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**. The changes in the manuscript have not yet been completed and the revision of the manuscript is currently still ongoing.

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. Their findings include a detailed evaluation of the main influencing factors on point sampling measurements, specifically for carbon dioxide (CO2) and black carbon (BC) measurements. 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.

However, the novelty of such work needs to be explored further. While the authors specify that this point sampling method is novel and/or surpasses the ability of many other (cited) studies on roadside emission measurement, this reviewer questions that assumption with the only notable differences coming in the use on light duty vehicles and the automation. It needs to be clear how this work contributes to the scientific knowledge on vehicle emissions measurements.

We thank the reviewer for the comments and suggestions. We summarized below the novelty of this work as we think there is more than "only" light duty vehicles and the automation:

1) We presented a software framework and a peak detection algorithm for automatic emission post-processing that is capable of delivering emission factors for thousands of vehicles. Comparable software frameworks have not been published in literature yet.
2) Until yet, very limited particulate matter data of real driving emissions of light duty vehicles were published. In Europe, especially particle number emissions are nowadays of interest due to the latest introduced emission standards. In Europe, no study exists which show PN or BC emission trends of the general light duty vehicle fleet. We provide the capability to do that und we show first results.
3) We presented a measurement system which is capable of determining emissions of
    - different vehicle types (light duty vehicles, heavy duty vehicles, …);
    - low-emission vehicles that meet the latest European emission standards;
    - vehicles in rather dense traffic situations with a distance between the vehicles down to 3 s.

   Up to now, literature PS studies mainly measured HDVs. The used setups restricted measurements to rather large distances (> 7-10 s) between the vehicles because of the slow response time of the used instruments. This limited the application to certain types of vehicles or low traffic areas.

   In addition, the setup enables automatic post-processing of the emission data. Commonly in PS studies only a camera was used for vehicle identification. These are often not able to capture all vehicles in dense traffic because they are too slow. Automatic post-processing of the emissions of individual vehicles is not possible if the emissions cannot be attributed to an individual vehicle. We use light barriers for exact pass time determination in addition to an ANPR camera.

To make the software framework also for other institutions available, we decided to publish the software framework along with the submission of the revised manuscript.

We revised the description of the peak detection algorithm in section "2.2.2 Emission event processing" to provide more detailed and a better structured information.

In addition, we provide in the revised manuscript more information on the accuracy of the emission event detection by comparing emission distributions including overlapping plumes compared to a more "conservative" parameter setting of the peak detection algorithm that minimizes plume overlaps.

Further exploration of the thousands of measurements made could help to enhance the novelty of this work by highlighting new or potential trends in vehicular emission such as emission control technology deterioration as mentioned in the introduction of this study. Additional findings, revisions and additional review would need to be completed prior to acceptance and publication.

We thank the reviewer for this input. It is also in our interest that further explorations are carried out on the measured data. This paper focuses on the further development and exploration of the PS method and is not intended to discuss emission trends in detail which we plan to publish in another work in the near future including high emitter identification. We extended in the revised version the methodological section by providing detailed information on the peak detection algorithm including plume separation and background determination. In addition, we publish the software framework so that also others may use it. In this manuscript, we exemplarily show the performance of our system by presenting emission trends for different vehicle types (LDV + HDV) even to the latest Euro emission standards for NO_x, BC and PN. We also compared our results with several literature studies. For BC and PN, hardly any literature exists which shows todays real world emissions of light duty vehicles captured by remote emission sensing. We show for the first time remote emission sensing results concerning BC and PN emissions of the European light duty vehicle fleet.

Additional comments:

Introduction, L 36.

"…by making wrong measurements."

Can this be further explained or cited?

During our measurement campaigns in Europe we had a lot of discussions with different institutions in different countries. We have been told, that in certain countries it is common that during PTI "wrong measurements" – manipulated measurements are performed e.g. the PTI exhaust measurement is performed with another vehicle located next to the vehicle which is tested. This bypasses the PTI measurement of the inspected vehicle.

