Design and evaluation of mobile monitoring campaigns for air pollution exposure assessment in epidemiologic cohorts

Magali N. Blanco,[a] Annie Doubleday,[a] Elena Austin,[a] Julian D. Marshall,[b] Edmund Seto,[a] Timothy Larson,[a,b] Lianne Sheppard[a,c]

[a]Department of Environmental and Occupational Health Sciences, School of Public Health, University of Washington, Hans Rosling Center for Population Health, 3980 15th Ave NE, Seattle, WA 98195

[b]Department of Civil & Environmental Engineering, College of Engineering, University of Washington, 201 More Hall, Box 352700, Seattle, WA 98195

[c]Department of Biostatistics, School of Public Health, University of Washington, Hans Rosling Center for Population Health, 3980 15th Ave NE, Seattle, WA 98195

Abstract

Mobile monitoring makes it possible to estimate the long-term trends of less commonly measured pollutants through the collection of repeated short-term samples. While many different mobile monitoring approaches have been taken, few studies have looked at the importance of study design when the goal is application to epidemiologic cohort studies. Air pollution concentrations include random variability and systematic variability, and we hypothesize that mobile campaigns benefit from temporally balanced designs that randomly sample from all seasons of the year, days of the week, and hours of the day. We carried out a simulation study of fixed-site monitors to better understand the role of short-term mobile monitoring design on the prediction of long-term air pollution exposure surfaces. Specifically, we simulated three archetypal sampling designs using oxides of nitrogen (NO_x) monitoring data from 69 California air quality system (AQS) sites: (1) a year-around, Balanced Design, (2) a Rush Hours Design,

and (3) a Business Hours Design. We used Monte Carlo resampling to investigate the range of possible outcomes (i.e., the resulting annual average concentration prediction) from each design against the "truth", the actual monitoring data. We found that the Balanced Design consistently yielded the most accurate annual averages; Rush Hours and Business Hours Designs generally resulted in comparatively more biased estimates and model predictions. Importantly, the superior performance of the Balanced Design was evident when predictions were evaluated against true concentrations but less detectable when predictions were evaluated against the measurements from the same sampling campaign since these were themselves biased. This result is important since mobile monitoring campaigns that use their own measurements to test the robustness of the results may underestimate the level of bias in their results. Appropriate study design is crucial for mobile monitoring campaigns aiming to assess accurate long-term exposure in epidemiologic cohorts. Campaigns should aim to implement balanced designs that sample during all seasons of the year, days of the week, and all or most hours of the day to produce generally unbiased, long-term averages. Furthermore, differential exposure misclassification could result from unbalanced designs, which may result in misleading health effect estimates in epidemiologic investigations.

1 Introduction

A large body of evidence links long-term exposure to air pollution to adverse health effects in humans, including mortality from cardiovascular outcomes and lung cancer.[1–6] Most such studies focus on criteria air pollutants such as $PM_{2.5}$ and NO_2, in large part to use available monitoring data to estimate exposures. While air pollution cohort studies may leverage data from exposure assessment campaigns that supplement the regulatory network data, few focus on ambient air pollution exposures that are not criteria air pollutants.[7–9]

Recently, mobile monitoring campaigns have been used to estimate long-term average air pollutant levels.[10–16] Mobile monitoring campaigns typically equip a vehicle with air monitors and collect samples while in motion (mobile sampling) or while stopped (stationary sampling). A single monitoring platform can thus be used to collect samples at many locations within a relatively short period of time, making it a time and cost-efficient sampling approach. Mobile campaigns are particularly well-suited for multi-pollutant monitoring of less frequently monitored traffic-related air pollutants that require expensive instruments or instruments that need frequent attention during the sampling period.

While many recent mobile monitoring campaigns have been leveraged to map air pollution exposures and link them to health studies, there has not been any literature on the appropriate design of a mobile monitoring campaign for application to epidemiologic cohort studies. Many mobile campaigns have been short, lasting from a few weeks to months and with few repeat visits to each location spanning one to three seasons.[12,17,18] Most campaigns have conducted sampling during weekday business or rush hours, ignoring the surrounding hours, when air pollution concentrations can be drastically different.[19,20] The design of these studies may not be adequate for long-term exposure cohort studies.

The goal of this paper is to shed light on the design of a mobile monitoring campaign for application to epidemiologic cohort studies. We carry out a set of simulation studies to better understand the role of mobile monitoring design on the prediction of long-term average surfaces. We use existing monitoring data from California to compare the primary, long-term site averages when all of the data are included to subsequent analyses utilizing subsets of the data. These data provide a unique opportunity to explore how temporal sampling strategies can influence the resulting estimated annual-average concentration. Our analysis requires having a long-term,

comprehensive set of measurement data, which therefore necessitates using fixed-site measurements rather than mobile measurements, to shed light on an aspect of study design for mobile monitoring.

2 Methods

2.1 Data

We simulate three sampling designs (see below) using hourly observations for oxides of nitrogen (NOx) collected during 2016 from regulatory Air Quality System (AQS) sites in California. NOx was selected since it is a spatially and temporally variable traffic pollutant with a strong diurnal pattern,[11,21,22] and it is measured at many regulatory monitoring sites in California, providing a large enough dataset for this analysis.[23]

We required that NOx observations meet various criteria to be included in this analysis. First, sites needed to have readings at least 66% of the time (5,797/8,784 hourly samples; 2016 was a leap year). Second, sites needed to have sampling throughout the year, such that data collection gaps were a maximum of 45 days long. Third, sites were required to have sampled for at least 40% of the time during various two-week periods that were used in two of our "common" designs (described below). This sample size ensured that we could sample during these periods without replacement. Fourth, sites were required to have positive readings (> 0 ppb) at least 60% of the time, thus ensuring that sites had sufficient variability in their concentrations and allowing us to model annual averages on the natural log scale. Finally, sites in rural and industrial settings (as determined by the US EPA)[24] were excluded.

2.2 Sampling Designs

We conducted simulation studies to characterize the properties of three sampling designs (Table 1, Supplementary Information [SI] Figure S1). Each design has a long- and a short-term sampling approach. Long-term approaches use all of the data that meet each design's definition to estimate site annual averages and are analogous to traditional, fixed-site sampling approaches

where sampling at a given location occurs over an extended period of time. Short-term approaches only collect 28 samples per site and are analogous to mobile monitoring campaigns that collect a few repeat samples per site. (The cut-off of 28 samples reflects our preliminary analyses showing that 28 hourly NOx samples are sufficient to estimate a site's annual average within about 25% error or less [SI Figure S2].) Each design has multiple versions where samples are collected at slightly different times. The various design versions are intended to reflect the bias produced if only certain times are included in the measurements. We simulated each short-term sampling approach 30 times (Monte Carlo resampling), and hereafter refer to each of these simulations as a "campaign" since each represents a potential mobile monitoring study.

