



Uncertainty associated with ambient ozone metrics in epidemiologic studies and risk assessments

Benjamin Wells¹ · Heather Simon¹ · Thomas J. Luben² · Zachary Pekar¹ · Scott M. Jenkins¹

Received: 20 December 2018 / Accepted: 12 February 2019 / Published online: 7 March 2019

© This is a U.S. Government work and not under copyright protection in the US; foreign copyright protection may apply 2019

Abstract

Epidemiologic studies relating ambient ozone concentrations to adverse health outcomes have typically relied on spatial averages of concentrations from nearby monitoring stations, referred to as “composite monitors.” This practice reflects the assumption that ambient ozone concentrations within an urban area are spatially homogenous. We tested the validity of this assumption by comparing ozone data measured at individual monitoring sites within selected US urban areas to their respective composite monitor time series. We first characterized the temporal correlation between the composite monitor and individual monitors in each area. Next, we analyzed the heteroskedasticity of each relationship. Finally, we compared the distribution of concentrations measured at individual monitors to the composite monitor distribution. Individual monitors showed high correlation with the composite monitor over much of the range of ambient ozone concentrations, though correlations were lower at higher concentrations. The variance between individual monitors and the composite monitor increased as a function of concentration in nearly all the urban areas. Finally, we observed statistical bias in the composite monitor concentrations at the high end of the distribution. The degree to which these results introduce uncertainty into studies that utilize composite monitors depends on the contributions of peak ozone concentrations to reported health effect associations.

Keywords Ozone · Composite monitor · Exposure surrogate · Exposure measurement error

Introduction

Background

Short-term exposure to ozone (i.e., exposure to ozone concentrations averaged over hours or up to several days) is associated with a range of health effects including morbidity (e.g., asthma exacerbations, respiratory-related hospitalizations) and mortality (U.S. EPA 2013). A weight-of-evidence characterization of health effects related to short-term exposure to ozone utilizes evidence from various scientific disciplines, including

experimental studies and epidemiologic studies. Experimental studies have the advantage of following a strict experimental design and evaluating fixed concentrations of ozone without the potential for confounding by other pollutants. Such studies have reported respiratory effects following short-term exposures to ozone concentrations at or above 60 ppb (e.g., Adams 2002, 2003, 2006; Schelegle et al. 2009; Kim et al. 2011), with some evidence for a threshold dose rate (McDonnell et al. 2012). However small sample size and the general exclusion of sensitive individuals limit their ability to examine more severe health endpoints and lower-incidence health effects (U.S. EPA 2013, section 6.2.1.1). By contrast, epidemiologic studies focusing on short-term exposure to ozone benefit from utilizing larger study populations and a range of real-world ambient ozone concentrations. However, epidemiologic studies are subject to additional uncertainties, including potential confounding by other pollutants (e.g., NO₂, PM_{2.5}) and other environmental factors (e.g., allergens, temperature), as well as exposure measurement error linked to the use of ambient measurements obtained from outdoor monitors (e.g., U.S. EPA 2013, sections 6.2.7 and 6.2.8). It is this last factor that informs the research presented here.

Epidemiologic studies focusing on short-term exposure to ozone and the most severe health effects (e.g., hospital

Electronic supplementary material The online version of this article (<https://doi.org/10.1007/s11869-019-00679-8>) contains supplementary material, which is available to authorized users.

✉ Heather Simon
simon.heather@epa.gov

¹ Office of Air Quality Planning and Standards, US Environmental Protection Agency, 109 T.W. Alexander Dr., Research Triangle Park, NC 27711, USA

² National Center for Environmental Assessment, US Environmental Protection Agency, Research Triangle Park, NC, USA

admissions, mortality) typically use a time-series or case-crossover design where the temporal pattern of ozone levels in an urban area is compared to the temporal pattern in the occurrence of a specific health outcome in that area using a statistical model. Such studies have traditionally used ozone concentration data collected from ambient air quality monitors as the basis for characterizing population exposure in an urban area (Smith et al. 2009; Zanobetti and Schwartz 2008). More recently, some researchers have begun to use fused data based on a combination of gridded photochemical model and monitor data to characterize health effects associated with long-term (annual) (Turner et al. 2016) and short-term (daily) (O'Lenick et al. 2017) population exposure. However, monitor data alone continue to be used in many studies of short-term exposure to ozone.

In epidemiologic studies of short-term exposure, researchers often aggregate ozone measurements across available monitors in an urban area into a single time series. We refer to this aggregate time series as a “composite monitor” (Smith et al. 2009; Zanobetti and Schwartz 2008). The use of a composite monitor has been justified on the grounds that short-term averages of ozone concentrations often have high spatial correlation within an urban area. The potential for ozone to display high spatial correlation at the urban level reflects the fact that it is formed through secondary (photochemical) processes rather than being directly emitted. Studies have documented spatial correlation of short-term averages of ozone concentrations at the urban level by directly comparing monitor distributions (Ito et al. 2007) and through simulations relating monitor concentrations to simulated ozone surfaces (Goldman et al. 2011). However, in both cases, the analysis of spatial correlation has focused on mean trends and did not look at the degree to which this correlation holds across the full distribution of ozone concentrations.

It has been well documented that ozone chemistry can occur rapidly on local scales and may result in strong concentration gradients either due to rapid depletion near large sources of nitrogen oxides (Simon et al. 2016) or in rapid episodic formation due to the presence of highly reactive volatile organic compounds in narrow plumes from industrial sources (Nam et al. 2006), which suggests that the spatial correlation may not hold in all cases. To our knowledge, researchers have not systematically examined the degree to which short-term averages of ozone concentrations across the full ambient distribution are spatially correlated within an urban area.

In this study, we examine the degree to which composite monitors capture the spatial variability of ozone concentrations measured at individual monitors throughout an urban area, with a focus on the performance of composite monitors on days when ozone concentrations are relatively high. Given the strong support from experimental studies for respiratory effects following short-term exposures to ozone concentrations at or above 60 ppb, the

performance of composite monitors at higher ozone concentrations is relevant to the interpretation of the results of epidemiologic studies examining short-term exposures. There is evidence of increased uncertainty in characterizing the effect of short-term exposure to ozone on mortality at higher concentrations (Smith et al. 2009). The relative paucity of data at the high end of the ozone concentration distribution within an urban area may contribute to this uncertainty. However, if correlations between individual monitor measurements are lower on high ozone days, then the use of a composite monitor to represent population exposure at the urban level could introduce additional uncertainty, which could in turn bias effect estimates toward the null hypothesis. The ability to quantify that bias associated with the use of composite monitors across ozone concentrations in studies of short-term exposure to ozone would aid in reducing such uncertainty and improve epidemiologic studies of health effects associated with short-term ozone exposure.

