

1 **Design and evaluation of mobile monitoring campaigns for air pollution exposure**
2 **assessment in epidemiologic cohorts**
3

4 Magali N. Blanco,^a Annie Doubleday,^a Elena Austin,^a Julian D. Marshall,^b Edmund Seto,^a
5 Timothy Larson,^{a,b} Lianne Sheppard^{a,c}
6

7 ^aDepartment of Environmental and Occupational Health Sciences, School of Public Health,
8 University of Washington, Hans Rosling Center for Population Health, 3980 15th Ave NE,
9 Seattle, WA 98195
10

11 ^bDepartment of Civil & Environmental Engineering, College of Engineering, University of
12 Washington, 201 More Hall, Box 352700, Seattle, WA 98195
13

14 ^cDepartment of Biostatistics, School of Public Health, University of Washington, Hans Rosling
15 Center for Population Health, 3980 15th Ave NE, Seattle, WA 98195
16
17

18 Corresponding Author: Magali N. Blanco
19
20

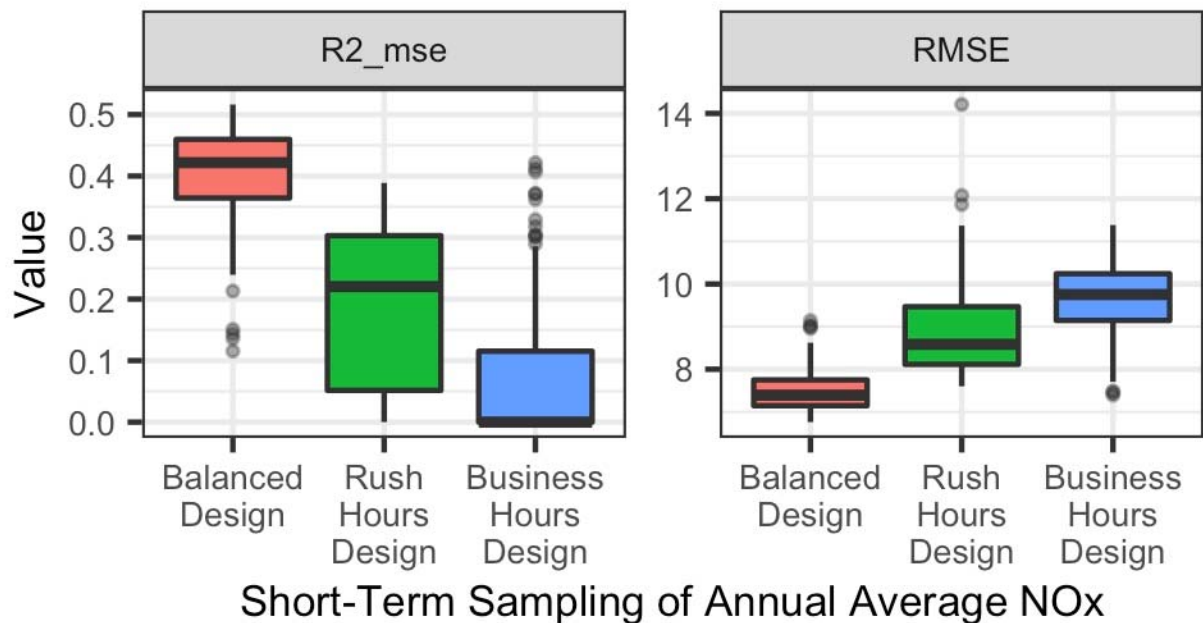
21 Competing Financial Interests: The authors declare they have no actual or potential competing
22 financial interests.

23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43

Abstract

Mobile monitoring campaigns to estimate long-term air pollution levels are becoming increasingly common. Still, many campaigns have not conducted temporally-balanced sampling, and few have looked at the implications of such study designs for epidemiologic exposure assessment. We carried out a simulation study of fixed-site air quality monitors to better understand how different mobile monitoring designs involving short-term stationary measurements at fixed locations impact the resulting exposure surfaces. We used Monte Carlo resampling to simulate three archetypal monitoring designs using oxides of nitrogen (NO_x) monitoring data from 69 regulatory sites in California: a year-around Balanced Design that sampled during all seasons of the year, days of the week, and all or various hours of the day; a temporally reduced Rush Hours Design; and a temporally reduced Business Hours Design. We evaluated the performance of each design's land use regression prediction model. The Balanced Design consistently yielded the most accurate annual averages; while the reduced Rush Hours and Business Hours Designs generally produced more biased results. A temporally-balanced sampling design is crucial for mobile monitoring campaigns aiming to assess accurate long-term exposure in epidemiologic cohorts.