Table 1. Simulated sampling designs used to estimate site annual averages.[1]

Design	Sampling Approach and Number of Samples[2]	Sampling Seasons	Sampling Days	Sampling Hours
Year-Around "Balanced" Design	Long-term (Max 8,784 (V1), 7,320 (V2), or 4,392 (V3) hourly samples per site x 1 campaign simulation)	Winter, spring, summer, fall	Mon – Sun	V1 (All Hours) V2 (Most Hours): 5 AM – 12 AM V3 (Truncated Hours): 6-9 AM, 1-5 PM, 8-10 PM
	Short-term (28 hourly samples per site x 30 campaign simulations)	7 samples per season	5/7 weekday samples; 2/7 weekend samples	Random hours according to V1, V2, or V3
Two-Season Weekday "Rush Hours" Design[4]	Long-term (Max 160 hourly samples per site x 1campaign	V4-5: winter & summer (2-wk period per season)	Mon – Fri	7-10 AM, 3-6 PM

	simulation)	V6-7: spring & fall (2-wk period per season)		
	Short-term (28 hourly samples per site x 30 campaign simulations)	14 samples per season		Random Rush Hours according to V4-5 or V6-7
Two-Season Weekday "Business Hours" Design[4]	Long-term (Max 180 hourly samples per site x 1 campaign simulation)	V4-5: winter & summer (2-wk period per season) V6-7: spring & fall (2-wk period per season)	Mon – Fri	9 AM – 5 PM
	Short-term (28 hourly samples per site x 30 campaign simulations)	14 samples per season		Random Business Hours according to V4-5 or V6-7

[1] There are three archetypal sampling designs, each with long- and short-term sampling approaches and multiple versions. Long-term approaches are analogous to traditional, fixed-site sampling, while short-term approaches are analogous to mobile monitoring campaigns. Short names for the sampling design appear in quotes.

[2] Long-term approaches have 1 campaign simulation (each includes all of the available data that meet that design's criteria), while short-term approaches have 30 campaign simulations (each with 28 samples). Maximum hourly samples per site varied because some sites had missing readings. Year 2016 was a leap year.

[4] See SI Table S1 for each version's exact sampling periods

The Year-Around "Balanced" Design represents an "ideal" sampling scheme: sampling is conducted during all seasons, days of the week, and all or most hours of the day. Version 1 collects samples during all hours of the day. Versions 2-3 reduce the sampling hours to reflect the logistical constraints of executing an extensive campaign: samples occur during most hours of the day (5 AM – 12 AM only; "Version 2") or during 6-9 AM, 1-5 PM and 8-10 PM

("Version 3"). Estimates from the long-term Balanced Design Version 1 are analogous to what might be collected from a traditional, year-around, fixed-site sampling scheme. For simplicity, we interchangeably refer to these as the "true" estimates or the "gold standard" hereafter, though we acknowledge that some error exists (e.g., due to missing hours or instrument accuracy).

The Two-Season Weekday "Rush Hours" and "Business Hours" Designs reflect common designs in the literature. Samples are collected either during summer and winter (Versions 4-5) or spring and fall (Versions 6-7). Sampling for each version occurs on weekdays during a two-week period each relevant season (See SI S1 for each version's exact sampling periods). Sampling is restricted to the hours of 7-10 AM and 3-6 PM (Rush Hours Design) or 9 AM – 5 PM (Business Hours Design). The short-term approach collects 14 random samples during each season.

2.3 Prediction Models

We estimated site annual averages from the data collected during each campaign. We log-transformed these before using them as the outcome variable in partial least squares (PLS) regression models, which summarized hundreds of geographic covariate predictors (e.g., land use, road proximity, and population density; see SI Table S2 for the covariates considered) into two PLS components (using the `plsr` function in the `pls` package in R). We evaluated the performance of each campaign using ten-fold cross-validated (CV) predictions on the native scale, incorporating re-estimation of the PLS components in each fold. The cross-validation groups were randomly selected and, importantly, fixed across all campaigns to allow for consistent model performance comparisons across design versions.

To best understand the role of design, we present results for annual average estimates, predictions, and model performance statistics. In descriptive analyses, we compare design-specific annual average estimates and predictions to the gold standard (long-term Balanced Design Version 1). We compare predicted site concentrations against predictions from the gold standard since epidemiologic air pollution studies often rely on predicted exposure, and the gold standard prediction represents the best possible prediction of annual-average concentrations that a study could hope to achieve. We complement this approach with model assessment evaluations of design-specific site predictions against two different references: an assessment against the true

averages, and a traditional model assessment evaluation against the respective design-specific annual average estimates. The traditional assessment compares the predicted exposures to the observed site measurements from which they were derived. This allows us to document the quantities that would normally be available from modeling the data measured from any specific campaign. We summarize the model performance in terms of cross-validated mean squared error (MSE)-based R^2 (R^2_{MSE}), regression-based R^2 (R^2_{reg}), and root mean squared error (RMSE). R^2_{MSE} reflects bias as well as variation around the one-to-one line. R^2_{reg} is based on the best fit line between the measurements and predictions, which adjusts for bias and slopes different than one, and is defined as the squared correlation between a measurement and a prediction. See SI Equations 1-3 for definitions.

We repeated these analyses for nitrogen dioxide (NO_2) and nitrogen monoxide (NO), adding a two ppb constant to all of the hourly NO readings before log-transforming to eliminate negative and zero concentration readings.

All analyses were conducted in R (v 3.6.2, using RStudio v 1.2.5033).[25]

3 Results

We included the 69 of 105 California AQS sites that met our data criteria (Figure 1, SI Figure S3). Sites were located in both urban and suburban settings, in residential and commercial areas.

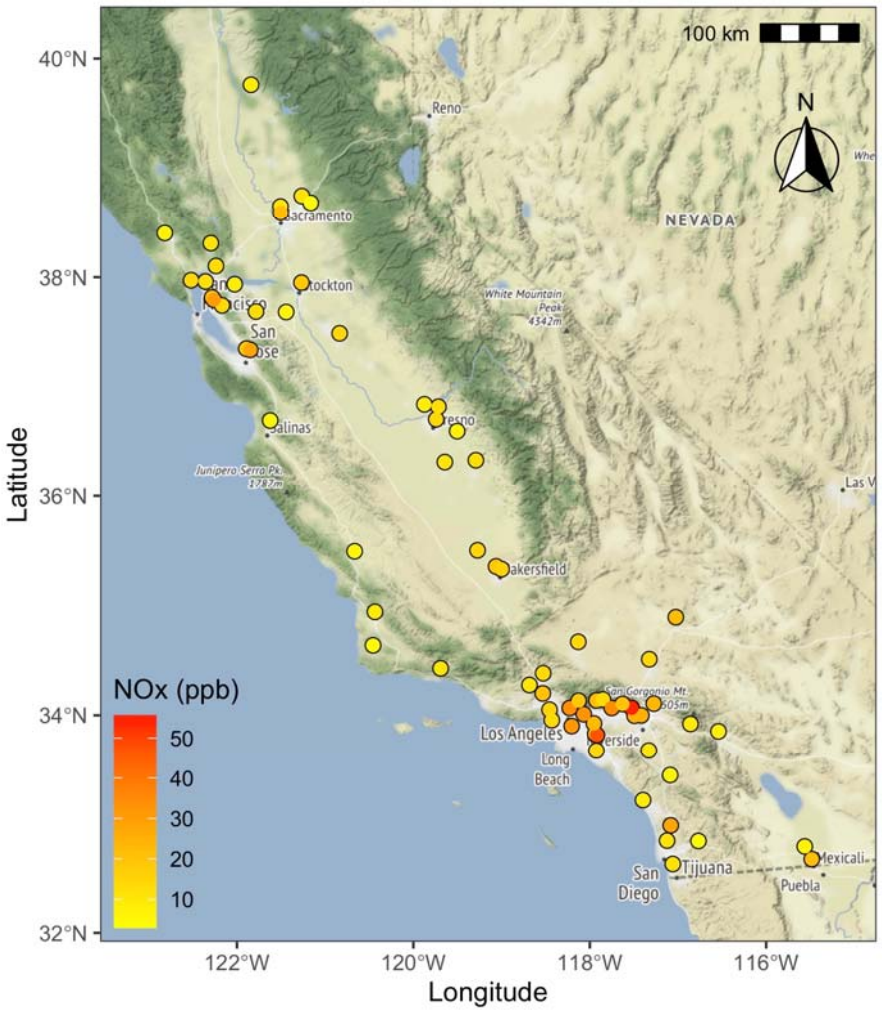

Figure 1. True annual average NOx measurements, as measured by the long-term Year-Around Balanced Design Version 1, for AQS sites included in this analysis (N=69).