This study explores the degree to which composite monitors used in ozone time-series studies are correlated with individual monitor distributions used in their derivation, with particular focus on the degree to which that correlation holds on higher ozone days. Additionally, we explore the degree to which composite monitors capture the upper-tail of the distribution of ozone concentrations measured at individual monitors.

Problem statement

Figure 1 shows examples of daily maximum 8-h ozone (MDA8) time series from two urban areas included in these analyses (i.e., Houston, TX and San Bernardino, CA) as well as the locations of individual ozone monitoring sites within each urban area. These time series demonstrate that while the composite monitor appears to generally capture very high and very low ozone concentrations measured at individual monitors in Houston during July 1996, there is much less correlation between individual ozone monitors (and consequently individual monitors with the composite monitors) in San Bernardino in August 1998. The San Bernardino time series shows many instances where different monitors measure peak ozone concentrations on different days (especially during the first half of the month), and as a result, the composite monitor time series is comparatively flat. The Houston example also shows an instance of greater spatial variability between monitors at the beginning of the month when ozone levels are higher, despite the composite monitor's apparent ability to capture high versus low ozone concentrations at individual monitors. Based on these observations, we hypothesize that, while individual ozone monitors and composite monitors may be reasonably correlated

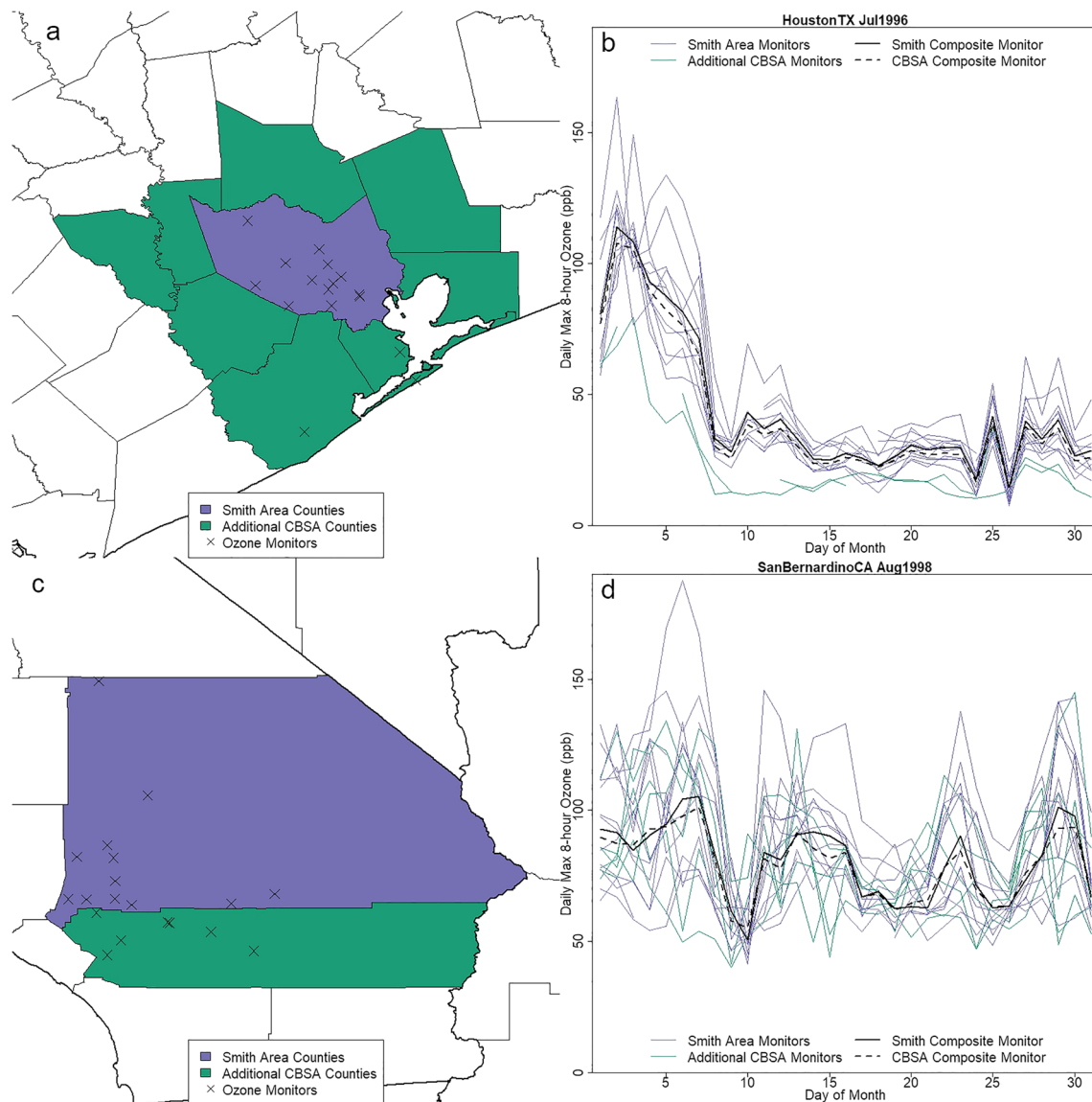


Fig. 1 Maps of ozone monitors (shown as x's) in the Houston (a) and San Bernardino (c) areas. Areas used in the Smith et al. (2009) study are shown in purple while additional counties included in the larger Core Based Statistical Area (CBSA) are shaded in green. One-month time series of 8-h daily maximum ozone values at individual monitors are

shown for Houston (b) and San Bernardino (d). The composite monitor time series created using ozone monitors located in the Smith study areas are shown with solid black lines while the composite monitor time series created using ozone monitors located in the CBSAs are shown with dashed black lines

based on comparison of mean trends, monitors are less well correlated at higher ambient ozone concentrations, which means that epidemiologic studies utilizing composite monitors could introduce uncertainty into effect estimates, particularly to the degree health effects are driven by higher ozone levels.

