ACP2019-173 by Vu et al.

Responses to the reviewers

General response: We thank both reviewers for providing detailed comments. We have addressed all the comments made and revised the manuscript accordingly.

Review 1

General comment: The major issue I see with this manuscript is in the lack of detail in model descriptions, evaluations, and data sources, all of which are lacking throughout the manuscript. I've laid out specific concerns below. Overall, a general lack of detail makes it difficult to trust the results and conclusions about the effectiveness of the various control actions.

Response: We agree with the reviewer that model description, evaluation and data sources are important in a scientific paper.

Exactly for this reason, we evaluated the model extensively in this work. In page 7 of the supplement, we have provided two figures (Figure S2 and S3; note that they are now Figure S3 and S4) to compare the model predicted variables with observed ones (i.e., for the 30% of the dataset that were not used for constructing the model). In page 7 of the supplement, we also provided the correlation coefficients between predicted hourly and observed concentrations for all the parameters. In Figure S3 and Figure 5, we provided the regression equations as well as the correlation coefficients. In page 3, line 109 to 111 of the original main text, we explained that "we firstly construct the RF model from a training data set (e.g., 70% of the all data available) of observed concentrations of a pollutant and its predictor variables and then validate the model by unseen data sets (testing data sets)". Furthermore, in Figure 5 of the original manuscript, we compared the model predicted monthly concentration of $PM_{2.5}$ by the RF model and the WRF-CMAQ model against the observed values. Therefore, the RF model results were evaluated against observations.

We have indeed calculated other parameters for model evaluation, for example RMSE, but we did not report it because the figures and the r2 already showed the good performance of the model. However, we respond in more detail below and have included more parameters in the revised manuscript.

Line 161-168: "Table S2, Figure S3-S4 and Section S3 provided information on the performance of our model using a number of statistical measures including mean square error (MSE)/ root mean square error (RMSE), correlation coefficients (r2), FAC2 (fraction of predictions with a factor of two), MB (mean bias), MGE (mean gross error), NMB (normalised mean bias), NMGE (normalised mean gross error), COE (Coefficient of Efficiency), IOA (Index of Agreement) as suggested in a number of recent papers (Emery et al. 2017, Henneman et al., 2017, and Dennis et al., 2010). These results confirm that the model perform very well in comparison with traditional statistical methods and air quality models (Henneman et al., 2015)".

The reviewer also questioned that there is a lack of detail on the data sources. We have explained in the original text that data were collected from the 12 national air quality monitoring stations in Beijing. In the revised manuscript, we made this clearer: "Hourly air quality data for six key air pollutants ($PM_{2.5}$, PM_{10}, NO_2, SO_2, O_3, and CO) was collected by 12 national air quality monitoring stations in Beijing by the China National Environmental Monitoring Network (CNEM). Hourly air quality data were downloaded from the CNEM website - http://106.37.208.233:20035. Since air quality data are removed from the website on a daily basis, data were automatically downloaded to a local computer and combined to form the whole dataset for this paper." All data are now available at https://github.com/tuanvvu/Air_Quality_Trend_Analysis (last access 5 June 2019).

With regards to the model descriptions, we did not generate this algorithm from scratch. We used the Grange et al. (2018) model as a basis. In the revised manuscript, we emphasized that in this work we modified the Grange et al. (2018) algorithm in order to understand the seasonal variation of air pollutants. We have revised our method section to make it clearer as below:

"A weather normalisation technique predicts the concentration of an air pollutant at a specific measured time point (e.g., 09:00 on 01/01/2015) with randomly selected meteorological conditions. This technique was firstly introduced by Grange et al. (2018). In their method, a new dataset of input predictor features including time variables (day of the year, the day of the week, hour of the day, but not the Unix time variable) and meteorological parameters (wind speed, wind direction, temperature and RH) is firstly generated (i.e., re-sampled) randomly from the original observation dataset. For example, for a particular day (e.g., 01/01/2011), the model randomly selects the time variables (excluding Unix time) and weather parameters at any day from the data set of predictor features during the whole study period. This is repeated 1,000 times to provide the new input data set for a particular day. The input data set is then fed to the random forest model to predict the concentration of a pollutant at a particular day (Grange et al., 2018; Grange and Carslaw, 2019). This gives a total of 1,000 predicted concentrations for that day. The final concentration of that pollutant, referred hereafter as weather normalised concentration, is calculated by averaging the 1000 predicted concentrations. This method normalises the impact of both seasonal and weather variations. Therefore, it is unable to investigate the seasonal variation of trends for a comparison with the trend of primary emissions. For this reason, we enhanced the meteorological normalisation procedure.

