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

Thank you to the authors and reviewers for their comments and revisions. However, both reviewers have indicated a continued need for substantial revisions and clarifications in order to make the manuscript acceptable for publication in ACP.
In particular, an acceptable version would require:
shortening and restructuring of a few sections (potential for where this could be achieved was highlighted by both reviewers in the two rounds of reviews).

Reply:
Thanks for your comments and suggestions! We have conducted a comprehensive revision of the paper and carried out additional computational experiments in order to answer all the questions being raised.

We have revised the paper structure, making lots of effort to shorten the paper into a neat form. In summary, the conducted revisions are outlined as follows:

Structure changes:

- *We have smoothed and shortened the description of Section I.*
- *The Air Quality System of Stockholm and meteorological forecast subsections are moved from the section of "methods" to a new section called "Background".*
- *Some results are moved to appendices to shorten the length of the main manuscript, e.g. some of Figures 7 – 14.*
- *We have merged the figures and created radar plots to simplify the presentation of the final results.*
- *Discussion section is revised to reduce the paper length.*

Additional content

- *We added two new subsections of data preprocessing and feature importance assessment in the methodology presentation.*
- *We added some explanations and results for hyperparameter tuning for each model.*
- *We added two new feature ranking methods (Permutation and SHAP) to satisfy the requirements of the reviewers*
- *We added some detailed analysis of feature importance in the result and discussion sections.*

Meanwhile, we complement necessary clarifications in the text in response to the concerns and highlighted comments of both reviewers. Please refer to the individual answer to each question of the reviewers.

additional analyses (e.g. addition to existing tables) contrasting different methods for measuring feature importances.

Reply:
As mentioned in the outline before, two additional methods were implemented into both the tree-based model (XGBoost and RandomForest) and LSTM model to thoroughly assess the stability of the feature's importance ranking and the temporal dependence of crucial features.

Subsection 3.5 was added to explain the methods of feature importance ranking. Tree-based models commonly employ three methods: MDI, permutation (Breiman, 2001), and SHAP (Lundberg et al., 2017). Also, LSTM models typically utilize gradient-based method, permutation, and SHAP (Shrikumar et al., 2017) for measuring importance.

Correspondingly, we add subsection 5.2 to discuss the results of feature analysis.

a test of how the selection of the COVID-affected year might influence estimated model performance

Reply:
Indeed, the COVID-19 pandemic had an impact on road traffic and pollutant emissions as a result of some restrictive regulations implemented (Sokhi et al., 2021; Torkmahalleh et al., 2021). However, the COVID restrictions in Sweden continued until February 2022, but our data in this study only encompasses observations up to December 31, 2021. So, the dataset is within entire duration of the COVID period, which makes it impossible to assess the effect of COVID-19 in this study. Nevertheless, since we will continue the model development in the longer run, the effect of COVID will be investigated when new data after 2022 is included in the model training and prediction. We expect this will be presented in our future work.

more details on the data pre-processing (and phrasing concerning its importance); if too long for the main text at least in the appendix
Reply:
We have added data preprocessing method in subsection 3.1, which describes the preprocessing method, and the effect of data interpolation.

a correct context and comparison of skill metrics with similar symbolic representations (r2-score is not the same as Pearson correlation)

Reply:

Thank you for the comment. There is a notation in the header of tables 3/4/5 in the previous version that "r" represents Pearson correlation, sorry for the misunderstanding !

We currently replace Pearson correlation with R-squared in the result tables, and the calculation formula can be found in subsection 3.6.

it is still not clear enough how/if/where hyperparameter tuning was conducted. The reviewers note that "This needs to be mentioned in the main text alongside the added list of actual parameters used. The LSTM was not subject to hyperparameter optimization. It is important for readers to understand that model and training choices are not arbitrary guess-work; it is a critical component for any model with the intention of deployment." and "they mention that they trained by setting different random number seeds to obtain optimal results. Searching over random seeds to find optimal results is likely to lead to overstating the performance of the model when it comes to using the model operationally or on other data?" Good answers to these points raised are critical to ensure the robustness of the results presented in the manuscript.
Overall, it would be necessary to comprehensively and convincingly address these concerns before the manuscript can be considered for publication. If necessary, analyses might need to be repeated/revised if e.g. the current hyperparameter tuning cannot be demonstrated to be robust (rotation of test vs. training years?). In addition, further minor comments by the reviewers would need to be addressed point-by-point.