3.1 Hourly Readings

Sites had on average (SD) of 8,090 (361) hourly readings, the equivalent of 337 (15) days of full sampling (See SI Table S3). Average (SD) hourly NOx concentrations were 16 (21) ppb (See SI Table S4). Sites had seasonal, daily, and hourly concentration patterns, with trends being more pronounced at some sites than others (See SI Figure S4-S6).

3.2 Annual Average Estimates

Across the 69 monitor locations, measured annual average concentrations (long-term Balanced Design Version 1), had a median (IQR) of 14 (10 - 21) ppb and ranged from 3-56 ppb. The short-term and long-term sampling approaches resulted in similar distributions of annual averages for different design versions. Figure 2 shows the long-term and a single short-term approach for each design. Overall, the long-term and short-term approach for each design version had very similar distributions. All of the Balanced Design versions resulted in only slight differences in their medians and IQRs. The Rush Hours Design versions generally resulted in slightly higher annual averages than the true averages, with some versions being more variable and having somewhat different distributions. The Business Hours Design versions resulted in annual averages that were generally lower than the true averages and less variable than the Rush Hours Design versions. See SI Table S5 for summary statistics. SI Figure S7 shows annual average estimates for all campaigns and pollutants.

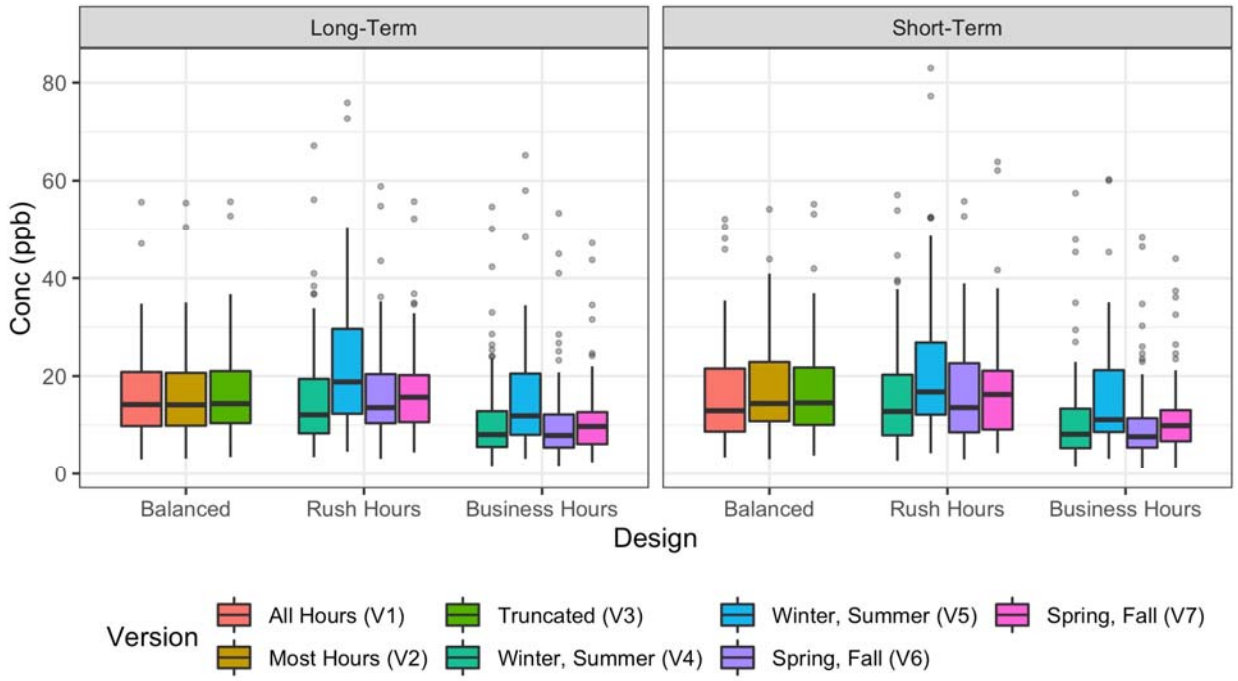

Figure 2. Distribution of NOx annual averages (N=69 sites) from different design versions. Showing the one campaign for each long-term approach and one example campaign for each short-term approach.

Figure 3 shows the site-specific distributions of annual averages across designs for short-term approaches relative to the true averages for a stratified random sample of 12 sites. Sites are stratified by whether their true mean concentration was in the low (<25th percentile), middle (25th-75th percentile) or high (>75th percentile) concentration category. The variation of averages across campaigns increases with concentration in all designs. Site-specific averages are similar to the true averages for all Balanced Design versions while there were multiple sites from the Business Hours Design versions with averages systematically lower. The Rush Hours Design versions also had many biased averages, although the direction of the bias varied by site and design version. SI Figure S8 shows these biases for all sites.