A related question is whether the composite monitor captures the range of ozone concentrations represented by the individual monitor values. While the composite monitor can never be expected to capture the highest concentration at an individual monitor, it is important that it captures the average “high” value across monitors. However, as demonstrated by the San Bernardino time series in Fig. 1d and the first week of

the Houston time series in Fig. 1b individual monitors often peak on different days. This disparity in the timing of peak ozone concentrations across an urban area could result in the composite monitor being systematically biased low across the upper end of the ozone distribution (e.g., the 98th percentile composite monitor value could be lower than the average of the 98th percentiles from the individual monitors). This could mute any statistical relationship derived between peak ozone levels and health effects and therefore impact the risk effect estimates in time-series epidemiologic studies. Thus, we additionally hypothesize that composite monitors show systematic bias at the high end of the ozone distribution when compared to the individual monitors.

Methods

We identified urban areas in the USA that have been used in previous epidemiologic studies (Smith et al. 2009; Zanobetti and Schwartz 2008) and have ozone monitoring networks that support generation of daily time series. We then evaluated the degree to which the composite monitor (in a given urban area) is correlated with its set of individual monitors, with particular emphasis on (a) whether that correlation holds across ambient ozone concentrations, and (b) whether the composite monitor captures the highest ozone measurements reported at the individual monitors. We also looked at the degree to which this correlation between composite and individual monitors is affected by consideration for (a) different short-term exposure metrics (1-h max, 8-h max, 24-h avg), (b) different definitions of the ozone season, (c) different urban spatial scales, and (d) whether this relationship has changed over time.

Data selection

We utilized ozone measurement data available in EPA's Air Quality System (AQS; <https://www.epa.gov/aqs>) to construct ozone time series for individual monitoring sites in select urban areas. We also used these data to construct composite monitor time series consistent with methodology used in a previous study (Smith et al. 2009). For our base dataset, we began by identifying urban areas that had been used in the Smith et al. (2009) study, which was also used as the basis for the risk modeling presented in the most recent Ozone Risk and Exposure Assessment (U.S. EPA 2014). We retrieved hourly ozone concentration data collected during April–October of 1996–2000, which corresponds to the last 5 years

of the time period used in the Smith study (1987–2000). We chose to use the later portion of the Smith et al. (2009) time period because the extent of the ozone monitoring network increased substantially during the 1990s, allowing for a larger and more consistent set of ozone monitors in our analyses. For each urban area, we selected monitors that were located within the counties identified in the Smith study that met certain data completeness criteria for this time period for use in our analyses. To meet the data completeness criteria, a monitor must have at least 14 days with at least 18 hourly measurements in every month during the April–October season, for each year in the 5-year period. Finally, we focused on 11 urban areas (listed in Table 1 below) that had at least 5 monitors meeting these criteria. This selection was intended to focus on areas where there were enough monitoring sites to make a correlation statistic meaningful (i.e., with too few monitors, comparing an individual monitor time series to its own composite monitor time series ceases to be meaningful). Note that these cities are not evenly spread through the country but are instead concentrated mostly in the West and South with the inclusion of one Midwestern city (Chicago). Maps of these 11 areas including monitor locations are shown in Fig. 1 (2 areas) and the Supplemental Information (Fig. S1, 9 areas).

In addition to the base dataset selected for the core analysis, we also explored the robustness of our conclusions to several choices defining the temporal and spatial nature of our dataset. We created three additional datasets for this purpose:

1. We selected monitors from the larger Core Based Statistical Area (CBSA) rather than the smaller core urban area used in Smith et al. (2009). There were an additional 21 areas including many additional areas in the Eastern

Table 1 Information regarding the 11 urban areas included in the base dataset

Area	Number of ozone monitors ^a	Population in 2000 (thousands)	Land area (km ²)	MDA8 ozone distribution in Smith et al., Apr–Oct, 1996–2000 dataset (ppb)					
				5th %	25th %	50th %	75th %	95th %	max
Bakersfield, CA	7	840	21,062	39	55	68	82	100	137
Birmingham, AL	5	658	2878	22	36	48	61	80	122
Chicago, IL	12	5195	2448	13	26	36	48	67	119
Dallas, TX	5	5038	7005	25	39	51	66	87	129
Fresno, CA	5	930	15,433	38	53	67	82	101	135
Houston, TX	9	4092	4414	19	30	42	61	90	164
Los Angeles, CA	12	9819	10,510	23	36	46	59	85	172
Phoenix, AZ	5	3817	23,826	36	48	57	66	79	106
San Bernardino, CA	8	2035	51,947	37	52	65	80	111	206
San Diego, CA	8	3095	10,895	30	40	47	56	74	142
Tucson, AZ	5	980	23,794	31	43	51	58	69	85

^a Only counts monitors with complete data during the time period of the dataset as described above

USA that met the 5-monitor minimum requirement when we allowed for monitors outside the immediate urban core.

2. We selected monitoring data that covered all months of the year instead of the shorter April–October monitoring season. Since some states do not operate some or all of their ozone monitors during the winter months, there were 5 fewer urban areas that met the 5-monitor minimum criteria.
3. We selected monitoring data from a more recent 5-year time period: 2011–2015. Since some states have changed the number of monitors and/or operating season between 1996 and 2015, this dataset included 4 additional areas that were not in the base dataset but did not include 1 area that was in the base dataset.

A map of the urban areas included in the base dataset and the three additional datasets is shown in Fig. 2.

Daily ozone metrics and composite monitor time series

For each monitor meeting the data completeness criteria listed in the previous section, we calculated three daily ozone metrics (1-h max, 8-h max, 24-h average). These metrics were calculated as follows:

1. 1-h max: maximum hourly ozone concentration reported on a given day. For days with fewer than 18 hourly measurements (75%), we did not calculate a 1-h max value.
2. 8-h max: maximum 8-h average ozone concentration for a given day. Moving 8-h averages were calculated for each

8-h period starting with 12:00 AM–8:00 AM, going forward to 11:00 PM–7:00 AM the next day. For 8-h periods with fewer than 6 hourly measurements (75%), we did not calculate an 8-h average. For days with fewer than 18 8-h averages (75%), we did not calculate an 8-h max value.

3. 24-h average: average of the 24 hourly concentrations reported on a given day. For days with fewer than 18 hourly measurements (75%), we did not calculate a 24-h average.

