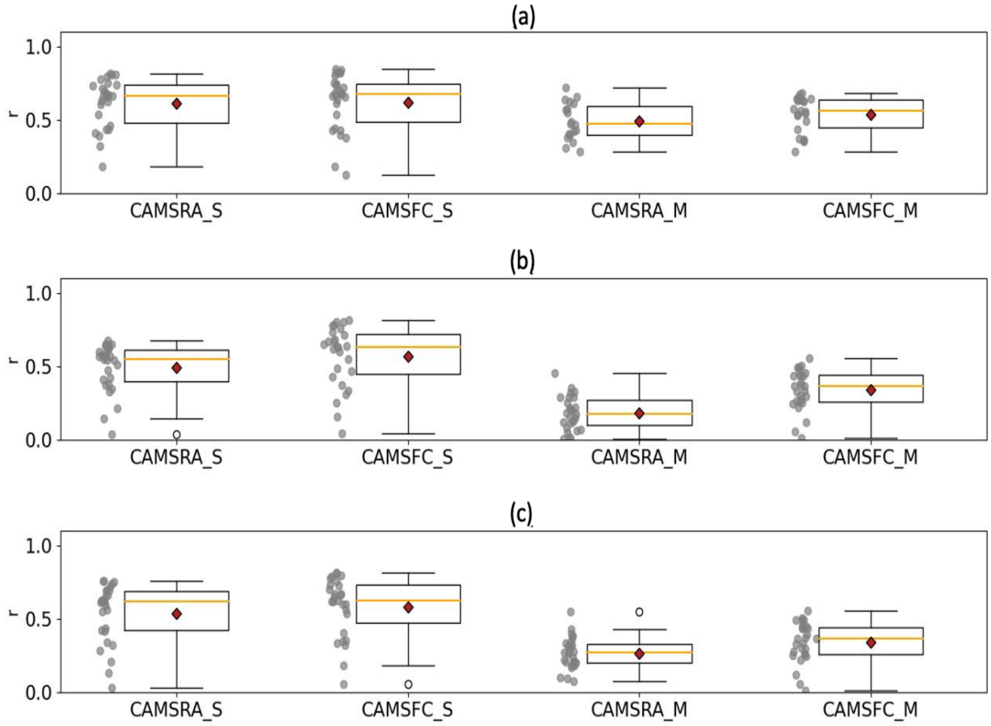


**Figure A9. The correlation (*r*) between measured $O_3$ and $O_3^{SD}$ by the (a) $MLR^{no\_lag}$, (b) $MLR^{no\_lag\ (expr1)}$, and (c) $MLR^{no\_lag\ (expr2)}$ models.**

- L282: "lagged $O_3$" please specify here (again) that this refers to lagged $O_3$ from actual observations, to help the reader understand.

  Sure, that was applied as

  L429-430: **In order to examine the effect of the CAMS products and lagged $O_3$ (from actual observations) on the LSTM model, we exclude the measured lagged ozone from the predictors of the LSTM model.**

- L301: typo 'products'

  That was corrected. This typo was also fixed in L9 and Table A3.

- L328: suggest to drop the sentence "These values…" - no need?

  That was modified.

  L507: **respectively, and increase to …**

- L364:”peroxides”->”proxies”

  Its sentences was deleted.

- L377-380: I’d expect here a comment on the implication of these findings, e.g. the importance of observed (lagged) $O_3$ as predictor (?) and/or the potential use cases of the methods as the authors have developed.

  Very good suggestion. So, the changes applied in the RM are as follow:

  L605: **That shows the importance of the observed (lagged) $O_3$ as a predictor in the LSTM.**

- Table 1: “single level” -> “surface level”

  Right. That was modified.

  Table 1: **surface level**

- Table A3: typo in units for UV

  That was corrected:

  Table A3: **J $m^{-2}$**