Improving correlations between land use and air pollutant concentrations using wavelet analysis: Insights from a low-cost sensor network

Naomi Zimmerman¹,*, Hugh Z. Li¹, Aja Ellis¹, Aliaksei Hauryliuk¹, Ellis S. Robinson¹, Peishi Gu¹, Rishabh U. Shah¹, Qing Ye¹, Luke Snell², R. Subramanian¹, Allen L. Robinson¹, Joshua S. Apte², Albert A. Presto¹

¹. Center for Atmospheric Particle Studies, Carnegie Mellon University, Doherty Hall, Pittsburgh, PA 15213, U.S.A
². Department of Civil, Architectural and Environmental Engineering University of Texas at Austin, 301E E Dean Keeton St C1700, Austin, TX 78712, USA
‡ Current address: Department of Mechanical Engineering, University of British Columbia, 2054-6250 Applied Science Lane, Vancouver, BC V6T 1Z4, Canada

Abstract

City-wide air pollution assessments have typically relied on a small number of widely separated regulatory monitoring sites or land use regression (LUR) models built using time-integrated samples to assess annual average population-scale exposure. However, air pollutant concentrations often exhibit significant spatial and temporal variability depending on local sources and features of the built environment. In 2016, the Center for Air, Climate, and Energy Solutions (CACES) Air Quality Observatory was launched at Carnegie Mellon University to better understand urban spatial and temporal pollution gradients on the <1 km scale. The specific goal of this study was to understand how the highly temporally and spatially resolved low-cost stationary sampler data could be linked to modifiable factors (such as land use characteristics). Measurements in Pittsburgh, PA, USA consisted of a staged deployment of 15 stationary air quality monitoring stations, which used a low-cost air quality monitor, the Real-time Affordable Multi-Pollutant (RAMP) monitor, for measuring CO, NO₂, O₃, and CO₂, the low cost Neighborhood Particulate Monitor for measuring PM₂.₅ and a higher cost instrument for measuring ultrafine particle concentration. The campaign was from August 2016-May 2017 and also included mobile sampling with reference-grade instruments in ~1 km² grids around the stationary monitors. The stations were deployed as a rural-urban-rural transect along the prevailing wind direction and in downtown urban locations with a range of modifiable factors, such as traffic, restaurant and population densities. Wavelet decomposition was used to separate the pollutant time series from the stationary samplers into short-lived (<2 h) pollution events, longer-lived events (2-8 h) and persistent enhancements (baseline changes >8 h) above the regional background. Compared to the non-decomposed total pollutant signal, the short-lived or persistent enhancement pollutant signals, which should come from local sources, were better correlated with covariates used in LUR model construction. For example, Pearson r between total vehicle

* Corresponding author. Tel: 1-604-822-9433; Fax: 1-604-822-2403
E-mail address: nzimmerman@mech.ubc.ca
counts in a 100 m buffer and NO₂ increased from 0.57 using the total pollutant signal to 0.83 using the persistent enhancement only. The findings from this study support building more accurate and higher time resolution (e.g., daily, hourly) LURs using low-cost sensors.

**Keywords:** Air quality, exposure, low-cost sensors, land use regression, urban emissions

**INTRODUCTION**

Traditional urban and regional air quality studies have typically relied on an existing network of regulatory monitoring sites located far from individual sources or used spatially-intensive but temporally-limited sampling campaigns to assess population-scale exposure (Ito et al., 2016; Jedynska et al., 2014; Wang et al., 2013). However, air pollutant concentrations often exhibit significant spatial variability depending on local sources and features of the built environment, which may not be well captured by the existing monitoring regime (Apte et al., 2017; Tan et al., 2014). Understanding this complexity through improved measurements, which in turn support spatially- and temporally-resolved models, is key to better characterizing and reducing population exposures to air pollution.

A significant fraction of air pollutant exposure research has highlighted the substantial health effects attributable to PM₂.₅ (particulate matter with a diameter <2.5 µm), including premature mortality, cardiovascular disease, lung cancer, and asthma (Brook et al., 2010; Pope et al., 2009; Raaschou-Nielsen et al., 2013). However, observed health effects may not be fully explained by PM₂.₅ mass alone. At the urban scale, epidemiological studies frequently demonstrate elevated risks of chronic and acute health effects for populations that reside near
major sources, such as roadways (Baumgartner et al., 2014; Hoek et al., 2002). This enhancement in risk near roadways is likely not solely attributed to PM$_{2.5}$; near-road concentrations of PM$_{2.5}$ are only modestly elevated (typically <1.5 times) above urban background concentrations (Apte et al., 2011; Karner et al., 2010). In contrast, other traffic-related air pollutants, including nitrogen oxides (NO$_x$), carbon monoxide (CO), black carbon, and ultrafine particles (UFP; particles <100 nm diameter) have much higher near-roadway concentration enhancements (>3-5 times) (Karner et al., 2010; Saha et al., 2018). These pollutants, or synergistic effects of exposure to multiple traffic-related pollutants, might explain some portion of the observed risk increment near roadways. However, the high intra-urban spatiotemporal variability of air pollutants means that urban background concentrations, such as those measured at regulatory monitoring sites, do not adequately represent population exposure for these species.

In this paper, the Center for Air, Climate, and Energy Solutions (CACES) Air Quality Observatory (AQO) at Carnegie Mellon University is introduced and initial findings are discussed. The overarching goal of the CACES AQO is to advance our understanding of the factors that influence variability in population exposure to air pollution. These factors include pollution sources, differences in the built environment, and atmospheric transformation. The CACES AQO was designed to achieve these goals with a combination of spatially and
temporally resolved measurements. The AQO uses a network of stationary, primarily low-cost, samplers operating at high time resolution to quantify temporally-varying intra-city pollutant gradients. These stationary samplers included the low-cost Real-Time Affordable Multi-Pollutant (RAMP) monitor (SENSIT Technologies) for measuring CO, NO₂, O₃ and CO₂, the low-cost Met-One Neighborhood Particulate Monitor (NPM) for measuring PM₂.₅ and a condensation particle counter (CPC, Aerosol Devices Inc). The stationary samplers are complemented with mobile sampling to investigate fine-scale (sub-km) spatial variations around each stationary sampler.