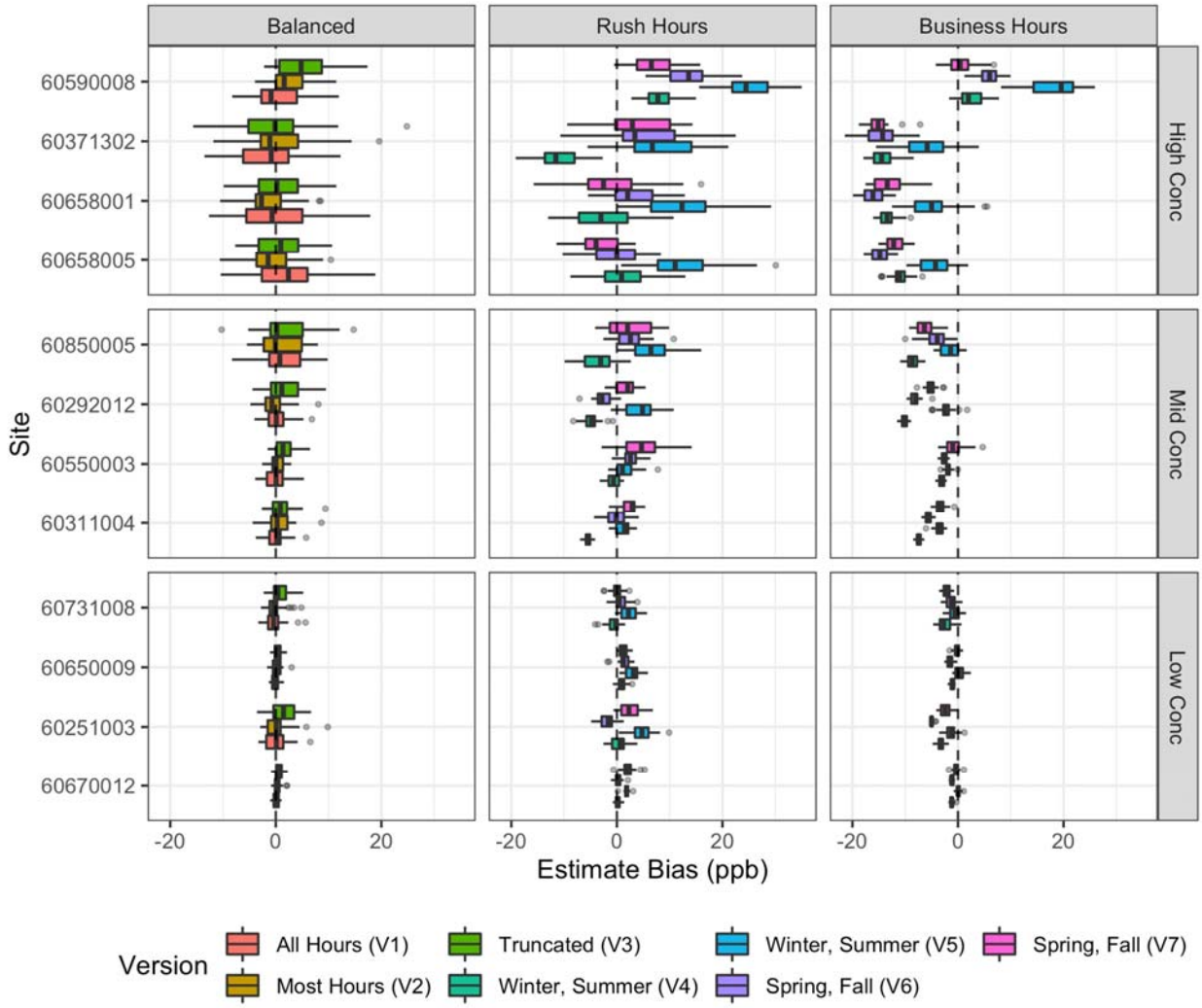

Figure 3. Site-specific NOx measurement biases for short-term designs (N = 30 campaigns) as compared to the true annual average at that site (long-term Balanced Design Version 1). Showing a stratified random sample of 12 sites, stratified by whether their true concentration was in the low (<25th percentile), middle (25th-75th percentile) or high (>75th percentile) concentration category and arranged within each stratum with lower concentration sites being closer to the bottom.

3.3 Model Predictions

The PLS model of the true annual average had a root mean square error (RMSE) of 7.2 ppb and a mean square error-based coefficient of determination (R^2_{MSE}) of 0.46.

We compared PLS model predictions from each short-term design to the gold standard model predictions. SI Figure S9 shows the relative standard deviations of predictions by design version, with 1 indicating that design predictions have the same standard deviation as the gold

239 standard model predictions. Overall, the Balanced Design predictions have similar variability to
240 those of the gold standard (range: 0.87-1.28), the Rush Hours Design predictions are more
241 variable (range: 0.90-1.74), and the Business Hours Design predictions are mixed: some less and
242 some more variable (range: 0.73-1.54). Figure 4 displays these comparisons as scatterplots and
243 best fit lines. The scatterplots show that there are a few sites, some of which have high leverage,
244 that have variable predictions in all designs. From the best fit lines, we observe that all short-
245 term Balanced Design versions resulted in the most accurate predictions on average, as indicated
246 by their overlapping general trends along the one-to-one line. The Rush Hours Design versions
247 were more likely to have a positive general trend, while the Business Hours Design versions
248 were more likely to have a negative general trend, indicating, for example, that higher
249 concentrations were more likely to be over- or under-estimated, respectively. However, there
250 was heterogeneity in this overall pattern across the Rush and Business Hours Design versions.
251 Furthermore, there was additional heterogeneity across individual campaigns. The SI contains
252 comparable figures comparing design predictions to the gold standard and additional figures for
253 NO and NO_2 (SI Figures S10-S13).
254

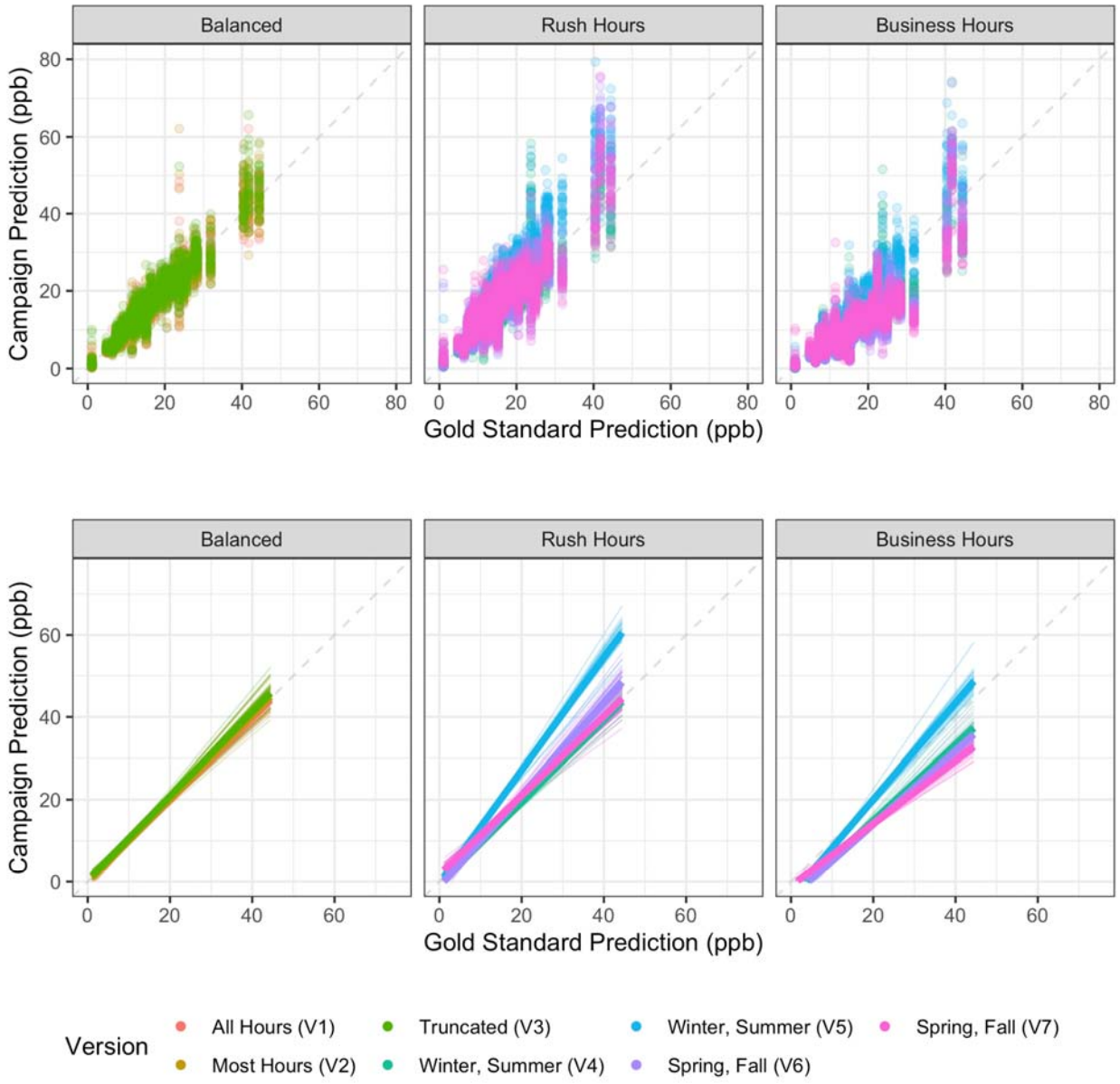

Figure 4. Scatterplots and best fit lines of cross-validated short-term predictions for 30 campaigns vs the gold standard predictions for NOx. Thin transparent lines are individual campaigns, colored by design version; thicker lines are the overall version trend. (One prediction is excluded for clarity from the Rush Hours Version 4 scatterplot at x=24 ppb, y=109 ppb [site 60731016] but included in the line plots.)

