March 27, 2024

Dear Prof. Rob MacKenzie - Editor of *Atmospheric Physic and Chemistry*,

We are submitting the revised manuscript (egusphere-2023-2968) titled "Identifying decadal trends in deweathered concentrations of criteria air pollutants in Canadian urban atmospheres with machine learning approaches" to *Atmospheric Physic and Chemistry* for possible publication.

We have carefully considered the comments provided by the the reviewers and revised the manuscript accordingly, as explained in the attached *Response to Reviewers* file. Detailed changes of the manuscript can also be found in the track-change version of the revised manuscript.

We hope you and the reviewers will find the paper meets the standard of this reputable journal

Sincerely,

Dr. Leiming Zhang, Senior Research Scientist
Air Quality Research Division, Science and Technology Branch, Environment and Climate Change Canada,
4905 Dufferin Street, Toronto, Ontario, M3H 5T4
Telephone: 647-956-8302
e-mail: leiming.zhang@ec.gc.ca

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: $\leq$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

Response to Referee #2

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 analyzes the long-term trends of air pollutant concentrations of NO2, SO2, CO, O3, (NO2+O3) and PM2.5 in 10 Canadian cities using observational concentrations and deweathered concentrations. The latter were generated with two machine learning methods. Correlation analysis was carried out to evaluate the association between concentrations and provincial level air emissions. Further, the impact of wildfire on PM2.5 air quality trend was investigated apparently by assuming all high concentrations were due to wildfires. The three datasets, the observed concentrations, and the deweathered concentrations by each of the two methods, yield similar trends in general. Therefore, the variation in weather conditions has a very small impact on the trends of annual mean concentrations, as expected. The long-term trends are consistent with results reported by other researchers with a few exceptions. The topic is relevant to the journal of ACP. The manuscript has a great potential to advance our knowledge in terms of the effectiveness of existing control mensurate in Canada and the directions of future mitigation policies. However, more in-depth analysis could strengthen the scientific contributions of the manuscript. Further, there are quite a few clarification issues. These are listed in the section below.

Response: We have added in-depth analysis and clarified vague statements in the revised manuscript, such as (1) adding a brief description of the advantages and limitations of the two machine learning (ML) methods in Introduction, (2) clarified input and output variables of the ML training and testing ins Section 2, (3) adding city-level point source emissions to help interpret the obtained trends of the pollutants in several places in Section 3, (4) adding a subsection (the new 4.1) to summarize perturbations due to varying weather conditions on the decadal trends of pollutants, among many other changes as specified in detail below.

Specific Comments

Abstract

Line 20, “on the time scale of 20 years or longer, the perturbation from varying weather conditions exerted a very minor influence on the decadal trends of original annual averages (within ±2%) in ~70% of the cases, and a moderate influence up to 16% of the original trends in the other 30% cases”. Those statistics were not presented in the main body. Further, leaping from the difference in the annual means between the observed and deweathered data to “influence on the decadal trends of original annual

averages" could be problematic.

Response: We have added a new subsection to address this issue, which reads:
"*4.1 Perturbations due to varying weather conditions on the decadal trends*
Perturbations due to varying weather conditions on the decadal trends of the studied pollutants are presented in detail in Section 3 above, and key findings are briefly summarized here. The perturbations are defined as the percentage differences between the trends of the original and deweathered annual average concentrations. In ~70% of the studies cases covering all the selected criteria pollutants in the ten cities, the perturbation due to varying weather conditions had an influence of within ±2% on the decadal trends of the original annual averages over the 20-year period. In the remaining cases, relatively larger perturbations were identified, but at most 16%, keeping in mind that a portion of the percentage differences between the trends of the original and deweathered annual average concentrations was likely caused by errors inherent from BRTs and RF predictions.

Specifically, in all the cases except CO in Quebec city (for which the calculated perturbation is 7% from BRTs and 12% from RF), at least one of the two machining leaning methods generated a perturbation of smaller than 5%. For example, the top three largest perturbations obtained from using one of the two machining leaning methods were all for SO_2, including 16% from RF in Winnipeg, 14% from BRTs in Montreal and 13% from RF from BRTs in Toronto; however, the corresponding perturbations from using another one of the two machining leaning methods were quite smaller (4%, 0.2% and 3%, respectively), indicating possible large methodology uncertainties. Thus, perturbations due to varying weather conditions should be generally small on the two-decade time scale in most cases."

