We thank Referee #1 for their helpful comments (in italics) and provide responses inline.

Overview:
The authors present results from a near-road motor vehicle emissions measurement campaign in Toronto, Canada. Concentrations of a variety of pollutants were measured 15 m from a busy roadway, and an algorithm was developed to (1) automatically identify vehicle emission plumes using CO2 data and (2) calculate fuel-specific pollutant emission factors for individual vehicle plumes. The authors identify a total of 103,000 individual plumes in their data set, though the number of plumes for which pollutant emission factors can be quantified is significantly lower due to instrument sensitivities and plume dilution. Results are used to investigate fleet-average emission factors, impacts of high-emitting vehicles, and the efficacy of local vehicle emission control regulations.

In general, I thought the authors present an interesting new technique for collection and analysis of near-road, high time-resolution air pollution data. The sample size of individual vehicle plumes considered here is certainly impressive and the authors use results for some interesting investigations. However, given the aim and scope of AMT, I would like to see a more thorough explanation of the algorithm used in this study for automated data analysis. This seems to be the novel contribution of the study in regards to measurement techniques, and the current treatment of the algorithm in the text left me somewhat confused and limited my confidence in presented results. I believe this issue should be addressed prior to consideration for publication in AMT. Specific comments are included below.

Specific Comments:
Lines 51-52: I found this paragraph to be slightly confusing, partially due to the use of "top-down" and "bottom-up" when referring to emissions modeling and emission factor characterization, respectively. This terminology indicates that emission models and emission factor characterization are two complimentary approaches for evaluating vehicle emissions; however this is not the case. I suggest deleting the terms "top-down" and "bottom-up". On a related note, it was not clear to me how real-world emission measurements can be used to "bridge the gap between laboratory and ambient observations". The authors should clarify what differences between laboratory and ambient measurements they seek to reconcile with their measurements.

"Top-down" and "bottom-up" terminology has been removed from lines 51 and 55 to improve clarity. The differences between laboratory and ambient measurements are described in lines 55 – 61, and are mainly a result of ambient effects on emissions post-tailpipe as well as the type and condition of vehicles within the fleet:

"Although engine dynamometer studies provide highly accurate measurements of tailpipe emissions in controlled laboratory conditions, they have two main limitations: (1) not being fully representative of real-world emissions (e.g., driving behaviour, ambient effects on emissions pre- and post-tailpipe, wear and age of vehicles in the in-use fleet) and (2) small vehicle sample size resulting from the restrictive cost and duration of experiments (Dallmann and Harley, 2010). Real-world measurement of in-use vehicle emissions presents a suitable method to bridge the gap between laboratory and ambient observations."

Lines 67: Remote sensing typically refers to a specific method for measuring pollutant concentrations in vehicle exhaust plumes employing cross-road spectroscopy. The authors incorrectly refer to all near and on-road emission measurement techniques (e.g. tunnel studies, plume capture, remote sensing) as remote sensing here. This indicates unfamiliarity with distinctions between methods used to evaluate vehicle emission factors in near-road settings (despite the large number of cited studies in this paragraph). I suggest the authors revise this paragraph to more clearly explain similarities and differences between real-world emission measurement techniques.

The term “remote sensing” has been removed from the text. While a comparison between the different real-world emission measurement techniques would be interesting, the authors feel that an in-depth review would add significant length and does not lie within the scope of this manuscript. Instead, we prefer to refer the reader to an excellent review compiled by Franco et al. (2013). Reference to this work and other relevant studies is included in the following lines:

“Real-world studies can include on-road and tailpipe measurements that provide representative EFs for in-use vehicles (Franco et al., 2013). Similar to engine laboratory studies, these measurements typically sample a limited number of vehicles. In contrast, the measure of vehicle emissions at the roadside or near-road region provide a much larger vehicle sample size over a relatively short period of time and have been successfully applied to tunnel (Pierson and Brachaczek, 1982;Miguel et al., 1998;Rogak et al., 1998;Kirchstetter et al., 1999;Hwa et al., 2002;Kristensson et al., 2004;Geller et al., 2005;Ban-Weiss et al., 2009;Ho et al., 2009;Dallmann et al., 2012;Mancilla and Mendoza, 2012;May et al., 2013), roadside (Bishop and Stedman, 1996;Sadler et al., 1996;Jimenez et al., 2000;Chan et al., 2004;Ko and Cho, 2006;Guo et al., 2007;Westerdahl et al., 2009;Phuleria et al., 2007), and near-road environments (Nickel et al., 2013;Liggio et al., 2012;Imhof et al., 2005;Ning et al., 2008;Ježek et al., 2014;Janhall et al., 2012).”