The distribution and density of major sources within the built environment is modifiable, and thus a critical objective of the CACES AQO is understanding the relationship between air pollution exposure and modifiable factors, such as traffic density. An approach for investigating this question is with land use regression models (LURs) (Beelen et al., 2013; Hankey and Marshall, 2015; Hatzopoulou et al., 2017; Hoek et al., 2008). LURs use spatial covariates as predictor variables in a multivariate regression to predict the concentration of different air pollutants (Ryan and Lemasters, 2008). The predictors for these models include parameters that are “modifiable” such as surrounding land use (e.g. commercial area in 100 m² radius) and traffic, and meteorological data within certain circular buffers (Habermann et al., 2015). LURs are typically constructed by simultaneously sampling at many locations that span the range of land uses or
“modifiable factors” in a region. Previously, cost-effective means of simultaneous sampling would involve filters and passive samplers, which are not temporally-resolved. This integrated (non-temporally resolved) data would include regional effects that should not be well predicted by local land use. By using real-time, high-spatial resolution sampling with low-cost sensors to span the range of modifiable factors in urban areas, we hypothesize that temporally-resolved LURs can be constructed that improve accuracy of prediction in complex urban areas and that these LURs may be more readily transferred to other urban areas by removing the influence of regional effects.

This manuscript describes the design of the AQO along with results collected in 2016-2017. We focus here on results from the network of stationary low-cost samplers; results from mobile sampling and comparison of mobile to stationary sampling are presented in other publications (Gu et al., 2018; Li et al., 2019). The primary goal of this study is to understand how the highly temporally and spatially resolved low-cost stationary sampler data can be linked to modifiable factors (such as land use characteristics). As part of this assessment, we used time frequency analysis of data from the stationary samplers to show that decomposing the signal to separate the regional background from variations occurring on timescales from <2 to >8 hours can improve correlations between pollutant concentrations and modifiable factors,
demonstrating the value of distributed time-resolved measurements that low-cost sensor networks enable

**METHODS**

**Measurement Sites and Sampling Period**

Stationary site sampling was conducted between August 3, 2016 and May 20, 2017. A 2-month deployment of 10 stationary monitoring sites (Fig. 1) was conducted between August-September 2016 along a transect which followed the prevailing wind direction (Fig. 1A, SW to NE) to determine how pollutant concentrations vary along the urban-to-rural gradient (~10-100 km), and to determine urban excess. Within the transect, sites were chosen to represent a broad range of land use classifications, including sites we generally classified as regional background, urban background, traffic-dominated, and restaurant dominated. Transect sites included: Carnegie, Beechview, Strip District, Hill District, Lawrenceville, Oakland, Highland Park, Aspinwall and Fox Chapel (Fig. 1). Following the 2-month deployment, the Hill District, Oakland, and Lawrenceville sites were left online, the remaining sites were returned to the Carnegie Mellon campus (Fig. 1: Oakland) for calibrations (see Section S3 of the Supporting Information), and one station was brought online in October 2016 in Downtown Pittsburgh (Fig. 1: Downtown #2). A second dedicated campaign was launched in January 2017 to focus on spatial variability of urban air quality; it included two additional downtown sites (Fig. 1: Downtown #1,#3), and sites in Mt.
Washington and Southside along with the ongoing measurements at the Downtown #2, Oakland, Lawrenceville and Hill District sites. These 8 sites remained online through May 20, 2017. After May 20, 2017, samplers were returned to the Carnegie Mellon campus for calibration prior to redeployment as part of a 50-node network in 2017-2018.

While the sites were generally classified as either rural background, urban background, traffic-dominated, restaurant-dominated or high traffic and restaurants, the sites were characterized in greater detail using land use covariates (modifiable factors). The chosen covariates were previously identified as significant in Pittsburgh (Tan et al., 2016) and include covariates for total vehicle counts, restaurant counts, vacant forest area, and total population density (Table S1). In Fig. 2, these covariates are provided for a subset of sites across all the general classifications to illustrate the extent of the covariate matrix. The full figure that shows the covariates for each site is provided in the Supporting Information (Fig. S1). For some covariates, such as industrial land use area or restaurant density, many sites have a value of zero.

A mobile laboratory equipped with reference grade monitors sampled at high time resolution (1 second – 1 minute) in ~1 km² grids around each stationary monitoring site (Li et al., 2019). Mobile sampling complements the stationary sampling by measuring the spatial gradients around each stationary site and around source areas such as highways and restaurant clusters. The spatially-resolved mobile samples can in turn inform a mechanistic understanding of
physicochemical pollutant transformations near sources. Details of the mobile sampling dates around each stationary sampler are provided in Table S3 of the Supporting Information.

Measurement Techniques

Central Site

A stationary central site was constructed on the Carnegie Mellon University campus in the Oakland neighborhood of Pittsburgh, PA (Fig. 1: Oakland). The site (40°26'31.5''N, 79°56'33''W) is located within a small (< 50 vehicles) controlled access, open air parking lot near the center of campus. It is equipped with reference-grade and research-grade instrumentation for measuring concentrations of NO$_2$, NO, PM$_{2.5}$ mass and composition, O$_3$, CO, and UFP with sub-hourly resolution. Instrument details including make, model, and time resolution are provided in the Supporting Information (Table S2). The central site is also equipped with a RAMP low-cost sensor system (the Oakland site), described below, so that data from this site are directly comparable to the other distributed monitoring locations. This site serves as our central reference site and is used to calibrate the low-cost sensors (see section S3 of the Supporting Information).