Reply:
Thank you for the comments! We agree that hyperparameter tuning is an important step in machine learning model construction. To clarify our work, we created subsection 3.4 on hyperparameter tuning. The optimal parameters and searching ranges for each model are shown in Appendix B.

To demonstrate the robustness and generalization ability of the models, we actually trained the model repeatedly 10 times by setting different random seeds. The results are evaluated using independent data and the mean value along with its 95% confidence interval is presented in Table 5/6/7. The findings reveal a quite small variation for the two tree-based models(XGBoost and RandomForest), whereas the LSTM model shows some fluctuation but less than 5% in most cases. This variance may be attributed to the stochastic initialization of weights, which influences the

subsequent trajectory of gradient descents and model adaptation in the training process.

We have addressed the concerns of the reviewers, point by point. Please refer to our answers to the reviewer's comments.

Replies to Referee #1

This paper has piece-meal improved – the track changes show that the most prevalent change was replacing MLs with ML models; I was expecting more significant changes or comments to be incorporated. There are several comments highlighted below made by the authors where it seems they misunderstand the importance of some of the ML methods used.

Reply:

Thanks for your suggestions and comments on this paper! We agree that the improvement in the first round is not comprehensive, and some of the points raised are missed. Furthermore, we have carried out a comprehensive revision and lots of additional computation experiments in the past few months to improve the manuscript and science behind it. The change we made can be summarized as follows:

Structure changes:

- *We have smoothed and shortened the description of Section I.*
- *The Air Quality System of Stockholm and meteorological forecast subsections are moved from the section of "methods" to a new section called "Background".*
- *Some results are moved to appendices to shorten the length of the main manuscript, e.g. some of Figures 7 – 14.*
- *We have merged the figures and created radar plots to simplify the presentation of the final results.*
- *Discussion section is revised to reduce the paper length.*

Additional content

- *We added two new subsections of data preprocessing and feature importance assessment in the methodology presentation.*
- *We added some explanations and results for hyperparameter tuning for each model.*
- *We added two new feature ranking methods (Permutation and SHAP) to satisfy the requirements of the reviewers*
- *We added some detailed analysis of feature importance in the result and discussion sections.*

[1] The explanation for Tables 2 and 4 being kept in the main text is fair, and the addition of bold-face makes it a lot easier to see the top-1 in these tables. However, the feature importance plots are not shown in the main text. If the authors misunderstood last time: The feature importance results are some of the main ML results of the paper and they need to be in the main text at the expense of some of figures 7 through 14. For example, show only Figure 8 and not both Figures 7 and 8, in addition to tables, because Figure 8 shows the skill-score metric. that you actually care about. You can move the other less important results showing many columns of bake-off results to the appendix / supplementary materials and reference them in the main text. That alone would represent a major improvement to the manuscript.

Reply:
Thank you for constructive suggestion! We have made new efforts to improve the presentation. We introduce radar charts to make it easier to present and compare performance metrics calculated from different models. In addition, for Figures 7-Figure 14, we merged some of the graphs to reduce space and moved the remaining results to the Appendix to shorten this manuscript.

[2] The authors state that "how to interpret feature importance has been a side-line topic for understanding the RNN model" and then "there are other approaches to measure feature importance such as (Zhou et al), but this is beyond the scope of our current study.
I do not understand these comments since the authors correctly note that different methods (and model parameterizations) can lead to different results. The same is also true for tree-based models, which is why I asked in the first round for the authors to provide at least one other method for those models, which they did not do. The application of these methods are not beyond the scope of this paper, these are key details that the authors missed.

Reply:
We didn't understand the comments precisely in the first round. Nevertheless, we had some internal discussions and finally implemented different feature ranking methods by following the suggestions of the reviewers.
We add subsection 3.5 in the main text to illustrate the feature ranking methods. We did thorough experiments to analyse the results, which are added in the sections of results and discussion.

[3] Since reviewer #2 also noted the permutation importance method, this needs to be shown alongside MDI for the tree models (why even the choice of MDI, why not fANOVA? These two alone commonly give different rankings). The permutation method can be applied to both the tree models and the RNN (as can SHAP). You can stack the results from MDI along with permutation results without having to add another figure in the paper. The conclusion from doing such an analysis will allow the authors to identify which top features the different methods have in common, as those are the robust ones that should be selected.

Reply:
Thank you for these detailed comments and good idea! Two additional methods were implemented into both the tree-based models (XGBoost and RandomForest) and LSTM model to thoroughly assess the robustness of the feature's importance ranking and the temporal dependence of crucial features.

