We thank the reviewers for their thoughtful comments. Below we detail our responses by line and include the edits we have made to the manuscript.

**Reviewer: 1**

*This analysis introduces an ad-hoc technique for analyzing the spatial patterns of air quality data from mobile monitoring. The goal is to determine so-called anomalies since the high anomalies may be driving the worst exposure events. Overall, while more can be done in this research field since it is data-rich and complex, I feel it is a novel analysis and one worthy publication.   
  
Two main points I think need to be addressed in the intro and/or discussion. 1) How the mapped anomalies might impact short-term and long-term exposure. 2) Are the anomalies detected purely high anomalies or are low anomalies possible?*

We thank the reviewer for their comments.

Concerning the first overall point, we added more context to the discussion. In particular, we write the following in L 432-441:

“Results from this algorithm could be incorporated into health assessment frameworks. Clustered anomalies could be grouped into source categories to facilitate simple exposure estimates from different sources. Apportioning anomalies to nearby sources and determining their frequencies would be an interesting approach to determining whether some sources are more harmful to health than other sources. Census-tract weighted probabilities of an anomaly could be employed in random walk simulations of cumulative air pollution exposure, providing a different metric to evaluate related health effects.27 Future work would focus on addressing serial dependency inherent in detected anomalies to develop probability-based exposure estimates, as well as the general development of a framework that relates health outcomes to the frequencies of these detected anomalies.”

Additionally, we acknowledge that anomalously low concentrations can be detected in this framework, though quite infrequently. We write the following:

“We observed the algorithms to capture plume behavior; the DBSCAN algorithm labeled 848 multivariate realizations with all pollutants lower than their respective 5th quantiles as plumes, or 0.07% of the total number of labeled anomalies.”  
  
*L54+ While Messier et al. was a purely spatial model, it is incorrect to say that LUR involves temporally aggregated data – there is nothing in the method that precludes the use of space-time data. Agree on the causality of LUR – not meant as a causal framework on its own. Disagree on the availability of land use variable datasets. The datasets used in LUR can typically be calculated for any location. Common sources include satellite imagery or distance variables, which are always calculatable. The more pressing issue of LUR is the availability of appropriate or sufficiently space-time resolution data. See Qi et al. in which they try to address this issue by using Google Street View images as inputs into LUR   
  
Qi, Meng, and Steve Hankey. "Using street view imagery to predict street-level particulate air pollution." Environmental Science & Technology 55.4 (2021): 2695-2704.*

We acknowledge these corrections and change those lines to the following:

“However, LUR models require spatiotemporal databases of sufficient temporal and spatial resolution to train on. While recent efforts have illustrated creative methods of creating these land use databases,3 use of these models is still limited through the availability of these databases. There is a need for the development of methods that can identify source influences in large mobile monitoring data sets at high time resolution without being subject to the availability of land-use variable databases.”

*Equation 1 has a typo – an empty box is shown where the epsilon should be*

The equation has been corrected.

*L166 – what is estimating via “statistical heuristics”? Seems ambiguous   
  
Section 2.4. I think the algorithm is an important development and as such should be included in the main manuscript (currently in SI). I would suggest a pseudo-code style algorithm (see example here https://www.overleaf.com/learn/latex/Algorithms). This could actually help shorten the section while being clearer and more concise because equations 3, 4, and 5 could all be included in it.*

We have included written pseudocode (presented as Figure 1) and have rewritten sections of the methods to address these changes.

L183-L199: “The algorithm estimates and MinPts parameters for daily time series in the campaign based on the number of points in each time series and its dispersion and subsequently performs DBSCAN using these estimated parameters. We define the MinPts parameter to be the product of the total number of points in the daily time series, n, and a fractional value parameter, . We set fval to 0.03 using the external validation set and describe the specific procedure in Section 2.6. We do not consider values of greater than 0.5 due to rapidly increasing computational cost and poor performance at higher values. After calculating MinPts, we determine using a k-nearest-neighbor (knn) distance ordering procedure in which the value of k was set equal to MinPts and in which a point is the kth nearest neighbor to another point if the distance between the two points is the kth shortest distance among all points. We construct an ordered knn distance set and determine the mean and standard deviation of the first 30 ordered distances, then define as the first distance that is greater than the mean plus 3 times the standard deviation of the subset of previously ordered distances. We iterate through the entire set of remaining distances, adding the current distance to the subset if it does not meet the criteria used to define . Once both and MinPts are determined, we run DBSCAN on the daily time series observations in which core points are labeled as normal and both border and noise points are labeled as anomalies.” *Figures 6 and 7. While North is assumed to be up, it is often helpful and common practice to include a north arrow. Additionally, it would be useful to have a scale bar and inset so as to know where Houston is in relation to other areas of Texas and the United States.*

Figures 6 and 7 have been updated to include a north arrow and inset.