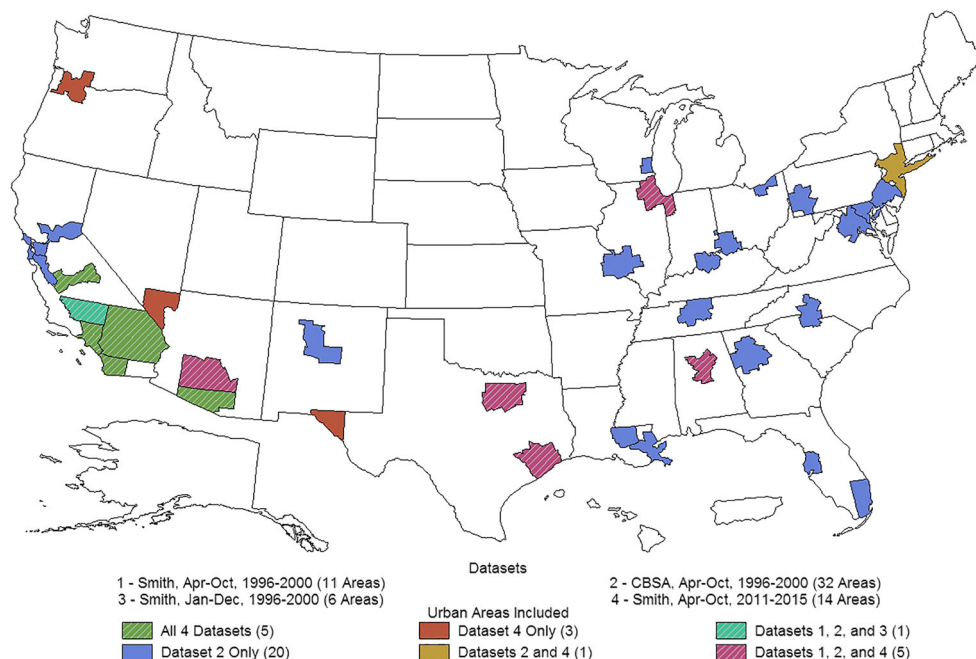
Following the data handling procedures used in the Smith study, composite monitor hourly averages were constructed for each urban area by averaging the hourly concentration values across all monitors meeting the data completeness criteria listed in the previous section. The three daily metrics listed above were then calculated using the hourly composite monitor values for each urban area. It should be noted that while the data completeness criteria were meant to minimize inconstancy across days in which data were included in the composite monitor, there are still some days with missing data from one or more monitors meaning that the composite monitor would be calculated from a subset of monitors in the urban area on those days.

Data analysis

Correlation between individual monitors and composite monitor

For each urban area, we calculated a single pooled Pearson's correlation coefficient (r) between the collective individual monitor time series, and the composite monitor time series.

Fig. 2 Map showing the locations of the urban areas in the various datasets used in the analyses. Dataset 1 is the base dataset used in the core analysis (includes all areas shown with gray hatch marks), and datasets 2, 3, and 4 are the three alternative datasets described above



This statistic was calculated using all days, then re-calculated after eliminating days where all monitors had 8-h max values below thresholds of 40, 50, 60, 70, and 80 ppb. There were at least 100 total days included in each calculation, with two exceptions for the 80 ppb threshold in Chicago (48 days) and Tucson (8 days). The total number of days in each urban area above each threshold in the base dataset is shown in the supplemental information (Table S1). In comparing correlations across metrics, thresholds or datasets, we used the Welch's t test (Welch 1947) to determine if the mean of the distribution of r values across cities were statistically different between metrics/thresholds/datasets with a significance threshold of $\alpha = 0.05$.

Individual monitor variance as a function of concentration

It is possible that any observed change in correlation detected in the threshold analysis described above is at least partially due to the fact that we are cutting off a portion of the observed variability in the range of ozone concentrations being considered and reducing the sample size in the process of looking across high days. However, a decrease in correlation at higher thresholds may also indicate that ozone values at individual monitors are less well correlated to each other and to the composite monitor on high days than on low days. If the latter assertion were true, we would expect to see an increase in the spatial variability between individual ozone monitors on high ozone days. Therefore, we performed an analysis examining the spatial variability across individual ozone monitors as a function of the composite monitor value.

We first fit a linear regression between the individual monitor values and the composite monitor. The residuals from this regression were then used to conduct a Breusch-Pagan test (Breusch and Pagan 1979) for the presence of heteroskedasticity, or non-constant variance, with a significance threshold of $\alpha = 0.05$. However, the Breusch-Pagan test is not able to determine whether the variance is increasing or decreasing as a function of the composite monitor. Therefore, we implemented a second step in the analysis which started by binning the composite monitor values into 1 ppb bins (e.g., the 20 ppb bin consisted of all values where the composite monitor was greater than or equal to 20 ppb and less than 21 ppb). We then took the standard deviation of all individual monitor values used to construct the composite monitor values in each bin. Finally, we fitted a weighted least squares regression of the standard deviations on the bins, with the number of individual monitor values in each bin as the weights. Figure 4 in the “Results” section provides an illustration of this methodology.

For each urban area, if the Breusch-Pagan statistic was significant ($p < 0.05$) and the slope of the weighted least squares line was positive, then we concluded that the variance in individual monitors increased as a function of the

composite monitor value. Finally, we conducted a paired t test to determine if there was a difference in the mean slope of the weighted least squares regression model between the base dataset and each of the three alternative datasets, using cities common to both datasets and a significance threshold of $\alpha = 0.05$.

Distribution of individual monitor values and composite monitor values

To compare the distribution of the composite monitor to those of the individual monitors, we first calculated the 50th, 75th, 90th, 95th, 98th, and 99th percentiles of the 8-h max metric at each individual monitor and the composite monitor for each urban area. Then within each urban area, we calculated the mean value across monitors for each percentile. For example, in Chicago, we averaged the 75th percentile values determined for each of the 12 monitors to come up with a single area-wide average 75th percentile value. We then took the difference between the composite monitor value and the average individual monitor value for each percentile within each urban area. To assess the presence of statistical bias, we used a one-sided Student's t test to determine if the mean difference in each percentile across the 11 urban areas in the core analysis was significantly different from 0 with a significance threshold of $\alpha = 0.05$.

Results

Correlation between composite and individual monitors by metric and concentration

Figure 3 shows boxplots of the pooled Pearson's correlation coefficients between individual monitors and composite monitors by ozone metric and by ozone concentration. Each boxplot represents the range of values calculated across the 11 urban areas (i.e., one value per urban area). The three leftmost boxplots show correlation coefficients when looking across all days for three metrics: 1-h max, 8-h max, and 24-h average. All correlation coefficients are high with mean r values of 0.87, 0.88, and 0.80 respectively. This indicates that the composite monitor generally does a reasonable job of capturing the urban-scale spatial variability represented by individual monitors on most days although we note that the correlation for the 24-h metric is skewed with substantially lower median correlation values. While the correlation coefficients for the 1-h max and 8-h max metrics were not significantly different from one another, the correlation values derived for the 24-h average metric were significantly different from the other two metrics with a p value of 0.04.