Distributed stationary sampling sites

Stationary multi-pollutant sampling was used to characterize spatial and temporal variations in pollutant concentrations of criteria air pollutants (CO, NO$_2$, O$_3$, PM$_{2.5}$) and ultrafine
particle (UFP) number concentration. Gaseous pollutants were measured with the Real-time Affordable Multi-Pollutant (RAMP) monitor, which was developed in a collaboration between Carnegie Mellon University and SenSevere LLC (now commercially available through SENSIT Technologies). The RAMP uses the following commercially-available electrochemical sensors from Alphasense Ltd: carbon monoxide (CO, Alphasense ID: CO-B41), nitrogen dioxide (NO$_2$, Alphasense ID: NO2-B43F), and total oxidants (O$_x$, Alphasense ID: Ox-B431). The unit also includes a nondispersive infrared (NDIR) CO$_2$ sensor (SST CO2S-A) which contains built-in T (method: bandgap) and RH (method: capacitive) measurement. Extensive details on the RAMP including a discussion on its calibration using random forest models (which account for pollutant cross-sensitivities, temperature and humidity effects) are available in Zimmerman et al. (2018) and summarized in the Supporting Information. All of the RAMP monitors used in this study were calibrated as part of Zimmerman et al. (2018), and some of the RAMP performance metrics from Zimmerman et al. (2018) have been provided in the Supporting Information. In general, RAMP mean absolute error was less than 20% for CO, CO$_2$ and O$_3$ and less than 30% for NO$_2$. Using the performance standards outlined in the US EPA Air Sensors Guidebook (Williams et al., 2014), the CO and O$_3$ monitors met Tier III (Supplemental Monitoring) standards, and the NO$_2$ monitor met Tier IV (Personal Exposure) standards.
The distributed low-cost monitors also included PM$_{2.5}$ as measured by the Met One Neighborhood PM$_{2.5}$ monitor. The PM$_{2.5}$ monitor uses a nephelometer for measurement, and requires a correction based on relative humidity and hygroscopic growth factor using estimates of the bulk particle composition (Cerully et al., 2015; Petters and Kreidenweis, 2007). This correction protocol is summarized in section S3 of the Supporting Information and further details have been published elsewhere (Malings et al., 2019).

Ultrafine particle (UFP) number concentrations were measured with water-based condensational particle counters (CPC, Aerosol Dynamics MAGIC CPC) (Hering et al., 2014). A total of 8 of 15 sites included UFP measurements (Beechview, Downtown #1-#3, Mt. Washington, Southside, Hill District, Fox Chapel). Prior to deployment, the MAGIC CPCs were co-located at the central site with a well-characterized butanol CPC (TSI 3776) for 3 weeks. Linear regression between the MAGIC CPCs and the butanol CPC were used to adjust the MAGIC CPC response to match the butanol CPC. Details of the CPC correction are provided in the Supporting Information.

The stationary samplers did not include any measurements of wind speed and wind direction, which may be a useful tool for correlating pollutant signals to specific sources or land use characteristics. The RAMP monitors can be equipped with an anemometer, however, there is no guarantee that wind speed and wind direction measured by the RAMP would be unaffected by
the RAMP mounting structure. Additionally, many RAMP monitors were located near-ground level and in urban street canyons where winds from any direction would enable measurements of nearby sources.

**Data Analysis**

The data first undergoes quality assurance protocols, which include removing any data with error codes, removing any spikes in concentration because of a power surge (i.e., power cycling typically causes an instantaneous spike in measurements) and removing any data in exceedance of the instrument detection limits (e.g., UFP concentrations >100,000 #/cm³, <5% of total UFP data collected). The RAMP gaseous data are then corrected using the random forest-based calibration algorithms developed from co-location with the supersite (Zimmerman et al., 2018), and PM$_{2.5}$ data is corrected using the RH measurement and estimated hygroscopic growth factors (Cerully et al., 2015; Malings et al., 2019; Petters and Kreidenweis, 2007). UFP concentrations are also adjusted based on the pre-campaign co-location with a reference CPC (TSI model 3776). Additional details on this data quality assurance program are provided in the Supporting Information (Section S3). All data from the distributed sites was recorded every 15 seconds and transmitted to the server every 5 minutes. The data was then down sampled to 15-minute time resolution to reduce the size of the data set and to reduce the noise from the gas phase and PM$_{2.5}$ sensor measurements.
The total pollutant signal at a site is affected by local sources, regional events, and the regional background. Here, we propose that two components of the time series are from local sources: short-lived (i.e., high frequency) spikes and low-frequency persistent enhancements above the regional background from local sources whose contribution to the total signal is more or less constant. To isolate the local and non-local components of the pollutant signal, the time series data at each site was decomposed using wavelet decomposition (Klems et al., 2010; Sabaliauskas et al., 2014). In wavelet decomposition, the time series data is passed through a high frequency and low frequency filter, which separates the data into the first level of detail (d_1) and the first level of approximation (a_1). The first level of approximation, a_1, is further decomposed into a second level of detail (d_2) and approximation (a_2). This process is repeated until the specified level of decomposition is reached. The original data are reconstructed by adding the final level of approximation (a_n) and the sum of each level of detail (d_1+d_2+…+d_n). The level of decomposition corresponds to changes on time scales of 2^n.

For the 15-minute resolution data, a 5-level wavelet decomposition was used to identify the low-frequency component of the time series data, which is equivalent to separating a baseline (a_5) that changes on the order of 8 hours (15 min x 2^5 = 8 h). This level of decomposition resulted in baseline concentrations that most closely mimicked the diurnal pollutant patterns of ozone. Ozone is a mostly regional pollutant with a well-established diurnal profile (peak
concentrations in the afternoon when sun has reached its apex and temperatures have peaked), making it an ideal choice for wavelet decomposition level selection. A second 3-level wavelet decomposition on the original time series data was then conducted to separate short-lived events. This 3-level decomposition is equivalent to separating out signals that change on the order of 2 hours (15 min x 2^3 = 2 h). For any given site, the difference between the total pollutant signal at time \( t \) and the 3-level decomposition baseline \( (a_3) \) was labelled as a short-lived event (Eq. 1). Longer-lived events were then defined as occurring on a timescale between 2 and 8 hours. The longer-lived event contribution was calculated as the difference between the 3-level decomposition baseline \( (a_3) \) and 5-level decomposition baseline \( (a_5) \) (Eq. 2). All wavelet decomposition used the wavelets toolbox in MATLAB following a similar method to that described in Klems et al. (2010) and Sabaliauskas et al. (2014).