In our algorithm, we firstly generated a new input data set of predictor features, which includes original time variables and re-sampled weather data (wind speed, wind direction, temperature, and relative humidity). Specifically, weather variables at a specific selected hour of a particular day in the input data sets were generated by randomly selecting from the observed weather data (i.e., 1988-2017 or 2013-2017) at that particular hour of different dates within a four-week period (i.e., 2 weeks before and 2 weeks after that selected date). For example, the new input weather data at 08:00 15/01/2015 are randomly selected from the observed data at 08:00 am on any date from 1st to 29th January of any year in 1988-2017 or 2013-2017. The selection process was repeated automatically 1,000 times to generate a final input data set. Each of the 1,000 data was then fed to the random forest model to predict the concentration of a pollutant. The 1,000 predicted concentrations were then averaged to calculate the final weather normalised concentration for that particular hour, day, and year. This way, unlike Grange et al., (2018), we only normalise the weather conditions but not the seasonal and diurnal variations. Furthermore, we are able to re-sample observed weather data for a longer period (for example, 1998-2017), rather than only the study period. This new approach enables us investigate the seasonality of weather normalised concentrations and compare them with primary emissions from inventories". (Line 171-204).

We provided the R code in the following website so that an experienced statistician will be able to test the model. https://github.com/tuanvvu/Air_Quality_Trend_Analysis

Specific comments and responses

1. **Comment:** abstract- "improved a novel machine learning-based random forest technique". How?

Response: In our study, we enhanced the weather normalisation technique using the random forest technique algorithm of Grange et al. (2018). We explained this in detail in the revised manuscript. Please see response to general comment above.

We have revised the text in the abstract to "applied machine learning-based random forest technique". (line 30 in the revised manuscript).

2. **Comment:** Line 75- "But they usually gave a poor fitting, suggesting a poor performance of the KZ filter model, or did not allow us to investigate the effect of input variables in neural network models (therefore it is referred as a "black- box" model): A poor fit does not necessarily reflect a poor performance; performance is dictated by the goals of the modeling, whereas fit is a measure of the ability to reproduce training data.

Response: The reviewer argued that "fit is a measure of the ability to reproduce training data". In our case, "fit" is a measure of the ability to reproduce testing data, rather than the training data. The training data are used to train the model. We agree that "performance is dictated by the goals of the modelling" but we do not think a model has a good performance if it failed to predict the testing data (e.g., observations). When modelling a time-series data set of pollutants, the performance of the model is usually evaluated by MSE (or RMSE) and R^2. Other parameters are also used, which are now included in a new table - Table S2 in the supplement to show the performance of our RF model.

We changed the sentence to "Among these models, the deep neural network models showed a better performance (i.e., higher correlation coefficient, lower root mean square error – RMSE) but did not allow us to investigate the effect of input variables". (line 84-87)

3. **Comment:** Line 79: Again, "performance" here is not defined. I recommend

Response: The reviewer wrote "I recommend" but we did not find what exactly the reviewer is recommending.

We explained in the revised manuscript that "performance" represents higher correlation coefficient, and lower root mean square error to make this clearer.

4. **Comment:** Line 79: Should mention the increased propensity of over-fitting with these models for completeness

Response: In this study, the over-fitting is checked by the testing data sets. The further investigation of over-fitting problem from the random forest algorithm is out of the scope of this study. We have discussed the over-fitting of decision tree models in the revised main text (Line 94-97): "Also, the decision trees models are prone to over-fitting, especially when the number of tree nodes is large (Kotsiantis, 2013). An over-fitting problem of a random forest model is checked by its performance using an unseen training data set".

5. **Comment:** Line 110: Recommend showing in Figure 1 that you used 70% of the data for training, 30% for model evaluation. In addition, I recommend reading Oreskes et al. (1994) for distinction between evaluation/validation on environmental datasets. Oreskes, N., Shrader-Frechette, K., & Belitz, K. (1994). Verification, Validation, and Confirmation of Numerical Models in the Earth Sciences. Science, 263(5147), 641–646.

Response: We followed the comment and added the information in the Figure 1. We also change the term “validation” into “evaluation”. Thanks for the recommended article. Oreskes et al. (1994) discussed the concept of model evaluation and validation in the Earth Sciences. In our specific case (regression modelling of a time series data sets), the valuation/evaluation of model are on cross-validation based on the out-of-bag technique and evaluation of the predicted concentration using a testing data set. Specifically, in the random forest algorithm that we applied, the algorithm used the out-of-bag technique: each decision tree is trained using a bootstrapped subset of observations. This means that for every tree there is a separate subset of observations (called OOB observations) not being used to train that tree. The model uses OOB observations as a test set to cross-validate the performance of the random forest. This is why we used the testing data set to evaluate the predicted values from models.

6. **Comment:** Line 95: “press.” has a period, whereas the other abbreviations do not.

Response: It is changed to pressure. We also removed abbreviations for other parameters.

7. **Comment:** Line 104: it => its

Response: We corrected it.

8. **Comment:** With a holdout analysis, there are many comparisons to be made beyond R^2 that tell us more about model fit. Many of the studies cited in the introduction include detailed evaluations, including with slope, intercept, and root mean square error. These should be included at the very least. There may be still other metrics that are informative for the evaluation in this particular application.