Introduction

A review of the "two popular machine-learning packages" (Line 167) should be presented, including the advantages over other methods, shortcomings or limitations, and applications in air quality studies.

Response: We have added a brief summary of the two packages in the revised introduction, which reads: "Machine learning techniques such as the random forest (RF) algorithm and boosted regression trees (BRTs) have been demonstrated to be a powerful tool to decouple impacts of emission changes and perturbations from varying weather and/or meteorological conditions, enabling the derivation of deweathered trends in air pollutants concentrations (Grange et al., 2018; Grange and Carslaw, 2019; Ma et al., 2021; Mallet, 2021; Shi and Brasseur, 2020; Wang et al., 2020; Munir et al., 2021; Lovric et al., 2021; Hou et al., 2022; Lin et al., 2022). 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 suffers 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)."

Method

The training of some machine learning methods requires input of known values of the dependent or output variable to facilitate the learning. For example, observed O3 concentrations are required to train the model to predict O3 concentrations using O3 precursor concentrations. Kindly specify all input variables in the training stage as well as the input and output during the testing stage. If deweathered air pollutant concentrations are required in the training phase, provide the sources of such datasets.
Kindly clarify that the training and testing were conducted for each site for each pollutant. It would be useful to have the performance matrix presented.
Kindly clarify 1) whether the performance presented are the results of the training phase or testing phase, 2) whether the trends presented are based on the training datasets, testing datasets, or the entire dataset, for the original, RF algorithm, and the BRTs, respectively.

Response: We have revised Section 2.2 to clarify these issues. All input variables were listed, and the training data sets were specified.

Fig 1 suggests that the observed concentrations were used in training and testing. There are two questions, 1) would a well trained and tested model be expected to predict concentrations with systematic and significant deviation for the observed values, 2) when the predicted concentrations deviate systematic and significantly from the observed values, is the bias due to a) a poorly trained model, b) uncertainty of the model predictions, c) influence of some factors, such as weather conditions, or d) some combination of causes listed in a)-c).

Response: The answer to question 1 above is "Yes", as has been widely reported in literature. As for question 2 above, the exact reasons for the bias in ML predictions are too complicated, and it is definitely important in developing the next generation of ML. However, our study here focuses on the application instead of the development of ML, and the detailed analysis on the ML performance issue is indeed beyond the scope of this study. We indeed need to make sure that the performance of the predictions meets the criteria values such as those set by USEPA.

Consequently, the use of regression slopes between the original and deweathered annual means to infer "perturbation from varying weather conditions" or "influence on the decadal trends of original annual averages" is debatable. As the authors pointed out,

the predictions of both models carry large uncertainties, and the agreement between the two models could be poor at times (Line 399). Therefore, some statements in the Abstract could be rephrased. Nonetheless, the 95% confidence intervals (CI) of each of the three annual concentration slopes, i.e., observed, RF-deweathered, BRTs-deweathered, could be used to determine whether the slopes are statistically different at 95% CI, and if yes in some cases perhaps derive a bound of the influence of the weather conditions on the long-term trend of original annual averages, with caution.

Response: We agree that the differences in the trends between the three sets of annual average concentrations could be largely due to the errors associated with the ML methods. We thus have added a detailed explanation on such issues in the revied section 4.1, as described in the response to the comment for the Abstract above.

We feel that it is very challenging to add the 95% confidence intervals in the trend analysis, which may not worth the great effort since we can already conclude that (1) trend differences due to varying weather conditions are mostly very small (<2% for 70% cases and <5% in most other cases) on the two-decade time scale, and (2) methodology uncertainties are mostly larger than the actual trend differences between the original and deweathered concentrations. We have added these details in the revised section 4.1.

Line 117, “To establish the relationship between air pollutants concentrations and emission reductions, the deweathered and original mixing ratios (or mass concentrations) of the air pollutants were correlated with the corresponding provincial-level emissions.” Did you mean the concentrations and emissions were correlated as seen in previous analysis, or correlation between concentrations and emissions was analyzed in this study? Kindly specify whether Pearson or Spearman correlation analysis was used and justify the method selection.

Response: We have revised the sentence as follows: “Pearson correlation analysis was further conducted for the deweathered and original mixing ratios (or mass concentrations) of the air pollutants against the corresponding provincial-level emissions”.