**Reviewer: 2**

*The study describes application of an unsupervised machine learning technique, DBSCAN, to mobile air quality monitoring data. Developing and characterizing new techniques for better parsing mobile monitoring data is an important goal, and of interest to many in the ES&T community. However, I found the paper was narrowly focused and not written in a style appropriate for the broad audience of ES&T. While machine learning techniques are important tools for this community, what are the overall science findings of the paper? I appreciate that different techniques for parsing the data are presented and compared, it is not totally clear that this more complicated approach truly exceeds the QOR approach described when it comes to scientific conclusions from the data. The authors describe clustering of high CO2 data, and of high BC/NOx/PM data, but do not explore the reasons why. Sounds like it might be related to fraction of heavy duty vehicles? More literature citations and exploration of this topic is warranted in my opinion. Little attention is paid to the details of the environmental dataset on which the analysis is based. How does this analysis differ from previous analyses of this dataset? Are any new insights possible, or does this technique just speed up analyses? What seems to be the main science finding, the comparison of different census tracts, almost seems like an afterthought in the current version of the manuscript.   
  
Too many important details are relegated to the supplement. Indeed, there are some interesting analyses there that are not described in the main body of the paper. In addition the manuscript assumes the reader is intimately familiar with machine learning and statistical techniques, as well as with the underlying dataset, which is hardly described in the manuscript. The writing is bogged down with references to figures and tables. This work has potential to be an important contribution, but the manuscript needs significant changes to achieve that goal. I appreciate the authors’ efforts to share the results of their work, but I would imagine this contribution either better suited to a specialty journal or with much more emphasis on science questions and findings if their target remains publication in ES&T.*

We thank the reviewer for their comments and have rewritten the manuscript to be more accessible to the readers of ES&T. We implement changes that focus on the implications of the machine learning work we’ve presented here, and why we think those implications would be of interest to the broader ES&T community.

In particular, we include the following lines.

L90-93: “The objective of this work is to establish a new method for detecting plumes in mobile monitoring time series, validate its performance, and use it to perform novel analysis that elucidates the impacts of different emission sources across census tracts in the Greater Houston area.”

L102-109: “We apply the algorithm to the Houston Mobile monitoring dataset to identify anomalies, which are then clustered into anomaly types linked to specific vehicle emission sources. We tabulate the number of these different anomaly types by census tract and derive anomaly frequencies, which are conceptualized as the probability of detecting a given anomaly type during the prescribed study period. In particular, we demonstrate differences in anomaly frequencies in census tracts across Houston, which can be used to tailor census-tract specific air monitoring regulation and enforcement strategies. We discuss the implications of the method, the results, and future directions for this research.”

L271-272: “This probability represents the chance of detection of a given anomaly during the campaign study period.”

L322-329: “Table 1, Figure 4, and Figure S5 contain descriptive statistics describing the contents of each cluster. The results are consistent with previously published emissions patterns associated with light and heavy-duty vehicles. Heavy-duty, diesel-powered vehicles emit more BC, NOx, and UFP per kilogram of fuel than light-duty vehicles.18–21 Additionally, loadings from the PCA biplot in Figure S5 when varimax rotated are consistent in split with those reported in Larson et al.; loadings are sequestered into BC/UFP-rich and CO2-rich factors which are attributed to light and heavy-duty vehicle activity. These loadings are given in Table S7.”

L346-347: “Results from these boxplots confirm that our clusters are linked to emissions from these different vehicle types.”

L396-400: “These disparities, and the presented evidence suggesting that the BC/UFP anomalies are closely related to heavy-duty vehicles, are consistent with previous modeling studies that show large contributions of heavy-duty vehicles to air pollution in Houston’s Ship Channel (HSC) neighborhoods and previous work pointing out elevated heavy-duty vehicle activity in the Ship Channel area.23,24”

L412-415: “We discuss the successful development of a new approach to detect plumes in mobile monitoring time series using an anomaly detection algorithm based on DBSCAN and use the resulting analysis to derive anomaly frequencies representative of different emission impacts in different Houston neighborhoods.”

L424-443: “While it is not definitive that this cluster type represents impacts from heavy-duty vehicles, for there is no observational evidence to connect those observations to those vehicle types directly, anomaly emission patterns are consistent with previously published studies analyzing emissions from light and heavy-duty vehicles (e.g. Larson et al.1 and references therein). Previous studies have also shown the large impacts of trucking on pollution in the HSC area and have raised environmental justice concerns with the burden of pollution from diesel-powered vehicle activity.23,24 Results from this work emphasize the need for additional investigation into the trucking activity in HSC neighborhoods, and, more broadly, illustrate how mapped spatial distributions of these anomalies can be used to inform regulatory activities.

