We thank Referee #2 for their helpful comments (in italics) and provide responses inline.

Wang and co-authors use a highly-time resolved measurement system located in a ground-floor laboratory in a building near a busy Toronto road to measure traffic-related plumes and derive fuel-based emission factors. Sampling periods are spread through the four seasons, beginning in December 2013 and running to September 2014. Plume detection was based on the rate of change of CO2 concentration. Over this period, the team identified over 150,000 plumes. Of these, 100,000 were of sufficient length (10s) and size (integrated CO2 area) for determination of the fuel-based emission factors for particle number (> 7 nm), CO, NO, NOx, black carbon, methanol, benzene, toluene, ethylbenzene, and xylenes. The technique recovers a distribution of the emission factors of the in-use vehicle fleet in Toronto. The plume detection rate (assuming sampling was conducted continuously for 3300 hours – the issue of downtime is not discussed) was about 45 plumes per hour.

1. The paper requires major revisions as I believe it is not in enough depth with respect to the methodological details and sensitivity of results to data analysis assumptions. And, for an article in a journal focusing on metrology technique, the paper has too much in terms of results and policy relevant statements. I believe a stronger emphasis on the technique and data analysis will serve the measurement technique and research group well in the long term. The results can be placed in a more focused article, with methodological details in the AMT paper. For example, at lines 113-117, the authors make clear that they have a strong interest in showing the applications of the method. I feel the method itself needs more exploration, and that the paper would be improved by more in depth methodological evaluation, and moving some of the applications (particularly the evaluation of the local emissions testing program) to a more policy relevant journal.

This reviewer believes the measurements are an important contribution, and that (because of the large number of plumes) a valuable quantification of fleet EF variability. The methodological and data analysis rigor that I believe are needed (and describe below) are not meant to be a barrier to publication, but rather to add value to the measurements so they can be widely used by the metrology, modeling, and policy communities.

The authors thank both reviewers in their request for more description and assessment of the method itself, including the plume identification algorithm and the emission factor calculation. We feel that the addition of more detailed descriptions of the EF calculation approach and the new sensitivity analyses have helped to strengthen both the paper and the overall method. These additions are discussed in response to specific comments below. While AMT is a method-focused journal, the authors find that the short discussion of the implications of these results will nonetheless be of interest to the readers.

2. At the paragraph starting at line 137, it is not clear which of these are from government statistics on the fleet, and which are from observations at the monitoring location. For the observations at the monitoring location, the method is not clear. If the method is video analysis, is daytime included as well as nighttime?

This paragraph was rewritten for clarification. The first part of the paragraph (lines 135 – 143) is derived from government statistics. The sentence starting on line 143 was measured at the site

using a dual radar system (SmartSensor HD), which previously was described in the Supplemental Information. Description of the SmartSensor HD has been moved into the main text to lines 185 – 191, and the paragraph in lines 134 – 145 has been rewritten as follows:

"From the 2009 Canadian Vehicle Survey conducted by Statistics Canada, the age breakdown averaged across all on-road vehicle types in Ontario is 45% below 5 years, 35% between 5-11 years, 14% between 12 – 18 years, and 6% older than 19 years: with a mean fleet age of around 7 years (StatsCan, 2010). Based on vehicle weight class, vehicles less than 4500 kg (i.e., cars, station wagons, vans, pickup trucks, small trucks) make up 97% of the vehicle fleet, with the remaining 3% being trucks heavier than 4500 kg (i.e., large pickup trucks, larger trucks, tractor trailers). Of the <4500 kg vehicle weight class, fuel type is dominated by gasoline (97%) compared to diesel (3%); the inverse is reported for the heavier weight class with 83% diesel and 17% gasoline usage (StatsCan, 2010). A similar fleet makeup is observed at the measurement site between vehicle classes (Figure 1c,d) with roughly 3 – 4% of the passing vehicles identified as trucks, with vehicle counting done using a dual radar system described in the next section."

3. The supplement has much less treatment of error, precision, uncertainty, and sensitivity analysis than I expected. These concepts have very limited discussion in the article itself. This seems a critical deficiency in a measurement-focused journal.