The use of “provincial-level emissions” (Line 120) in the analysis should be justified when there is a large variation of emission reductions among different cities in a province, such as Ontario. Note that the National Pollutant Release Inventory (National Pollutant Release Inventory - Canada.ca) could provide emission inventories at a smaller spatial scale.

Response: We have added Table S2 for the city-specific primary emissions to strengthen our analysis in the revision. However, only air pollutant emissions from point sources after 2002 are available from the database.

Line 129, “2.1 Monitoring sites and data sources”. Kindly specify 1) the averaging time

and unit of each pollutant. 2) when a large percentage of data is missing in a particular year, either pollutant concentrations or meteorological parameters, was that year considered in the trend analysis?

Response: The averaging time and unit have been added in the revised Table S1. We have also revised this part as follows: "Multiple monitoring sites exist in most cities, but only one urban background site was selected in each city mentioned above based on the following criteria: with the most complete dataset of the five selected criteria pollutants (NO_2, CO, SO_2, O_3 and $PM_{2.5}$), with the longest data record, and with valid data in each year (Table S1). In cases with a data gap longer than a year, e.g., in Quebec City, Halifax and Calgary, data at a nearby urban background site (within 1 km) were then used to fill the gap."

The deweathered concentrations were less affected by the data loss in a particular year because they are generally constant through a particular year. Moreover, the deweathered and original annual averages generally show consistent results, e.g., there were data loss for NO_2 mixing ratios in Halifax during a few months in some years, but the difference between the BRTs-deweathered and original annual averages was $\leq$3%.

Line 189, "The testing datasets were different between the RF algorithm and the BRTs." Kindly 1) justify the use of different testing datasets for the RF algorithm and the BRTs and the impact of this approach on the comparison of the performance of the two machine-learning methods. 2) whether the training datasets were the same or different between the RF algorithm and the BRTs.

Response: The software randomly divides the training and testing data sets and the user cannot control this. This point has been clarified in the revision. Normally, over one hundred thousand datasets are used for training for each pollutant in a city. Any random choice on using over one hundred thousand datasets for training of machine learning would not yield any detectable difference.

Line 265, "100th percentile", kindly explain this term, considering that the maximum concentration is still part of the sample. For example. the 90th percentile means 90% of the data points are below that value. Similarly, "95th-100th percentile PM2.5 mass concentration" (Line 519) needs explanation.

Response: The 100^{th} percentile is the maximum value in a particularly year and this has been defined in the revision. With the definition, it is fine to use 95th-100th percentile PM2.5 mass concentration.

Line 282, kindly justify the use of "The M-K analysis" instead of other trend detection methods.

Response: We have added the justification, which reads: "The M-K trend test is a non-

parametric test applicable to any type of data distribution and is employed…"

The attribution of all high PM2.5 concentrations to wildfire seems speculative or qualitative. The authors may want to clearly state the assumptions or use wildfire database and airmass directions to identify concentration data points under heavy influence of wildfires.

Response: The influence of large-scale wildfires on time series of $PM_{2.5}$ mass concentrations is very unique and obvious, i.e., a rapid increase to over 100 μg m^{-3}, lasting for tens or several days, as shown in Fig. S1ab in the original manuscript. In the last two decades, there were no other natural and anthropogenic sources in Canada that could lead to such rapid increase and long duration of $PM_{2.5}$ mass concentration. We agree that our approach might miss out some moderate and small wildfire contributions and our estimation could be considered as the lower limit value for all wildfires, as clarified in the revision.

The wildfire database suffers from the weakness in distinguishing smoldering combustion and flaming combustion. The former normally yields a much higher contribution to $PM_{2.5}$ mass concentration than the latter, but it is less detectable than the latter. The authors would like to use the pattern of time series of $PM_{2.5}$ mass concentration to identify the influence of large-scale wildfires, although we respect the reviewer's arguments.

Line 576, "Thus, O3 data with mixing ratios lower and higher than 40 ppb were analyzed separately below, with the former case representing net O3 sinks occurring in the atmospheric boundary layer and the latter one representing net O3 sources occurring therein (Table 3)." Kindly provide citations to support the classification of net source or net sink, i.e., 39 ppb being net sink and 40 ppb being net source. Atmospheric reactions of O3 suggest that both production and consumption occur at urban centers and a short distance downwind. Further, O3 concentrations vary significantly with season and city. Perhaps O3 concentrations collected at a background site could be used instead. Alternatively, the use of city specific median value to obtain two concentration levels in each city could be considered instead of a fixed value of 40 ppb.

