

Title: Elucidating multipollutant exposure across a complex metropolitan area by systematic deployment of a mobile laboratory

Ilan Levy, Cristian Mihele, Gang Lu, Julie Narayan, Nathan Hilker and Jeffrey R. Brook

We thank the reviewer for his/her valuable comments and suggestions. Below are our responses to each comment in blue text.

General comments:

This study presents some interesting data and a potentially useful analysis but, in my opinion, it is not publishable in its current form.

First of all the authors should take a clear decision whether they will focus this paper on the exposure issues only or whether they want to discuss it also in relation to applying the new instrument in future epidemiological research. I suggest focusing the paper on the exposure issues only.

We agree with this suggestion and clarify this intent in the revised paper. The paper's original objective is to explore how a mobile lab equipped with multiple instruments to measure a large number of pollutants can be used to gain new insights about spatial patterns of long term exposure within cities. For this purpose, however, there are two fundamental challenges of using a mobile lab. The first challenge, which is obvious and one reason mobile labs have not frequently been used for long-term exposure studies, is that a mobile lab cannot be measuring at more than one location at a time. The second is that only a limited number of days can be studied because it is costly and demanding to conduct mobile measurements of multiple pollutants with the demanding suite of instruments we employ for a relatively long time period. Thus, the main objectives of this paper are to evaluate an approach to address these challenges and then to explore features of spatial variations in multipollutant concentrations based upon the data acquired through this approach. We have modified the text in the introduction and elsewhere as needed to make the objectives much clearer to readers. Specifically, we hypothesized that a deployment strategy could be developed to reduce the impact of these challenges and implemented and evaluated this strategy in this paper. The strategy we proposed is to systematically repeat a driving route through a wide range of urban micro-environments, and critically to cover this route in multiple seasons and randomly varying the times that the lab visits all the different locations on the route, but limited to daytime periods. We previously conducted a simulation study using hourly monitoring data in multiple Canadian cities to assess the relationship between the number of visits and the error in estimating long term averages. This study suggested that our deployment strategy could work and thus we conducted the study presented in this paper to assess this issue.

So far, I didn't learn from this paper how to use the data in an epidemiological study. In my opinion the only way for using the data in epidemiological research is the estimation of the long term average for specific locations (such as annual averages) and using the estimates for further modelling, for example LUR modelling. However, the estimation of annual averages for this approach could be also done by satellite monitoring sites (as already done in the past) and there is only a limited necessity to change the methodology. Thus, the exchange of the satellite monitoring sites by CRUISER would be nice, but it is not really crucial.

The objective of this paper is not to demonstrate use of the data directly in assigning exposures in epidemiological studies. In addition to evaluating the hypothesis we highlight above another purpose of this paper is to introduce the dataset and how it was obtained and how well it represents long term averages laying the foundation for further studies. We have published one of these studies using our unprecedented multipollutant measurements to learn more about spatial correlations among pollutants and how NO₂, a commonly used indicator of spatial patterns in long-term exposure, relates to other pollutants. That knowledge helps inform interpretation of epidemiological studies relating long term intra-urban exposure patterns using selected indicator pollutants in univariate models. We are also currently developing Land-Use Regression models from our dataset and because of the number of pollutants measured at all the same locations we can develop, evaluate and may be able apply models for more pollutants than previously possible. We have modified the text in the introduction and elsewhere as needed to better insure that readers are not expecting that the spatial data we present are to be used directly in epidemiological studies.

Moreover, the estimation of the annual averages could be conducted by CRUISER only in near-road environments and not in urban background locations (where the study population may also live).

This is a potential concern with mobile lab deployment and we have already acknowledged this in the paper and point out why it is not an issue. Our driving route routinely took CRUISER to the middle of neighbourhoods where there was no other traffic around and thus, we are confident that among our population of points we have excellent representativeness of urban background locations. In Levy et al. (2013) we presented our observed distributions of NO₂, UFP, BC, OM and HOA average concentrations among measurement locations (road segments) and in Brook et al (2013) we present histograms of NO₂, BC and UFP. These clearly show that we have data covering from the lowest levels in Montreal (the cleanest neighbourhoods mostly experiencing levels similar to the regional, rural background) to the peak concentration areas (near the port and highways) and that the mode of the distribution is near the typical urban background reported from the city's monitoring network which includes sites located to be population-representative.