Response: Figure 5 and Figure S3 in the original supplement (now becoming Figure S4) have already showed information on some of the information suggested. In the revised manuscript, we provided more parameters, including the RMSE and other parameters recommended in the papers suggested by the reviewer (comment 17) in the supplement in Table S2.

9. **Comment:** sample => samples

Response: We corrected it.

10. **Comment:** Line 140-150: Was this a separate random forest model from the initial model described in the “Random Forest (RF) model development” section?

Response: No. In the revised manuscript, we re-wrote the section to make this clearer. In our study, we applied the RF which was already built using R codes from Grange et al. (2018). Their codes were originally based on the R package “ranger” by Wright et al. (2018) (https://github.com/imbs-hl/ranger)” Please see response to general comment above.

11. **Comment:** Line 152: This statement (“only either data (MET data) sets were re-sampled”) directly contradicts the statement in the paragraph above.

Response: This appears to be a misunderstanding. We have re-written the whole section to make this clear. Please see response to general comment above.

12. **Comment:** Lines 162-8: Please state what you are regressing using the Theil-Sen estimator

Response: It is the concentration of a pollutant after weather normalisation. The Theil-Sen estimator is usually used for long-term trend analysis of a pollutant. We used this estimator to find the slope of the concentration trend of a pollutant. We modified the text to make it clear. (Line 207-208)**: "**The Theil-Sen regression technique was performed on the concentration of air pollutants after meteorological normalisation to investigate the long-term trend of pollutants".

13. **Comment:** Lines 207-210: The conclusion that this evidence indicates a robust model requires more exploration. What about the meteorology from 1998-2013 would result in the 2μg m 3 increase in detrended PM2.5 in 2017?

Response: We are unable to understand the question. We did not mention in any part of our model "2μg m 3". Thus, we cannot directly respond to this comment. We compared the model predicted concentrations against the observations (test dataset) in Figure S3 and S4, which showed the performance/bias of the model. Matrices for model performance are also shown in Table S2. We've revised the section to avoid confusion (Line 279-282):

"When meteorological conditions were randomly selected from 2013-2017 (instead of 1998-2017) in the RF model, the normalised level of $PM_{2.5}$ in 2017 was 60 μg m^{-3}, which is 1 μg m^{-3} difference to that using 1998-2017 data. This difference is due to the variation of the long-term climatology (1998-2017) to the 5 year period (2013-2017)"

14. **Comment:** Line ~220: This could also indicate that formation/deposition/reaction of PM10 and NO2 are affected differently than the other pollutants. From the evidence provided, it is difficult to fully embrace the claim that PM10 and NO2 were affected by sources that were not controlled. Figure 2 presents no evidence relating to dust events that I can see.

Response: We agree and revised this to:
"The Action Plan also led to a decrease in PM_{10} and NO_2 but to a lesser extent than that of CO, SO_2 and $PM_{2.5}$, indicating that PM_{10} and NO_2 were affected by other less well controlled sources or different atmospheric processes". (Line 292-294).

15. **Comment:** Line 223: Figure 3 does present differences between urban/rural/suburban, but there is no information on how many sites and their location. I recommend including a map so that distance to roadways/industries/spatial representativeness can be determined

Response: Site information is given in Shi et al. (2019). However, to make this clearer, we've added a figure and a Table S1 in the supplementary to show in detail the different type of sites (Figure S1).