Lines 102-103: From a statistical perspective, is it possible to estimate the number of individual vehicle emission measurements needed to accurately estimate the fleet-average EF?

A sensitivity analysis was added to the supplementary text on how the fleet mean EFs were affected by random subsampling of the dataset, with the changes in the mean plotted in Figure S4. The mean did not change by more than 10%, however once random subsampling was below 5% of the dataset, the mean EFs began to deviate upwards to 14% for certain pollutants. The lines added were in the supplementary information as follows:

“The number of individual plume measurements required to accurately estimate the fleet mean EFs was tested with a sensitivity analysis using random subsampling of the dataset was conducted. This was done by decreasing the fraction of the data used to calculate the mean (Figure S4) with 100% being 103,000 plumes to 95, 75, 50, 25, 10, 5, 1, 0.5, and 0.1%. With a random subsampling with a 10% fraction of the data, the mean EFs deviated at most 6%. However, with 1% and 0.1%, the deviation in mean EFs were upwards of 19% and 87%, respectively. Although the objective of this analysis was to evaluate the fleet mean EF, it also shows that the variability of EFs across the fleet population can be estimated with >10% or >10,000 plumes for these pollutant measurements."

Lines 211-213: Please include the actual number of stable ambient periods considered in the calculation of instrument sensitivities. Did sensitivities remain constant in each of the 4 seasonal sampling periods? Also, please explain how EFDL values for each species were calculated and include EFDL values in Table 1.

Determining the effective sensitivity for each instrument, which is used as the minimum threshold for the emission factor determination, was calculate based on periods of non-vehicle influence. To ensure that these periods were representative across the entire measurement campaign, 2 – 3 minute periods were chosen at the beginning, middle, and end of each campaign. The time series of all pollutant measurements were visually inspected to ensure there were no spikes or plumes, and subsequently verified by video recordings of the roadway. The difference between the maximum and minimum points of each vehicle-free period was used to determine a mean minimum threshold value for each pollutant, which corresponded to around 5 – 6σ in each case. These values are included in Table 1 as the effective sensitivity of each instrument and a description has been added to the text at lines 223 – 235:

"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)."

Section 2.3: Given the aim and scope of this journal, I expected to see a more complete description of the automatic plume identification and integration method included in the main text. It seems to me that the novel aspect of this work in regards to atmospheric measurement techniques is the algorithm used for the analysis of high time-resolution data collected at a near-road sampling site. Unfortunately, the description of the algorithm is limited to six lines in the main text, and I remain somewhat confused about how (1) vehicle plumes were identified, (2) how integration start and end points were identified, and (3) how values for the two criteria for a "successful" plume capture were determined (plume width greater than 9 seconds and average delta [CO2] greater than or equal to 5 ppm).

The description of the method has been expanded, including a sensitivity analysis of various changes to the plume identification algorithm and the resulting effects on the mean EFs. Figure 2 has been changed to indicate the inflection points marking the beginning and end of the identified plume, and Figure 3 has been moved from the Supplementary Information to the main manuscript as a flow diagram outlining how the data was processed. The main criterion of a successful plume capture was based on the time series of CO_2. A sensitivity analysis was performed using this criterion to find an optimal capture while avoiding erroneous plume identification (i.e., small and/or shallow slope increases in the CO_2 time series, very short spikes). While this criterion was adjusted, visual confirmation of successfully captured plumes was also performed. A new description has been added in lines 237 – 264:

"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."