This problem doesn't exist for the satellite monitoring sites, which could be located at almost all relevant locations. Furthermore, I have severe doubts whether the described design really allows the estimation of annual averages as concluded in this study (see specific comments).

Satellite sites or saturation monitoring has been quite successful for producing exposure models (LUR), in fact a model exists for Montreal for NO₂. Attempting to characterize average concentrations at multiple points for subsequent LUR model development using a mobile lab represents a different approach and one we are exploring, not as an alternative, but as a complement. In terms of cost, satellite sites are generally much more attractive. However, our mobile lab has the advantage of being able to reliably measure many more pollutants than can be deployed simultaneously with the satellite site or saturation monitoring approach. For example, while this approach has been able to operate at a relatively large number of sites for passively measured pollutants (e.g., NO₂, NO_x, VOC) they have not been able to cover as many sites with active samplers for PM_{2.5}, for example. As a result, those studies have had to make temporal adjustments to harmonize the sets of measurements taken in different time windows and that step adds a variable amount of uncertainty which is likely one of the reasons that ESCAPE PM models, for example, have considerable variations in R² among cities (Eeftens et al., 2012). While our mobile lab approach presents different challenges, such as discussed above, one unique aspect is that we are able to consider the complex mixture thus asking more in depth questions about what the limited number of pollutants considered in the satellite site approach might represent. In addition, preliminary results indicate that these data are promising for LUR model development. We have modified the text in the introduction to try to make these different strengths and weaknesses clearer to readers.

Testing whether our design can estimate annual averages is one of the primary objectives of the paper. We have made this clearer in the introduction in the revised paper. We prefer to present the data and allow readers to judge whether the design is successful based upon the facts provided.

With respect to short-term epidemiological studies, I don't see any possibility for application of the data in such studies. If the authors really want to postulate the using of CRUISER in epidemiological studies, a clear description of how to use the data in which studies is needed. In this case, also a deeper discussion of the current stage of exposure assessment in epidemiological research is needed.

Please see our response to the comment above. Indeed the data are not likely to be directly useful for an epidemiological study. The richness of the multipollutant data we have obtained can help inform epidemiological studies based upon intra-urban variations and, as with the satellite or saturation monitoring data, the CRUISER data can potentially be used to develop exposure models.

We have expanded the Introduction to give some more information about the different epidemiological study designs, but mainly just to clarify the points we have raised above and the context with which readers should interpret our findings.

Going too deeply would add length and the reviewer had also suggested that this paper is too long.

The problem of air pollutant variability between and within a city is well known in the epidemiology and it was evaluated in many studies (Jerret et al., 2005, Marshall et al., 2008, Brauer 2010, Boogaard et al., 2011, Cyrus et al., 2012, Eeftens et al., 2012). While the small scale variability is well characterized for some pollutants (especially for PM₁₀ or PM_{2.5}), it is not for ultrafine particles. This is the reason that no long-term studies on UFP and health were conducted until now. It shows clearly that the epidemiologists are aware about the necessity of sufficient characterization of large and small scale temporal and spatial variability for all air pollutants under study. However, some sentences in the manuscript suggest rather the opposite: "Nearby microenvironments may have a wide range in average pollution levels varying by up to 300 %, which may cause large misclassification errors in estimating chronic exposures in epidemiological studies". Without any further evaluation and discussion such sentences are misleading and should be deleted.

The reviewer is correct and there is no argument that epidemiologists are aware of the spatial and temporal variability in pollution levels and that large contrasts were already shown in other studies. One of the messages that the manuscript is trying to convey is that existing monitoring networks and even saturation campaigns (i.e., use of tens or even over a hundred passive samplers for developing Land Use Regression models) in a complex urban region cannot capture the full complexity that exists in pollution levels within a large city with multiple emission sources. Even the use of Land Use Regression models at a spatial resolution of 5 meters only characterize exposure levels uncovered from the initial saturation campaigns based upon the sites selected and then utilize the correlations found between those data and the available predictors (i.e., available GIS data for the specific city) to approximate the true exposure variability.