Synopsis: Air pollution mobile monitoring campaigns rarely conduct temporally balanced sampling. We show that this results in biased annual average exposure estimates.



44

45

46 1 Introduction

47
48 A large body of evidence links long-term exposure to air pollution to adverse health
49 effects in humans, including mortality from cardiovascular outcomes and lung cancer.¹⁻⁶ An
50 increasing number of studies are using mobile monitoring campaigns to assess average long-term
51 air pollutant levels.⁷⁻¹³ Mobile monitoring campaigns typically equip a vehicle with air monitors
52 and collect short-term samples while in motion (non-stationary sampling) and/or while stopped
53 (stationary sampling). The focus of this analysis is on the latter mobile monitoring design. A
54 single monitoring platform can be used to collect samples at many specified locations within a
55 relatively short period of time, making it a time and cost-efficient sampling approach. Mobile
56 campaigns are particularly well-suited for multi-pollutant monitoring of less frequently
57 monitored traffic-related air pollutants that require expensive instruments or instruments that
58 need frequent attention during the sampling period.

59 And while a few studies have investigated the number of sampling locations and repeat
60 samples needed to improve the resulting exposure surfaces from mobile monitoring
61 campaigns,^{14,15} to the best of our knowledge, none have considered the importance of conducting
62 temporally-balanced sampling when the goal is estimation of a long-term average for
63 epidemiologic application. This is particularly relevant since many pollutants, particularly those
64 related to traffic, experience strong diurnal and seasonal concentration trends.^{16,17} Collecting
65 limited or unbalanced sampling may thus be sufficient to answer questions surrounding peak
66 concentrations or source identification, but it may produce biased long-term estimates and be
67 inadequate for epidemiologic applications.²² In general, many mobile monitoring campaigns
68 have been short, lasting from a few weeks to months and with few repeat visits to each location

69 spanning one to three seasons.^{9,18–20} Most campaigns have conducted sampling during weekday
70 business or rush hours, ignoring the surrounding hours, when air pollution concentrations can be
71 drastically different.^{8,17,21} Furthermore, short-term mobile monitoring campaigns often collect
72 non-stationary (mobile) measurements, which can be much shorter in duration than stationary
73 campaigns (e.g., a few seconds per road segment vs minutes or hours per stop location). It is an
74 open question whether these shorter sampling times along with the platform’s increased
75 proximity to immediate vehicle sources (e.g., in a traffic queue while stopped at a traffic signal)
76 may produce more biased and less precise exposure surfaces when compared to short-term
77 stationary monitoring.^{12,23–26}

78 The goal of this paper is to shed light on the temporal design of a short-term stationary
79 mobile monitoring campaign for application to epidemiologic cohort studies. We carry out a set
80 of simulation studies to better understand the role of mobile monitoring design on the prediction
81 of annual average surfaces. We use existing monitoring data from California to compare the
82 primary, annual site averages when all of the data are included to subsequent analyses utilizing
83 subsets of the data. These data provide a unique opportunity to explore how short-term stationary
84 sampling strategies can influence the resulting estimated annual-average concentration. Our
85 analysis requires having a long-term, comprehensive set of measurement data, which therefore
86 necessitates using fixed-site measurements rather than mobile measurements, to shed light on an
87 aspect of study design for short-term stationary mobile monitoring.

