We gratefully thank the reviewer for carefully reading and providing feedback to our manuscript. Below we provide our point-to-point responses to the reviewer's comments. The comments by the reviewer are marked in **black**, responses are marked in **red** and changes to the manuscript are indicated in **blue**.

Other changes and comments:

- We adapted formula A1 in the Appendix to reflect the calculations used in the TUG-PDA rather than the typical calculation used in the literature. We added the text: "*In our approach (see Eq. A1), we use different start (t_1, t_3) and stop times (t_2, t_4) for the pollutant and CO_2 integration. Similarly, the BG values are determined independently ($[P]_{tP0}$, $[CO2]_{tCO20}$).*"
- Minor adaptions were done in lines 332, 774-776 and we added acknowledgements.
- If the manuscript is accepted, we would suggest to the editor that the appendices that are not necessary for the understanding of the main part are moved to a supplementary file. This is in our opinion: Appendix B, E, G and H

Referee 1

This study developed and demonstrated a point sampling method to automatically measure emissions from a large-scale of individual vehicles. In this works, the authors present their system that can be used for particulate matter (PM) and gas emissions measurements, which is notably independent of vehicle type. They find that when using their peak detection algorithm (TUG-PDA), they can separate vehicle-specific emissions down to a spacing of just a few seconds between vehicles. In this study, they present initial findings from the use of this method that collected ~100,000 vehicle records from several measurement locations, mainly in urban areas. When compared to equivalent remote sensing measurements, the authors found good agreement even with the newest standards which are harder to capture due to their lower emissions and the current remote sensing abilities. This paper is well written and organized.
This manuscript presents novel work on the development of a plume detection system. The authors have done a lot of work to respond to and update their work based on the last round of revisions. With that being said, if the authors are able to update their work with the minor revisions listed in this report, this manuscript should be accepted for publication.

Line 19: Define NOx here instead of in line 21

Many thanks for this suggestion. NO_x is now introduced in line 19.

"Of specific interest are nitrogen oxide (NOx) …"

Line 47-48: "Other PM metrics such as PN or BC cannot be accurately determined using these systems…"
This point needs to be further clarified for the reader.

We thank the reviewer for this comment. Open-path RES systems only provide PM estimates (PM mass or opacity). They do not measure BC or PN. We revised the description accordingly.

We replaced "*Other PM metrics such as PN or BC cannot be accurately determined using these systems*" with "*Other PM metrics such as PN or BC are not measured by these systems as they only give PM estimates (Knoll et al., Under review).*"

Table 1 /Table 2: Because Table 2 is not referenced in the text and is just hanging in the section that is placed in, it's recommended to add to Table 1 or putting it in the Appendix and referring to it in the main text.

Table 2 is referenced in the text in line 203: "*The TUG-PDA searches around the vehicle pass time (default window: -1 s to 6 s) for a sequence of positive concentration gradients above a defined threshold (see Table 2)*". Both referees in the previous review asked about the thresholds used for the peak detection, which is also an important information for the reader in our opinion.

We moved Table 2 directly to 1c (line 207) where it is mentioned to make it clearer to the reader.

2.2.1 Pre-processing:
This section has a lot of technical information that does not fit well into the bulk of the manuscript. It's recommended to simplify to the following:
1. Raw data from the instruments are time aligned
2. The CO2 is the default time resolution (keep statement about if instruments have large response time differences)
3. Outliers are filtered (state metrics for this) and measurements are smoothed

We are grateful to the reviewer for the suggestion of a simpler and more compact presentation of the pre-processing steps.

2.2.1 Pre-processing has been revised to (line 175 – 182): "*Prior to the actual emissions calculations, three main steps are taken to prepare the raw instrument data.*

1. *Time series data from the different instruments are time-aligned based on manual pollution peaks taken during the measurement campaign (e.g. with a lighter). We align the concentration time series data to the vehicle passes which cause the fastest response (e.g., from vehicle with tailpipe on the same side as the sample extraction).*
2. *The time resolution of the CO2 and pollutant data is equated (default time resolution of 0.5 s) and the CO2 and pollutant data sets are combined into a composite data set.*
3. *The time series data are then smoothed with a rolling Gaussian filter (default window size 5 samples) to reduce the dependence on short variations and outliers. If instruments with large differences in response times ($\Delta t > 2$ s) are used, the response function of the instruments must be aligned.*"

Lines 194-201:
This section leading up to describing the algorithm do not add value to the methods section and ends up being more distracting. The audience should understand the concept of PS by this part of the paper and therefore, there is no need to provide these detailed ideas. It's recommended to cut or slim to one introduction statement on why you

developed the algorithm.