We feel it is instructive to point out the magnitude of the concentration differences we have observed and certainly do not dwell on it in the paper. The existence of such differences is obviously not surprising, but pointing some of them out does help exemplify what is seen with our deployment approach of CRUISER and the advantage of using a mobile campaign over saturation measurements. However, we agree that is not necessary or appropriate to imply that this translates into exposure errors in epidemiological studies and so have deleted such statements.

The manuscript is very long and it is difficult for the reader to catch the main messages. The whole manuscript should be definitely shortened. Some parts of the results section should be moved to the method section (see specific comments).

We have done our best to streamline the manuscript and improve its organization while clarifying its key points. We have moved several paragraphs from the results to the Methods section and to a Supplemental Material along with Figure 2.

The authors state that in this study a number of hypotheses can be explored, for example:

- (1) measurements taken by a monitoring network are not representative of all areas within a city and underestimate maximum exposures;
- (2) predictions from numerical air quality models at fine grid resolution cannot account for the variability in pollution levels existing within a neighbourhood scale. Both hypotheses are trivial and don't need any further exploration.

We agree that these hypotheses are obvious and testing them is of limited interest. They were intended to be general, but have clearly not served the paper well. Thus, we have replaced them with the more important, we feel, hypothesis and objectives described in our responses to the comments above.

Specific comments:

Abstract:

Page 31586, line 10: it is not true that 23 pollutants were measured: 20 pollutants were measured, 3 were calculated (please correct)

To make the Abstract easier to read, the sentence in the Abstract was only changed so as to avoid the word "measurements":
"Mobile data of 23 air pollutants was analyzed at high resolution in Montreal"

In the Measurements section the wording was changed to:
"Measurements of 20 different species were taken simultaneously from the CRUISER platform throughout the campaign at time resolutions ranging from 0.5 second to 2 minutes. Three additional species were derived from the measurements."

Page 31586, lines 14 -17: "This approach allowed linkage of the mobile measurements to the network observations and to generate average maps that provide reliable information on the typical, annual average spatial pattern" this sentence is not true (see comments below)

Please see our comments below

Page 31586, lines 19 -23: "Nearby microenvironments may have a wide range in average pollution levels varying by up to 300 %, which may cause large misclassification errors in estimating chronic exposures in epidemiological studies" this sentence is not true (see general comments)

As mentioned above, the second part of the statement was deleted.

2.2 Measurements

Speed correction: apart from the vehicle speed also wind speed should influence the airflow in the inlet. How did the authors adjust for it? What was the R² for the linear regression between PM (corrected) and PM (original) as stated in the equation?

Measurements of wind speed and direction were not taken while the vehicle was moving, so its effect on PM cannot be estimated. It is assumed, however, that since CRUISER is always changing direction while driving in the city, the wind speed would sometimes be subtracted from the driving speed and sometimes added to it, so that on average it should have a small net effect.

The speed correction is now described in more detail in the Supplemental Material and in our response to the other reviewer's comments. We now present the scatter plots from which the correction factors were calculated along with further statistical analysis.

2.3 Mobile measurement strategy

Page 31592, lines 19-20: The number of measurements per km of road (more than 2000) is really very impressive. But what does it mean? Given that the CRUISER travelled with an average speed of 25 km h⁻¹, it needed about 144 seconds per km and consequently 144 every second measurements were conducted. It means that for achieving of the (apparently) huge number of 2000 observation, only 14 measurement days (or trips) were needed (2000/144=13.9). The authors should consider how to express the number of observations per point or per route in a more common way, for example how many times the route (including the specific road segment) was completed.

The number of measurement days on which each road segment was sampled is indeed an important measure which could aid the readers evaluate the amount of information imbedded in the results. However given the limitations of the length and trying not to burden the readers we chose to report this only in the text. In the Methods section we report:

"There were 11, 17 and 6 mobile measurement days in the winter, summer and autumn, respectively, with 2-13 hours on each day (median of 9 hours)."

