Response to Referee #1

We greatly appreciate the reviewer for providing constructive comments, which have helped us improve the paper quality. We have addressed all of the comments carefully, as detailed below. The original comments are in black and our responses are in blue.

Overall comments

This manuscript presents a trend analysis of various air pollutants in 10 Canadian cities over a (at maximum) 30-year period. The trend analysis has been conducted in a way that controls for weather over the analysis period and a second component of the data analysis involves exploring outliers, that are wildfire events, and their influence on the trends observed. The results of the analysis are what is expected and are in line with observations gathered from other urban areas with developed economies. Namely, NOx (NO2 is focused on here), CO, and SO2 mole fractions have declined over the analysis period with CO and SO2 becoming less of an "optional issue" in modern times. The reduction of NOx (specifically NO) has however produced the urban O3 rebound where the reduction of NO has resulted in increasing or stable trends of O3. PM2.5 trends are more mixed because of location-specific features of primary emissions and secondary generation processes at local and regional scales. In this sense, the manuscript does not contribute too much new to the literature with respect to processes or mechanisms, but the results are likely important to the Canadian cities in question. I defer to the editor to make a judgement on this novelty point. The manuscript has been well-written and constructed.

Response: We agree with the reviewer that processes and mechanisms used for interpreting the generated trends are mostly available from literature. We would like to point out that the innovation of the study does not limit to only exploring new processes and mechanisms. The new investigation methods of the trends and associated drivers also contribute to innovation. For example, this study for the first time analyzed O_3 trends separated at a level above and below 40 ppb, a threshold value to assess the impact of O_3 on ecosystems and an approximate value of O_3 derived from the stratosphere invasion. This approach better characterized the O_3 trend at unhealthy levels and accurately evaluated the generation of O_3 in the troposphere. This study also applied our previously developed identical-percentile autocorrelation analysis method to quantify the perturbations from extreme events such as large-scale wildfires on large percentile $PM_{2.5}$ concentrations and confirmed that unpredictable large-scale wildfires were overwhelming or balancing the impacts of emission reductions on $PM_{2.5}$ in western Canada. The new findings from this study would better service the future air quality protection in Canada and the whole North American where suffered from large-scale wildfires.

My general feedback is as follows. The authors have used two closely related methods to conduct their meteorological normalisation, and these two methods more or less produce the same result. I think it would be useful to add an explicit method comparison objective to the study and add a paragraph to the results or discussion section addressing the similarities and differences between the methods. The manuscript only considers mean slopes (as determined by Mann-Kendall tests) for the trend analysis when exploring the change over time for pollutants and locations. I would like to see these trend slopes plotted alongside the non-normalised and normalised concentrations, at least for one example. It should also be acknowledged and discussed further that collapsing a time series into a single mean slope value misses other changes over time which may also be important when considering the introduction of air quality management policies, especially immediately after a policy change. The manuscript lacks an in-depth site or location-specific interpretation of the results because the focus is placed on stating the trends observed. It seems that many of the anomalies, for example, Hamilton's SO2 (a port city), could be further explained by city-specific interpretation. I am not familiar with these cities, but the manuscript would be far stronger if more site and city-specific information were evaluated and added to the discussion. The authors have used an approach called the identical-percentile autocorrelation method to both determine outlier events (that are driven by wildfires) and their influence on the trend. I believe the text in the methods needs to be addressed further because I cannot follow clearly how and why this approach is conducted.

Response: The similarities and differences between the two machine learning techniques have been well documented by Grange et al. (ACP, 18, 6223–6239, 2018), which has been cited in our study. In the revised manuscript, we have provided a brief summary of the key points regarding the similarities and differences between the two methods in Section 2, which reads: “The advantages and limitations of RF algorithm and BRTs have been described in detail in earlier studies (Grange et al., 2018; Grange and Carslaw, 2019). Briefly, BRTs method is fast to train and make prediction, but suffered heavily from overfitting, which may result in unreliable predictions. RF method can control the overfitting, but yields a poor prediction for outliers in large percentiles. Thus, using two methods with different strengths and weaknesses, although their predictions are similar in many ways, can constrain methodology uncertainties and better evaluate perturbations due to varying weather conditions than using only one method, as has been demonstrated in our earlier study (Lin et al., 2022).”