Figure 3 demonstrates that when looking only across high ozone days, the correlation between individual monitors and

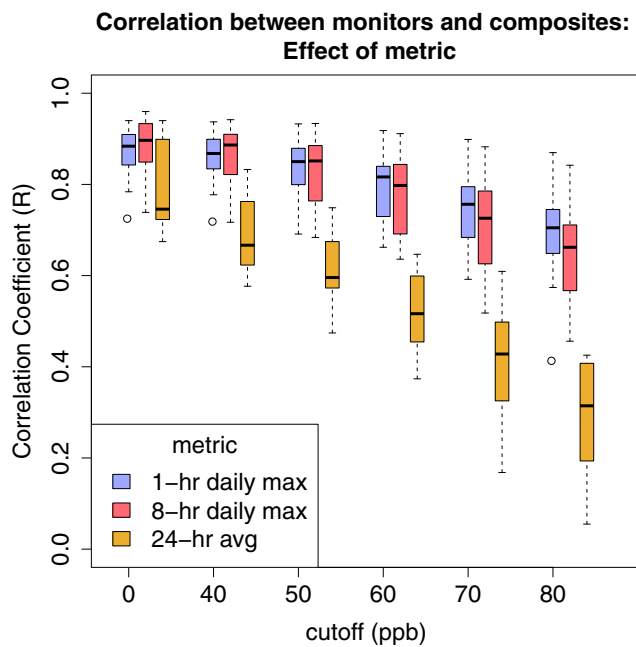


Fig. 3 Boxplots showing the correlation between individual monitor ozone values and the composite monitor values across 11 urban areas in base dataset by ozone metric and concentration threshold. The boxes represent the interquartile range across cities with the median value shown as the horizontal line. The whiskers extend to 1.5 times the interquartile range. Outliers are shown as circles

the composite monitor decreases. For instance, while the mean correlation value for the 8-h max metric across the 11 urban areas was 0.88 using all days, but this value drops to 0.71 when looking only at days where one or more monitors measured concentrations above 70 ppb (the level of the current US ozone standard). The decrease in correlation at higher concentration thresholds is significant for the 8-h max and 1-h max, but it is especially pronounced for the 24-h average metric which has a mean correlation value of only 0.40 using the 70 ppb threshold.

To evaluate whether our results were impacted by size of the urban area, season, or years of data, we conducted a sensitivity analysis using three additional datasets (shown in Figs. S2–S4 in the Supplemental Information). Using the Welch's *t* test, we saw no statistical difference in the distribution of correlation coefficients between any of these datasets and our base dataset either when using a common set of urban areas or when expanding to include additional cities in the other datasets. The sensitivity analysis that examined size of the urban area can additionally provide some insight into the impact of geographic representativeness of this dataset. As mentioned previously, the base dataset was heavily concentrated in the Southern and Western regions of the USA. When including the larger CBSA (dataset 2) instead of only the urban core, there were many additional areas with at least 5 monitors

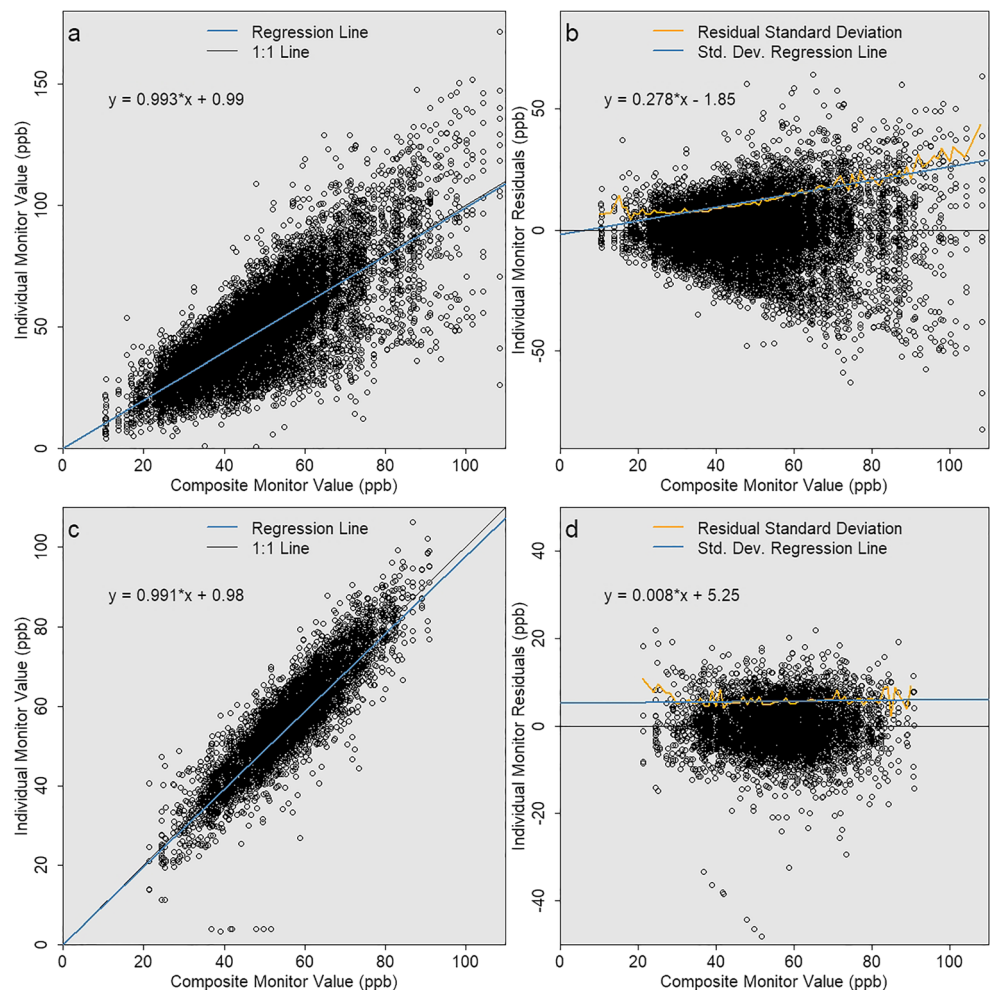
meeting our data completeness criteria including 4 new areas in the Northeast/Mid-Atlantic, 6 new areas in the Midwest, 5 new areas in the Southeast, 1 new area in the Southwest, and 3 new areas in California (see Fig. 2). Comparison of the red and orange bars in Fig. S2 shows that the results are largely unchanged between the original 11 areas and the new dataset with more representative geographic coverage. Finally, the sensitivity which examined the impact of using 2011–2015 data (Fig. S4) shows that while the general patterns are similar across time, the distributions are much tighter with more recent data.