A sensitivity analysis has been included, which addresses the points in reviewer comment #8 and tests how shifting the criteria and parameters affects the mean emission factors derived for the different pollutants. A summary of the percent differences in mean values (Table S1) is included in the supplementary information, with the results and discussion of the analysis added to lines 237 – 264 of the text. This gives the reader a range of the potential errors associated with the method, and the instrumental uncertainties are included in Table 1 as effective sensitivities and uncertainties.

"A sensitivity analysis was conducted on the effects of changing various capture criteria (i.e. minimum plume duration and CO_2 response, shifting start and end times of the plumes) and is summarized in the Table S1 of the Supplementary Information. The largest changes are -10% for the mean CO EF when decreasing the minimum CO_2 response threshold. Benzene (C_6H_6) EFs were also impacted by changing in the minimum CO_2 response, with a change of up to -8% with increasing response threshold. A change in minimum plume duration had little effect, mainly due to the overlap with the minimum CO_2 response criterion being more stringent than minimum plume time. However, in the extreme cases where the minimum CO_2 response threshold was set to zero to nullify its effect, changing the minimum plume time to 30 seconds resulted in higher values on average for the pollutant EFs. By lowering the minimum plume time to 1 second, the opposite effect is observed, mainly due to the inclusion of very small plumes having zero EF values. Neither of these criteria would ever be set for a 1 Hz CO_2 signal, as a minimum of 30 seconds removes nearly 70% of the plumes whereas a minimum of 1 second would consider even the smallest increase in the CO_2 trace to be a plume.

Considering the plume start and end times, increasing or decreasing the plume period by 4 seconds results in a change in mean EF values of 0.2 – 10%, depending on the pollutant, which is further described in the Supplementary Information (Table S1). Increasing or decreasing the plume period by 10 seconds results in a change in mean EF values of 0.2 – 18%, again depending on the pollutant (Table S1). However, in both cases, BT or zero EFs can affect the mean EF as a result of the minimum criterion used. Thus, a better way to look at the effects on the absolute values of EFs would be to focus on above threshold (AT) EFs only. In this case, increasing/decreasing the plume periods by 4 and 10 seconds results in only a 0.2 – 8% and 0.2 – 10% difference in mean EFs, respectively. The effect on mean EF as a function of random subsampling of the dataset was also tested. With a subset containing 10% of the original dataset, the percent difference of the mean EFs ranged between 0.3 – 6.3% depending on the pollutant with the highest being for CH_3OH. At a 1% and 0.1% fraction of the data, the mean EFs deviated upwards of 20% and 90% respectively. Further discussion of the sensitivity analyses are included in the Supplementary Information."

4. The paper mentions validation of the PTR-MS using standard gases and refers the reader to Wang et al. (2005). Does a 2005 paper really quantify the errors in a measurement taken in 2013/2014? Can the uncertainty and detection limit of the instrument be stated, if not in the paper, in the supplement?

The original citation of Wang et al. (2005) referred specifically to the method description and errors associated with the off-line GC-MS analysis of VOCs. However, due to potential confusion, this has been removed completely because it does not provide any information on the PTR-TOF-MS calibration itself. A six level calibration curve was performed on-site for the PTR-TOF-MS for the relevant VOCs. The integrated GC-MS VOC data were used only to validate the mixing ratios reported by the PTR-TOF-MS.

5. Is there any intercomparison information or electrometer calibration available to quantify the CPC accuracy?

Inter-comparison with auxiliary particle counting and sizing instrumentation (API 651, FMPS 3091, TSI 3788) co-located at the field measurement facility was used to verify agreement between instruments. This included an electrometer based FMPS but this instrument alone was not treated as a particle counting measurement standard. Rather, as not absolute standard for particle number counting is available, comparisons with multiple instruments using different operating principles was used. The assumption underlying this approach is that it is highly improbable that a number of different instruments would all drift in such a way as to continue to agree with each other.