We thank the reviewer for this suggestion. From importance is the statement that the CO_2 and pollutant emissions are separately processed because this is one of the main differences from previous PS studies.

We shortened the introduction to (line 184): "*We have developed a dedicated algorithm, TUG-PDA, which separates the measured emissions and assigns them to the by-passing vehicles.*"

Line 211: "There must be either at least two gradients"
Which two gradients are being referred to? As in two pollutants need to be rising and having a gradient? Please clarify in the text.

We wrote (line 208-214): "The TUG-PDA searches around the vehicle pass time (default window: -1 s to 6 s, highlighted in Fig. 4) for a sequence of positive concentration gradients above a defined threshold (see Table 2) of the processed analyte (visualized in Fig. 4). The thresholds were determined based on a large number of manual reviews of TUG-PDA results. There must be either at least two gradients or one very large gradient (> 10 times the threshold) above the threshold."
The thresholds are defined in Table 2.

We shortened and clarified the description and added "defined thresholds" to make it easier for the reader to understand (line 202-205): "*The TUG-PDA searches (default window: -1 s to 6 s) for a sequence of data points with positive concentration gradients above a defined threshold (see Table 2) of the processed analyte around the vehicle pass time (visualized in Fig. 4). There must be either at least two data points of the analyte with a gradient above the threshold or one data point with a very large gradient (> 10 times the threshold).*"

Figure 3:
This figure can go into the appendix. It is too much to take in and understand in the main text. Also, it's recommended to add step numbers to make it easier to follow the diagram.

We thank the reviewer for this comment.

We added step numbers in the figure and also in the descriptive text. We also restructured the descriptive text (added 1b, 1c and 4) to make it easier to follow and relate it to the figure. We have also clarified in the caption and text that certain processing steps are only carried out for pollutant or CO_2 emissions: "*Specific processing steps are only applied to CO_2 (3b) or to pollutants (1b, 1e, Stop 4, Stop 5, 3c, 3d, 3.1).*"
As suggested, we moved the detailed flow chart to the appendix.
We drew an overview version of the figure, which we included in the main text, with only the main processing steps to make it easier for the reader to understand (see figure below).

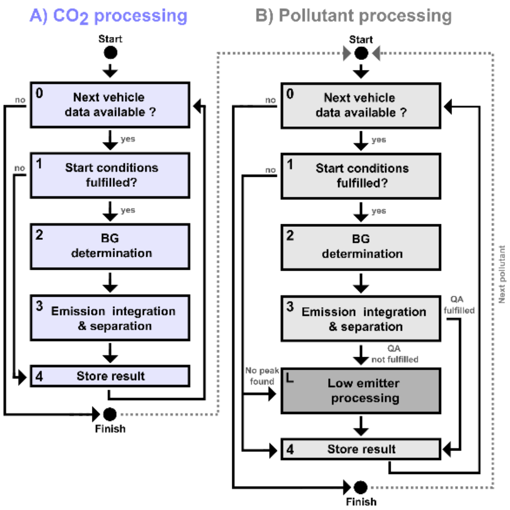

Figure 3. Emission event processing - flow charts of the peak detection algorithm (TUG-PDA). CO2 and pollutant (e.g., BC, PN, NOx) emissions are processed separately. The algorithm is applied first to CO2 (left) and then to the individual pollutant emissions (right). A detailed flow chart can be found in the Appendix (Fig. C1).

Figure 4:
It would be helpful to have the axes go lower than background so the reader can see how the background level is determined and potentially used as a stop (while recognizing that this specific example is not due to background but because of a passing vehicle)

We thank the reviewer for this suggestion.

We adjusted Figure 4 so that the signal is clearly visible down to the background. The background is also subtracted from the areas shown, similar to Figure 5.

Line 269:
Another recent paper on plume detection was published and should be referenced here. https://www.sciencedirect.com/science/article/pii/S1364815222003000

We thank the reviewer for providing this interesting paper.