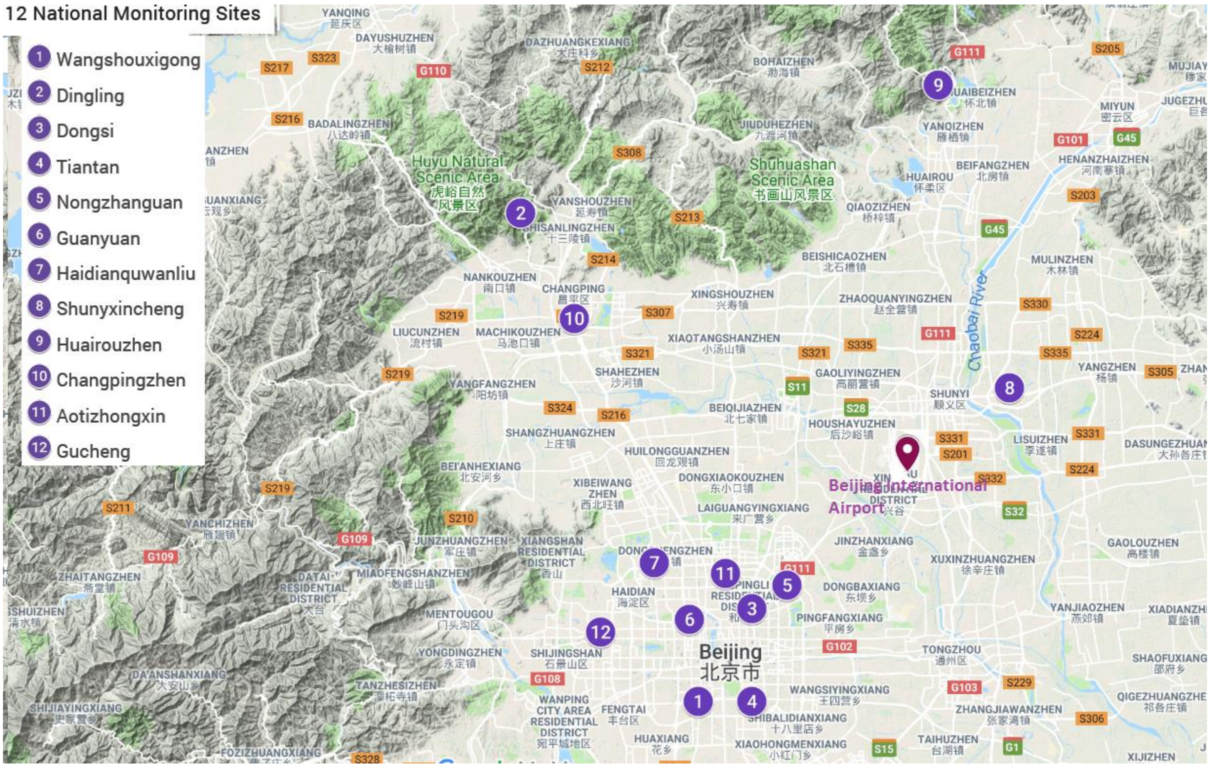


Figure S1. Map of 12 monitoring stations in Beijing.

We were not sure why the reviewer mentioned industrial sites. There is no industrial site in Beijing so we were unable to include this in the figure.

16. Comment: Line 230: This evaluation is difficult to interpret. Are the average WRF-CMAQ values calculated in the same grid cells as the monitors? Presumably, CMAQ modeling used emissions for year 2017 (state this explicitly if so), what about years 2013 and 2016 make them reasonable comparison years for detrended PM2.5?

Response: WRF-CMAQ modelling has been described in Cheng et al. (2018). The average WRF-CMAQ values were calculated for the whole of Beijing. Yes, the CMAQ modelling used the emissions for year 2017. This is now clarified in the text (Line 119-120): "Monthly emission inventories of air pollutants were from Multi-resolution Emission Inventory for China **(**http://www.meicmodel.org/), and for the whole Beijing region".

The 2013 year was chosen because it is the start-year of the Action Plan. 2016 was chosen to see the immediate effect of the 2017 measures in comparison the year before. More detailed explanation is given in Cheng et al. (2018).

17. Comment: Line 241-247: For model evaluation, I recommend including the recommended statistics from extensive publication on appropriate evaluation approaches like in Emery et al. 2017, Henneman et al., 2017, and Dennis et al., 2010. Emery, C., Liu, Z., Russell, A., Talat Odman, M., Yarwood, G., & Kumar, N. (2016). Recommendations on Statistics and Benchmarks to Assess Photochemical Model Performance. Journal of the Air & Waste Management Association. Dennis, R., T. Fox, M. Fuentes, A. Gilliland, S. Hanna, C. Hogrefe, J. Irwin, S.T. Rao, R, Scheffe, K. Schere, D.A. Steyn, and A. Venkatram. 2010. A framework for evaluating regio- nal-scale numerical photochemical modeling systems. J. Environ. Fluid Mech.10:471–89. doi: 10.1007/s10652-009- 9163-2. Henneman, L. R., Liu, C., Hu, Y., Mulholland, J. A., & Russell, A. G. (2017). Air quality modeling for accountability research: Operational, dynamic, and diagnostic evaluation. Atmospheric Environment, 166(2017), 551–565.

Response: Thanks for these recommended articles. We provided an additional table (Table S2) to include the parameters recommended in these publications.

18. Comment: Line 259: Please define the term "based line"

Response: The "baseline" of a pollutant (except for ozone) was the defined as the lowest concentration of air pollutants in the summer (the summer concentrations) – please see line 334-336: "On the other hand, the "baseline" SO_2 concentration – minimum monthly average concentration in the summer (Figure 2) – also reduced somewhat during the same period."

19. Comment: Line 280: This contradicts the statement above that buffered changes in NO2 are due exclusively to sources that were not controlled

Response: The sentence was changed to: "The different trends between SO_2 and NO_2 indicate that other sources (e.g. traffic emissions, Figure S9) or atmospheric processes have a greater influence on ambient concentration of NO_2 than coal combustion. For examples the chemistry of the $NO/NO_2/O_3$ system will tend to "buffer" changes in NO_2 causing non-linearity in NO_x-NO_2 relationships." (Line 356-360).