Subsection 3.5 was added to explain the methods of feature importance. Tree-based models commonly employ three feature ranking methods: MDI, permutation(Breiman, 2001), and SHAP(Lundberg et al., 2017). Also, LSTM models typically utilize gradient-based method, permutation, and SHAP(Shrikumar et al., 2017) for measuring importance.

The results of four feature importance methods were moved to the main text, as shown in Figures 9 and 14, and more results are displayed in Appendices E and H. In addition, we use a model as an example and discuss the correlation between features and time-dependent properties based on TreeSHAP in subsection 5.2.

[4] The authors stated "there are issues such that the gradient-based method calculates temporally varying rankings making the rankings of features dependent on the testing set". Correct. It would be a nice addition for the authors to investigate the September 20th case, again, why the large drop? Do the feature importance methods tell us what's going on?

Reply:
Thank you for the suggestion! We have tried the idea to explain why the peaks cannot be captured such as the case of 20th Sept mentioned by the reviewer, and to see if we can explain the gaps by top-ranked features. Unfortunately, we didn't find any solid explanation. Nevertheless, the top-ranked features do help understand the models better, and some features show interesting temporal patterns. We have added some analysis in subsection 4.1.2 / 4.2.2 and a discussion about the temporal dependency of the top-ranked features in subsections 5.2.

[5] The authors made the comment "In the pre-processing process, outliers, such as negative pollutant measurements, are identified and removed and furthermore, standard methods, such as interpolation, are applied to handle missing values in the data. Given the current length of the paper, adding such details will not benefit the paper's readability. But we have added a description in subsection 2.1 as follows: …" Imputation / managing missing data are critical details regarding the data preparation where the authors seem to diminish their importance with these comments.

Reply:
We added a more detailed subsection in the methodology part, subsection 3.1, which describes the missing data case, the preprocessing method, and the effect of data interpolation.
[6] The authors now state in the manuscript: "Due to the temporal correlation of the air pollutant concentrations, the principal assumption of cross-validation is not

satisfied. Therefore, to preserve the time-dependent property, “TimeSeriesSplit” was chosen as the cross-validation strategy. … The value of parameter k is set as at 5.”
I think what the authors mean to say is that cross-validation via random sampling is not satisfied and that a time-ordered strategy is required (cross-validation is being applied along the time coordinate). You need to say that you used sklearn when “TimeSeriesSplit” is mentioned for the first time, readers who are not familiar with sklearn will have no idea what it means. That an ensemble of size 5 was created means you can put the error-bars on the tables as I originally asked in the first round. You also noted that an ensemble was created via different seed choices. RNNs are usually very sensitive to initial weight choices. Whichever ensembling method you choose to report the error bars, state it clearly in the table captions.

Reply:
Thank you for the comments! We have revised the paper accordingly. To demonstrate the robustness and generalization ability of the models, the training process repeats 10 times with different random seeds. The results are evaluated on independent test set, and the mean values along with its 95% confidence interval are shown in Table 5/6/7. The results reveal quite small variation in the two tree-based models(XGBoost and RandomForest), whereas the LSTM model presents a bigger variance but less than 5% in most cases.

[7] The coefficient of determination, R2, is defined as 1 - \sum_i (y_i - f(x_i))^2 / \sum (y_i - `<y>`)^2. When used as a metric, R2=0 is the baseline model that predicts `<y>`; R2>0 is more skill-full relative to that baseline, <0 less skill-full than the baseline. Pearson correlation is not a measurement relative to a baseline, there can be high-correlation while at the same time poor skill relative to the simple baseline. Table 2 has “r = Pearson correlation” – in general the coefficient of determination is not the square of the Pearson score.

Reply:
Thank you for the comment! There is a notation in the header of tables 3/4/5 in the previous version that “r” represents Pearson correlation, sorry for the misunderstanding !

The formula for all metrics is summarised in subsection 3.6. We currently replace Pearson correlation with R-squared value in the result tables but for comparison, it is still shown in the radar plot.

[8] The authors mention that they performed a grid-search of the hyper-parameters. This needs to be mentioned in the main text alongside the added list of actual

parameters used. The LSTM was not subject to hyper-parameter optimization. It is important for readers to understand that model and training choices are not arbitrary guess-work; it is a critical component for any model with the intention of deployment.