No constant site was used as the regional background, as changing wind directions or local sources may affect the representativeness of a given site as the regional background. Instead, the regional background at a given time \( t \) was defined as the minimum 5-level baseline concentration \( (a_5) \) across all the active sites (Eq. 3), except for \( O_3 \), which can be depleted near urban traffic sources and thus the regional background was calculated as the median baseline concentration \( (a_5) \) across all the actively sampling sites (Eq. 4). The regional backgrounds were then used to calculate persistent enhancements at each site, calculated as the difference between
the baseline from the 5-level decomposition and the regional background (Eq. 5). Using this
definition, the site defined as the regional background at a given time step has zero persistent
enhancement.

While there was no constant regional background site, sites labelled as either rural or
urban background (Fig. 1) were the regional background for the vast majority of the time series
data (e.g., 88% of the time for CO, 70% of the time for NO\textsubscript{2}). Additionally, by calculating the
regional background using the 5-level decomposition baseline, the data is inherently de-noised,
limiting the impact of noise on selection of a regional background.

\textbf{Short – lived event} \( (t) = \text{Total Signal}(t) - a_3(t) \) \hspace{1cm} (1)

\textbf{Longer lived event} \( (t) = a_3(t) - a_5(t) \) \hspace{1cm} (2)

\textbf{Regional background} \( (t) = \min_{\text{all sites}} [a_5(t)] \) \hspace{1cm} (3)

\textbf{[O\textsubscript{3} only] Regional background} \( (t) = \text{median}_{\text{all sites}}[a_5 (t)] \) \hspace{1cm} (4)

\textbf{Persistent enhancement} \( (t) = [a_5(t)] - \text{Regional background}(t) \) \hspace{1cm} (5)

To illustrate a typical decomposition, 48 hours of CO data from the Downtown #3 site are shown
in Fig. 3. In Fig. 3, we observe a roughly constant persistent enhancement of 100 ppb above the
regional background.

\textbf{RESULTS AND DISCUSSION}

\textbf{Average Concentrations – Stationary Sampling}

The total pollutant concentrations across all sites and all measurement campaigns are
summarized in Fig. 4, which shows box plots of CO, CO\textsubscript{2}, NO\textsubscript{2}, O\textsubscript{3}, PM\textsubscript{2.5} and UFP. While
measurements at different sites were in some cases made in different seasons (e.g., many sites on the rural-urban-rural transect were only sampled from August-September 2016), we used the Oakland site (which operated continuously from August 2016 to May 2017) to assess seasonal variability. At the Oakland site, average concentrations over many months did not vary substantially by season and thus the data set was used as a whole to assess large- and fine-scale spatial patterns. The most significant seasonal variability was for ozone (median concentration at the central site: 18 ppb in fall 2016 vs. 28 ppb in spring 2017), and details of the seasonal analysis are provided in the Supporting Information. A limitation of the seasonal assessment was that it was limited to one site. A site-by-site assessment of seasonal variability will be addressed during an ongoing longer-term deployment of an expanded CACES network.

Fig. 4 shows that all of the measured pollutants are spatially variable. In general, near-source sites such as Downtown #1-3 (high traffic and restaurants) and Aspinwall (adjacent to a highway) have the highest concentrations. However, the spatial patterns are not identical for all pollutants, as described in more detail below. This reflects the mix of emissions from the sources that impact the various sampling sites.

For CO and CO$_2$, mean concentrations were highest at the sites with high traffic volumes (e.g., Downtown #1-3, Aspinwall), as expected. Restaurant density alone was not a significant contributor to CO concentration; the restaurant-impacted Southside and Strip District sites are
only modestly elevated relative to the urban background sites for CO, and do not have a significant CO₂ enhancement. The highest CO and CO₂ concentrations were observed at the Aspinwall site, which was located <15 m from a major highway in Pittsburgh.

The impact of nearby traffic on concentrations at the Aspinwall site was also evident when comparing the mean NO₂ and O₃ concentrations. In very near-road environments, fresh NO emitted by vehicles reacts with O₃ to form NO₂, resulting in slightly decreased O₃ concentration. At the Aspinwall site, NO₂ concentrations were approximately 3x the background concentration, and on average, O₃ concentrations were suppressed by 5 ppb.

Higher mean NO₂ (1-2x background) combined with slightly lower O₃ was also observed at two of the Downtown (high traffic, high restaurant) sites. Within the Downtown sites, NO₂ concentrations were lower at Downtown #2. Downtown #2 was characterized by similar total vehicle counts, but 27-48% fewer diesel vehicles on the nearest road than Downtown sites #1 and #3 (Fig. 2). The RAMP monitor at Downtown #2 was also deployed approximately 2 months prior to the Downtown #1 and #3 sites, which may explain the lower O₃ at this site compared to the other downtown sites. CO is similar at Downtown #2 and the other Downtown sites. This suggests that variations in fleet composition over very small scales (100s of meters) can influence the observed mix of air pollutants. In this case, a higher fraction of diesel vehicles, which emit more NOₓ than gasoline vehicles but have similar emissions of CO (May et al., 2014; Wang et al.,
leads to higher NO\textsubscript{2} for Downtown sites #1 and #3 than at Downtown #2, but similar CO at all three sites.

Across the entire transect, there is an average urban excess of approximately 2 µg/m\textsuperscript{3} of PM\textsubscript{2.5} for the urban background sites relative to the rural sites. The average urban excess increases to near 5 µg/m\textsuperscript{3} in the source-impacted Downtown #1 and Downtown #2 sites. At the Downtown #3 site, PM\textsubscript{2.5} concentrations were on average almost 3x the regional background. This site was located across a pedestrian alley from a restaurant equipped with both a wood-fired oven and a deep fryer, and the restaurant exhaust vents were within 20 m of the sampling location. The high average PM\textsubscript{2.5} concentration at this site was largely due to the wood-fired oven; beginning around 12PM local time when lunch service began at the restaurant, PM\textsubscript{2.5} concentrations would sharply increase.