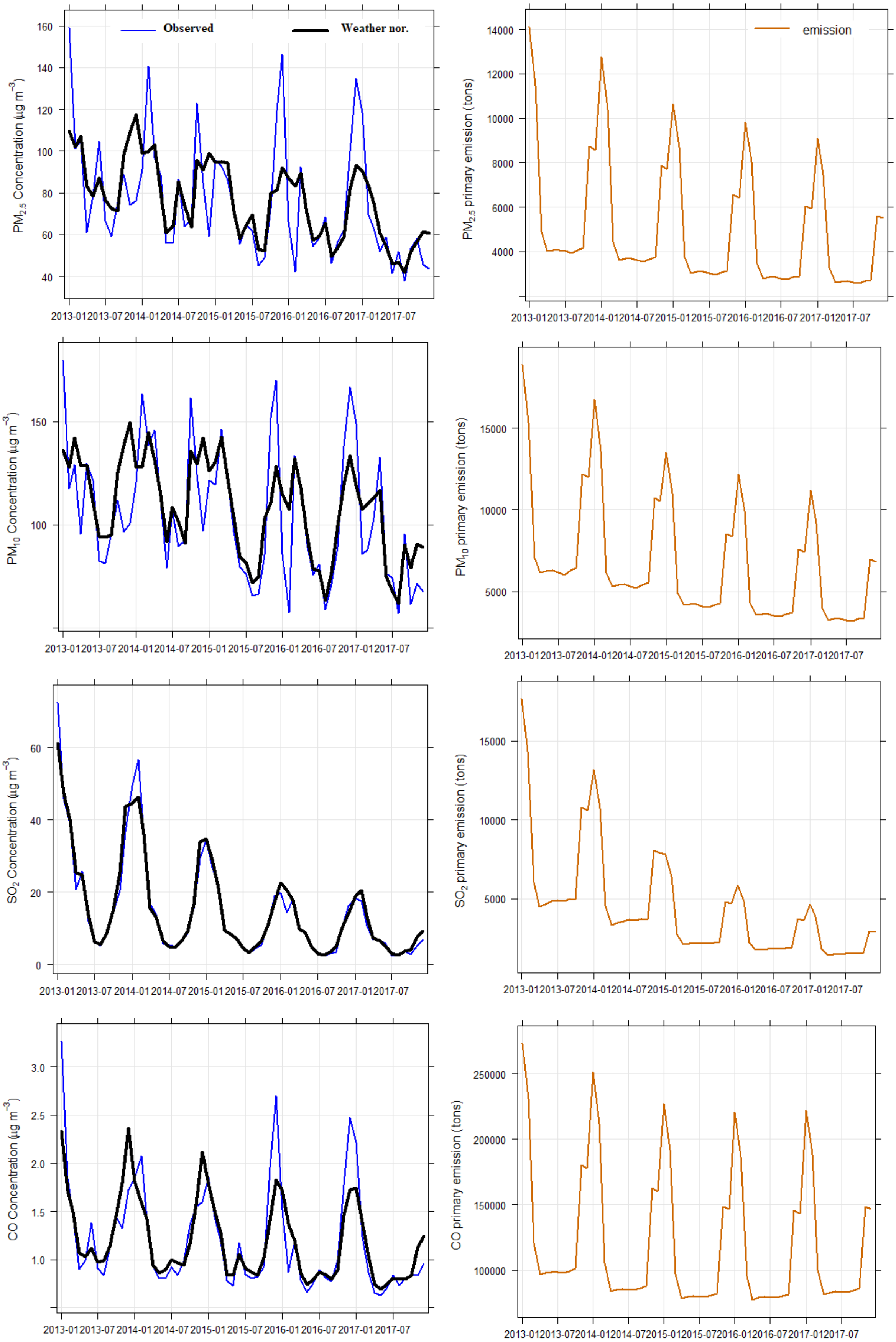
20. Comment: Line 330: Please elaborate on which data would improve this study.

Response: We refer to detailed information on the implemented policies such as the start/end date of air pollution control actions. It is now included in the main text. (Line 413-415).

21. Comment: Figure 2: I recommend including separate plots for emissions and concentrations. Plots with two vertical axes can lead to information manipulation (it is not clear, for instance, why an SO2 concentration of 40ppb corresponds to an emissions level of 2 kilotons). It would be useful to include correlations between detrended emissions and concentrations. Further, I recommend extending all vertical axes to values of 0.

Response:

We plotted the figures (see below) as suggested. We can easily replace the figure with the following ones. However, we felt that it is harder to compare the observed concentration, weather normalised concentration and primary emission in these new figures. Therefore, we suggest that it would be better to plot the primary emissions and concentrations in a single figure for a comparison.