In another part of the Methods we also mention that the entire east or west route was covered on each day. This should give the readers a clear view of the number of times each road segment was covered in each of the seasons.

I assume that the huge number of observation was achieved only for pollutants measured every second. The number of observation for pollutants measured every 2 minutes is much smaller. It should be indicated in the manuscript.

For instruments with a time resolution longer than 1 second, the value was repeated over the entire measurement period and then joined with the 1second GPS data so as to allocate the measurement to its spatial reference. Therefore, we indeed have fewer unique measurements than the 2000+ measurements per segment displayed in Figure 1c and 1d for the best case scenario. For example, for the GRIMM Dust

monitor reporting PM every 6 seconds, each 6-seconds value was repeated for the six 1-second GPS locations. In the worst case of the AMS, the 120 second measurement was repeated 120 times. However, as the six second and even the 120 second resolution measurements overlap with the road segments differently each time the route is driven we still have good potential to spatially resolve the concentration patterns to finer scales than the distance typically covered by 120 seconds would imply. However, the number of unique measurements and the spatial resolution is clearly expected to be better as the time resolution of the measurements improves from 120 seconds to 1 second, although the magnitude of this improvement is dependent upon many factors that are difficult to characterize and were beyond the scope of this paper. These issues are briefly described in the paper in the third paragraph of the Measurements section:

“The data were then combined to one dataset with the time increment set to one second, and instruments with greater time intervals were given repeating values to reflect the more-integrated sampling. Although the spatial allocation of the measurements with longer time intervals is not as refined as for those with the one second time resolution, the road segments are sampled differently each time the route is driven and therefore the multiple repetitions of the route have good potential to spatially resolve the concentration patterns to finer scales than the distance traveled at these times would imply. “

2.4 Spatial analysis

I have doubts whether a spatial analysis for a given study area could be done based on measurements done not simultaneously and without any adjustment on the temporal variation. First of all, if the measurements are conducted at different times of the day, the results should be different even for the same site, due to the diurnal pattern of air pollutant concentrations. Furthermore, the day-to-day variation of all air pollutants is (in almost all areas around the world) very strongly influenced by meteorological conditions. I assume that this is the case also for Montreal. Therefore measurements conducted at different days and different times of the day are not comparable without any adjustment for the temporal correlation. This adjustment could be done by using a reference site operated continuously during the whole study period. This approach was already very often used and it is described sufficiently in the literature.

Temporal variability during the mobile measurements is an important issue and one that we recognized as we developed our deployment strategy (described above and below). Clearly these stem from the basic fact that a mobile lab cannot take simultaneous measurements at multiple locations. However, the nature of the temporal variations that our mobile lab deployment is potentially impacted by in estimating longer term average concentrations is much different than that being adjusted for with the “approaches already very often used” pointed out by this reviewer. In those studies they are attempting to combine multiple saturation or satellite measurement campaigns taken in the same city but during completely different time periods (often sequentially). In such cases they are reasonably justified in using a reference site to account for larger scale temporal differences

between these periods mainly due to meteorology and/or seasonality. Clearly, in those cases such an adjustment is necessary. We do not have this same issue because by design the mobile lab data from every location came from the same days.

However, there are limitations arising from temporal mis-matching because a mobile lab like CRUISER can only be operated in a specific study region (due to operational costs) and it cannot take simultaneous measurements at more than one location. As we discussed above, evaluating whether or not this issue can be accounted for is one of the key goals of this paper and motivated our deployment strategy (described below). Our revisions to the manuscript in the Introduction and Methods were intended to make these points clearer to the readers.