We adjusted in the revised manuscript: "…by manipulating measurements."

Methods, 2.1 Measurement Setup, L93.

“Using light barriers restricts the measurement location to single-lane roads or roads with islands between the lanes. Alternatively, vehicle detection can be performed with radar, video, or LiDAR systems.”

What does this limitation have on the type of vehicles able to be measured with this system?

Do the other sampling options for vehicle tracking listed have the same capabilities but better capture for more road types. If so, why were they not used? Please sure explain the impacts of this sampling method especially with regards to vehicle population and potential bias.

Light barriers limit the application to single lane roads or roads with islands between the lanes. With other vehicle detection technologies such as Radar or LiDAR our presented approach could be applied to multilane roads. With the simple sampling setup (sample extraction from the side or middle of the road) only the outer lanes could be monitored without structural changes to the road. The application to multilane roads (without islands) has not yet been tested.

There is in general no limitation for specific types of vehicles. The limitation in case of vehicles is the exhaust pipe position. In Europe, most vehicles (passenger cars, trucks, motorcycles) have the exhaust pipe at the bottom of the rear or on the side. The tailpipe points straight back or down. A significant share of the motorcycles have the tailpipe pointing upwards which impedes the measurement (at low heights). This is also described in “3.4.1 Fleet composition and capture rate”. We showed in the application example (Fig. 12-15) that we can measure the most common vehicle types.

Methods, 2.1 Measurement Setup, L110.

“In general, the closer the sample inlet is to the emission source (tailpipe) the smaller the dilution and the higher the capture rate are.”

Also discussed in 3.2.1 Sampling Position and eventually mentioned on L463.

Though true, in the schematic, diagram, and later sections, the sampling inlet is located near the ground, what was the capture rate for vertically oriented tailpipes? How did that influence your sample population? Is there potential for this method to be adjusted to capture all tailpipe orientations?

In Europe, there are practically no vehicles with vertical exhaust pipes, with the exception of non-road mobile machinery. Therefore, we have no statistics for such LDVs or HVDs. We see with our sampling setup a lower capture rate for motorcycles whose exhaust pipes are often sloped upward (and which have a lower exhaust flow rate → described in “3.4.1 Fleet composition and capture rate”). With our sampling approach at the roadside or center of the road, vehicles with vertical tailpipes are not captured. If vehicles with vertical tailpipe should be measured, the sample extraction should be done from the top (e.g. from bridges), as shown in several publications referred to in our manuscript.

Methods, 2.2.2 Emission event processing, L187.

"At the same time, care must be taken to ensure that the CO2 plume detected of the passing vehicle is related to the pollutant emission detected. Therefore, checks are implemented which compare the duration of the integrated CO2 and pollutant data and verify if the areas overlap appropriately."

Generally, the procedure of peak identification and peak alignment with passing vehicles needs to be further clarified. Can you explain more on how you know that CO2 has returned to baseline to meet the conditions outlined? Does the end of the pollutant peak only rely on another vehicle pass being detected? If so, what does it mean with regards to truly capturing the extent of a CO2 peak? Other works cited have specific quantitative assumptions for the rise above baseline for pollutants and CO2 as well as for the return to baseline after a vehicle passes, can more quantitative information like this be provided?

We don't necessarily wait until the CO_2 concentration has returned to the baseline (background). This is only one of the conditions defined in the algorithm which stops the plume integration. There are three mainly three criteria defined which stop the integration of a plume:

- The concentration level is below the determined BG concentration
- Another vehicle passed and the concentration gradient is again rising
- The maximum defined plume duration is reached. 25 s are used in this study.

In addition, two further criteria are defined which cross-check that the areas of CO_2 and pollutant agree (difference between the integrated areas and the stop time of the integration are not allowed to exceed defined values).