Observed
Weather nor.
$PM_{2.5}$ Concentration ($\mu g\ m^{-3}$)
160
140
120
100
80
60
40
emission
$PM_{2.5}$ primary emission (tons)
14000
12000
10000
8000
6000
4000
PM_{10} Concentration ($\mu g\ m^{-3}$)
150
100
PM_{10} primary emission (tons)
15000
10000
5000
SO_2 Concentration ($\mu g\ m^{-3}$)
60
40
20
0
SO_2 primary emission (tons)
15000
10000
5000
CO Concentration ($\mu g\ m^{-3}$)
3.0
2.5
2.0
1.5
1.0
CO primary emission (tons)
250000
200000
150000
100000
2013-01
2013-07
2014-01
2014-07
2015-01
2015-07
2016-01
2016-07
2017-01
2017-07

The reviewer asked us to include correlations between detrended emissions and concentrations. We emphasise here that emissions cannot be detrended. They are based on bottom-up estimates which have nothing to do with meteorology. We tried to extend all vertical axes to 0, but they make the figure less readable (e.g., the temporal trends are hard to see).

22. Comment: Figures S4 and S5 require more description. What are Variable Importance and Variable Interactions?

Response: This has been added to the description in Figure captions.

23. Comment: Where is the emissions data from? What locations?

Response: We have added to the revised text: "Monthly emission inventories of air pollutants were from Multi-resolution Emission Inventory for China **(**http://www.meicmodel.org/), and it is for the whole Beijing region" (Line 119-120). The MEIC emission inventory is internationally recognized as the leading inventory for China.

24. Comment: I recommend moving much of the information on the regulations from the supplement to the main text body. I recommend using consistent language to refer to the weather normalised concentrations. At points in the manuscript, figures, and tables, these values are referred to as detrended, "Nor."

Response: We moved the key information on regulations into the main text. We use the term "weather normalised concentration" and change the "Nor." and "detrend" in Table 1 and Figure 2 to "model".

Review 2:

1. **Comment:** The authors note the use of met data from Beijing Airport. How representative is this data of all sites studied? I'm a little concerned this forms an important factor in determining the general applicability of the model. As the paper by Grange and Carslaw 2019 shows, the selection of wind directions, for example, can have significant impact on model fidelity if a site is affected by specific geography.

Response:

Airport met data are most representative of regional scale meteorology of the whole city. Because the meteorological measurements at each site are seriously affected by very local influences, it is not meaningful to compare the meteorology with that at the airport. Air pollution in the Beijing area is a regional phenomenon (Shi et al. 2019). We found very high correlations between air pollutant concentrations measured from 12 monitoring sites (Shi et al. 2019).

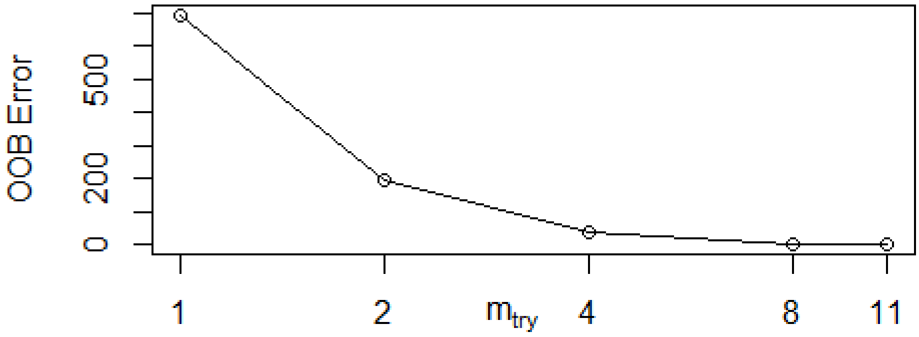
In Grange & Carslaw's paper, they also used the surface met data from the airport using the "worldmet" package. Regarding the selection of wind directions, Grange & Carslaw (2018) also noted that "Interestingly, wind direction was often a relatively unimportant variable (Fig. 4). This may be due to daily wind direction averages not contributing much information gain in the model because the aggregation period results in the metric representing atmospheric motion rather poorly".

2. **Comment:** Rather than referring to variables 'such as', please be specific in all cases.

Response: It is corrected!

3. **Comment:** You state that the 'regression model is an ensemble-model which consists of hundreds of individual decision tree models'. Please clearly state the number and how hyperparameters were derived.

Response: It is given in the SI (Section 3, Figure S1). The number of trees is 200, the minimum size of terminal nodes (Nodesize) is 3 and the variables randomly sampled for splitting (Mtry) the decision tree is 4. Mtry can be estimated based on the OOB error (as in the figure below). The number of trees and modesize was determined by RMSE and R^2. It is found with the tree numbers larger than 150 and the nodesize of 3, the RMSE is minimum and stable. A larger number of trees and nodesizes lead to little improvement in R value and RMSE, but it significantly increases the computation time. Another way we optimize the Mtry and nodesize is by a trial and error method, in which we vary the Mtry from 3 to 10 and number of trees from 20 to 500 to find the dependence of the error on the values of Mtry or number of trees.



4. **Comment:** You state you used ’e.g. 70% of the all data [correct - of all the data]’. Is this an example or is this the actual training portion you used? I think this is clarified later on but please refrain from vague statements in describing any model development workflow.