6. The acronym ADL is used in the supplement without being spelled out.

The description to this abbreviation was added to the main text (line 261) and the supplementary information (line 143). In order to avoid confusion of the technical definition of "detection limit" of an instrument (i.e., 3σ) and our ADL minimum threshold value used as a criterion for a plume capture, the abbreviations using "detection limit" have been changed. We now use the terms above threshold (AT) and below threshold (BT) when referring to emission factors.

7. In the supplemental section, the integral of the CO2 time series is reported as having units of ppmv. However, the time integral of a CO2 time series will have dimensions of mixing ratio seconds.

We agree that the time integral of a CO_2 time series has dimensions of ppmv-s.

8. At line 195, how was 10 seconds and 5 ppm selected as the threshold. How sensitive are the results to this threshold? I believe a sensitivity analysis is needed, with the results recalculated for alternate settings (e.g. 12 seconds and 4 ppm; 8 seconds and 6 ppm, etc.). Alternately, can the error or uncertainty embodied in the arbitrary selection of 10 sec or 5 ppm be quantified and then propagated through the analysis by traditional error propagation / error analysis techniques.

Our threshold was 5 ppmv-s, which can be equated to many combinations of average mixing ratio and duration. The issue of sensitivity to the 10 s additional criterion was further addressed in response to reviewer comment #3.

9. What is the accuracy or uncertainty associated with the selection of the baseline? (line 207)

This was evaluated in two ways. Firstly, we compared the integration baseline defined by the algorithm with a baseline chosen by the user manually for the same plume. The emission factors between these two methods only differed by 8% on average. Secondly, the uncertainty was tested in a sensitivity analysis by shifting the plume start and end time, which would shift the baseline up and down. This affected the resulting integrated value and resulting mean EF at most by 10%.

10. What is the formula for the effective sensitivity (line 210 and table 1)

The formula for effective sensitivity has been added to the main text in lines 221 – 222.

$$Effective\ Sensitivity = \overline{(\mathrm{P}_{max} - \mathrm{P}_{min})} \qquad (2)$$

$$EF_{min} = \left(\frac{Effective\ Sensitivity}{\Delta[\mathrm{CO_2}]+\Delta[\mathrm{CO}]}\right) w_{\mathrm{C}} \qquad (3)$$

Additionally, a description has also been added in lines 223 – 235 of the text:

"The effective sensitivity for each instrument was determined based on period of stable signal with no vehicle influence across all measurements over 2 – 3 minutes. Three of these periods within each measurement campaign were chosen from the beginning, middle, and end of the campaign. The mean difference between the maximum and minimum signal during these stable periods (Eq. 2), corresponding to roughly 5 – 6σ, were used to set a minimum threshold for the plume capture criterion for each pollutant measurement respectively (Table 1). Emission factors calculated from captured plumes, those that had detectable CO_2, but pollutant signals lower than

the instrument sensitivity were classified as below threshold (BT) EFs and calculated using the effective sensitivity value (Eq. 3)."

11. The time resolution in table 1 – is that the reporting frequency of the instrument? Another important temporal variable in the method is the degree of temporal "smearing" that occurs; in other words if a sharp step change is applied at the inlet of the monitoring system, how sharp is the reported signal from each of the instruments?

The time resolution in Table 1 is the reporting frequency of the instrument. In most cases this is also the minimum measurement frequency of the instrument with two exceptions: 1) the PTR-TOF-MS actually measures at 10 Hz, however this signal is far too noisy and is set to a reporting frequency of 2 seconds that is internally averaged by the instrument; and 2) the CO monitor was set to a reporting frequency of 10 seconds, however it was discovered that the measurement frequency is actually internally averaged to 60 seconds and this cannot be changed by the user. Thus the signal seen in Figure S1 for CO does appear to have the described "smearing". The impact of this effect is discussed in response to comment #22. Significant "smearing" was not observed for the other instrumentation.