Expanding upon the temporal adjustment issue further, we point out that in our study the mobile measurements covered a wide and heterogeneous region while measuring multiple pollutants. Each pollutant, as we show in our results, has different emission sources such as roads for TRAP (e.g., nitrogen species, UFP, HOA) or oil refineries for Benzene and SO₂. Moreover, each source has its own distinct temporal variability, with roads having a well-documented morning and afternoon peaks and the refineries having multiple processes within the plant (e.g., transportation of oil products to and from the harbor by trucks or stacks emitting pollutants at different elevations above ground), each with a different temporal behavior. As a result, a realistic temporal adjustment that would truly accounts for nature of CRUISER's data is not feasible. This is because a reference site for temporal adjustment, if this is the approach taken, should be representative of the temporal variability over the entire domain at the scale relevant to our data. For example, if we chose an urban background site (i.e., a site distant from main roads and traffic sources) for the diurnal variability in TRAP it would not be representative of the diurnal emissions near busy traffic or for sources that vary differently, while if we chose a road-side site, it would not be representative of the conditions in the residential neighborhoods or of area impacted by different sources. Instead, we hypothesized that we could account for the complex temporal factors and other sources of variability that affect mobile lab data by establishing three main criteria for the deployment approach. (1) To cover the entire route on each day to minimize the impact of meteorological variability from synoptic scale systems. (2) Randomize the times when CRUISER visited the different parts of the route to minimize the impact of diurnal variability related to local scale meteorology (e.g., mixing heights) and emission rates (e.g., rush hour). (3) Maximize the number of times the entire route is covered over multiple time periods spread throughout the year, including some weekends.

We hypothesized that our criteria listed above would yield a set of location-specific averages that were not biased by temporal factors and as best as is realistically feasible we evaluated these averages against the actual annual averages to test this hypothesis. As indicated above, a simple reference site adjustment is not applicable to our type of data. We also point out that central site temporal adjustment used for correcting sequential satellite site datasets also involves considerable assumptions about this approaches' effectiveness, which likely varies

by city and time period of the study and which should be evaluated in greater detail.

3.1 Representativeness of the mobile measurements

Obviously different measurement methods were used by CRUISER and VdM. If so, a direct comparison of the measurement (side by side) is needed before and after the study period. A strong correlation is needed for any further comparison. Without such site by site comparison the interpretation of the results is somewhat crucial, as the authors stated on page 31595, lines 18-19 or page 312596, lines 16-17.

A side-by-side comparison before and after each of the measurement periods in each of the three seasons would be ideal for having a reliable reference between CRUISER and the VdM AQ sites. Clearly, this was not physically possible. There are too many instruments on CRUISER, which could not be moved into a VdM site to be all operating on the same inlet. However, such a comparison was not necessary and the approach we took was different, based upon the reality of monitoring networks and mobile lab designs and constraints. This approach met our study's intended needs of documenting that, as is prudent, the mobile data are "ground-truthed" to the long term network. This serves two purposes. First and most importantly, when we compare the annual average estimates from the CRUISER data with the actual annual averages, which are only available from the VdM network, to evaluate how well our deployment approach works for obtaining representative, we have some understanding of what magnitude of difference could be due to the different measurement systems utilized and more importantly if there is a bias. Second, so when we report concentrations in other locations any future comparisons of those values to what the monitoring network has traditionally reported can be considered in light of differences potentially arising from the different measurement systems used. Ideally, such data would come from long term comparisons using the same sample inlets, but in practice this is rarely, if ever feasible for multipollutant mobile lab measurement efforts. This is because a side-by-side comparison of the type suggested by this reviewer would have to be done at multiple monitoring sites and over weeks per site to be fully representative of the variability in meteorological and emission conditions, covering, for example, weekdays vs. weekends, windy days vs. stagnant winds, etc. Such a procedure will take a long time, add to the costs of study and, as mentioned above, is not physically possible. Another important physical constraint, if CRUISER were actually able to park extremely close to the VdM sample inlet, is power supply for CRUISER when it is parked for long periods which is not available at the AQ monitoring sites. Last, there is also a concern about leaving the lab parked overnight in an unprotected location. Nonetheless, we believe our approach to side-by-side comparison for shorter independent time periods (10-30 minutes for each "visit") is representative of the conditions during the measurement days and serves our intended purpose of providing the necessary context for subsequent comparisons.

We have briefly mentioned some of these issues in the revised manuscript. More importantly, to reduce the manuscript's length and to avoid distracting readers from the key points we have moved Figure 2 to the Supplemental Material (now Figure SM-D1).