Section 2.3: An important question related to my comment above is the how data streams from each of the instruments were time-aligned in order to assure that plume start and end times derived from CO2 data were appropriate to use for other species. For example, were any offsets made to the data to account for different instrument response times?

A sensitivity analysis was used to determine the lag time associated with each instrument relative to the CO_2 measurement, because this is the reference variable. The sensitivity analysis involved shifting the time series of each pollutant in 1 second increments within a ±30 second range. Each time-shifted series was then correlated with the CO_2 time series to find the most appropriate lag time. As all instruments had relatively high time responses and the residence time from the inlet to the instrument was short, the variations in lag times were never higher than 30 seconds and in most cases were between 0 – 10 seconds. After each pollutant time series was corrected for lag, it was also visually inspected to ensure the plumes matched well temporally. A description of this process has been added to lines 193 – 199:

"All pollutant measurement data were first time corrected to match with the CO_2 signal. This was done by a sensitivity analysis that involved shifting the time series of each pollutant in 1 second increments within a ±30 second range. Each time-shifted series was then correlated with the CO_2 time series to find the most appropriate lag time. After each pollutant time series was corrected for lag, typically by 0 to 10 s, it was also visually inspected to ensure the plumes matched well temporally."

In the Supporting Information, the authors state that a 20 second buffer was added on each side of the plume length for CO integrations due to the lower time response of the CO instrument. How does this impact CO EF calculations for successive plumes separated by less than 20 seconds (as shown in the right panel of Figure S1)? It seems as if this would result in integration of the same selection of CO data for two separate plume events.

The authors would like to thank the reviewers for this comment. The impact of this effect has been addressed in response to reviewer #2 in comment #22.

Lines 207-208, SI Individual EF analysis section: In lines 207-208, the authors state that background concentrations are set to the minimum concentration level at the beginning or end of a plume. In the supporting info, the authors describe a slightly different method for determining background concentrations for BC, which involves averaging lower value points to determine a minimum value. This is a relatively minor difference, but in relation to my comment above on Section 2.3, these inconsistencies make it difficult to understand how the algorithm developed by the authors actually evaluated EFs for exhaust plumes.

The BC background correction described in the SI was originally performed in an attempt to take into account the effect of instrumental noise. However, we subsequently found that the minimum threshold approach was more effective for correct plume identification. This sentence has now been removed from the supplementary text to improve clarity and does not affect the conclusions in any way.

Section 3.1: For the 103,000 successful plume captures, please include descriptive statistics (range, mean, median, etc.) for $\Delta[CO2]$, dilution ratio, and plume width.

This information has been added to lines 270 – 276 for plume width in terms of time, $\Delta[CO_2]$, integrated amount as mg C m^{-3}, and dilution ratio.

"The length of the plumes ranged from 10 to 197 seconds, with a mean and median duration of 57 and 53 seconds, respectively. In stagnant conditions, plumes have been observed to last upwards of three minutes, however these plumes are infrequent and make up less than 0.1% of the total captured plumes. The mean and median (range) of $\Delta[CO_2]$ was 10 and 7 ppmv (3 – 330 ppmv), the integrated amount was 192 and 128 mg C m^{-3} (50 – 3675 mg C m^{-3}), and the dilution ratio calculated by assuming 140,000 ppmv (14%) tailpipe emission of CO_2 was 15500 and 14000 (424 – 49970)."

Lines 238-251: The authors compare particle number EFs calculated from measurements made at 3 m and 15 m from roadside for 287 vehicle plumes as a method validation step. This is a useful comparison; and I suggest that, in addition to comparing the mean PN EF calculated at 3 m and 15 m, the authors include a scatter plot showing the comparison of PN EFs calculated at 3 m and at 15 m for the 287 individual plumes. This would provide an indication of how effective the authors' methods are in quantifying EFs for individual plumes. This plot can be included in the supplementary information.