12. At line 224, QA measures are mentioned, but it is not clear what the QA measures are. From figure S2, I gather that the QA is to test if the peak is sufficiently large. This falls short of what most readers think of in terms of quality assurance – which usually involves some combination of tests relative to independent challenges such as replicate measures, a test of monitor response to a calibration standard, zero check, leak check, etc.

The term "quality assurance" has been changed throughout the text to "plume capture criteria", as it is a set of criteria and thresholds that define as successful capture of a specific plume, which is already described in the methods of the text. Quality assurance was also conducted on each instrument before and after each campaign in the form of leak tests, calibrations and zeroing.

13. The validation methods mentioned at line 234 are important, but this is the first mention of them in the paper.

The need for validation of the method is outlined at the end of the introduction section in lines 104 – 108. Interpretation and discussion of the sensitivity analysis of the method has been added at the end of the Experimental Methods section addressed in reviewer comment #3, in addition to validation done using the test vehicle driven by the measurement site discussed in Section 3.1.

14. What is "capture efficiency" as used at line 253

Capture efficiency is the number of captured plumes divided by the number of total recorded passes of a test vehicle. This has been added to lines 298 – 306:

"During the same nighttime tests, a passenger car was driven past the measurement site multiple times in order to better test the capture efficiency of the method. The capture efficiency in this context represents the number of plumes that passed the capture efficiency divided by the total number of passes recorded from the tests. Based on CO2 plumes that passed the minimum

capture criterion, for roadside versus near-road inlet measurements, the capture efficiency at night averaged at 70% and 46%, respectively. Drive state was also investigated, where the vehicle cruised pass, was set to idle in line with, braking at, or accelerating from the inlet. Idling in line with the measurement inlet resulted in a slightly higher capture efficiency of 52% compared to cruise, accelerating, and braking (44%)."

15. I find the report of the experiments in the paragraph starting at line 253 incomplete and confusing. Is there a diagram of this test, and clear explanation of what is being controlled, what is being measured / calculated, etc?

A description of this has been added in lines 292 – 296 to better explain the experiment, however the set up was the same as the ambient measurements of the fleet that would pass the site, with the exception that the vehicle passing is known and logged. For testing capture efficiency the only measurement required was CO_2. If the integrated amount of CO_2 is above our threshold value, this event is considered a captured plume and marked as captured. Because an observable CO_2 increase was not always recorded for every test vehicle pass, a capture efficiency lower than 100% is possible.

16. At line 264, it seems there is confusion between a high concentration event, and a high emission factor.

This analysis focused on high concentration events and any mention of high emission factors have been removed from this paragraph to avoid confusion. The text has been changed in lines 312 – 316:

"These individual plumes were then classified as having no truck influence (cars only) or having at least one truck (trucks with or without cars) using video recordings of the passing vehicles. Of the 152 plumes examined, 88 had an influence from trucks, and 62 were from cars only periods. Truck-influenced periods had higher NO_x, BC, and PN EFs on average, whereas car plumes had higher CO and BTEX EFs (Figure 5)."

17. At line 276, what constitutes a "well-tuned" engine for the purposes of this study? This is, as used, an unnecessarily vague term.

This part of the data analysis and corresponding description has been removed from the text to improve flow and clarity.

18. I do not see the purpose of Figure 1e.

Figure 1e and associated interpretation has been removed from the main text.

19. The results from testing against the direct injection engine are discussed but not shown graphically or quantitatively. In the current manuscript, it is stated that a test was done, and that the result was within 14% of a value in the literature. This is to be accepted as method validation. I believe significantly more detail is needed to make this into something that lends credibility to the measurement technique. Comparing to only one emission factor, and not

discussing its representativeness of accuracy / QA – lessens the impact of this section. Not presenting any tabular or graphical data from the drive by plumes also lessons the impact. I recommend this be strengthened or deleted. And since comparison to a different vehicle is always needed (I don't believe there are dynamometer tests and drive by tests in Toronto of the same vehicle), removal of this is probably prudent.