Variance in individual monitors as a function of the composite monitor

Figure 4 shows two examples of the variance analysis, described previously in the “Methods” section, applied to daily maximum 8-h values in Los Angeles and Phoenix. The left-hand panels show scatter plots of the individual monitors versus composite monitor values for Los Angeles and Phoenix with the linear regression used to conduct the Breusch-Pagan test in blue. The right-hand panels show residual plots derived from the linear regressions in the left-hand plots for the two cities, with the binned standard deviation values in orange and the weighted regression line in blue. The top two panels show a clear pattern of increasing variance in the individual monitor values as a function of the composite monitor value in Los Angeles, as evidenced by the increasing scatter about the regression line in the left-hand panel and increasing scatter about the regression line in the right-hand panel. Conversely, the bottom two panels do not indicate any clear trend in variance as a function of the composite monitor in Phoenix, as evidenced by the relatively constant levels of scatter over the entire range of concentrations.

The full set of results for all urban areas in the base dataset and the three alternative datasets are shown in Tables S2–S5 in the Supplemental Information. To summarize this information, starting with the 11 areas in the base dataset, 9 areas showed increasing variance in individual monitor values as a function of the composite monitor (as illustrated by the Los Angeles panels) and 2 areas (Phoenix and Tucson, AZ) did not show significant heteroskedasticity. Looking at the alternative dataset which used the CBSA instead of the smaller urban core study area, 29 of 32 areas showed increasing variance, while the remaining 3 areas (Albuquerque, NM; Phoenix and Tucson, AZ) did not show significant heteroskedasticity. The paired *t* test did not show a significant difference in mean slope for the 11 areas in the base dataset ($p = 0.88$). For the alternative dataset which used full-year data in place of the April–October season, 5 of 6 areas showed increasing variance while the remaining area (Tucson, AZ) showed decreasing variance as a function of the composite monitor. The

Fig. 4 Scatter plots of the composite monitor and individual monitor daily maximum 8-h values for Los Angeles (a) and Phoenix (c) with regression lines shown in blue. Residual plots for Los Angeles (b) and Phoenix (d) with binned standard deviations shown in orange and regression lines shown in blue



paired t test showed a significant decrease in mean slope compared to the base dataset for the 6 areas included in both datasets ($p = 0.007$), indicating that the increase in variance is smaller when including data from fall and winter months which typically have lower ozone concentrations. Finally, all 15 areas in the alternative dataset looking at more recent monitoring data (years 2011–2015 instead of 1996–2000) showed increasing variance, but the paired t test did not indicate a significant difference in the mean slope ($p = 0.43$). It is also noteworthy that all areas which did not exhibit increasing variance as a function of the composite monitor were located in the desert Southwest region of the USA.

Degree to which composite monitors capture high-end peak ozone levels

The left-hand panel of Fig. 5 shows the cumulative ozone distribution for individual monitors and the composite monitor in Houston, TX. The symbols on the far right of the distribution show the maximum ozone concentration for individual monitors and the composite

monitor. It is clear in this figure that the composite monitor maximum value is at the lower end of the maximum ozone values observed across individual monitors. At the 90th percentile, the composite monitor value appears biased low compared to the values across individual monitors. The right-hand panel of Fig. 5 shows the differences between the composite monitor distribution and the mean of the individual monitor distributions. Again, the boxplots show the distribution of these differences across urban areas (one data point for each area). All of these differences are statistically different from zero except for the 50th percentile which has a difference of 0.26 ppb and a p value of 0.06. The average difference is larger for the higher percentiles and reaches 4.18 ppb at the 98th percentile and 5.4 ppb at the 99th percentile (differences are even larger for some individual areas). This shows that across the 11 urban areas in our core analysis, the upper end of the ozone distribution represented by the composite monitor is systematically biased low compared to the upper end of the ozone distribution measured at individual monitors.

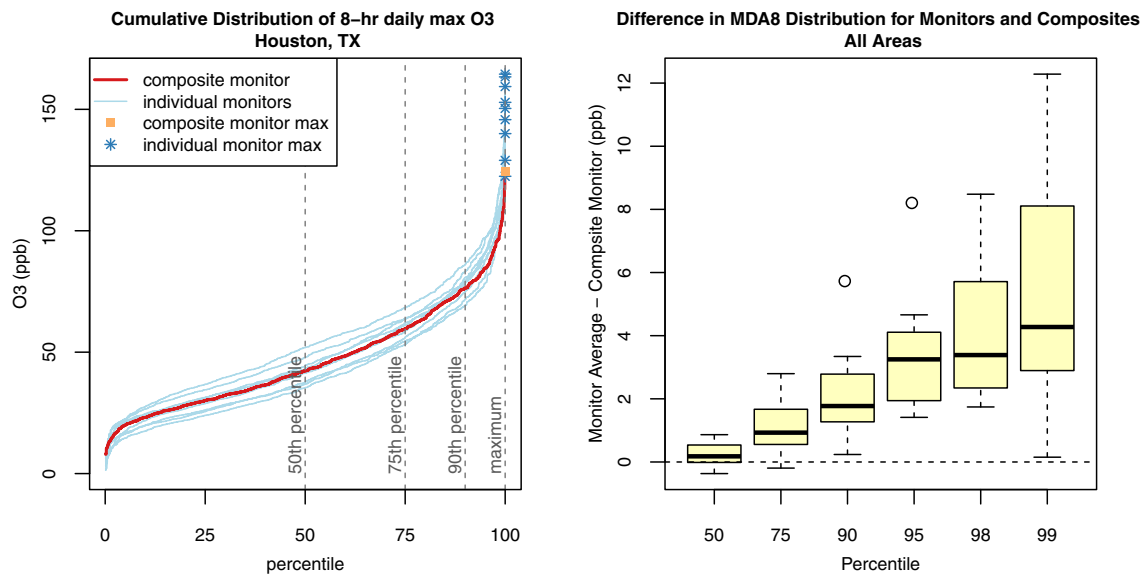


Fig. 5 Example showing distribution of ozone concentrations for individual monitors and composite monitor in Houston, TX (left). Boxplots showing the average difference between the individual monitors and composite monitor at various distributional percentiles for

the 11 areas in the base dataset (right). The boxes represent the interquartile range across cities with the median value shown as the horizontal line. The whiskers extend to 1.5 times the interquartile range. Outliers are shown as circles