We referred to it in line 262: "*An open source mobile air quality dashboard, including a real-time peak detection algorithm was published by Kelly et al. (2023).*"

Section 2.2.3:
This section can be revised to be more direct about exactly what this method does. Also, generally, the methods section still seems quite lengthy. It is recommended to highlight the specific and unique points that applies to this PDA system.

We thank the reviewer for this suggestion. We described our methods in detail such that a reader can understand and duplicate our work which was also suggested by reviewer 2 (major revision): "But if the algorithm will not enter the public domain, then I think the manuscript requires major revision to be a more complete methods paper that could be independently duplicated by others."

We revised section 2.2. Data analysis (as suggested by the reviewer, see other comments) to be more understandable and more direct about our method.
We revised section 2.2.3 to be more direct about our method (line 271 to 280): "*Once the ERs of passing vehicles have been determined the measurement results are combined with the vehicle's technical data. Several details from the vehicle technical data are required during the emission analysis to calculate EFs and to perform further statistical analysis. Necessary fields for our post-processing are:*

- *The fuel type (e.g., gasoline, diesel) to calculate fuel-based EFs.*
- *The CO_2 emissions measured during the type-approval process of the vehicle model are required to calculate the distance-related EFs.*
- *The European emission standard class is used to classify vehicles according to their emission limits.*
- *The vehicle category is used to perform detailed evaluations for specific vehicle types.*

With the help of our local partners, we obtained the necessary technical data from the government authorities. The captured license plates are pseudo-anonymized to respect privacy rules."

Lines 306-311:
These sentences make the TUG-PDA sounds like it hypothetically can get down to a detection of 3 s but it is not able to in many cases, which contrasts a lot of the high level take aways from this paper. It is suggested to reword in the more active and present tone to express exactly what the system can do and what factors are able to be adjusted for varying situations.

We thank the reviewer for this suggestion.

We rephrased the text to be specific what the TUG-PDA is capable and what not (line 299-307): "*The TUG-PDA resolves emissions down to a small distance (default: 3 s) between vehicles, if the time between the vehicles is large enough (greater than 3 s) and if a dedicated CO_2 peak from the vehicle is observed. Several tests are implemented to determine whether the emissions really come from the current vehicle or are caused by interference from previous vehicles or another source. If other influences are observed, the distance between the vehicles is too small, or overlapping plumes cannot be separated, the measurement is invalid and the emissions for the vehicle cannot be determined. Plume separation can be tuned using several parameters such as gradient thresholds (Table 2), the minimum time allowed between vehicles or the minimum number of samples required as used in the software. This can be very useful for instruments with different response times and for locations with dense traffic to obtain a sufficient number of measurements. Restricting measuring to low-traffic areas would severely limit the application.*"

Figure 5a:
For V4, the BC integrated areas seems to be cut off early. Though it may be following the rules, the BC plume both starts later and ends later than the corresponding CO2 plume and therefore, should be integrated to basically be a time shifted version of the CO2 plume instead of being cut off which would lead to an underestimation of the EF (this is discussed later on when the author states "The median EFs were 19% lower than in cases without interference"

Because of this, I believe it is very important for the authors to directly respond to the previous comment from Referee 2: Have you characterized how different fractions of peaks captured and assumed baseline concentrations impact resulting emission factors, using the subset of peak events that were 100% isolated with all pollutants starting and ending at the background condition? Authors: We thank the reviewer for this suggestion. We have not evaluated that in detail, but that is a very interesting suggestion for future investigations.

The authors have done some interesting analysis already that answers some of the referee's questions but they should go a step further and address how this ultimately would effect a TUG-PDA users emissions output.

We thank the reviewer for this comment and for pointing this out. For V4, the algorithm underestimates the BC area because the BC background concentration is determined too high. We fully agree that with visual inspection you can see that the BC area is underestimated. It is not so easy to have an automated plume separation that can handle all possible cases of what the emissions look like and accurately calculate the EF. There are some trade-offs or inaccuracies and these will be addressed in the future. We also addressed that in the manuscript between line X1 and X2: *"In the current implementation of the TUG-PDA, the BG determination for overlapping plumes is done by calculating an average value between the median concentration directly between the overlapping plumes and a common BG when no vehicle is passing. This is a simple estimation and entails deviations from the actual situation. This can be seen, for example, in Fig. 5a) for vehicle V4. The BC background is overestimated. This results in a too small integrated area (BC4) and thus underestimated emissions."*