Response: Only O_3 mixing ratios measured at remote sites under conditions with negligible natural and anthropogenic sinks can be used to estimate the O_3 derived from stratosphere, e.g., the observations during polar night at Alert and during the spring at Kejimkujik of Nova Scotia (Lat. 44.433611, Log. -65.205833, the data is available at www.canada.ca/en/environment-climate-change/services/environmental-indicators/air-quality.html.). The stratospheric input with negligible natural and anthropogenic sinks is quite stable at 40 ppb (Barrie et al., 1988, and references there in). Moreover, 40 ppb has been widely used as a threshold value to evaluate the impacts of O_3 on ecosystems (e.g., AOT40) (this information has been added in the revised manuscript).

Results

When reporting model performance in the main body, the units of the statistical metrics should be included when applicable.
When presenting r or R2 values, p-values should be included in the plots.

Response: The related information were added in the revised manuscript where applicable.

Line 403, “The increased uncertainties led to the difference between the RF-deweathered and original SO2 mixing ratios being up to 16% in Winnipeg.” Kindly clarify how these uncertainties were quantified for each pollutant in each city.

Response: This part has been revised as follows: “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 S4. 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.”

Discussion

This section is a mixture of method, results, and discussion. A consolidation of all methods or results in perspective sections could improve the readability of the manuscript.

Response: In the Discussion Section, we try to deepen the analysis, which sometimes needs additional analysis results. We could not find additional methods description in the section, except the justification for the choice of O3 mixing ratios ≥40 ppb and <40 ppb.

Conclusions

Limitations of the study could be included.

Response: We are not aware of any additional limitations except the methodology uncertainties that have been addressed in various places in Sections 3 and 4.

The reviewer did not find any conclusions on
1) “the perturbations from varying weather conditions on the observed mixing ratios” or on the long-term trends (Line 104)

Response: We have added section 4.1 for a detailed summary and we have revised the abstract for a brief summary on this finding. We feel that there is no need to repeat such a statement in the Conclusion section since there are already too many materials in this section summarizing the other major results. Simply repeating materials already presented in the Abstract in the Conclusion section may not be a good practice in our opinion.

2) whether the deweathered datasets yield any trends which are statistically different from that by the original dataset,

Response: See our response above related to this comment.

3) the benefit of employing two machine learning methods, and
4) whether the deweathering process is recommended in trend analysis of air quality.

Response: The benefit of employing two machine learning methods has been well reported in the literature and is presented in Introduction. However, BRTs suffers from overfitting while RF has the weakness in predicting the outlier in large percentiles. Here, we used both methods for trend analysis and presented the deweathered results, which were generally consistent with the original trends for over 20-year data analysis in Canada, but not for all the cases. We feel that the results could be similar or quite different when the datasets are in different lengths (such as with only 10-year data or less) or in different countries. A general recommendation should not be made solely based on results from just one study.

Overall, the scientific contribution and policy implication of the manuscript could be strengthened by considering the following, Incorporation Canadian perspectives, perhaps a map showing the locations of the 10 cities could aid the discussion of regional or transboundary inputs, if any.
Providing city specific information, such as site classification, major emission sources of each pollutant in each city, proximity to major point sources, emissions of such point sources from NPRI. Tidying up the interpretation of statistical results. Offering more reasoning, for example, whether the small influence of weather conditions on the 2-3 decades trends of air quality is expected and why; reasons of large discrepancy in emission trend and concentration trend, such as the decreasing trends in NO2 concentrations when NOx emissions were increasing during the same period (L352), and no trend in CO concentrations when CO emissions were increasing during the same period (L381). Including more in-depth analysis, e.g., whether the deweathered datasets yield any trends which are significantly different from that by the original dataset and why, and whether the two deweathered datasets yield similar or different trends, and why.

Response: We have addressed most of the concerns the reviewer listed here, as can be seen from our detailed response above. However, we admit that we do not have the

capacity to address everything raised here.

Other Clarification Issues

Citations seem missing at times, e.g., L46, “CAAQS”; L132, NAPS; L282, The M-K analysis.

Response: We are not sure what the reviewer meant here. These acronyms have been defined either in earlier places or on these lines and references have been provided where applicable.