The pollutant for which there was the most significant spatial variability was UFP, which has highly dynamic behavior (Klems et al., 2010; Padró-Martínez et al., 2012; Sabaliauskas et al., 2012; Zimmerman et al., 2016). The number, size distribution and composition of ultrafine particles evolves rapidly in space and time, owing to complex processes of dilution, nucleation, condensation, evaporation, and coagulation (Saha et al., 2018). As such, UFP exposure in near-source regions differ fundamentally from those experienced by populations who breathe urban background air. UFP concentrations at Downtown #3 (near the wood-fired oven) were on average
almost double those at sites less than 500 m away. The average UFP concentration at the restaurant-dominated (but low traffic) Southside site was as high or higher than Downtown #1 and #2, but this site has similar CO and NO$_2$ concentrations as urban background sites. This suggests that restaurant emissions disproportionately affect spatial characteristics of UFP. In separate manuscripts, Robinson et al. (2018) and Ye et al. (2018) use the mobile sampling data show that restaurant emissions generate distinct plumes of high PM mass and number concentrations, echoing the results from Downtown #3.

**Decomposition into Persistent enhancement, Short-lived events, and longer-lived events**

As defined in Equations (1-5), the total pollutant time series at the stationary sites were decomposed into persistent enhancement (> 8 h), longer-lived event (2-8 h) and short-lived event (<2 h) components. For the stationary sites, the average persistent, longer-lived and short-lived enhancements above the regional background are summarized in Fig. 5. What is immediately apparent from Fig. 5 is that the bulk of the enhancement at a given site is “persistent” in nature (i.e., not varying significantly over time). Of the total enhancement above background, the longer-lived and shorter-lived event enhancements accounted for less than 20% of the total concentration for all pollutants. However, the role of short-lived enhancements may be damped compared to using higher time resolution measurements. The 15 minute time resolution used here is insensitive to individual plumes, for example from nearby diesel vehicles, that last ~1 minute.
or less (Tan et al., 2014). While it is difficult to identify exact sources of the persistent
enhancements or short-lived events due to limited site data beyond land use covariates, the
hypothesis is that the persistent enhancement occurs due to sources that emit ‘persistently’ (e.g.,
restaurants open for prolonged periods, industry, etc.). While these persistent sources are likely
not emitting constantly, the prolonged period of emissions combined with the impact of street
canyons and meteorology (e.g., night time inversions) may prevent pollution from ever fully
clearing out from the site. Short-lived enhancements are likely from sources that are not emitting
as consistently, or from local sources that pass the sampling site (e.g., a series of particularly
heavy emitting vehicles, or brief periods of intensive cooking during common meal times).

Across the different sites, Downtown #3, adjacent to the high emitting restaurant,
continues to stand apart from most other sites, despite having similar land use characteristics to
the adjacent Downtown sites. As expected, this site has the largest contribution from “short-lived”
events across all pollutants, and a large “longer-lived” enhancement in PM$_{2.5}$, presumably from
the periods of intensive cooking combined with the street canyon. The wavelet decomposition
approach taken here also allows for the separation of positive short- and longer-lived O$_3$ events
from negative persistent enhancements due to scavenging of O$_3$ by NO to form NO$_2$ in the near-
road environment. Most sites with high traffic density (Downtown #1-3, Carnegie, Aspinwall)
exhibited some degree of negative persistent enhancement for O$_3$. The Mt. Washington and
Beechview sites had the largest positive persistent O₃ enhancements. These two sites had the highest elevation; elevation has been previously shown to have a positive correlation with O₃ concentration (Malmqvist et al., 2014; Wolf et al., 2017) possibly due to stronger winds and no formation of shallow boundary layers (Stedman and Kent, 2008). The separation of these components may enable better correlations between modifiable factors / land use covariates, as it inherently separates the role of vehicle emissions and other sources from the regional or urban background. Additionally, while rural background sites such as Chartiers Valley and Fox Chapel had the smallest persistent enhancements across all pollutants, the average persistent enhancement is never zero, i.e., all sites with some nearby human activity will at some points have some small enhancement above the regional background.

For a high-traffic, high-restaurant site (Downtown #1) and an urban background site (Mt. Washington), the diurnal patterns of the total signal, persistent enhancement and short-lived events for CO and UFP are shown in Fig. 6. Due to its small overall contribution (Fig. 5), the longer-lived event enhancements were excluded from the diurnal analysis. Longer-lived events are likely from local sources that are not permanently located by the site. Some examples include emissions from roadwork and construction, or from local events (sporting events, arts and culture). A more detailed assessment of what sources result in short-lived vs. longer-lived contributions using observations at the sites may help pinpoint these sources in future studies.
From Figs. 2 and 6, we observe that the downtown sites never reach the regional background. At night when human activity is minimized (i.e., most of the population is asleep), a reasonable hypothesis is that a downtown site should be similar to the background. However, in this analysis, even at night there is a non-negligible persistent enhancement. One possible explanation for this is that despite low human activity, nighttime inversions combined with street canyons reduce the rate at which air pollutants are cleared away from a site and this pollution is never fully cleared before human activity increases again in the morning. Even at the urban background (Mt. Washington) site, there is a persistent enhancement during overnight hours, though it is smaller than for Downtown #1. This may suggest that even in urban background locations, human activity and associated emissions never reach the same low levels observed in sparsely populated rural background locations.

For both sites, the total concentration shows an increase in CO and UFP in the morning, corresponding to rush hour. After falling from the morning peak, UFP remains roughly stable for the remainder of the day, whereas CO decreases until approximately 4PM, then it begins to climb again, presumably due to the combination of evening rush hour and a nighttime inversion. Additionally, the total signals for CO at the two sites are clearly separated from the regional background, while for UFP sometimes Mt. Washington essentially at regional background levels.
This reinforces the complexity and extremely high temporal and spatial variability of UFP relative to CO.