Figure 5 shows site-specific comparisons of predictions across 30 short-term campaigns relative to the gold standard predictions for a stratified random sample of 12 sites in order to characterize relative bias (see SI Figure S14 for all sites). Overall, the short-term Balanced Design predictions had a median (IQR) bias of 0.2 (-1 – 1.4) ppb relative to the gold standard

265 predictions (see SI Table S7 for details). All Balanced Design predictions were very similar to
266 the gold standard predictions, though some sites frequently had larger biases. The Rush Hours
267 and Business Hours Design versions were more likely to consistently produce biased site
268 predictions, with a median (IQR) bias of 1.2 (-1.2 – 4) ppb and -3.8 (-6.6 – -1.4) ppb,
269 respectively. While the Rush Hours Design versions generally resulted in higher predictions
270 across sites (with some inconsistency across versions for a few sites), the Business Hours Design
271 versions resulted in predictions that were both lower and higher than the gold standard
272 predictions across sites. There were also a few sites that tended to have more biased and/or more
273 variable predictions relative to the gold standard across all designs. We observed similar patterns
274 when looking at estimate (rather than prediction) biases (See Figure 3, SI Figure S8).
275

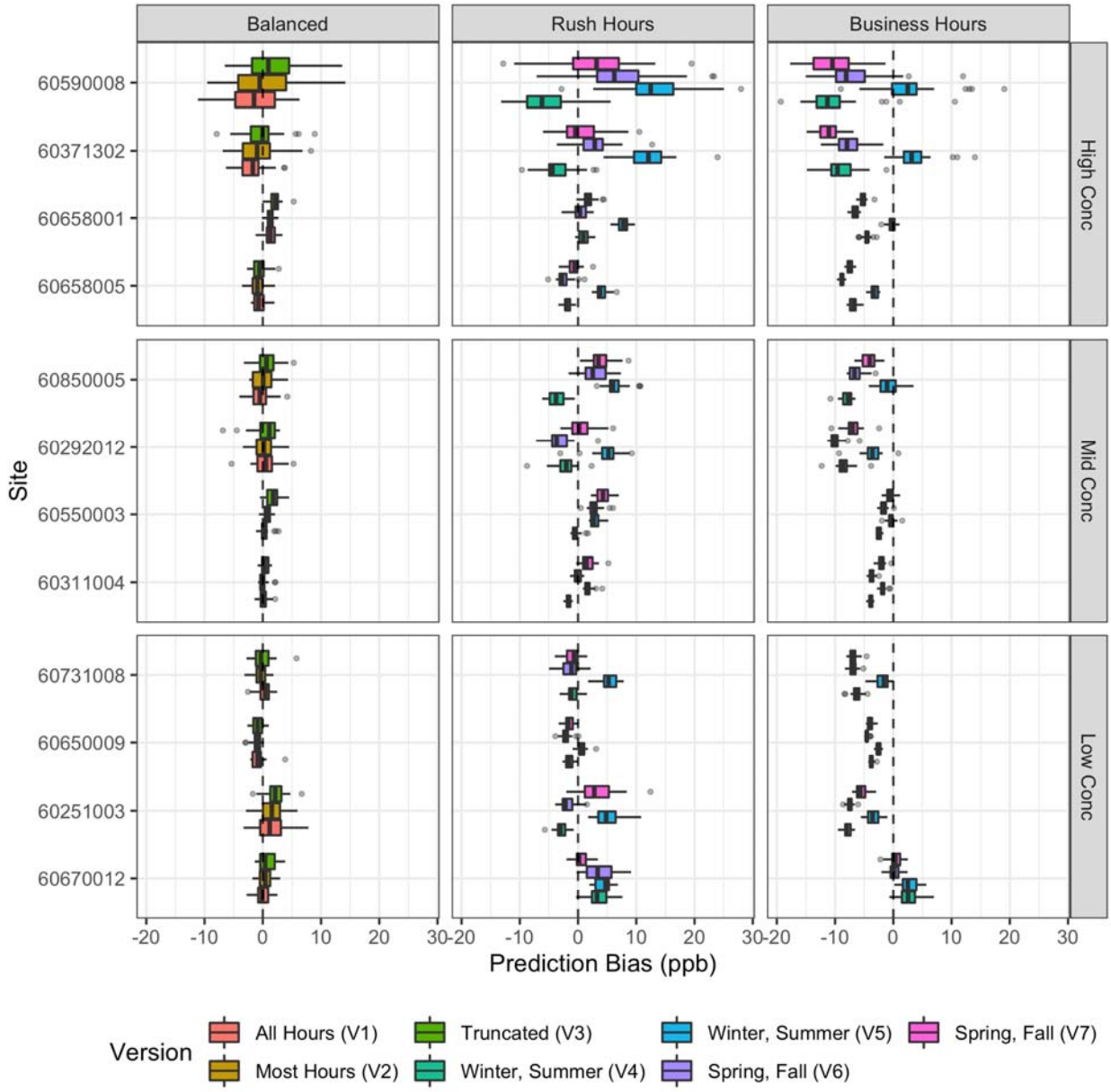

Figure 5. Site-specific NOx prediction biases for short-term designs (N = 30 campaigns) as compared to the gold standard predictions (long-term Balanced Design Version 1). Showing a stratified random sample of 12 sites, stratified by whether true concentrations were in the low (Conc < 0.25), middle (0.25 ≤ Conc ≤ 0.75) or high (Conc > 0.75) concentration quantile and arranged within each stratum with lower concentration sites closer to the bottom.

3.4 Model Assessment

Figure 6 shows the out-of-sample prediction performances relative to the observations from the true averages (left column) and the specific design (right column), for both the long-term and short-term approaches. The boxplots quantify the distribution of performance statistics across all 30 short-term campaigns while the squares show the performance for the long-term approach of the same design version. When assessed against the true averages, all the Balanced Design versions generally perform better than either the Rush Hours or Business Hours Design versions with higher CV R^2_{MSE} and CV R^2_{reg}, and lower CV RMSE estimates. This is particularly apparent for the long-term approach. Furthermore, within design the performance for the long-term approach is better than the majority of the short-term campaigns. There is considerable heterogeneity in performance across the Rush Hours and Business Hours Design versions. In contrast, when assessed against observations from the same design, as would typically be done in practice, the role of sampling design on prediction performance is not as evident. The superior performance of the Balanced Design is not as apparent, and some of the Rush Hours and Business Hours Design versions appear to perform better. There are also a few campaigns that show poor performance, even under the Balanced Design. SI Figure S15-S16 show similar results for NO_2 and NO, with NO showing more variability and some lower performing statistics.