L95, “but most modeling results suffer from large uncertainties, which could exceed annual average changes of the simulated pollutants.” Kindly clarify. Did you mean, “but most modeling results suffer from large uncertainties which could exceed changes in annual means of the simulated pollutant concentrations”?

Response: Agree and corrected.

L99, “weather and/or meteorological conditions”, kindly specify weather conditions and meteorological conditions, or perhaps choose one.

Response: The specific meteorological conditions (variables) that are needed for ML training are specified in section 2.2 and may not be needed here for easy reading.

L104, L14, L306, L309, L601, “the perturbations from varying weather conditions on the observed mixing ratios”, The reviewer is unsure about the “perturbations from varying weather conditions”, perhaps “perturbations due to varying weather conditions”, “perturbations from normal weather conditions”?

Response: The “perturbations due to varying weather conditions” is more suitable. Corrected throughout the manuscript.

L105. “criteria air pollutants”. Kindly specify whether it is “some criteria air pollutants”, or “all criteria air pollutants”.

Response: We changed to “some criteria air pollutants”.

L130, the list of the 10 cities should include the provinces.

Response: We feel this may not be necessary. For example, most of international readers would know Toronto instead of Ontario. If they don’t know Toronto, the inclusion of Ontario does not help anyway. The provinces might be added after the small cities, which are lack of international visibility.

L145, kindly provide 1) a list of monitoring stations in each of the 10 cities that are within 1 km, and 2) a list of sites with one or more years of unfilled missing data. Occasionally, long-term monitoring stations are relocated within a city. Was that encountered in this study?

Response: Fig S1 have listed all sites and the period of missing data used in this study. For all these sites in each city, the information can be found https://open.canada.ca/data/en/dataset/1b36a356-defd-4813-acea-47bc3abd859b. The website has been presented in the manuscript. It is very odd to list those sites, where the data were never used in this study. For some pollutants in some cities, no data are available to fill the missing data, such as no NO_2 data in Calgary since 2007 and Halifax since 2017, as shown in Fig S1a.

Line 147, "SO2, CO, NOx and PM2.5 emission data", kindly specify 1) the reporting time period, e.g., annual or monthly, 2) the types of emissions included/excluded, such as residential wood burning and wildfires.

Response: The detailed information can be found in https://www.canada.ca/en/environment-climate-change/services/environmental-indicators/air-pollutant-emissions.html. The website has been presented in the manuscript.

Line 158, the category of "AQHI between 4-6" should be provided.

Response: We believe the current information is clear enough, i.e., "Outdoor activities may be reduced at AQHI between 4-6 for certain population with some health issues."

Line 166, "2.2 Statistical analysis", data sources of "meteorological parameters" are better placed in section 2.1 Monitoring sites and data sources.

Response: We prefer to keep it this way since the meteorological data described here is for ML input. We feel keeping the software package and its input descriptions together is easy to read.

Line 171, "(hour, day, weekday, week and month)", kind specify each parameter, e.g., hour (0-23), day (1-365 or 366), week (1-52), month (1-12).

Response: It is really unnecessary because the software package is in public domain. The readers should follow the software protocol, especially for input data format, rather than the limited information presented here. This manuscript is already very long and we try to minimize the not-so-critical information.

Line 185, "Nevertheless, good performance can still be achieved in the present study mainly because of multi-decade length of the datasets". Kindly provide evidence that a

large dataset would lead to a good performance or rephrase.

Response: We believe it is a common sense for ML that more available data for learning, the better results would be obtained for their predictions. This also explains why the extreme values in large percentiles were usually poorly predicted by the ML methods because of their low occurrence frequencies, as shown in Fig 1.

Some results are in the Method section, e.g., Line 201-215, which could be better placed in the Results section.

Response: These materials are not trend analysis results generated form this study, but rather an illustration of the calculation procedure and should be presented in the Method section.

Some methodology descriptions are in the Results or Discussion sections, e.g., Line 563-577, which could be better placed in the Method section.

Response: See our response above for a similar comment.

Line 292, “Fig. 3a and b show decadal variations in the original annual averages of NO2 mixing ratios...” The reviewer could not find “b” being NO2 concentrations. Similarly, not both “Fig. 3c, d” (Line 463) are PM2.5 concentrations.

Response: The order of Fig 3a-d has been corrected.