In the revised manuscript, we have added the trend slopes plotted alongside the non-normalised and normalised concentrations for NO_2 as an example in the Supporting Information (Fig S3c).

We have added city-level primary emissions in Hamilton to support our analysis.

We agree with the reviewer that collapsing a time series into a single mean slope value misses other changes over time, some of which may also be important. The identical-percentile autocorrelation method that we developed earlier and adopted in this study, is for the purpose of handling this issue. The method can be used to clearly identify whether the large percentile values follow the general trend or not. If the large percentile values always follow the general autocorrelation trend, the mean slope can well reflect the long-term trend. If the large percentile values deviate largely from the general autocorrelation trend, the mean slope is not insufficient to reflect the long-term trend. In the latter case, the trend of large percentile values deviated from the general autocorrelation trend should be analyzed further along with that extracted from the mean slope.

We have checked multiple times the Method section describing the identical-percentile autocorrelation method and feel that sufficient details have been presented. We believe that readers can easily pick up the approach and understand its performance through a simple test using their own data. We agree that it is not easy to fully capture the method by a quick reading without any test due to the many steps involved in this method. Doing a test is the only way for a deep understanding of the method, especially from steps 3 to 5.

Specific comments

Line 13. Why has (NO2 + O3) been used over Ox in the manuscript? Is this not a standard abbreviation used in the atmospheric sciences?

Response: We have replaced NO_2 + O_3 with O_x throughout the whole manuscript.

Line 16. Replace including with the: "...methods, the random forest algorithm..."

Response: Corrected as recommended in the revised manuscript.

Line 61. Use the superscript notation for 60 ug m-3

Response: Corrected as recommended.

Line 63. Stress that this is an issue for all areas of the world.

Response: We have rewritten this sentence to make this clearer, which reads: “An urgent issue for all areas of the world is to overcome challenges to further lower ambient NO_2, O_3 and $PM_{2.5}$ concentrations in order to meet the WHO 2021 AQG.”

Line 82. Elemental carbon rather than element carbon.
Line 90. Ozone to O3.
Line 97. Remove "The" at the start of the sentence.
Line 100. Both Grange et al., 2018 and Grange and Carslaw, 2019 are usually cited together here.
Line 109. Ox?

Response: All of the above technical corrections have been addressed in the revised manuscript.

Line 120. Should a method comparison objective be added regarding the two decision tree methods that are implemented?

Response: We have added such a description as mentioned in our response to general comments above.

Line 145. This logic might be questionable. If observations are not accessible for a site, using the nearest site might not be a good substitute because there could be a change of site type, generally a shift from urban background to urban traffic or vice-versa. If this were to happen, the time series no longer represents the same monitoring conditions. In a related question, why was the analysis not conducted at a site level? Were the time series generally not continuous among the cities for the analysis period?

Response: The analysis was conducted at a site-level, but we assume that the site can represent a typical urban environment (urban background) of the specific city in which the site is located. In other words, it is the city-level pollution that should be focused. For a decade long observation, a short-term shut-down of some sites is common. In each city, we normally selected an urban background site with no data loss for over one year. In Quebec, Halifax and Calgary, no sites can meet the criteria. In this case, the observations at two neighboring sites within 1 km for monitoring urban background air quality were used. This has been clarified in the revision.

Line 173. Was dew point an important variable for training and prediction? I would expect not if relative humidity and temperature were included. Was an importance analysis conducted?

Response: The reviewer is right; dew point is not an important variable. The software automatically generates the importance ranking of input variables, but the Julian day generally ranks the top three, as reported by Grange et al. (ACP, 18, 6223 –6239, 2018). We have no idea on the role of Julian day and did not include the importance results in this study.

Line 184. The ERA5 reanalysis global model product provides these additional variables and could be included in future analyses.

Response: We revised the sentence as: “These additional meteorological parameters were not included in the present study and could be included in the future analyses.”

Line 201. These figures show that the two methods produce more or less the same result and relates to by general comment above. The scatterplots show different observations due to the different sets for training and testing sets. This is probably worth a note in the caption.