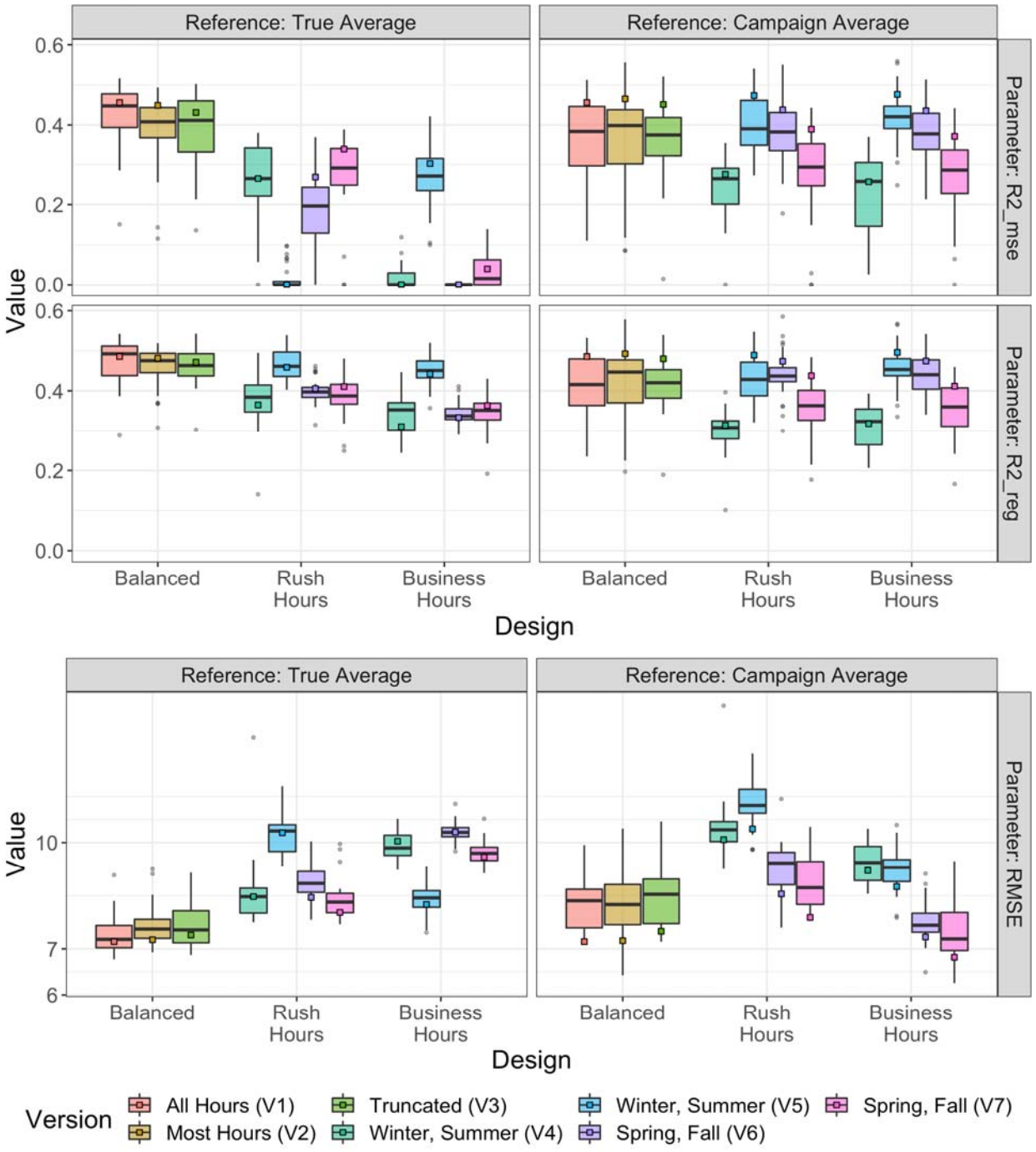

Figure 6. Model performances (MSE-based R2, Regression-based R2, and RMSE), as determined by each campaign's cross-validated predictions relative to: a) the true averages (long-term Balanced Version 1), and b) its respective campaign averages. Boxplots are for short-term approaches (30 campaigns), while squares are for long-term approaches (1 campaign).

3.5 NO and NO_2

We found similar results for NO and NO_2 (See the SI).

4 Discussion

In this paper we have used existing regulatory monitoring data to deepen our understanding of the importance of mobile monitoring study design for application to epidemiologic cohort studies. Others have shown that short-term data can be used to estimate long-term averages.[11,12] What has been missing from the literature until now, however, is the impact of mobile monitoring study design on the accuracy and precision of long-term exposure estimates and model predictions, particularly when the goal is to produce predictions for an epidemiologic study. Our results indicate that for designs with a sufficient number of short-term samples (about 28 or more), the design rather than the sampling approach (short- vs long-term) has the largest impact on the estimated long-term averages. We focus the rest of this discussion on the short-term approaches for each design, which resemble mobile monitoring, though the long-term approaches produced similar results.

In terms of specific design, we found that all of the Balanced Design versions resulted in similar annual averages as the true averages (long-term Balanced Version 1), while the Rush Hours and Business Hours Design versions were more likely to result in more biased and more or less variable annual average estimates. Specifically, the Rush Hours Design was more likely to overestimate, while the Business Hours Design was more likely to underestimate site averages. This result was likely because the Balanced Design captured much of NOx's temporal variability by allowing for samples to be collected during each season, day of the week, and all or most times of the day, all periods during which meteorology and traffic activity patterns impact air pollution concentrations (SI Figure S4-S6). The Rush Hours Design, on the other hand, was restricted to two sampling seasons and was more likely to sample during high concentration times of day and days of the week. The Business Hours Design had similar limitations though it was more likely to sample during low concentration times.

We found a similar pattern with the predictions: similar predictions across all Balanced Design versions, while most of versions in the Rush Hours tended to overpredict and those in the Business Hours tended to underpredict. However, this varied by design version, suggesting that the particular four weeks of sampling are an important source of heterogeneity in the results. The predictions were more variable for all Rush Hours Design versions and one Business Hours Design version (SI Figure S9). One Business Hours Design version was less variable, while two versions were about the same relative to the gold standard predictions.

The similarity in annual averages and predictions across all of the Balanced Design versions suggests that campaigns with slightly reduced sampling hours (for example, due to logistical constraints) should to a large degree still produce unbiased annual averages at most sites. On the other hand, campaigns that follow more temporally restricted sampling designs such as the Rush Hours and Business Hours Designs may produce systematically biased results, with the degree and direction of error being heavily impacted by the sampling window that happens to be selected.

At the site level, we saw that while any individual study campaign had the potential to produce biased estimates and predictions, the Rush Hours and Business Hours Designs were more likely to do so than the Balanced Design. The direction and magnitude of the bias varied by site and depended upon the sampling design and the typical seasonal, day of week, and time of day patterns of pollution at that site. This suggests that a simple correction factor (e.g., the ratio of the true annual average concentration relative to the resulting concentration from a given design) is unlikely to appropriately adjust for bias at the site level. Given that higher concentration sites were more likely to have greater degrees of bias and variation (Figure 4 – Figure 5), non-balanced designs may misrepresent some sites more than others and lead to differential exposure misclassification in epidemiologic studies. Thus, while non-balanced design may be appropriate for non-epidemiologic purposes including characterizing the spatial impact of traffic related air pollutants during peak hours for urban planning and policy purposes, these could be misleading in epidemiologic applications.

In this study we were able to evaluate prediction model performance against the true annual average exposure as well as against the observations typically available for model performance assessment. Performance assessment against the true averages indicates that the Balanced Design is clearly the best, and that there is little degradation in performance across

design versions. This means that it is possible to design high quality mobile monitoring studies that accommodate some measure of logistical feasibility, for example, by not requiring sampling in the middle of the night. In contrast, the performance of the Rush Hours and Business Hours Designs is comparatively worse, indicating that the logistically appealing approach that samples only four weeks during two seasons, during daytime hours, and only during weekdays is inadequate for providing high quality estimates of annual averages. Further, the performance of these designs varies considerably and unpredictably depending upon the specific pair of two-week periods that were selected for sampling. Additionally, comparison of the two R^2 estimates (R^2_{MSE} and R^2_{reg}) indicates that not all of their poor performance is due to the inability to predict the same value as the truth (R^2_{MSE}), but due to systematic bias in the design.

Further, it is notable that the standard approach to model assessment, comparing model predictions to observations collected during the sampling campaign, doesn't clearly reveal the superior performance of the Balanced Design or the inherent flaws of the Rush Hours and Business Hours Designs. In fact, some of the Rush Hours and Business Hours Design versions perform better than the Balanced Design when evaluated against the campaign's observations. This is because the evaluation doesn't take into account that the observations are biased because of the sampling design.