Discussion

Overall, our results demonstrate that ozone concentrations from composite monitors are good indicators of the area-wide ozone concentrations they are meant to represent. This is true across different geographic regions, across ozone metrics, and across most of the distribution of measured ozone concentrations. These findings are consistent with earlier studies examining mean trends in monitored ozone concentrations in the urban context (Goldman et al. 2011; Ito et al. 2007). As illustrated in Fig. 3, the overall correlation between individual monitors and the composite monitor is quite high (i.e., most values are greater than 0.9). However, we note that the correlation between individual monitors and composite monitors is substantially reduced on higher ozone days (e.g., days with at least one monitor with ozone measurements > 70 ppb) suggesting that there is the potential for composite monitors to contribute increased exposure measurement error on high ozone days. This reduction is particularly pronounced for the 24-h metric. Our finding that the correlations between individual and composite monitors decrease at higher ozone levels is corroborated by our results showing a significant increase in variance in ozone measurements at individual monitors as the composite monitor value increases for most urban areas, as illustrated in the plots for Los Angeles presented in Fig. 4. However, we did observe different degrees of spatial variability across geographic locations in these results with the subset of study areas found in the desert southwest not showing this pattern of increased variance, as illustrated in the plots for Phoenix presented in Fig. 4.

Our findings further suggest that, in addition to reduced correlation between composite monitors and individual monitors at higher concentrations, composite monitors may not perform as well in capturing peak ozone levels. The results illustrated in Fig. 5 suggest that high-end composite monitor values for a specific percentile are systematically lower than the average values across individual monitors for that same percentile. This demonstrates that the composite monitor does not capture the full range of concentrations observed at individual monitors.

Our results suggest that epidemiologic studies of short-term exposure to ozone that use composite monitors may be subject to exposure measurement error, particularly at higher concentration levels. If morbidity and mortality associated with short-term exposure to ozone are largely driven by high ambient ozone concentrations, then exposure measurement error associated with capturing patterns of exposure on high-ozone days (due to the use of composite monitors) could result in the effect estimate generated being biased toward the null. This is of particular concern for epidemiologic studies using the 24-h metric. The ability to quantitatively estimate the bias introduced by using composite monitors to assign exposure in epidemiologic studies of short-term exposure to ozone would reduce uncertainties associated with these studies. By identifying and characterizing this issue, including real-world, data-driven examples, we have provided an initial step toward this goal.

The degree to which the use of composite monitors to assign exposure results in exposure measurement error depends in part on the spatial and temporal patterns of

ambient ozone concentrations and on the behavior of the populations being studied. With regard to spatial and temporal patterns, NO_x emissions controls can both increase and decrease ozone across a single urban area (Simon et al. 2016). In many cases, local ozone titration can deplete ozone near large sources of NO_x emissions. The highest ozone concentrations occur downwind after the high NO_x concentrations have mixed in with VOC precursors and after there has been sufficient time of ozone formation reactions to occur. Consequently, NO_x reductions may lead to localized increases in ozone, often in densely populated urban cores, while also decreasing ozone concentrations at the highest downwind monitor in the area (Cleveland and Graedel 1979; Sillman 1999; Xiao et al. 2010; Kelly et al. 2015; Simon et al. 2016). This divergent impact of emissions controls within an urban area may help explain some of our results.

With regard to the behavior of populations, previous studies have shown that mobility can impact personal exposure to pollutants and in turn can impact measurement exposure error (Marshall et al. 2006; Setton et al. 2011; Dias and Tchepel 2014; Steinle et al. 2015). Given that finding, if individuals in the study population have limited mobility, and assuming those individuals are not uniformly distributed across the study area, there is increased potential for the use of composite monitors to introduce error into exposure estimates. In contrast, even if a study area does have zones of higher ozone and zones of lower ozone, more mobile populations that combine time spent across multiple zones would be expected to have a time-averaged exposure profile that is reasonably approximated by a composite monitor metric. Thus, if the study population has relatively high mobility, then the use of composite monitors may introduce less exposure measurement error.

As noted earlier, researchers have recently begun to supplement composite monitor-based time-series epidemiologic studies with exposure assessment based on modeled ozone predictions. While the research presented here is mainly relevant in interpreting potential error associated with time-series studies using composite monitors to assign exposure, it does have implications for epidemiologic studies that use models to assign ozone exposure. Specifically, if our research identifies the potential for substantial exposure measurement error at higher ozone concentrations, then an argument could be made for the use of alternate methods of creating spatially varying ozone surfaces or modeled ozone exposure in epidemiologic studies, since these conceptually could reduce this source of exposure measurement error.

These findings also have implications for risk assessments conducted using effect estimates obtained from these types of epidemiologic studies. Namely, if exposure measurement

error related to the use of composite monitors is found to impact effect estimates (e.g., bias them toward the null), then risk estimates generated using those effect estimates will be biased in a similar fashion. Furthermore, if a risk assessment is intended to characterize reductions in risk associated with reducing ozone on high days, then additional uncertainty on higher ozone days would have a relatively greater effect on the risk estimates generated. The fact that some emissions controls which are necessary to reduce peak ozone concentrations can also lead to ozone increases at times and locations where ozone concentrations are low (US EPA 2014; Simon et al. 2016) makes it especially important to accurately characterize the magnitude of ozone impacts on health outcomes for different portions of the ozone distribution.

The results of our study suggest the importance of future research exploring the degree to which health effects associated with short-term exposure to ozone are driven by peak ozone concentrations. Should we find that health effects are driven by short-term (peak) ozone concentrations, then, in cases where time series include a substantial number of high days, the potential for introducing exposure measurement error into time-series studies through the use of composite monitors would be increased and effects estimates could be biased toward the null.

There are other factors that should be considered in interpreting our findings. First, the degree and nature of any exposure measurement error associated with use of composite monitors may vary geographically, as evidenced by the lack of heteroskedasticity in the southwest. While, our base dataset contained only 11 urban areas, with a geographic bias toward the southern and western parts of the USA, sensitivity analyses looking at a wider set of areas showed consistent results with this analysis. Second, our results suggest that the degree to which exposure measurement error results from the use of composite monitors does depend on the exposure metric used, with the 24-h average metric having relatively greater disconnect between composite monitors and individual monitors.