By separating the signal into persistent enhancements and short-lived events, we see that for Mt. Washington, an urban background site, the persistent enhancement is roughly constant throughout the day for both UFP and CO. The short-lived events increase gradually starting at 4 AM, likely due to traffic emissions during morning rush hour. Following this gradual rise, there is a sharp spike in CO and UFP enhancements in Mt. Washington around 8-9AM. Given that these sharp increases occur for both UFP and CO, this suggests a real effect vs. noise, and may represent the peak of the morning rush hour or emissions from an unidentified source near the sampling location. The short-lived enhancement of UFP at Mt. Washington increases again at noon, possibly due to the influence of increased restaurant emissions, which were shown in Fig. 4 to disproportionately affect UFP concentrations.

For the high-traffic, high-restaurants downtown site, the persistent enhancement in UFP and CO is elevated between 7AM and 12PM during the rush hour. There is also a small bump in the persistent enhancement around 3-5PM that is likely associated with the afternoon rush hour. The short-lived enhancement of CO only increases during morning rush hour, whereas for UFP four spikes are seen at approximately 7AM, 1PM, 4PM and 6PM. A potential explanation for these spikes include morning traffic (7AM spike), and restaurant emissions associated with lunch.
(1PM), happy hour specials (4PM) and dinner (6PM). This hints at the significance of comparing decomposed pollutant signals to modifiable factors (such as land use characteristics) to improve our understanding of the driving factors behind air pollutant concentrations.

To explore this hypothesis, the average total pollutant signal (‘Average’), persistent enhancement (‘Persistent’), short-lived event enhancement (‘Short-lived’) and combined persistent and short-lived event enhancements (‘Short-Lived + Persistent’) for all pollutants across all sites were regressed against the 8 modifiable factors described in Fig. 2, Fig. S1 and Table S1.

A summary of the Pearson correlation coefficients by pollutant signal and modifiable factor is shown in Fig. 7. In Fig. 7, there are several correlations between modifiable factors and pollutants that are improved by separating the enhancements from the regional background. One exception is industrial land use area, which did not benefit from signal decomposition for any pollutant. This is likely due to the poor stratification of this modifiable factor; of the 15 stationary sites, only five had non-zero industrial land use areas. Similarly, of the pollutants analyzed, only UFP did not benefit from the decomposition into different time signals; correlations were essentially unchanged with this analysis. We hypothesize that this was due to the limited number of sites measuring UFP (only 8 / 15) and the fact that there were no uniquely traffic-dominated sites with UFP measurements. A larger number of sites may be needed, with greater variation in
land uses, to observe differences in correlations for UFP. An alternative interpretation is that the
persistent and short-lived enhancements of UFP were relatively uniform across all sites (e.g.,
uncorrelated with land use); however, this cannot be concluded with much confidence due to the
small number of sites. The data collected as part of Saha et al. (2019) will be used to explore this
in greater detail in future work.

Commercial land area correlations are improved for CO₂ (0.31 for short-lived vs. 0.08
average), NO₂ (0.6 for short-lived vs. 0.39 average), O₃ (0.46 for short-lived vs. -0.26 average)
and PM₂.₅ (0.67 for persistent vs. 0.52 average). This suggests that the bulk of CO₂, NO₂, and O₃
from commercial land use is from short-lived events, whereas increased commercial land area
consistently increases PM₂.₅ over longer time scales, in keeping with our understanding of PM₂.₅
as a more regional (i.e., less dynamic) pollutant (Wu et al., 2015). Restaurant density is also
better correlated with short-lived events of several pollutants (CO₂, NO₂, O₃). Restaurants
emissions are time-variant in nature (higher during common meal times), thus this improved
correlation is also expected. Given recent studies on the increasing role of restaurant emissions in
the urban pollutant landscape (Chen et al., 2012; Laurent et al., 2014; Robinson et al., 2018),
improving our predictive capacity of the air pollution impact of high restaurant density is critical.

Vehicles are another important emission source in urban areas. The only vehicle-related
correlation that improved upon signal decomposition was that persistent enhancements in NO₂
are better correlated with vehicle volume than average concentrations. This echoes the decomposed CO concentrations shown for Downtown #1 in Fig. 6, where both persistent and short-lived CO concentrations increased during traffic rush hours. However, the decomposed CO signals did not have improved correlations with either total vehicle traffic or diesel vehicle traffic. This may be a result of vehicle emissions dominating the total CO concentration, rather than just local enhancements over the background. Similarly, decomposed NO₂ concentrations do not have an increased correlation with diesel traffic relative to the total concentration, even though diesel vehicles are a major source of NOₓ emissions.

The analyses shown in Fig. 7 demonstrate that isolating the different time-frequency components of stationary site data can provide better correlations with modifiable factors such as land use covariates, which are used in LUR building. This follows the logic that if a regional background is common among all sites, including the regional background or regional events in LURs effectively dampens any relationships between land use and observed concentrations. As such, we suggest that when using time-resolved air pollutant data to build LURs or other spatial models, a better approach would be to construct the LUR for one or all of the enhancements, and then layer this on top of a regional background. Additionally, this study highlights the need for high time resolution air pollutant data when constructing LUR models. This study suggests that
low-cost sensor networks, such as those introduced in the CACES AQO, are a promising way to collect the high spatial and temporal resolution data needed to support these models.

**CONCLUSIONS**

The CACES Air Quality Observatory (AQO) aims to improve our understanding of the relationships between modifiable factors (such as land use), concentrations of multiple pollutants, and air pollution exposure. In this paper, the CACES AQO was introduced and results were provided from an extensive deployment of a network of 15 stationary, mostly low-cost, monitoring stations complemented by mobile sampling. From this work, substantial site-to-site variability was identified, even among sites of similar general pollutant class (high traffic, high restaurant) and less than 500 m apart (e.g., Downtown #1-3).

To enhance our understanding of the relationship between temporal trends and modifiable factors, the air pollutant concentrations were decomposed into persistent enhancements above the regional background, longer-lived event enhancements (2-8 h) and short-lived event enhancements (<2 h). While the dominant component of the total enhancement above background was the persistent enhancement (~80% across all pollutants), the correlations between modifiable factors and pollutant signals were the highest when comparing short-lived event enhancements to restaurant density and commercial area. By correlating short-lived events in lieu of average concentrations to land use covariates, correlations with commercial area
improved substantially for CO$_2$, NO$_2$ and O$_3$. For PM$_{2.5}$, the correlation was most improved by using the persistent enhancement. This suggests that the bulk of CO$_2$, NO$_2$, and O$_3$ from commercial land use is from short-lived events, whereas increased commercial land area consistently increases PM$_{2.5}$ over longer time scales.