Results from this algorithm could be incorporated into health assessment frameworks. Clustered anomalies could be grouped into source categories to facilitate simple exposure estimates from different sources. Apportioning anomalies to nearby sources and determining their frequencies would be an interesting approach to determining whether some sources are more harmful to health than other sources. Census-tract weighted probabilities of an anomaly could be employed in random walk simulations of cumulative air pollution exposure, providing a different metric to evaluate related health effects.27 Future work would focus on addressing serial dependency inherent in detected anomalies to develop probability-based exposure estimates, as well as the general development of a framework that relates health outcomes to the frequencies of these detected anomalies.”

*Detailed comments   
TOC art: TOC art should be understood by a casual reader. This image is only meaningful to a reader who has read the entire manuscript.*

The TOC art has been updated to be more accessible. *Lines 62-63: “dimension reduction techniques to describe patterns in the data would likely return results that are weighted towards data in which source influences are largely negligible.” What do you mean by source influences? I don’t understand this sentence as currently written.*

We have rewritten this to the following.

L55-62: “If a mobile monitoring campaign were conducted focusing largely on residential areas with brief excursions into traffic congested areas, such as highways, performing PCA or other dimension reduction techniques to describe patterns in the entire dataset would likely return results that are weighted towards residential areas with negligible source influences. This type of analysis generates solutions in which there is a demarcation between a majority of points with little source influence and a smaller subset of source-influenced points elevated in all pollutants, which is not compelling if one’s objective is to determine the specific sources affecting the measurements.” *Line 118: Is there a compelling reason why the time series data was given at 1 s time resolution when it meant interpolating ~90% of the NOx observations?*

We have written this portion of the text to the following.

L128-136: “In the original data set, NO and NO2 were taken on a 5-s time resolution, while CO2, BC, and UFP were all at 1-s resolution. To perform analysis at a finer temporal resolution, as well as to address missing data, we use monotone Hermitian splines to impute missing measurements up to a 6-s time gap. While previous mobile monitoring studies have fused 5-s data with 1-s data by repeating the same 5-s measurement each second across the entire interval,8,10 we argue that using splines provides a better estimate of missing 1-s information in this context. Previous studies have focused on preserving the spatial meaning of concentration plotted on maps at very fine spatial intervals; here, we are more interested in estimating temporal variations in missing concentrations, and splines are suitable tools to do so for brief, 6-s intervals.” *Line 130: What is the TigerLINE census? Would be helpful to describe what this is without the reader having to go look up the reference.*

We have added the following.

L141-142: “Using road shapefiles available through the TigerLINE road database,11 we assign road categories to each of our points based on their respective latitude and longitude coordinates.”

*Line 133: I assume the self-sampling contamination problem exists because the sampling vehicle is itself an emission source. However, this isn’t something I would know from the sparse information about the sampling campaign previously provided.*

We have rewritten this methods section to be more explicit about the data used in this analysis. The section has been rewritten as follows.

L113-150: “Data were collected during the Houston mobile monitoring campaign and are described in detail elsewhere.8,9 The campaign’s objective was to measure air pollution on a very fine spatial scale in 35 different census tracts across the Greater Houston area in a 9-month timespan. Two Google Street View cars were driven through these census tracts systematically to evaluate spatial differences in the concentrations of 7 pollutants. In this analysis, we restrict the set of analyzed pollutants to be BC, CO2, UFP, and NOx. Here, we do not consider PM2.5 and O3 due to the influence of secondary processes. A table providing the instruments used to measure each respective pollutant is given in Table S1. BC, CO2, and UFP measurements were taken on 1-s time resolution, while NO and NO2 measurements were taken on 5-s time resolution. With the addition of logged GPS coordinates from each car, the campaign generated a massive spatiotemporal dataset spanning millions of observations across the 9-month span.