In summary, our results continue to demonstrate that composite monitors are relatively good indicators of the area-wide ozone concentrations they are meant to represent. To the extent that time-series studies are conducted in locations where most days have higher ambient ozone concentrations, there is more potential that the use of composite monitors could result in exposure measurement error that would likely result in bias of the effect estimate toward the null. Future research that evaluates the degree to which short-term (peak) exposures to higher ambient ozone concentrations (e.g., 60–80 ppb) are responsible for reported associations with mortality and serious morbidity outcomes would be informative. Additional research that utilizes more detailed activity profiles and more spatially and temporally refined ambient ozone fields to derive more representative and detailed ozone exposure profiles could address this issue.

Acknowledgements The authors would like to thank Pat Dolwick, Jennifer Richmond-Bryant, Elizabeth Naess, Richard Wayland, James Hemby, Jackie Ashley, and Michael Koerber at the US Environmental Protection Agency (EPA) for their thoughtful review and comment on this article. Although this paper has been reviewed by the US EPA and approved for publication, it does not necessarily reflect the US EPA's policies or views. The authors declare no competing financial interests.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

References

- Adams WC (2002) Comparison of chamber and face-mask 6.6-hour exposures to ozone on pulmonary function and symptoms responses. *Inhal Toxicol* 14:745–764
- Adams WC (2003) Comparison of chamber and face mask 6.6-hour exposure to 0.08 ppm ozone via square-wave and triangular profiles on pulmonary responses. *Inhal Toxicol* 15:265–281
- Adams WC (2006) Comparison of chamber 6.6-h exposures to 0.04–0.08 ppm ozone via square-wave and triangular profiles on pulmonary responses. *Inhal Toxicol* 18:127–136
- Breusch TS, Pagan AR (1979) A simple test for heteroskedasticity and random coefficient variation. *Econometrica* 47(5):1287–1294
- Cleveland WS, Graedel TE (1979) Photochemical air pollution in the Northeast United States. *Science* 204:1273–1278
- Dias D, Tchepel O (2014) Modelling of human exposure to air pollution in the urban environment: a GPS-based approach. *Environ Sci Pollut Res* 21(5):3558–3571
- Goldman GT, Mulholland JA, Russell AG, Strickland MJ, Klein M, Waller LA, Tolbert PE (2011) Impact of exposure measurement error in air pollution epidemiology: effect of error type in time-series studies. *Environ Health* 10:61
- Ito K, Thurston GD, Silverman RA (2007) Characterization of PM_{2.5}, gaseous pollutants, and meteorological interactions in the context of time-series health effects models. *J Expo Sci Environ Epidemiol* 17: S45–S60
- Kelly JT, Baker KR, Napelenok SL, Roselle SJ (2015) Examining single-source secondary impacts estimated from brute-force, decoupled direct method, and advanced plume treatment approaches. *Atmos Environ* 111:10–19
- Kim CS, Alexis NE, Rappold AG, Kehrl H, Hazucha MJ, Lay JC, Schmitt MT, Case M, Devlin RB, Peden DB, Diaz-Sanchez D (2011) Lung function and inflammatory responses in healthy young adults exposed to 0.06 ppm ozone for 6.6 hours. *Am J Respir Crit Care Med* 183:1215–1221
- Marshall JD, Granvold PW, Hoats AS, McKone TE, Deakin E, Nazaroff WW (2006) Inhalation intake of ambient air pollution in California's South Coast Air Basin. *Atmos Environ* 40:4381–4392
- McDonnell WF, Stewart PW, Smith MV, Kim CS, Schelegle ES (2012) Prediction of lung function response for populations exposed to a wide range of ozone conditions. *Inhal Toxicol* 24:619–633
- Nam J, Kimura Y, Vizuete W, Murphy C, Allen DT (2006) Modeling the impacts of emission events on ozone formation in Houston, Texas. *Atmos Environ* 40(28):5329–5341
- O'Lenick CR, Chang HH, Kramer MR, Winquist A, Mulholland JA, Friberg MD, Samat SE (2017) Ozone and childhood respiratory disease in three US cities: evaluation of effect measure modification by neighborhood socioeconomic status using a Bayesian hierarchical approach. *Environ Health* 16:36
- Schelegle ES, Morales CA, Walby WF, Marion S, Allen RP (2009) 6.6-hour inhalation of ozone concentrations from 60 to 87 parts per billion in healthy humans. *Am J Respir Crit Care Med* 180:265–272
- Setton E, Marshall JD, Brauer M, Lundquist KR, Hystad P, Keller P, Cloutier-Fisher D (2011) The impact of daily mobility on exposure to traffic-related air pollution and health effect estimates. *J Expo Sci Environ Epidemiol* 21:42–48
- Sillman S (1999) The relation between ozone, NO_x and hydrocarbons in urban and polluted rural environments. *Atmos Environ* 33:1821–1845
- Simon H, Wells B, Baker KR, Hubbell B (2016) Assessing temporal and spatial patterns of observed and predicted ozone in multiple urban areas. *Environ Health Perspect* 124:1443–1452
- Smith RL, Xu B, Switzer P (2009) Reassessing the relationship between ozone and short-term mortality in U.S. urban communities. *Inhal Toxicol* 21:37–61
- Steinle S, Reis S, Sabel CE, Semple S, Twigg MM, Braban CF, Leeson SR, Heal MR, Harrison D, Lin C, Wu H (2015) Personal exposure monitoring of PM_{2.5} in indoor and outdoor microenvironments. *Sci Total Environ* 508:383–394
- Turner MC, Jerret M, Pope CA et al (2016) Long-term ozone exposure and mortality in a large prospective study. *Am J Respir Crit Care Med* 193:1134–1142
- U.S. EPA (2013) Integrated Science Assessment (ISA) of Ozone and Related Photochemical Oxidants (Final Report, Feb 2013). U.S. Environmental Protection Agency, Washington, DC, EPA/600/R-10/076F, 2013
- U.S. EPA (2014) Health risk and exposure assessment for ozone (final report, Aug 2014) U.S. Environmental Protection Agency, Research Triangle Park, EPA/452/R-14-00-004a
- Welch BL (1947) The generalisation of students problem when several different population variances are involved. *Biometrika* 34:23–35
- Xiao X, Cohan DS, Byun DW, Ngan F (2010) Highly nonlinear ozone formation in the Houston region and implications for emission controls. *J Geophys Res* 115:D23309. <https://doi.org/10.1029/2010JD014435>
- Zanobetti A, Schwartz J (2008) Mortality displacement in the association of ozone with mortality: an analysis of 48 cities in the United States. *Am J Respir Crit Care Med* 177:184–189