Response: It is the actual training portion we used. It is now updated in the text.

5. **Comment:** It is customary to combine a single random sampling strategy with K-folds [e.g. 5] validation. Has this been used? If not, why?

Response: No, in our study, we used out-of-bag (OOB) score estimation instead of the K-folds for model cross-validation. In the random forest algorithm which we used: each decision tree is trained using a bootstrapped subset of observations. This means that for every tree there is a separate subset of observations (called OOB observations) not being used to train that tree. The model can use OOB observations as a test set to cross-validate the performance of the random forest. The learning algorithm compares the observation's true value with the prediction from a subset of trees not trained using that observation, and calculates the overall score as a single measure of a random forest's performance.

6. **Comment:** If random sampling, how do you know if using different initial seeds in any random number generator leads to better or worse results? I can’t see any code sharing so can’t check this - please see a further comment on this.

Response: We have already considered this and used the function set.seed before running the RandomForest function to test the reproducibility. The result is almost the same. The code is available on: https://github.com/tuanvvu/Air_Quality_Trend_Analysis/blob/master/R/Air_Quality_Weather_Normalised_Trend.R

7. **Comment:** The authors talk about an ’enhanced’ normalisation procedure. Please explain more clearly how this is different from the original paper by Grange et al 2018. I will admit, that paper isnt as clear as it could be, but they do provide the model base. As far as I can tell, both studies only re-sample weather data.

Response: The concept of weather normalisation is similar and was introduced by Grange et al. (2018). Both studies re-sample the weather data, but we did it in a different way.
In Grange et al. (2018), both the weather and time predictor features (except the Unix date) were randomly generated from the original data set of predictor features as the following code:

```
"# Randomly sample observations
n_rows <- nrow(df) #df is original data set
index_rows <- sample(1:n_rows, replace = replace)
# Transform data frame to include sampled variables
df[variables] <- lapply(df[variables], function(x) x[index_rows])”
```

It means the seasonal, weekend/week, hour and weather data are also re-sampled.
In our study, only weather data were re-sampled. The advantage is that we can now see the seasonal effects. We revised the text to:

“In our algorithm, we firstly generated a new input data set of predictor features, which includes original time variables and re-sampled weather data (wind speed, wind direction, temperature, and relative humidity). Specifically, weather variables at a specific selected hour of a particular day in the input data sets were generated by randomly selecting from the observed weather data (i.e., 1988-2017 or 2013-2017) at that particular hour of different dates within a four-week period (i.e., 2 weeks before and 2 weeks after that selected date). For example, the new input weather data at 08:00 15/01/2015 are randomly selected from the observed data at 08:00 am on any date from 1st to 29th January of any year in 1988-2017.” (Line189-196).

8. **Comment:** Also there is no discussion of classification into back trajectories, for example, or estimated boundary layer heights etc. If these products are not used, how is this study an enhancement?

Response: Thank you for the suggestions. We did add the back trajectories into the model, but it did not improve the model's performance. Therefore, we have not included this in the model. We now added a sentence in the Supplement to make this point clearer (Line 107-108, SI).

We used the hourly data sets as input variables in our study. Estimated hourly boundary layer heights from models, e.g., WRF-Chem are highly uncertain. Using such uncertain data will cause unpredictable uncertainty in our results. Our RF model performed very well already, with existing input variables.

9. **Comment:** In some ways I struggle to see how section 2 'weather normalisation' is significantly different from the Grange et al approach. If they are different, they need clearly stating why - perhaps even with a visual workflow/table for each - and a comparison on findal data products. The title of the paper leads me to believe this is a new technique.

Response: Please find our response to comment 7. We clarified that we did not create a new technique. We applied the random forest model and only enhanced the "weather normalisation technique". However, the key point of this work is that we can now look at applications of the method to evaluate the air quality trends in Beijing, including seasonal variations.

10. **Comment:** line 104 - concentrations of an air pollutant and it[s] predictor variables - please correct

Response: It is corrected.

11. **Comment:** line 116: 'These time variables' - do you mean parameters that vary with time or the time variable?

Response: We mean the time variables (features): date of year, hour of the year and week/weekend. This is now modified.

12. **Comment:** line 119 [equation with no label] - what is the significance of year 'i'? Is this defined on, say, the Unix epoch?

Response: Yes, it is. It is corrected to ith year (i from 2013 to 2017).

13. **Comment:** line 134: 'To validate the model for unseen data sets, a test data set which represents 30% of entire data sets[set] is input into the random forest model which has been constructed from training data sets.' This is a confusing statement. The test and training sets refer to both features and predicted variable. Thus, only features are 'input into the model'? Please re-phrase this. In fact, I would suggest you consider using the term 'features' when referring to variables to which you are fitting the model.