In Figure 2 some scatter plots of CRUISER's vs. VdM measurements are shown. Some scatter plots are showing surprising low correlation between the measurements. For example, the R^2 for PM_{2.5} measurements is 0.60. Given that the CRUISER was operated in close proximity to the AQ sites, it is very low. In our network we observed R^2 of 0.90 for both traffic and urban background sites located 3-4 km apart from each other. I wonder that in the scatter plot for SO₂ also values below the limit of detection (1 ppb?) are displayed.

The reviewer is correct that these correlations are not perfect, even though for some pollutants (NO, NO_x and O₃) they are fairly high ($0.81 < R^2 < 0.85$). We address this issue in detail in Section 3.1 and some potential causes for the lower correlations are given, such as differences in the instruments, different elevations above ground, etc. A larger number of such co-measurements may yield higher correlations, but the more important purpose of the plots is to demonstrate if there are systematic differences or bias as in subsequent figures and tables we compare CRUISER averages, which are hypothesized to represent long-term (e.g., annual) averages, to the actual annual averages, which are only available at multiple sites from the VdM monitoring network. As noted above, to save space and keep readers focused on the main points in the paper we have moved Figure 2 into the Supplemental Material.

The detection limit in Table 1 refers to the original 10 seconds data measured by the instrument. In the scatter plot, however, each point is the mean of multiple measurements taken over a longer time period of 10-30 minutes. Detection limits are much lower for these averaging times and the data shown are above these values.

Page 31597, lines 13-26: The whole paragraph should be moved to the method section.

This section was moved to the Methods as suggested (Section 2.4).

Page 31598, lines 1-10: It is difficult for me to believe that the annual averages could be estimated based on very few and rather short term measurements – some studies on this issue were already published and support this finding. Cyrus et al. (2006) showed that “monthly means” based on 6-7 measurements distributed over a two week measurement period for each month substantially over- or underestimate the “true” monthly mean values.

Although it is true that Cyrus et al (2006) found the monthly means of a small sample to over/ under predict the reference measurement, they also report that the annual mean of the samples was within 10% of the annual mean of the reference. Given the differences in sample sizes between the two measurements (83 vs. 342

for the sampled and daily measurements, respectively), this result is rather encouraging. We also previously examined this issue (Xu et al, 2007), highlighted this work in the original manuscript and developed our deployment approach on those concepts.

The issue of the representativeness of sampling methods for evaluating chronic exposures can be viewed in context of sample theory – the greater the sample the better the accuracy. However in the context of real life, a greater sample has its costs, whether financial, in labor or others. This balance between accuracy and costs is eventually determined according to the abilities and needs of every research. In this paper we are reporting our findings about the accuracy of the sampling strategy we applied, in the hope that other studies will be better informed to make the decision for themselves whether to aim for greater accuracy at greater costs. It is also worth pointing out that much longer time windows can be covered cost-effectively with passive samplers for some pollutants (e.g., NO₂, NO_x) and this has served as the basis for much of the previous work on characterizing and modeling spatial patterns within cities. However, other trade-offs are an issue and in this case, it is the accuracy of the passive measurement technique. So a greater portion of the year could be covered, but the underlying measurement error is greater than what can be obtained with the instrumentation on board CRUISER or other similar mobile labs.

The requirement in this study is that “typical days” should be chosen for the measurements (page 31597, lines 28-29). How to find it? What is the definition of the “typical days”. What is “typical” for winter and what for summer season, which days are “typical”: rainy, sunny, with low wind speed or rather stormy? I see that the differences between the estimated annual means and the “true” annual means in Montreal are not very big. However, what is the reason for it and could it be expected also in other cities around the world. May be the rather low concentrations (and probably low day-to-day or season-to-season variation) make it possible for Montreal, but in this case the authors should discuss the unique situation in their study region. It might be also helpful to see the time series of the pollutant under study for the whole year 2009 (with indicated time periods of CRUISER measurements).

The measurement days were not selected a-priori based on any conditions so that they will be representative of typical conditions, except the selection of three seasons during which the measurements were done. As stated above, one of our criteria was to maximize the number of times the entire route was covered over multiple time periods spread throughout the year, including some weekends. We expect that during these days it is likely we capture a range of days, some typical, other atypical, if ‘typical’ can actually be defined.