These findings suggest that a more complex LUR that leverages different time resolutions in data may yield more accurate and more transferable results. The high temporal and spatial resolution of low-cost sensor networks are promising candidates for constructing these more complex models. Since completing this component of the CACES study, the network has expanded to 57 low-cost monitoring sites which will be used to assess the enhanced predictive power of an LUR constructed using different temporal resolutions. Such an LUR will be the focus of future work.

Some limitations of the study included incomplete ultrafine particle concentration data (i.e., at only some sites) and an inability to assess the seasonal differences at each site (seasonal analysis was limited to the continuously operated Oakland site). Ongoing work to address these limitations include ongoing mobile sampling around the 15 sites discussed in this study, and at many of the new sites as part of the expanded 50+ site network. Additionally, expansion of the ultrafine particle data set has been undertaken by rotating the CPCs to different sites. This
additional data is available in Saha et al. (2019). From continuous measurement from the 50+ site network we will also be better able to assess seasonal impact on site concentrations.

Ultimately, we believe that the lessons learned as part of the CACES AQO will inform the design of future approaches for measuring population exposure to air pollution, and in turn inform urban design and the siting and control of emissions sources. Additionally, the measurements collected as part of the CACES AQO will both corroborate and complement the results of national-scale modeling activities and LUR development conducted in other ongoing CACES research projects. Furthermore, the measurements generated in this work will produce a rich dataset for understanding how differences in modifiable factors affect the spatial distribution of numerous source-resolved pollution constituents and the resulting gradients in exposure.

ACKNOWLEDGEMENTS

This publication was developed under assistance agreement RD83587301 awarded by the U.S. Environmental Protection Agency as part of the Center for Air, Climate, and Energy Solutions. It has not been formally reviewed by EPA. The views expressed in this document are solely those of authors and do not necessarily reflect those of the Agency. EPA does not endorse any products or commercial services mentioned in this publication. Additional funding was provided by the
Heinz Endowments [Grants E2375 and E3145]. N. Zimmerman’s funding was provided by the Government of Canada NSERC Postdoctoral Fellowship [PDF-487660-2016].
REFERENCES


Sabaliauskas, K., Jeong, C.-H., Yao, X., Evans, G.J., 2014. The application of wavelet decomposition to quantify the local and regional sources of ultrafine particles in cities. Atmos. Environ. 95, 249–257. https://doi.org/10.1016/j.atmosenv.2014.05.035


FIGURE CAPTIONS

**Fig. 1:** A) Map of the stationary measurement sites in the Pittsburgh, PA region. The sites are labelled by neighborhood, with D1 = Downtown #1, D2= Downtown #2, D3= Downtown #3 and colored by general classification. General classifications of the sites are also provided in the list to the right of the map. The boundary for the city of Pittsburgh is shown in light brown. B) Prevailing wind direction in Pittsburgh, PA.

**Fig. 2:** Modifiable factors for a subset of the 15 stationary sampling sites. Bars are colored by the general site classifications of regional background (green), urban background (blue), traffic-dominated (grey), restaurant-dominated (purple) and high traffic and restaurants (red). The number to the right of the bars indicate the site ranking relative to the 15 stationary sampling sites (e.g., for Annual Average Daily Diesel Traffic on Nearest Road, Downtown #1 is 2nd highest, Fox Chapel is lowest, ranked 15).
Fig. 3: Sample of the wavelet decomposition for CO at the Downtown #1 site. A 3-level wavelet decomposition is used to separate short-lived events (yellow, <2 hr) from the overall signal. A 5-level decomposition is used to separate the regional signal (purple) and persistent enhancement (blue), which corresponds to changes in baseline occurring on the time scale of >8hrs. The persistent enhancement is isolated by taking the 5-level decomposition baseline and subtracting the regional signal. Longer-lived events (red, 2-8 hr) are determined from the difference between the 3-level decomposition and the persistent enhancement (see Equations 1-5).

Fig. 4: Box plots of the total pollutant concentrations (no decomposition) across the 15 stationary sites and entire measurement campaigns. Sites are ordered from upwind to downwind along the rural-urban-rural transect. Boxes are colored by the general site classification introduced in Figure 1. Only 8 sites had measurements of ultrafine particle concentration, those without are denoted ND for no data. For the box plots, whiskers = 10th and 90th percentile, box edges are 25th and 75th percentiles, the line is the median, and the circle markers are the averages.

Fig. 5: (a-f, left axis): Average pollutant concentrations above the regional background separated into persistent enhancement (blue), longer-lived events (red, 2-8h events), and short-lived events (yellow, <2h events) across the 15 stationary sites and entire measurement campaigns. (a-f, right axis, grey marker): The ratio of the total enhancement (persistent + short-lived + longer-lived) to the regional background. Only 8 sites had measurements of ultrafine particle concentration, those without are denoted ND for no data. For ozone, the persistent enhancement could be positive or negative due to the influence of local traffic on O3 concentrations.

Fig. 6: Sample diurnal patterns from two sites (Downtown #1 and Mt. Washington) and two pollutants (top: UFP, bottom: CO) for total signal (left), persistent enhancement (middle) and short-lived events enhancement (right). The total signal panel also shows the regional background (black trace).

