Post-process correction improves the accuracy of satellite $PM_{2.5}$ retrievals manuscript - Reply to Adam Povey, Supriya Mantri and Laura Horton

Andrea Porcheddu, Ville Kolehmainen, Timo Lähivaara, Antti Lipponen

We would like to thank you for reading the manuscript and giving comments. Below we answer to each of the comments.

1 Answers

1.1 The introduction was clear with good motivations and references but, in line 25, the reference to the WHO air quality guideline does not specify the time frame. Is it an annual average?

Yes, it is an annual average.

1.2 Is it possible to give any additional reasoning behind the use of hourly downscaling of daily PM averages? This strikes us as one of the more consequential choices in this methodology and neglects complexities such as the diurnal cycle. (We aren't questioning the authors' judgement, nor disagreeing; merely curious as we wouldn't have thought of this.)

We are simply unrolling 24h averages given every hour, so the hourly measurements information is contained in each 24h average.

1.3 In section 2.3 the authors mention they use MERRA-2 reanalysis variables as an input for the model and provide a lengthy list in the appendix. Have all of the available variables been used as an input or are there some variables not included? It may be helpful for those designing similar algorithms if you also mention which variables were not included and, if so, why not?

A table with all the features used in the training of the model is available at the end of the manuscript. We have used all the MERRA-2 reanalysis variables mentioned in the table. We have also carried out SHAP analysis that showed us that all MERRA-2 input variables were informative and had non-negligible effect on the model output.

1.4 Could you provide a reference for the RH equation on line 93? Or is it at standard temperature/pressure?

The RH equation is based on the Clausius-Clapeyron equation. We have now mentioned the Clausius-Clapeyron equation and included a reference to a paper using the formula and having more information about the equation in the revised manuscript.

1.5 In the NASA Black Marble Night Light section (2.6.2), the authors could include a reference to the data set used, especially a DOI so we can distinguish which of the four available datasets was used.

We have added a reference to the dataset in the revised manuscript.

1.6 We found Sections 3.3 and 3.4 quite opaque. They could be improved by adding more specific details about the process, such as the equations used as an input to the correction model. A flow chart of the steps used in the approach would make it more clear for an audience which is less familiar with machine learning and neural networks.

The sections have been improved adding some more equations and explanations of their content.

1.7 Those of us with little experience with neural networks did not understand what Figure 2 wished to convey and those of us used to neural networks felt Figure 2 adds little to the text; it could be removed.

Thank you for the suggestion but we decided to keep the figure since we think it could be informative.

1.8 At line 183, it would be helpful to have an understanding of what is meant by "slightly better"?

We modified the text from "slightly better" to "better".

1.9 Figure 3 could be improved as it is hard to distinguish between lines, perhaps using a filled histogram of stacked bars.

We modified the plot with thicker lines and no transparency for the colors.

1.10 We are curious if the authors considered any methods to amplify the availability of high PM2.5 observations, such as data augmentation? Our understanding was that the balancing of training data is an important step in constructing a neural network to recognise rare events and we would value the authors' opinion.

We have tested augmenting the data using CTGAN (Conditional Tabular Generative Adversarial Network) and TVAE (Tabular Variational Auto Encoder) (Xu et al., 2019) but didn't observe any improvement.

1.11 From line 223, what do you mean by "the fully learned approach were less accurate than with the post correction approach"? What metric of accuracy was used and how significant was the difference? This would help guide our own efforts in neural network generation.

We calculated different metrics including R^2, RMSE and MAE. In particular, for the post-correction approach we obtained R^2=0.55, RMSE=6.2 $\mu g/m^3$ and MAE=4.26 $\mu g/m^3$, while for the fully-learned approach we obtained R^2=0.51, RMSE=6.5 $\mu g/m^3$ and MAE=4.50 $\mu g/m^3$.

1.12 In Figure 4, are the authors certain about the plotting of panel B? Compared to panel A there appears to be some duplication of points up the y-axis. For example, there are three points at the extreme right of (A) but over ten in (B). This effect is not exhibited in panels D-F.

Yes, we are sure about the plotting. We checked further and we found no problems.

1.13 In Figure 5 (comparison of the uncorrected and corrected methods at the ground stations), do the authors have any understanding of why there is a large discrepancy between OpenAQ and both satellite estimates for most of the sites (i.e. the blue dots do not overlap the red line in 5/9 cases shown). Is it because of some local source (e.g. roads or small industrial buildings) in close proximity of the stations that isn't present in Madrid?

We looked at the correlations between input features and errors and we did't find any single feature that could explain the errors for the shown satellite overpass. Please note that the monthly averages for the test station in Paris are well aligned with the ground based data.

1.14 In Figures 5 and 6, could the dots representing sites and arrows indicating one be made substantially larger and outlined with a colour not in the plot (such as green or blue)? Our older member had failed to see them on his own.

The figures have been modified now, we chose green colour and larger arrow/font.

1.15 In Figure 7 bottom left (Hull Freetown) why is the uncertainty so large April? What are the possible reasons for high PM2.5 in February?

We don't have any explanation for the large uncertainty in April: we looked at correlations between errors and input data but we didn't find anything interesting. For what regards the high $PM_{2.5}$ in February, it could be caused by the lower boundary layer height in winter months (as other causes like house heating, etc.).

1.16 If practical, it would be interesting to add an appendix highlighting a few ensemble members for Fig 5 and/or 6. This would demonstrate if the smoothness of the fields shown at the top right of those figures is due to the action of the neural network or due to the median filter over the ensemble.

The smoothness of the fields is not due to the median filter since the same smoothness can be seen in the figure regarding the not-corrected method (where no median filter is applied).

1.17 It would be interesting to hear if training the model over an area with higher PM2.5 levels, such as South Africa or India, and then testing over central Europe improves the model's performance, particularly for higher PM2.5 values.

That's our plan for future studies, we are curious to see if gathering more data from all around the world could improve the modelling in Europe.

References

Xu, L., Skoularidou, M., Cuesta-Infante, A., and Veeramachaneni, K.: Modeling Tabular data using Conditional GAN, 2019.