The reviewer is correct in commenting that there is not enough description or data information of this separate test in the results to have a significant influence in terms of the validation of the method. Although the dynamometer tests we conducted were did not test the exact same engine, both engines were from the same manufacturer and are the same model and year. It was not expected that the two engines would exhibit exactly the same EFs, but being in the same range is promising in terms of the method. As a response to the reviewer comment, this data and interpretation has been removed from the text and will be further expanded upon in a future publication, which will focus on these dynamometer tests.

20. The size cut of the CPC (and thus of the particle number emission factor) needs to be mentioned more prominently, since the intercomparison of particle number emission factors is only relevant when the size cut of the measurement is known. It should appear at line 180 (methods section) and in all tables and figures than include PN measurement or PN emission factor.

This has been added in the tables and included in the instrumentation part of the methods section.

21. How much downtime occurred during each measurement campaign? If there was significant downtime, was the remaining active sampling time concentrated in particular time periods such as certain days of the week or nighttime periods?

There was not a lot of down time across the four campaigns, however there were two weeklong periods of downtime that did occur in the spring and summer campaigns as a result of one instrument failing (May 17 – May 23, 2014 & July 28 – Aug 4, 2014). Because one criterion of the analysis requires all instruments to have been running, these segments were completely removed from the data analysis. Because the entire period was removed, there is no bias in terms of time of day effects on the data.

22. In the supplemental section, it can be seen that the CO plumes start before the CO2 plumes. However, it seems the plume start time is determined by the CO2 time series. The CO instrument seems to be smearing out the CO over time. If this is not accounted for the CO emission factors will be biased low. This is a key issue in the method (possibility for temporal mismatch and temporal broadening of the peaks across different instruments). It needs to be clearly explained/ explored in the methods and results sections – possibly with support in the supplemental section. Alternate methods such as allowing different plume start and end times for different pollutants need to be considered. Sensitivity analysis of the results to the algorithm start and end time selection method is also needed.

The authors would like to thank both reviewers for this comment in realizing the difference in time response of the CO measurement. This has been addressed by altering the method to reflect

the lower time response of the CO instrument. Although the method worked well for the other instruments, that all had similar time responses, with CO_2 measured at 1 Hz and CO averaged at 60 seconds, the broadening of the CO signal was an issue in terms of affecting adjacent plumes. In a new method, specific to CO, the same identification and emission factor calculation was used, however CO_2 was averaged to 60 seconds to match the CO time response, and plumes were identified using the averaged signal. The maximum of each CO plume was recorded and then matched with plumes identified using CO_2 at 1 Hz, where all other plumes that did not have a matching CO plume signal were given a zero or below threshold EF value. As a result of this reanalysis, the 10 – 20 second buffer was no longer required. In terms of instrument response time, the other measurements were visually inspected to ensure this broadening was not an issue. Although the mean CO EF value did decrease by nearly 33% from 14 g kg^{-1} to 10 g kg^{-1}, this was mainly due to an increase in zero value BT EFs that were previously adjacent to plumes that had large detectable CO. It is important to note that the above threshold EFs (included in Table S3 of the supplementary text) only decreased slightly from 41 g kg^{-1} to 39.9 g kg^{-1}, thus there was a much less significant change in terms of the actual CO EF values. This reanalysis of CO EFs did not change the conclusions of the paper, although the decrease in number of CO plumes did increase the contribution from the top 25% emitters from a 93% contribution to almost 100%. The percent contribution of vehicles that surpassed the Drive Clean limit for CO decreased from 60% to 54%, with fewer plumes exceeding this limit. A description of this change was added to lines 214 – 220 of the text:

"For CO, the response time of the instrument is much lower than all the other instruments with plumes visibly wider than the other traces (Figure S1). Thus for CO the plume identification had to be altered slightly where the same identification algorithm was applied to identify each plume, but CO_2 was averaged to 60 seconds to match with the CO signal. The plume identified at this lower time resolution was then matched with the nearest plume identified using CO_2 at 1 Hz (Figure S2). This was only required for calculating EFs for the CO signal, which was measured at a lower time response than other pollutants."

Minor comment 23. Line 143 "observation is observed"

Text changed to "A similar fleet makeup is observed…"