Fig. 7: A Pearson correlation coefficient heat map between modifiable factors and the average concentrations, the persistent enhancement, the sporadic enhancement, and the combined persistent and short-lived enhancements. The correlation between the average concentration and the modifiable factors was consistently increased by correlating against one of the decomposed signals. The modifiable factor labels are described Fig. 2, Table S1, and Fig. S1. The * notation denotes a p-value <0.05.
FIGURES

Fig. 1

General Site Classification:
- Traffic/Diesel-dominated:
  - Carnegie
  - Aspinwall
- Restaurant-dominated:
  - Southside
  - Strip District

High Traffic, Restaurants:
- Downtown #1 - #3

Urban Background:
- Beechview
- Mt. Washington
- Hill District
- Lawrenceville
- Oakland
- Highland Park

Rural Background:
- Chartiers Valley
- Fox Chapel

Fig. 2

Vacant forest area (m²) in a 1000 m buffer

Commercial land use area (m²) in a 300 m buffer

Industrial land use area (m²) in a 300 m buffer

Annual Avg. Daily Diesel Traffic on Nearest Road

Legend:
- High Traffic, Restaurants
- Traffic/Diesel Dominated
- Urban Background
- Restaurant Dominated
- Rural Background
Fig. 3

Fig. 4
Fig. 7

Pollutant Signal Component

Land Use Covariates

Vacant Forest Commercial Elevation Industrial Diesel Population Density All Vehicles Restaurants

<table>
<thead>
<tr>
<th>Pollutant Signal Component</th>
<th>Vacant Forest</th>
<th>Commercial</th>
<th>Elevation</th>
<th>Industrial</th>
<th>Diesel</th>
<th>Population Density</th>
<th>All Vehicles</th>
<th>Restaurants</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average</td>
<td>-0.54*</td>
<td>0.33</td>
<td>-0.47</td>
<td>0.39</td>
<td>0.66*</td>
<td>0.01</td>
<td>0.63*</td>
<td>0.36</td>
</tr>
<tr>
<td>Persistent</td>
<td>-0.56*</td>
<td>0.3</td>
<td>-0.44</td>
<td>0.48</td>
<td>0.65*</td>
<td>-0.04</td>
<td>0.89*</td>
<td>0.18</td>
</tr>
<tr>
<td>Short-Lived</td>
<td>-0.52*</td>
<td>0.3</td>
<td>-0.48</td>
<td>0.39</td>
<td>0.64*</td>
<td>-0.16</td>
<td>0.77*</td>
<td>0.34</td>
</tr>
<tr>
<td>Short-Lived + Persistent</td>
<td>-0.56*</td>
<td>0.31</td>
<td>-0.45</td>
<td>0.47</td>
<td>0.66*</td>
<td>-0.06</td>
<td>0.89*</td>
<td>0.21</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Pollutant Signal Component</th>
<th>Vacant Forest</th>
<th>Commercial</th>
<th>Elevation</th>
<th>Industrial</th>
<th>Diesel</th>
<th>Population Density</th>
<th>All Vehicles</th>
<th>Restaurants</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average</td>
<td>-0.45</td>
<td>0.29</td>
<td>-0.37</td>
<td>0.29</td>
<td>0.59*</td>
<td>0.36</td>
<td>0.57*</td>
<td>0.24</td>
</tr>
<tr>
<td>Persistent</td>
<td>-0.45</td>
<td>0.23</td>
<td>-0.38</td>
<td>0.41</td>
<td>0.59*</td>
<td>0.16</td>
<td>0.83*</td>
<td>-0.02</td>
</tr>
<tr>
<td>Short-Lived</td>
<td>-0.08</td>
<td>0.0*</td>
<td>-0.31</td>
<td>-0.1</td>
<td>0.47</td>
<td>-0.05</td>
<td>0.01</td>
<td>0.37</td>
</tr>
<tr>
<td>Short-Lived + Persistent</td>
<td>-0.45</td>
<td>0.27</td>
<td>-0.39</td>
<td>0.4</td>
<td>0.61*</td>
<td>0.15</td>
<td>0.81*</td>
<td>0.01</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Pollutant Signal Component</th>
<th>Vacant Forest</th>
<th>Commercial</th>
<th>Elevation</th>
<th>Industrial</th>
<th>Diesel</th>
<th>Population Density</th>
<th>All Vehicles</th>
<th>Restaurants</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average</td>
<td>-0.28</td>
<td>0.52*</td>
<td>-0.28</td>
<td>-0.11</td>
<td>0.59*</td>
<td>-0.15</td>
<td>0.18</td>
<td>0.51</td>
</tr>
<tr>
<td>Persistent</td>
<td>-0.28</td>
<td>0.67*</td>
<td>-0.38</td>
<td>-0.2</td>
<td>0.6*</td>
<td>-0.16</td>
<td>0.29</td>
<td>0.44</td>
</tr>
<tr>
<td>Short-Lived</td>
<td>-0.3</td>
<td>0.44</td>
<td>-0.31</td>
<td>-0.09</td>
<td>0.68*</td>
<td>-0.31</td>
<td>0.28</td>
<td>0.37</td>
</tr>
<tr>
<td>Short-Lived + Persistent</td>
<td>-0.31</td>
<td>0.63*</td>
<td>-0.38</td>
<td>-0.17</td>
<td>0.67*</td>
<td>-0.23</td>
<td>0.3</td>
<td>0.45</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Pollutant Signal Component</th>
<th>Vacant Forest</th>
<th>Commercial</th>
<th>Elevation</th>
<th>Industrial</th>
<th>Diesel</th>
<th>Population Density</th>
<th>All Vehicles</th>
<th>Restaurants</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average</td>
<td>-0.56</td>
<td>0.75*</td>
<td>-0.72*</td>
<td>0.13</td>
<td>0.87*</td>
<td>-0.4</td>
<td>0.72*</td>
<td>0.63</td>
</tr>
<tr>
<td>Persistent</td>
<td>-0.57</td>
<td>0.58</td>
<td>-0.55</td>
<td>-0.11</td>
<td>0.8</td>
<td>-0.76</td>
<td>0.57</td>
<td>0.45</td>
</tr>
<tr>
<td>Short-Lived</td>
<td>-0.48</td>
<td>0.64</td>
<td>-0.59</td>
<td>0.12</td>
<td>0.61*</td>
<td>-0.39</td>
<td>0.62</td>
<td>0.46</td>
</tr>
<tr>
<td>Short-Lived + Persistent</td>
<td>-0.51</td>
<td>0.66</td>
<td>-0.61</td>
<td>0.02</td>
<td>0.82*</td>
<td>-0.52</td>
<td>0.65</td>
<td>0.54</td>
</tr>
</tbody>
</table>