In this work, we create a multivariate dataset consisting of the aforementioned four air pollution variables at 1-s time resolution, along with corresponding latitude/longitude coordinates and timestamps that span 277 separate days of sampling for a total of 5,301,507 observations. The BC data were smoothed with a 10-s time window to limit the effects of noise on subsequent analysis. In the original data set, NO and NO2 were taken on a 5-s time resolution, while CO2, BC, and UFP were all at 1-s resolution. To perform analysis at a finer temporal resolution, as well as to address missing data, we use monotone Hermitian splines to impute missing measurements up to a 6-s time gap. While previous mobile monitoring studies have fused 5-s data with 1-s data by repeating the same 5-s measurement each second across the entire interval,8,10 we argue that using splines provides a better estimate of missing 1-s information in this context. Previous studies have focused on preserving the spatial meaning of concentration plotted on maps at very fine spatial intervals; here, we are more interested in estimating temporal variations in missing concentrations, and splines are suitable tools to do so for brief, 6-s intervals. Total imputed percentages for each pollutant were 1.06%, 80.0%, 80.0%, 0.42% and 0.49% for BC, NO, NO2, CO2, and UFP respectively; 90.1% of NOx realizations had at least one imputed measurement. Any multivariate realization with at least one missing observation in a variable not imputed was excluded otherwise. Days in which the cars operated had to possess a minimum of 600 measurements to be included in the analysis. Using road shapefiles available through the TigerLINE road database,11 we assign road categories to each of our points based on their respective latitude and longitude coordinates. To be consistent with Miller and Actkinson et al.,8,9 we restrict our analysis to points with logged latitude/longitude coordinates on primary, secondary, local, and private roads, as well as ramps and service drives because these are roads typically relevant to an individual’s exposure.11 To account for GPS error, we remove logged GPS coordinates whose nearest neighbor distance to a TigerLINE shapefile point is more than 30 m. Additionally, we observed evidence of the vehicles sampling their own exhaust when driving to and from dead ends in a previous analysis of the dataset.8 Because we do not want to characterize our own individual vehicle’s emissions, we remove points less than 30 m from a dead end in a road.”

*Line 136: A bit odd that credit wasn’t given to Ester et al. prior to this point, even though DBSCAN was described.*

We have included a citation earlier in the manuscript. *Line 144: the equation doesn’t fully show up in the PDF version I have*

We have fixed this equation.

*Line 159: would be helpful to see visual plume flagging as a supplementary figure*

We have replaced Figure S1 with a plume flagging example. *Line 176: if I understand correctly, you’re constraining the number of MinPts to be <3% (0.03) of the whole dataset?  If so, might be helpful to state this in plain language*

*Line 179: would be helpful to have a simpler explanation here, especially for readers that aren’t familiar with knn clustering. Points are put in order by the value of a variable? In figure 1, what is the unit for nearest neighbor?*

We have rewritten this methods section to be more approachable for readers. See rewritten L183-199 above. *Line 237: “available in base”? Base what?*

We have rewritten this line to the following:

L245-246: “We perform k-means clustering on the extracted anomalies in R using the kmeans function available in R’s base package.13” *Line 276: I don’t believe the average ES&T reader will know what confusion matrices are. Can you please give a brief explanation here? That will also help with comprehension of figure 3*

We have rewritten lines describing the confusion matrix to the following:

L288-296: “Because this baseline agreement is so high, we create confusion matrices to probe sources of agreement and disagreement between each algorithm’s predicted anomalies and the validation set labeled anomalies and display them in Figure 3. Confusion matrices compare how an algorithm categorizes points with the points’ true categories. In our work, confusion matrices tabulate the number of points that a given algorithm labels as normal or anomaly that are correspondingly labeled as normal or anomaly in the validation set. As an example from Figure 3, there are 397,035 points that the DBSCAN algorithm labeled “Normal” that were also “Normal” in the validation set.” *Lines 303-311: The writing of this paragraph can be much improved. Right now it just reads like a laundry list of analyses. Instead of telling what each figure/table shows, it would be much more helpful if results from the data were provided in typical sentences, with references to figures and tables to provide evidence for the assertions. Furthermore, the authors assume that the readers have a rich knowledge of their dataset that they cannot have unless they read all the other papers that describe the dataset.*

We have rewritten this paragraph. Please see lines L322-329 *Table 1: please keep the order of cluster descriptions the same as in the text. It appears to me that the “third cluster” referred to in the text (lines 302-303) corresponds to the “2nd cluster” in the table.*

We have rewritten to the following:

L319-323: “We cluster detected anomalies using R function kmeans, which consistently yields one cluster rich in CO2 concentrations (“CO2 Cluster”), another cluster that contains lower (but still higher than their non-anomaly counterparts) concentrations of all four pollutants for both QOR and DBSCAN derived anomalies (“Transition Cluster”), and a third cluster rich in BC/NOx/UFP (“BC/UFP Cluster”) concentrations.”

*Line 330: What are AADT counts? I note that it is defined later on (in figure 5 caption), but should be defined in the text*

We add a definition for AADT in the following lines of the text:

L257-259: “Extracted roadway variables from the shapefile included Annual Average Daily Traffic Counts (AADT), Truck Annual Average Daily Traffic Percentage (TRUCK\_AADT\_PCT), and the number of all trucks in AADT (AADT\_TRUCKS).”

*Table 2: Would be helpful to have some socioeconomic variables here as well and/or information about the relative proportion of sources in each tract, such as % of industrial land.*

We have added square mileage and the number of industrial facilities/square mile to the census tract table, in addition to updating it to be more relevant to the manuscript.