As suggested, we characterized how different fractions of the plume influence the resulting EFs. We added in section 3.1 (line 350 to 354): "*We also looked at how accurate the EF can be calculated using only a fraction of the plume. Therefore, we selected only plumes without interference from other vehicles and calculated EFs using the TUG PDA when the algorithm used only a fraction of the plume in the interval between 3 s and 23 s. Similarly to the investigation shown in Fig. 6, we found that when only a fraction of the plume is used that the EFs are underestimated. The median underestimation for an early cut-off at 3 s is 27 %. The deviation decreases with increasing fraction of the plume (see Appendix Fig. D2).*" We added in Appendix D (line 695 to 699): "*Figure D2 shows how accurate EFs can be calculated when using only a fraction of the plume. Therefore, the algorithm selected 82 plumes that were not affected by emissions from other vehicles. The average plume length of this selection was 18 s and 30 of the plumes were longer than 25 s. The full distribution using the algorithm's defined maximum plume length of 25 s is shown in Figure D2 on the left. Figure D2 (right) shows the deviation from the full plume when only a fraction between 3 s and 23 s is used. The median deviation is maximum at 3 s with 27 % and decreases steadily with increasing plume fraction.*"

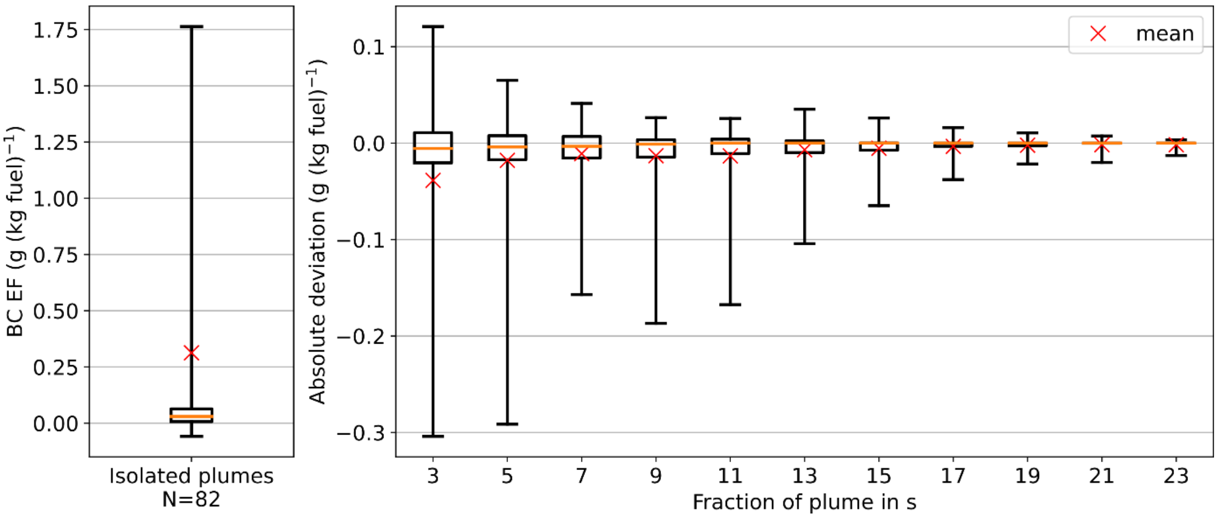

Figure D2. Deviation when using only a fraction of the plume to calculate EFs compared to using the entire plume. Left: Distribution of EFs of plumes without interference from other vehicles. Right: Deviation from full plume (25 s) using only fractions between 3 s and 23 s.

You can also see that the influence on the resulting EFs is relatively small (not included in the manuscript as Figure D2 should contain all this information):