It is notable that the performance of our short-term campaigns was fairly consistent with, though generally slightly worse than, the performance observed in the long-term campaign for each design version (Figure 6). However, occasionally there was an "unlucky" short-term campaign with meaningfully poorer performance than the other campaigns of the same design. This is true even for the Balanced Design versions where 1-2 of the 30 campaigns (~3-6%) had notably worse performance as quantified by R^2. It may be possible that this result is driven by a few high-leverage outlier sites that impact the prediction model performance. In practice, mobile monitoring study investigators are likely to investigate high-leverage sites and address their influence in their prediction modeling.

Our study focused on short-term campaigns with 28 repeat samples per site. We did not consider campaigns with fewer or more visits. As evident in SI Figure S2, the percent error in estimating the annual average from fewer than 25 visits skyrockets, suggesting that site estimates will be considerably noisier in mobile campaigns with few repeat visits, regardless of the study design. Prediction model performance is thus likely to decrease as the number of visits per site

399 decrease. Logistically, it is also difficult to achieve balance in sampling over time across season,
400 day of week, and time of day with fewer than 28 samples per site. Furthermore, we note that this
401 study focused on a few generalizable, common designs in the literature, though many other
402 approaches have been taken. We expect that the variety of mobile campaign designs that have
403 been implemented will all produce slightly different results.

404 In putting these results in context, it is important to recognize that in this simulation study
405 we are using existing regulatory monitoring data that has been through extensive quality
406 assurance and quality control processes to approximate the data that might be collected by
407 mobile monitoring campaigns. For instance, we used hourly averages to approximate much
408 shorter-term sampling durations (e.g., a few minutes or less) that would be collected during a
409 mobile campaign. Shorter duration sampling will affect the noise in the data, to an amount that
410 depends on the environment (e.g., temporal patterns in the concentrations of the pollutant being
411 measured) and the instruments. (For comparison, however, our additional evaluations of minute-
412 level data suggest that the decrease in percent error in going from two-minute to hour-long
413 samples is at most a few percent.) Further, our study took place throughout California, a large,
414 geographically diverse area with varying climate profiles.[26] This may explain the moderate
415 model performance of the gold standard campaign. While such a large sampling domain would
416 be challenging for a real-world mobile monitoring campaign, the overall conclusions of this
417 study are likely generalizable given that traffic-related air pollution concentrations in generally
418 exhibit temporal patterns that vary by location. While we observed moderate model
419 performances, campaigns with smaller, less spatially heterogenous study areas may see higher
420 performances.

421 There are several other differences between our study and a typical mobile monitoring
422 study. By definition, mobile monitoring campaigns collect non-stationary (mobile)
423 measurements, which are subject to jostling inherent with on-road sampling, even if some of the
424 sampling occurs while the vehicle is stationary. Further, mobile samples may be less precise than
425 what we observed from fixed, regulatory monitoring sites due to differences in instrumentation,
426 instrument quality, and maintenance. Mobile sampling platforms are more likely to be
427 immediately near another vehicle (e.g., in a traffic queue while stopped at a traffic signal) than a
428 fixed-site monitor. Another distinction is that while we sampled measurements within sites at
429 random, mobile campaigns typically sample from sites along a fixed route or in a designated

area. This induces some spatial correlation in the mobile monitoring results that is not part of our simulations. Furthermore, we did not consider the importance of the distribution of sampling locations in this study, which is particularly relevant when the exposure assessment goal is an epidemiologic application. Selecting sites that are representative of the target cohort's residence locations will ensure the spatial compatibility assumption is met, which is an important way to reduce the role of exposure measurement error in epidemiologic inference.[27] This consideration is especially relevant for mobile monitoring near major sources (e.g., airports, marine activity, and industry),[11,12,28–34] which may or may not represent a study cohort of interest.

Our evaluation focused on NOx, NO, and NO_2, which are quickly and moderately decaying air pollutants (concentrations reach background levels approximately 400-600 m from roadway sources).[21] Campaigns that measure these pollutants may be more susceptible to sampling design than campaigns that measure less spatially- and/or temporally-variable pollutants such as $PM_{2.5}$.[21] We selected NOx, NO, and NO_2 because these are often measured in short-term campaigns, and data for these pollutants are more widely available. Non-criteria pollutants, for example ultrafine particulates (UFP), however, have also received increasing attention in recent years given their emerging link to adverse health effects.[7,35–37] Still, high-quality information about their spatial distribution is essentially absent, and most studies have implemented short-term mobile sampling approaches[34] that may not be temporally balanced and potentially be misleading.

An important next step in this work is to understand whether the differences in exposure estimates that we observed across study designs have a meaningful impact on epidemiologic inferences. This is of particular interest considering that year-around, balanced designs are resource-intensive and rare, while shorter, more convenient campaigns are more common in the literature. More research is needed to better understand how and whether unbalanced mobile monitoring campaigns may contribute high quality exposure assessments for epidemiology. Regardless of design, we expect that the predictions from all of the campaigns will result in both classical-like and Berkson-like error.[27,38–40] Specifically, the predictions capture only part of the true long-term exposure (Berkson-like error), while the parameters in the prediction model are inherently noisy (classical-like error). However, these measurement error methods have not to date considered exposure assessment study design, beyond considering the importance of spatial compatibility, i.e., that distribution of monitoring locations is the same as the distribution of

participant locations. Our work suggests that deeper understanding of the role of exposure assessment design on epidemiologic inference is an important area of research.

4.1 Conclusions and Recommendations for Mobile Monitoring Campaigns

Mobile monitoring study design should be an important consideration for campaigns aiming to assess long-term exposure in an epidemiologic cohort. Given the temporal trends in air pollution, campaigns should implement balanced designs that sample during all seasons of the year, days of the week, and hours of the day in order to produce unbiased long-term averages. Nonetheless, restricting the sampling hours in balanced designs, for example due to logistical considerations, will still generally produce unbiased estimates at most sites. On the other hand, unbalanced sampling designs like those often seen in the literature are more likely to produce biased long-term estimates, with some sites being more biased than others. And while predictions from these restricted designs may at times perform similarly to balanced designs (or, more problematically, may erroneously *appear* to perform similarly when evaluated against measurements which are themselves biased samples), this performance may strongly depend on the exact sampling period chosen and may thus be difficult or impossible to anticipate prior to conducting a new sampling campaign. Furthermore, the differential exposure misclassification that may result from these designs may be problematic in epidemiologic investigations. Finally, studies that implement unbalanced sampling designs may have hidden exposure misclassification given that both the observations and model predictions may be systematically incorrect. By implementing a balanced sampling design, campaigns can thus increase their likelihood of capturing accurate long-term exposure averages.

5 Funding

This work was funded by the Adult Changes in Thought – Air Pollution (ACT-AP) Study (National Institute of Environmental Health Sciences [NIEHS], National Institute on Aging [NIA], R01ES026187), and BEBTEH: Biostatistics, Epidemiologic & Bioinformatic Training in Environmental Health (NIEHS, T32ES015459).

Research described in this article was conducted under contract to the Health Effects Institute (HEI), an organization jointly funded by the United States Environmental Protection Agency (EPA) (Assistance Award No. CR-83998101) and certain motor vehicle and engine manufacturers. The contents of this article do not necessarily reflect the views of HEI, or its sponsors, nor do they necessarily reflect the views and policies of the EPA or motor vehicle and engine manufacturers.

