Response letter for reviews of "Improving 3-day deterministic air pollution forecasts using machine learning algorithms"

Referee comments in black and replies in *blue, italics.*

Replies to Editor

The authors have implemented substantial changes to address the reviewer comments so that the manuscript is now ready for publication subject to technical corrections. In particular, the authors are asked to:

- check for effective ways to shorten the main part of their manuscript, e.g. through re-wording or by moving sections to the Supplementary Material. While it might be important to demonstrate certain points, especially in response to valid comments made by the the reviewers, the authors could re-consider whether a pointer to the Supplementary might suffice in a few cases. Further efforts to tighten the presentation where possible would enhance accessibility. Candidates could be the interpolation/missing data section (Figures 3 and 4), which could be moved to the appendix, as could be the part on hyperparameter tuning.

Reply:

Thank you so much for your decision on accepting the paper subject to minor revisions! We have conducted some revisions of the paper to shorten the main text according to your suggestion.

Specifically, we moved the example of the interpolation result, Figure 3, and the example of the hyperparameter tuning result, Figure 6, to Appendix B and C, respectively.

- re-read the manuscript and check for grammatical errors and typos. While there will be copyediting, this will reduce the risk of carrying over errors into the published version of the manuscript, especially those that are hard to spot for copyeditors. For example, I noticed several misspellings such as 'learin_rate'. In addition, there should always be spaces separating words from abbreviations (e.g. line 24 "exPlanations(SHAP)")

Reply:

Thank you! We did a comprehensive copyediting and corrected the grammatical errors and typos we found.

Please also consider brief explanations (maybe a couple of sentences each) on the following points mentioned in the letters to the editor:

- Being more specific on the exact imputation techniques used would add helpful details.

Reply:
The essence of interpolation is to find the samples that are most correlated with the missing sample and replace the missing value with these correlated samples. In the context of time series data, the samples at time "t" are strongly correlated with time "t ± 1". At the same time, they are highly correlated with samples at time "t-p" and "t-2p", where "p" is the data's periodicity.

Considering the above two points and our prediction scheme, detailed in section 3.2, we adopted the historical average interpolation based on the periodicity. Subsequently, the missing value at time t is substituted by the average of the available data from two preceding periods (i.e., "t – p" and "t – 2p") as well as their adjacent values (i.e., "t - p ± 1" and "t - 2p ± 1").

- Discussing limitations around still struggling to capture high concentrations would add useful context.

Reply: Thank you for the comments ! Our models indeed face challenges when predicting high-concentration values. Below, we outline the key limitations and potential reasons:

1. Sample Imbalance:
High-concentration peaks have minor occurrence in the sample, and they are not as well-represented in our training dataset as low or normal values. Such imbalance may introduce bias in the prediction models, leading to reduced performance in predicting high-concentration values.

2. Model Limitations:
While the machine learning models have effectively improved the predictions by the deterministic model, they are still not explicitly designed to capture extreme events. The model's architecture, feature selection, or training process may have limitations in handling high pollutant concentrations.

3. External Factors:
High pollutant concentrations can be influenced by external factors, such as sudden changes in meteorological conditions, industrial incidents, or instantaneous emission sources. These factors either do not belong to our feature space or are not adequately captured by our models due to their unpredictability, leading to reduced prediction accuracy for high-concentration values.

In our future work, we therefore plan to explore methodologies for refining our model architecture and enhance the model capability of predicting high pollutant concentrations.

Finally:
- Figures 9 and 14 should be updated to have different colours for MDI/gradient methods.

Reply:
We have modified all figures for feature ranking, including Figure 7, Figure 12 and the ones in the Appendix, also MDI for blue points and Gradient for purple points, respectively.

- The radar plots are difficult to read. For example, the bottom of Figure 11 is cut off, but more importantly the axis in the title is said to be e.g. 0 - 1, but the axis on the radar plot is labelled in %s. It would be simpler if these were labelled with the numbers, rather than percentages. This would also remove the need for declaring the axis range in the title.

Reply:
We updated all radar plots, including Figure 5, Figure 8 and Figure 9. Numbers are labelled on each axis to make it easier to read.

<u>Replies to Referee #1</u>

Thank you for submitting the revised manuscript, which addresses most of my comments, including the values of hyperparameters and the details of the tuning, the inclusion and explanation of different feature importance methods and a clear explanation of how the model was trained re different seeds. I am happy to accept that the study is limited to coronavirus years, and looking at different years will be a part of future work. I still have a couple of minor comments on the figures
1. Figures 9 and 14 should be updated to have different colours for MDI/gradient methods.

Reply:
Thank you for your valuable suggestions and efforts in previous rounds of reviewing.

We have modified all figures for feature ranking, including Figure 7, Figure 12 and the ones in the Appendix, also MDI for blue points and Gradient for purple points, respectively.

2. The radar plots are difficult to read. For example, the bottom of Figure 11 is cut off, but more importantly the axis in the title is said to be e.g. 0 - 1, but the axis on the radar plot is labelled in %s. It would be simpler if these were labelled with the numbers, rather than percentages. This would also remove the need for declaring the axis range in the title.

Reply:
We updated all radar plots, including Figure 5, Figure 8 and Figure 9. Values are now labelled on each axis to make it easier to read and understand.

Replies to Referee #2

N/A