What we present in Section 3.1 is an analysis of the measurement days to see whether they may be considered “representative” of the annual averages in 2009. We have changed the text to make this clearer:

“The VdM average among all the driving days was calculated to determine if, collectively, the driving days in each season were atypical of the annual averages

(ratios C/D and B/D in Table 2). Table 2 shows that on the selected driving days NO_x tended to be higher on average by 18%, compared to the 2009 daily averages. However, the overall ranking among the sites during these days was similar to the annual pattern (Figure 2). For the other pollutants (CO, PM_{2.5}, SO₂, O₃, NO and NO₂) the average difference between the study period VdM observations and the annual average among the sites with measurements were 9%, 16%, -1%, -23%, 28% and 12%, respectively (Table 2). The fact that this comparison is between VdM data (i.e., there are no methodological differences in the measurements) implies that the measurement days when CRUISER was driving in Montreal were somewhat representative of the long term averages. They tended to be biased high, except O₃, which was biased low. However, in terms of combustion pollutant levels (NO_x), the average high bias was 18% for the period, while for NO₂, which is often of most interest as an exposure indicator, the bias was smaller, at 12%. These differences indicate that collectively the days selected for driving were reasonably representative of what Montreal typically experiences.”

3.2 Intra-urban variability observed by CRUISER

We know very well that the concentrations of air pollutants in the vicinity of strong local sources are elevated. The whole section is showing that and could be significantly shortened.

In the revised version of the paper the main objective is to present a methodology for taking mobile measurements that will be representative of long term exposures. Although some of the things we show in Section 3.2 are not new for the research community, they demonstrate how a mobile lab can be used for this purpose. However addressing this comment and a comment by the other reviewer we have shortened Section 3.2, removing the more anecdotal findings and keeping only the main results.

Table 2: A/B and A/C might be more interesting for the reader as C/D and B/D (those relationship could be calculated for any monitoring network, without CRUISER).

As stated above, one of the main objectives of the revised manuscript is to evaluate the ability of a mobile laboratory to take highly resolved measurements in an urban environment that are representative of the long term exposures. For this purpose we first evaluate in section 3.1 how well the specific days on which the measurements were performed are representative of the 2009 annual averages at the different sites, comparing VdM daytime and daily averages to VdM’s 2009 averages (i.e., ratios C/D and B/D in Table 2). We then compare CRUISER’s measurement days to the VdM 2009 annual averages (ratio A/D). CRUISER’s measurements are also compared to the VdM’s concurrent measurements in Figure SM-D1. We therefore chose not to add the A/B and A/C ratios trying not to burden the readers with too many details, making the paper harder to follow.

Figure 3: The differences between the “daytime averages” and “daily averages” are really very small. Taking into consideration the mostly common diurnal pattern of air pollutants, it is somewhat surprising. Are there any explanations for it?

The “daytime averages” were calculated from the measurements taken between 09:00 and 20:00, which cover the times CRUISER was driving. Indeed the differences between the “daytime averages” and “daily averages” are small for most pollutants with the exception of ozone which has higher values during the daytime, as shown in figure SM-A1. Although we did not investigate this thoroughly, one reason for this behavior, we expect, is the fact that most of the reference VdM sites are located away from the emission sources and represent the urban background. Therefore they do not have as dramatic daytime peaks in pollution levels usually observed next to emission sources (i.e., mainly the morning and afternoon peaks near roads) and do not show a significant diurnal pattern that might cause differences between the daytime and daily averages.

4 Discussion

Page 31606, lines 14-16: It might be interesting for readers from other countries to get to know the requirements for monitoring site location in other parts of the world. So for example clear criteria for siting of the measurement stations are provided by the EU. With respect to the protection of human health, all Member States are required to provide data on the areas with highest concentrations (hot spots) as well as on those being representative for the exposure of the general population (urban background). All parts of the discussion and conclusions related to exposure assessment in epidemiological studies should be corrected by any expert working on this field.