Line 370, “Halifax (90-92%)…”, kindly clarify the meaning of the ranges.

Response: This sentence has been revised as follows: “The original and deweathered annual averages of CO decreased …”

Line 376 and other places including tables, the term “grand total and transportation emissions” is confusing. Kindly clarify whether there is one item, i.e., “grand total including transportation”, or two items, i.e., “grand total excluding transportation and transportation emissions”. Similarly, “total grand” in some tables.

Response: Rerevised as: “grand total emissions and transportation emissions” everywhere.

Line 440, “The increased O3 mixing ratio values likely reflected the lower limit resulted from the reduced titration reaction between O3 and NO (Simon et al., 2015; Xing et al., 2015).” Kindly rephrase.

Response: The sentence is revised as: “The increased O_3 mixing ratio was likely caused by the reduced titration reaction between O_3 and NO, considering the reduced

photochemical formation of O_3 in the troposphere."

Line 555, seasonal averages, kindly specify which months are classified as each of the four seasons.

Response: Added as recommended.

Line 669, kindly clarify whether 1) "decrease in NO2 during the last 2-3 decades varied by 37%-62%", or "decrease in NO2 during the last 2-3 decades ranged 37%-62%", and 2) "37%-62%" are among the three datasets or among the 10 cities.
Tables should be referenced when reporting results.

Response: The range (37%-62%) is among the 10 cities. The sentence has been revised as: "The overall percentage decrease in NO_2 during the last 2-3 decades among the 10 cities ranged from 37% to 62%, and the annual decreasing rates varied from 0.31 ppb year $^{-1}$ to 0.74 ppb year $^{-1}$."

Fig 1, kindly clarify what is being predicted by the models, deweathered concentrations or observed concentrations. If the former, kindly clarify the source of observed deweathered concentrations. If the latter, kindly justify the need of those models.

Response: The caption is 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."

"Fig. 2. Correlations between hourly PM2.5 concentration in a single year and its 22-year average in each hour in Edmonton." Did you mean, "Fig. 2. Correlations between hourly PM2.5 concentration in a single year and 22-year average PM2.5 concentration in each hour of a year in Edmonton"? Furthermore, the reviewer could not find "percentile series" in the "Left column".

Response: Corrected.

"Fig. 4 Deweathered hourly mixing ratios of O3 (left column) and NO2+O3 (right column) at levels ≥40 ppb in five eastern Canadian cities." These bar charts seem to suggest little variability among hourly concentrations within any of the years.

Response: Yes, the deweathered hourly mixing ratios reflect the average of 1000 runs in different meteorological conditions and should have little variability within any of the years, except those accident events.

Fig 6. Kindly clarify the pollutant studied and which factor the “perturbation contribution” refers to.

Response: It should be Fig. 5, and it is for $PM_{2.5}$ (information added)

Editorial Suggestions

Typos, syntax errors, and awkward word choices could be corrected. For example,

Line 44, “human health and the Environment”, maybe “human health and the environment”.

Line 62, “95% cities”, perhaps “95% of cities”.

Line 113, “accurately quantify”, suggest considering a more conservative term such as “better quantify”.

Line 130 and other places, “Quebec”, “Quebec City” could be more appropriate.

Line 163, “British Columbia Province”, could be replaced with “British Columbia” or “the province of British Columbia”. Similarly, “Alberta province” (Line 354).

Line 172, “ambient temperature”?

Line 192, “coefficient of determination (R^2)” could be more appropriate.

Line 206, “reasonably well reproduced”, kindly rephrase.

Line 211, “good predictions”?

Line 292, “Fig. 3a and 3b”?

Line 342, “strong correlations”?

Line 366, “mixing ratios again the original ones varied from 0.97 to 1.03”, kindly rephrase.

Line 378, “nearly” could be replaced with “approximately”.

Line 399, “regional transport on the continental scale”, kindly rephrase.

Line 410, “large discrepancy”?

Line 595, “In the cases with O3 mixing ratios □ 40 ppb”?

Line 617, “dominantly contributed to the population-weighted exposure to PM2.5 in northern Canada (59%) and western Canada (18%)”. kindly rephrase.

Line 678, “By only considering”?

Line 693, “caused AQHI to a level of above 10”, perhaps “elevated AQHI to a level of above 10”.

Response: All of the above and similar places have been corrected as recommended.