Unfortunately this could not be easily done as it was impossible to lag correct all the plumes due to effects of different wind speeds and direction between the roadside and the building. Coupled with the sheer number of plumes, during rush hour when this test was conducted, a mean value of each distance was used in place of a scatter plot in Figure 5. This is a better test for the fleet mean EFs calculated using this method. However, during the nighttime drive-by tests where the distance measurements were conducted with less traffic, plumes from individual vehicles passing by were easily identifiable at both 3 m and 15 m. These cases were used for the EFs of individual plumes and included in Figure 5 as a scatter plot. A description has also been added to lines 291 – 295:

"Individual plumes were also matched based on specific vehicles observed at the site, where the same test was conducted at night during low traffic periods where the plume travelling from the roadway could be directly associated with the plume that reach near-road building inlet. In this case, the average ratio of ten different plumes were 1.012 ± 0.253% ($R^2 = 0.94$)."

Lines 264-278: The authors select a subset of 152 plumes that had "elevated pollutant concentrations" to examine the effect of individual vehicle types (e.g. truck vs. passenger car). The selection criteria used is vague and seems to be somewhat arbitrary. For example, was an individual plume selected only if concentrations of all species were elevated? Given that this analysis only involves ~0.1% of the authors data set of captured plumes, I feel that the selection criteria should be more objectively defined and explained more clearly in the manuscript.

The purpose of this approach was to manually choose a subset across the campaigns based on a typical user's visual evaluation of the data. This approach will be inherently subjective, and highlights the subjectivity of a user manually picking plumes by visual means versus an automated method that would attempt to capture as many plumes as possible, irrespective of peak magnitudes. Although the data from this manual analysis was used to identify cars vs. trucks as a means of validating a select subset of plumes for manual versus automated calculation, it was not designed to be representative of the entire vehicle fleet. Thus we note in the text that these plumes are likely more representative of the higher emitting vehicles in the fleet.

Line 306: It's not clear why the CO emission factor reported here is more representative of LDVs than CO EFs reported in cited studies. This wording implies other studies exclude low-

emitting vehicles from fleet-average EF calculations, which is not the case for tunnel and remote sensing studies.

The sentence was meant to be a commentary on the method used in this study, where if low-emitting vehicles (i.e., zero CO EF value plumes) are not included, then our mean CO EF would be significantly higher. As the reviewer stated, this should not a problem for integrated measurements typical of tunnel studies. This is more a cautionary statement for anyone who may explore a similar individual plume analysis rather than an integrated approach. This section in lines 346 – 352 has been changed accordingly:

"Inclusion of those low-emitting vehicles resulted in a fleet mean CO EF within the range of those previously reported for LDVs (Dallmann et al., 2012;Wang et al., 2009;Park et al., 2011;Dallmann et al., 2013;Hu et al., 2012;Bishop et al., 2011). Plumes containing detectable CO were not as frequently observed relative to other pollutants like NO_x. If zero value BT EFs are removed from calculating the mean the resulting mean CO EF is 39.9 g kg-fuel^{-1}, which is within the upper range of LDV CO EFs previously reported."

Minor Comments:
Lines 24-25: Suggest changing BC and PN emission factor units from mg kg-1 and kg-1 to mg (kg fuel)-1 and (kg fuel)-1, respectively.

These have been changed respectively, into mg kg-fuel^{-1} and # 10^{14} kg-fuel^{-1}, and for per kg units in the text as kg-fuel^{-1}.

Line 36: This sentence seems to imply there were only 3.7 million deaths globally in 2012. I suspect this refers to mortality attributable to air pollution. I suggest rephrasing the sentence to make this connection clear for the reader.

Text has been changed to "On-road motor vehicles are one of the largest contributors to air pollution in urban environments (Franco et al., 2013) and are thus one of the major sources underlying the estimated 3.7 million air quality related deaths in 2012 worldwide (WHO, 2014)."

Figure 1d: Please include units for wind speed and the radial axis.

Wind speed and radial axis units added to Figure 1d.

Lines 143-147: How were diurnal profiles for LD and HD vehicle activity at the site obtained? Suggest mentioning methods in the text.

This information was included in the supporting information, but was moved into the main text in Section 3.1 lines 185 – 190 to describe the SmartSensor HD dual radar system used to count passing vehicles.

Lines 184-185: Please include units for BC conversion factor.

Units added to line 188 as m^2 g^{-1}.