Reply:
Yes, we agree that hyperparameter tuning is an important procedure for machine learning model construction. We add a subsection, 3.4, on hyperparameter tuning in the main text to meet the requirements of both reviewers. We show that we did the hyperparameter optimization for all models(XGBoost, RandomForest and LSTM). To shorten the presentation, the parameter selection range and optimal parameters are added in Appendix B.

Replies to Referee #2

Thank you to the authors for responding to reviewer comments.

Further comments:

1. Additional details on the data splitting is now provided, and temporal splitting has been used to train, validate and test the models. Thank you to the authors for adding this.

Reply:
Thank you for notifying this!

2. Thank you also for including the details on hyperparameters in 2.4. Currently however, it simply says that the 'default parameters' were used for the tree-based methods. The actual values of these parameters should be included, perhaps in the appendix.

Reply:
Thank you for the comments! We added a subsection, 3.4, on hyperparameter tuning in the main text to meet the requirements of both reviewers. We show that we did the hyperparameter optimization for all models (XGBoost, RandomForest and LSTM). To shorten the presentation, the parameter selection range and optimal parameters are added in Appendix B.

3. I think that the fact that the study uses data from coronavirus years should be highlighted. It may be the models are able to perform well on these years, but not on other years, even with further training? This is mentioned is the author response, but is not in the manuscript currently.

Reply:
Yes, unfortunately, our current study is based on the data until the end of 2021. We added a short discussion in subsection 3.1 as follows:

"It should be noted that there are several studies showing the impact of the COVID-19 pandemic on pollutant emissions as a result of some restrictive regulations (Sokhi et al., 2021; Torkmahalleh et al., 2021). The COVID-19 pandemic in Sweden

commenced in January 2020 and continued until February 2022, so the majority of the data is collected during this pandemic period."

Since we will continue the model development in the longer run, the effect of COVID will be investigated when new data after 2022 is included in the model training and prediction. We expect this will be presented in our future work.

4. Thank you for including feature importance details. It should be mentioned that different feature importance methods can provide different importances, and there are many options available for this.

Reply:
We have implemented two additional methods for both tree-based models (XGBoost and RandomForest) and LSTM model to thoroughly assess the stability of the feature's importance values and the temporal dependence of crucial features.

Subsection 3.5 was added to explain the methods of feature importance. Tree-based models commonly employ three feature ranking methods: MDI, permutation(Breiman, 2001), and SHAP(Lundberg et al., 2017). Also, LSTM models typically utilize gradient-based method, permutation, and SHAP(Shrikumar et al., 2017) for measuring importance.

The results of four feature importance methods were moved to the main text, as shown in Figures 9 and 14, and more results are displayed in Appendices E and H. In addition, we use a model as an example and discuss the correlation between features and time-dependent properties based on TreeSHAP in subsection 5.2.

5. In the authors' response, they mention that they trained by setting different random number seeds to obtain optimal results. Searching over random seeds to find optimal results is likely to lead to overstating the performance of the model when it comes to using the model operationally or on other data? This is also not mentioned in the manuscript.

Reply:
Thank you for the comment! To demonstrate the robustness and generalization ability of the models, the models are trained 10 times with different random seeds. The

results are evaluated on independent test data, and the mean values along with its 95% confidence interval are shown in Table 5/6/7.

References

Breiman, L.: "Random forests." Machine learning 45, 5-32, 2001.

*Lundberg, S.M. and Lee, S.I.: A unified approach to interpreting model predictions. *Advances in neural information processing systems*, *30*, 2017.*

*Shrikumar, A., Greenside, P. and Kundaje, A.: Learning important features through propagating activation differences. In *International conference on machine learning* (pp. 3145-3153). PMLR, 2017.*

*Sokhi, R.S., Singh, V., Querol, X., Finardi, S., Targino, A.C., de Fatima Andrade, M., Pavlovic, R., Garland, R.M., Massagué, J., Kong, S. and Baklanov, A.: A global observational analysis to understand changes in air quality during exceptionally low anthropogenic emission conditions. *Environment international*, *157*, p.106818, 2021.*

Torkmahalleh, M.A., Akhmetvaliyeva, Z., Darvishi Omran, A., Darvish Omran, F., Kazemitabar, M., Naseri, M., Naseri, M., Sharifi, H., Malekipirbazari, M., Kwasi Adotey, E. and Gorjinezhad, S.: Global air quality and COVID-19 pandemic: do we breathe cleaner air?, 2021.