Response: It is re-phrased in the model evaluation line 145-147: "As shown in Figure 1, the whole data sets were randomly divided into: 1) a training data set to construct the random forest model and 2) a testing data set to test the model performance for unseen data sets. The training data set comprised of 70% of the whole data, with the rest as testing data". We changed the "variable" to "predictor features" as suggested.

14. **Comment:** line 140: 'A weather normalisation technique predicts the concentration of an air pollutant at a specific measured time point but with various meteorological conditions

(termed as “weather normalised concentration”).’ Do you mean to state that this technique predicts the concentrations of an air pollutant as a function of meteorological factors alone?

Response: It is not so, because it is also a function of the time variables. If a new weather condition is inputted to the model, it can predict the concentration of a pollutant in a certain time period.

15. Comment: line 142: ’Both time variable (month, hour) and meteorological parameters, except the trend variable were re-sampled randomly and was added into the random forest model as input variables to predict the concentration of a pollutant’. This is a confusing statement when referred to ’adding’. What do you mean by adding? On top of preexisting variables?

Response: “add” here means input. This is now updated: “A weather normalisation technique predicts the concentration of an air pollutant at a specific measured time point (e.g., 09:00 on 01/01/2015) with randomly selected meteorological conditions. This technique was firstly introduced by Grange et al. (2018). In their method, a new dataset of input predictor features including time variables (day of the year, the day of the week, hour of the day, but not the Unix time variable) and meteorological parameters (wind speed, wind direction, temperature and RH) is firstly generated (i.e., re-sampled) randomly from the original observation dataset. For example, for a particular day (e.g., 01/01/2011), the model randomly selects the time variables (excluding Unix time) and weather parameters at any day from the data set of predictor features during the whole study period. This is repeated 1,000 times to provide the new input data set for a particular day. The input data set is then fed to the random forest model to predict the concentration of a pollutant at a particular day (Grange et al., 2018; Grange and Carslaw, 2019). This gives a total of 1,000 predicted concentrations for that day. The final concentration of that pollutant, referred hereafter as weather normalised concentration, is calculated by averaging the 1000 predicted concentrations.”. (Line 171-184).

16. Comment: Section 3.4 Please explain why, in a few cases, normalised values are higher than original.

Response: As we discussed in Figure 4, if the weather during that month is more favourable for the dispersion of air pollutants, the normalised values will be higher than the observed concentration.

17. Comment: Section 3.5 ’Our results confirmed that the “Action Plan” has been highly effective’. Please define ’highly effective’.

Response: We’ve updated this to “’Our results confirmed that the “Action Plan” has led to major improvement in air quality.”

18. Comment: Code/data availability: The current paper has no statement on this. The authors need to meet the current data and code sharing standards provided by Copernicus: https://www.atmospheric-chemistry-and-physics.net/about/data_policy.html https://peerj.com/articles/cs-86/ Indeed, there are currently many uncertain aspects of this study which could be resolved by clear code sharing and documentation.

Response: They are now available at: https://github.com/tuanvvu/Air_Quality_Trend_Analysis

19. Comment: There are a number of grammatical issues throughout the paper:

Response: A senior co-author has re-checked the grammar throughout the manuscript.

Reference:

In-Depth Study of Air Pollution Sources and processes within Beijing and its Surrounding Region (APHH-Beijing), Shi, Z., Vu, T., Kotthaus, S., Harrison, R.M., Grimmond, S., Yue, S., Zhu, T., Lee, J., Han, Y., Demuzere, M., Dunmore, R.E., Ren, L., Liu, D., Wang, Y., Wild, O., Allan, J., Acton, W.J., Barlow, J., Barratt, B., Beddows, D., Bloss, W.J., Calzolai, G., Carruthers, D., Carslaw, D.C., Chan, Q., Chatzidiakou, L., Chen, Y., Crilley, L., Coe, H., Dai, T., Doherty, R., Duan, F., Fu, P., Ge, B., Ge, M., Guan, D., Hamilton, J.F., He, K., Heal, M., Heard, D., Hewitt, C.N., Hollaway, M., Hu, M., Ji, X. Jiang, R. Jones, M. Kalberer, F.J. Kelly, L. Kramer, B. Langford, C. Lin, A.C. Lewis, J. Li, W. Li, D., Liu, H., Liu, J., Loh, M., Lu, K., Lucarelli, F., Mann, G., McFiggans, G., Miller, M.R., Mills, G., Monk, P., Nemitz, E., O’Connor, F., Ouyang, B., Palmer, P.I., Percival, C., Popoola, O., Reeves, C., Rickard, A.R., Shao, L., Shi, G., Spracklen, D., Stevenson, D., Sun, Y., Sun, Z., Tao, S., Tong, S., Wang, Q., Wang, W., Wang, X., Wang, X., Wang, Z., Wei, L., Whalley, L., Wu, X., Wu, Z., Xie, P., Yang, F., Zhang, Q., Zhang, Y., Zhang, Y. and Zheng, M., Atmos. Chem. Phys., 19, 7519–7546, 2019