The following gradients are set as default values in the software environment and were used in this study:

- CO_2: 6 ppm / s
- BC: 4 (µg /m³) / s
- PN: 4000 (#/cm³) / s
- NO_x: 12 ppb /s

We have rewritten the section "2.2.2 Emission event processing" by providing more detailed and better structured information on the plume detection and separation. Several quality assurance (QA) measures are applied in the software which are described in the revised manuscript:

QA measures applied during the start condition of the peak detection algorithm:

"

- ***Vehicle distance:*** *First, when a new vehicle pass is fetched, it is checked whether the distance to the next vehicle pass is sufficient (≥ 3 s). If this is not the case, the processing for the current vehicle is stopped and the algorithm proceeds to the next vehicle. At this small spacing, there is a large uncertainty that emissions will be attributed to the wrong vehicle due to differences in the sampling delay between vehicles.*
- ***Interference:*** *The detected gradient (plume) must not be from a previous vehicle. The processing is skipped if a rising gradient (start condition) from the previous vehicle in a pre-defined time frame (default: 5 s) is found and the plume was not processed yet.*
- ***Pollutant vs CO_2 start time:*** *The pollutant plume must start in a pre-defined window compared to the CO_2 plume (default: -1 to 8 s).*

"

QA measures applied after the integration:

“

- ***Duration:*** *The duration of the integrated plume must have a minimum length (default: 3 s).*
- ***Plume strength:*** *The CO_2 area must be greater than a defined minimum concentration (default: 80 ppm s).*
- ***CO_2 vs pollutant:*** *The CO_2 and pollutant areas must overlap by at least 50 percent.*
- ***CO_2 vs pollutant:*** *The CO_2 area must not be longer than a pre-defined factor compared to the pollutant area (default: pollutant area / CO2 area > 0.6)*

“

We also replaced Figure 3 with a precise flow chart of the peak detection algorithm. We have added Table 2, which gives the default values for emission gradient thresholds. We will also provide recommendations for parameter tuning of the peak detection algorithm.

Conclusion, L559.

“The core of this software is the TUG-PDA, which determines and separates vehicle emissions down to a distance of 3 s between the vehicles, if appropriate instruments are used.”

Does this software have potential to be adapted fit the sampling behaviors of a range of instrumentation? The following bullet point is helpful to understand instrument requirements but if others were to try to adapt or replicate this work, is it fully instrument limited? Thinking back to the example provided comparing the AE33 and the BCK. How could one use the AE33 with this method/software? If this is outside the scope of this work, that is fine but please acknowledge.

The software is not developed for specific instruments. We developed the software for modularity and extensibility, such that new instruments can easily be integrated. Each instrument is defined in a separate Python class. New instruments can be added by copying an existing instrument class and adapting the code for the new instrument. In general, any instrument that provides time series data can be used. Parameters such as “gradient thresholds” (for plume start detection), “minimum time to next vehicle” or the “minimum number of required samples for a valid measurement” must be tuned for instruments with different characteristics (sensitivity, response time).

We provided recommendations for instruments to get the best possible PS results (see Table 1). Instrument characteristics such as response time (a large response time requires a larger distance between the vehicles) and sensitivity (a low sensitivity makes it difficult to quantify emissions of vehicles with low emissions / new emission standards) have a great influence on the results. Parameters of the software framework can be adjusted for different instruments. The AE33 can be used with this software framework for locations with low traffic density or for sampling specific vehicles (e.g. trucks with vertical tailpipe when sampling from the top). The rather large response time (~ 7 s) of the AE33 restricts the application to traffic situations with a distance between the vehicles of more than 7 – 10 s.

We added in section “2.2 Data analysis”: “*The software has not been developed for specific instruments and in general any measurement device that provides continuous measurement data can be integrated. However, we strongly recommend to consider the recommendations in Table 1 when selecting instruments to achieve the best possible results.*”

We added in section “3.3.3 Instrument characteristics”: “*However, the Aethalometer can be used for PS, as has already been demonstrated in several studies. The traffic density must be low enough (distance between vehicles greater than 7-10 s) or if only certain types of vehicles (e.g. HDVs with vertical exhaust pipes) are measured, which naturally entail a greater distance between exhaust plumes.*”