6 References

1. Hoek, G. *et al.* Long-term air pollution exposure and cardio- respiratory mortality: a review. *Environ. Health* **12**, 43 (2013).
2. Kampa, M. & Castanas, E. Human health effects of air pollution. *Environ. Pollut.* **151**, 362–367 (2007).
3. Rückerl, R., Schneider, A., Breitner, S., Cyrys, J. & Peters, A. Health effects of particulate air pollution: a review of epidemiological evidence. *Inhal. Toxicol.* **23**, 555–592 (2011).
4. Schwartz, J. Air pollution and daily mortality: a review and meta analysis. *Environ. Res.* **64**, 36–52 (1994).
5. Chen, H., Goldberg, M. & Villeneuve, P. A systematic review of the relation between long-term exposure to ambient air pollution and chronic diseases. *Rev. Environ. Health* **23**, 243–298 (2008).
6. Pope, C. A., Dockery, D. W. & Schwartz, J. Review of epidemiological evidence of health effects of particulate air pollution. *Inhal. Toxicol.* **7**, 1–18 (1995).
7. Weichenthal, S. *et al.* Within-city Spatial Variations in Ambient Ultrafine Particle Concentrations and Incident Brain Tumors in Adults. *Epidemiology* **31**, (2020).
8. Weichenthal, S. *et al.* Long-term exposure to ambient ultrafine particles and respiratory disease incidence in in Toronto, Canada: a cohort study. *Environ. Health* **16**, 64 (2017).
9. Downward, G. S. *et al.* Long-term exposure to ultrafine particles and incidence of cardiovascular and cerebrovascular disease in a prospective study of a Dutch cohort. *Environ. Health Perspect.* **126**, 127007 (2018).

10. Hankey, S. & Marshall, J. D. Land Use Regression Models of On-Road Particulate Air Pollution (Particle Number, Black Carbon, PM2.5, Particle Size) Using Mobile Monitoring. *Environ. Sci. Technol.* **49**, 9194–9202 (2015).
11. Apte, J. S. *et al.* High-Resolution Air Pollution Mapping with Google Street View Cars: Exploiting Big Data. *Environ. Sci. Technol.* **51**, 6999–7008 (2017).
12. Hatzopoulou, M. *et al.* Robustness of Land-Use Regression Models Developed from Mobile Air Pollutant Measurements. *Environ. Sci. Technol.* **51**, 3938–3947 (2017).
13. Patton, A. P. *et al.* Spatial and temporal differences in traffic-related air pollution in three urban neighborhoods near an interstate highway. *Atmos. Environ.* (2014) doi:10.1016/j.atmosenv.2014.09.072.
14. Van den Bossche, J. *et al.* Mobile monitoring for mapping spatial variation in urban air quality: Development and validation of a methodology based on an extensive dataset. *Atmos. Environ.* (2015) doi:10.1016/j.atmosenv.2015.01.017.
15. Kerckhoffs, J. *et al.* Comparison of ultrafine particle and black carbon concentration predictions from a mobile and short-term stationary land-use regression model. *Environ. Sci. Technol.* **50**, 12894–12902 (2016).
16. Xie, X. *et al.* A Review of Urban Air Pollution Monitoring and Exposure Assessment Methods. *ISPRS International Journal of Geo-Information* vol. 6 (2017).
17. Weichenthal, S. *et al.* A land use regression model for ambient ultrafine particles in Montreal, Canada: A comparison of linear regression and a machine learning approach. *Environ. Res.* **146**, 65–72 (2016).

18. Minet, L., Gehr, R. & Hatzopoulou, M. Capturing the sensitivity of land-use regression models to short-term mobile monitoring campaigns using air pollution micro-sensors. *Environ. Pollut.* **230**, 280–290 (2017).
19. Batterman, S., Cook, R. & Justin, T. Temporal variation of traffic on highways and the development of accurate temporal allocation factors for air pollution analyses. *Atmos. Environ.* **107**, 351–363 (2015).
20. Saha, P. K. *et al.* Quantifying high-resolution spatial variations and local source impacts of urban ultrafine particle concentrations. *Sci. Total Environ.* **655**, 473–481 (2019).
21. Karner, A. A., Eisinger, D. S. & Niemeier, D. A. Near-roadway air quality: Synthesizing the findings from real-world data. *Environ. Sci. Technol.* **44**, 5334–5344 (2010).
22. Riley, E. A. *et al.* Multi-pollutant mobile platform measurements of air pollutants adjacent to a major roadway. *Atmos. Environ.* **98**, 492–499 (2014).
23. US EPA. Air Quality System (AQS). *US Environmental Protection Agency* https://www.epa.gov/aqs (2019).
24. US EPA. AirData Pre-Generated Data Files. *US Environmental Protection Agency* https://aqs.epa.gov/aqsweb/airdata/download_files.html (2019).
25. R Core Team. R: A Language and Environment for Statistical Computing. *R Foundation for Statistical Computing* https://www.r-project.org (2019).
26. Li, L. *et al.* Ensemble-based deep learning for estimating PM2.5 over California with multisource big data including wildfire smoke. *Environ. Int.* **145**, 106143 (2020).
27. Szpiro, A. A. & Paciorek, C. J. Measurement error in two-stage analyses, with application to air pollution epidemiology. *Environmetrics* (2013) doi:10.1002/env.2233.

562 28. Dodson, R. E., Houseman, E. A., Morin, B. & Levy, J. I. An analysis of continuous black
563 carbon concentrations in proximity to an airport and major roadways. *Atmos. Environ.* **43**,
564 3764–3773 (2009).
565 29. Riley, E. A. *et al.* Correlations between short-term mobile monitoring and long-term passive
566 sampler measurements of traffic-related air pollution. *Atmos. Environ.* **132**, (2016).
567 30. Austin, E. *et al. Mobile ObserVations of Ultrafine Particles: The MOV-UP study report.*
568 (2019).
569 31. Hudda, N., Gould, T., Hartin, K., Larson, T. V. & Fruin, S. A. Emissions from an
570 international airport increase particle number concentrations 4-fold at 10 km downwind.
571 *Environ. Sci. Technol.* **48**, 6628–6635 (2014).
572 32. Lack, D. A. & Corbett, J. J. Black carbon from ships: a review of the effects of ship speed,
573 fuel quality and exhaust gas scrubbing. *Atmospheric Chem. Phys.* **12**, (2012).
574 33. Kozawa, K. H., Fruin, S. A. & Winer, A. M. Near-road air pollution impacts of goods
575 movement in communities adjacent to the Ports of Los Angeles and Long Beach. *Atmos.*
576 *Environ.* **43**, 2960–2970 (2009).
577 34. Riffault, V. *et al.* Fine and Ultrafine Particles in the Vicinity of Industrial Activities: A
578 Review. *Crit. Rev. Environ. Sci. Technol.* **45**, 2305–2356 (2015).
579 35. Kilian, J. & Kitazawa, M. The emerging risk of exposure to air pollution on cognitive decline
580 and Alzheimer ' s disease e Evidence from epidemiological and animal studies. *Biomed. J.*
581 **41**, 141–162 (2018).
582 36. Lane, K. J. *et al.* Association of modeled long-term personal exposure to ultrafine particles
583 with inflammatory and coagulation biomarkers. *Environ. Int.* **92–93**, 173–182 (2016).

584 37. US EPA. Integrated science assessment (ISA) for particulate matter (final report, Dec 2019).
585 *US Environ. Prot. Agency* (2019).
586 38. Gryparis, A., Paciorek, C. J., Zeka, A., Schwartz, J. & Coull, B. A. Measurement error
587 caused by spatial misalignment in environmental epidemiology. *Biostatistics* **10**, 258–274
588 (2009).
589 39. Szpiro, A. A., Sheppard, L. & Lumley, T. Efficient measurement error correction with
590 spatially misaligned data. *Biostatistics* **12**, 610–623 (2011).
591 40. Sheppard, L. *et al.* Confounding and exposure measurement error in air pollution
592 epidemiology. *Air Qual. Atmosphere Health* **5**, 203–216 (2012).
593