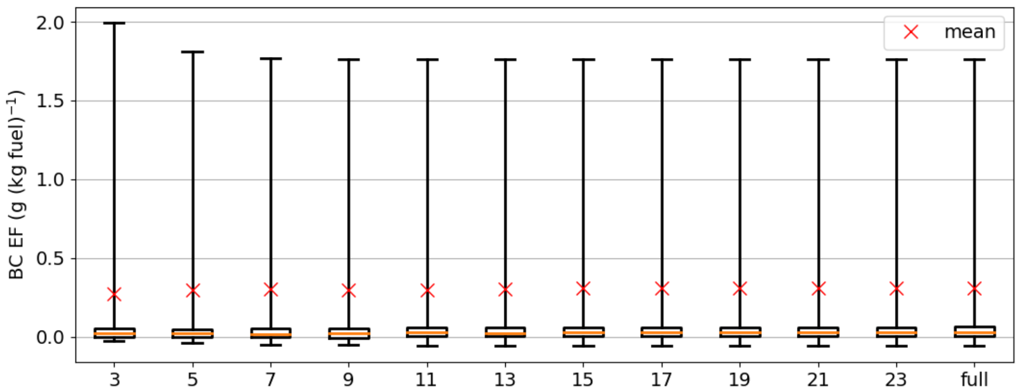

Figure 6:
Define what the whiskers (Confidence intervals? What percentage?) are in this box and whisker plot

We thank the reviewer for pointing this out.

We added in the caption of Figure 6: "*The whiskers represent the 2.5 and the 97.5 percentiles.*"

Line 366:
These laboratory measurements will need to be further defined especially with respect to determining limits of detection for BC as this is something that is not typically done.

We thank the reviewer for pointing this out.

We added the following information (line 367): "*We characterized the BCT and the AE33 in the laboratory for properties relevant for PS (see Table 1). A miniCAST soot generator (Jing*

Ltd, Model 6204 Type B) was used as the particle source. The instruments measured in parallel downstream of a catalytic stripper which removed volatile compounds (Knoll et al., 2021)."

Line 378:
This sentence and the paragraph need to be clearer about what differences between the two instruments and their performance are. This sentence I believe is applied to only BC emissions from the AE33 vs from the BCT. The BCT measurements can be separated for the two passing vehicles while the. Measures from the AE33 are not able to be seperated. Is that correct?

We thank the reviewer for this suggestion. Yes this sentence is only applied to the comparison between AE33 and BCT. It is correct that for the shown example the emissions can be separated using the BCT but not using the AE33 because of the slower response time (BCT: 0.9 s, AE33 7 s).

We added at the beginning of the section that the comparison is about BC (line 364): "*For our study we selected two instruments, the custom-designed BCT and the Aethalometer AE33 (Magee Scientific), for their applicability in determining BC emissions using the developed TUG-PDA.*"
We clarified what the differences are in the mentioned paragraph (line 380 to 385): "*The emissions captured for the two vehicles overlap, but they can be separated using the BCT. In contrast, the AE33 response time is much slower and the maximum concentration is reached after the second vehicle (V2) has passed by. In this case it is not possible to separate the BC emissions of the two vehicles using the AE33. This example illustrates the importance of choosing instruments with a fast response time when measuring in dense traffic. Individual characteristics (see Table 1), such as the response time, that do not meet the requirements severely limit the application.*"

Section 3.2.1.
The focus of this section is specifically to compare two BC instruments while the bigger scope of this work is to developing the PDA. I think that this section should be remove or added to the appendix or framed as a case study to emphasize the importance of understand the instruments used with the PDA.

We thank the reviewer for this comment. In "3.2 Factors influencing point sampling measurements" we evaluate different influences on PS such as instruments characteristics, measurement location, sampling position and meteorological influences. As the selection of instruments with appropriate characteristics is of great importance in PS, we prefer to leave this section in the main text. If the editor prefers we will move it to the Appendix.

We added at the beginning of Section 3.2.1 that this is about a case study to evaluate instruments for their applicability using the TUG-PDA (line 364): "*For our study we selected two instruments, the custom-designed BCT and the Aethalometer AE33 (Magee Scientific), for their applicability in determining BC emissions using the developed TUG-PDA.*"

Line 389:
Please edit to be clearer. What is VSP?

We introduced VSP in the method section (line X1) and we described it in Appendix G.

We added the reference to the appendix (line 393): "*(VSP, see Appendix G).*"

Figure 8.
What is roadside?
Also, the fits do not have any statistical information and therefore do not add any meaning takeaways. Edit figure to clearly show the message the authors wish to convey.
Is there a strong statement that supports the placement to be in the middle in order to captures the higher levels of CO2?

The wording "roadside" (roadside measurements) is often used in the literature as synonym for "point sampling" when sampling from the side of the road. Reviewer 2 (major revision) wanted us to show a second trend line for roadside measurements.

Table 3:
Needs to be moved up in the text to where is it referred to. Do not leave it dangling at the end.

We thank the referee for this suggestion. In the layout of the final paper, we will make sure that the table does not hang at the end.