Response: Figure caption has been revised as: “Fig. 1. Performance evaluation of the predicted NO_2 hourly mixing ratios by BRTs and RF algorithm against those observed in Halifax during 1996-2017. Red lines represent linear regression, and color bar reflects data number density. Note that different observational data sets are shown between (a) and (b) because the inputs for the two packages (BRTs and RF) are randomly divided into two groups for training and testing.”

Line 226. I think this section needs to be revised for simplicity. I do not understand how this approach isolates and quantifies the effect of the extreme events. Could a simple outlier test suffice?

Response: We do not think that a simple outlier test works. We believe that readers can easily pick up the approach and understand its performance through a simple test using their own data, according to the details presented in the manuscript. Please also see our response above related to this point.

Line 282. Was block bootstrapping used or the Mann-Kendal trend tests?

Response: The Mann-Kendal trend tests were used as clarified in the revision.

Line 294. Please consider line plots for this type of plot. It is understandable however if points work better.

Response: Line plots were even worse according to our test. We tried small sized markers in the revision.

Line 311. How was 5% determined? To get a robust uncertainty measurement, a number of tests would need to be run and compared to a ground truth?

Response: We have revised the sentence for clarify, which reads: “indicating that the uncertainties in the slope associated with the RF-deweathered averages can be as large as 5% (8% minus 3%) because of its poor prediction for large outlier values.”

Line 315. Please consider plotting the values of the trends together too. This would give a good graphical comparison among all the different time series.

Response: An example has been added in the revised SI (Fig S3c).

Line 357. Can you conclude CO and SO2 are no longer an issue across most of Canada's urban areas?

Response: Yes, this has been clarified in the revision.

Line 402. Do you have an explanation of why the two closely related algorithms had a larger difference in this particular case?

Response: The large difference is associated with high concentrations of large percentiles. In the end of this paragraph, we have added: "The increased uncertainties led to the difference between the RF-deweathered and original SO_2 mixing ratios being up to 16% in Winnipeg, based on the slope of 1.16 listed in Table S3. The difference between the BRTs-deweathered and original SO_2 mixing ratios was, however, only 4%, suggesting that the perturbation due to varying weather conditions might be within 4-16%. Again, the RF algorithm suffers from the weakness in predicting large outlier values."

Line 416. I am not familiar with Hamilton, Ontario, but a quick look shows this city is a port city. This feature would probably explain the observed anomalous SO2 behaviour when considering other cities.

Response: City-specified air pollutant emissions have been added to interpret the trend, which reads: "Such a large discrepancy indicates that the reduction in SO_2 emission in Hamilton likely substantially lagged behind the average provincial level. This is indeed the case since SO_2 emissions from registered facilities in Hamilton (Table S2) fluctuated around $8.67\pm1.75*10^3$ tons year^{-1} during 2002-2009 and then increased to $1.14\pm0.13*10^4$ tons year^{-1} during 2010-2018."

Line 422 and 452. Ox?

Response: Corrected.

Line 515. The same question as line 402, is there an explanation for this behaviour between the decision tree algorithms?

Response: The cause has been explained in our response above. We have revised the sentence to: "However, the RF method seemingly failed to learn the wildfire signals and missed predicting the spikes associated with largely increased natural emissions because of its inherent weakness."

Line 530. Add "a": "Note that a large....". It would be useful to state that this indicates a high variability.

Response: revised as recommended.

Line 552. How much higher?

Response: ≤0.3%. This has been added in the revision.

Line 555. It sounds like high AQHI values are found in all seasons. Perhaps stating the frequency per season would be a clear way to present these differences.

Response: There is no clear trend in other seasons, except a bit more frequent in summer. It might be unnecessary to highlight the seasonal difference in the other three seasons.

Line 569. Replace "were" with "are".

Response: Corrected.

Line 609. I like this section and is an important component of the study where this source of air pollutants will be more of an issue moving into the future. I would like some clarity on the method however so it can be better understood how these conclusions were made.

Response: Please see our Response to the general comments related to the identical autocorrelation method.

Line 643. Add this statement to the conclusions too.

Response: Revised as recommended