The topic of monitoring sites locations is an important one and we agree that some of the results presented in this work are relevant for this issue. However discussing this is beyond the scope of this paper and would make the paper even longer than it already is. We refer this reviewer and readers of these interactive discussions to Craig et al., 2008, where one of the present authors (Brook) has discussed this topic. Readers may also find the series of articles accompanying Brook et al. (2013) of interest in the context of North American directions in monitoring air quality associated with hot spots due to traffic emissions.

We believe that the revised manuscript is less focused on stating implications for epidemiological studies to avoid controversy. The results of this paper are intended to be informative for those aiming to conduct and interpret the results of epidemiological studies assessing the potential impacts of intra-urban air pollution gradients as opposed to being directly useable in subsequent epidemiological work.

Literature

Boogaard et al.: "Contrast in air pollution components between major streets and background locations: Particulate matter mass, black carbon, elemental composition, nitrogen oxide and ultrafine particle number" *Atmospheric Environment* 45 (2011) 650-658

Brauer: "How Much, How Long, What, and Where Air Pollution Exposure Assessment for Epidemiologic Studies of Respiratory Disease" *Proc Am Thorac Soc* Vol 7. pp 111– 115, 2010

Cyrus et al.: "Evaluation of a sampling strategy for estimation of long-term PM_{2.5} exposure for epidemiological studies" *Environmental Monitoring and Assessment* (2006) 119: 161–171.

Cyrus et al.: "Variation of NO₂ and NO_x concentrations between and within 36 European study areas: Results from the ESCAPE study" *Atmospheric Environment* 62 (2012) 374-390.

Eeftens et al.: "Spatial variation of PM_{2.5}, PM₁₀, PM_{2.5} absorbance and PM_{coarse} concentrations between and within 20 European study areas and the relationship with NO₂ - Results of the ESCAPE project" *Atmospheric Environment* 62 (2012) 303-317.

Jerret et al.: "A review and evaluation of intraurban air pollution exposure models" *Journal of Exposure Analysis and Environmental Epidemiology* (2005) 15, 185–204.

Marshall et al.: "Within-urban variability in ambient air pollution: Comparison of estimation methods", *Atmospheric Environment* 42 (2008) 1359–1369.

References:

Brook RB, I Levy, C Mihele, 2013: From near-road to urban background: lessons learned from mobile lab monitoring. *EM Magazine*, July 2013

Craig L., Brook J.R., Chiotti Q., Croes B. et al. (2008) Air pollution and public health: A guidance document for risk managers. *Journal of Toxicology and Environmental Health, Part A*, 71, 588–698.

Eeftens, M., Beelen, R., de Hoogh, K., Bellander, T., Cesaroni, G., Cirach, M., Declercq, C., Dèdelè, A., Dons, E., de Nazelle, A., Dimakopoulou, K., Eriksen, K., Falq, G., Fischer, P., Galassi, C., Gražulevičienė, R., Heinrich, J., Hoffmann, B., Jerrett, M., Keidel, D., Korek, M., Lanki, T., Lindley, S., Madsen, C., Mölter, A., Nádor, G., Nieuwenhuijsen, M., Nonnemacher, M., Pedeli, X., Raaschou-Nielsen, O., Patelarou, E., Quass, U., Ranzi, A., Schindler, C., Stempfelet, M., Stephanou, E., Sugiri, D., Tsai, M.-Y., Yli-Tuomi, T., Varró, M. J., Vienneau, D., Klot, S. von, Wolf, K., Brunekreef, B. and Hoek, G.: Development of Land Use Regression Models for PM_{2.5}, PM_{2.5} Absorbance, PM₁₀ and PM_{coarse} in 20 European Study Areas; Results of the ESCAPE Project, *Environ. Sci. Technol.*, 46(20), 11195–11205, doi:10.1021/es301948k, 2012.

Levy I, C Mihele, G Lu, J Narayan and JR Brook, 2014: Evaluating Multipollutant Exposure and Urban Air Quality: Pollutant Interrelationships, Neighborhood Variability, and Nitrogen Dioxide as a Proxy Pollutant. Environmental Health Perspectives