88

89

90 2 Methods

91

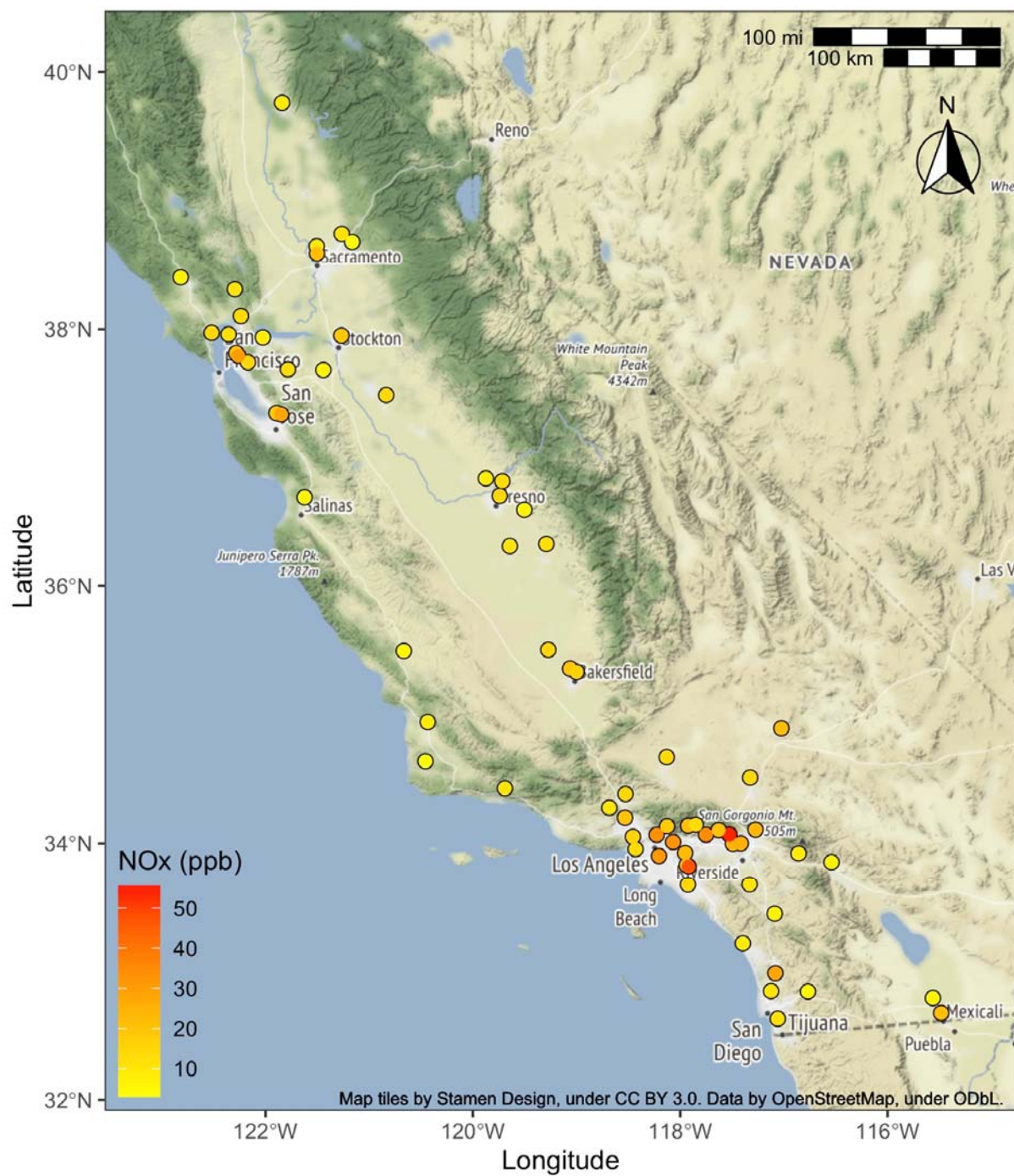
92 2.1 Data

93

94 We simulate three sampling designs (see below) using hourly observations for oxides of
95 nitrogen (NO_x) collected during 2016 from regulatory Air Quality System (AQS) sites in
96 California. NO_x was selected since it is a spatially and temporally variable traffic pollutant with
97 a strong diurnal pattern,^{8,27,28} and it is measured at many regulatory monitoring sites in
98 California, providing a large enough dataset for this analysis.²⁹

99 We included 69 of 105 California AQS sites that met various criteria (**Error! Reference**
100 **source not found.**, SI Figure S3). First, sites needed to have readings at least 66% of the time
101 (5,797/8,784 hourly samples; 2016 was a leap year). Second, sites needed to have sampling
102 throughout the year, such that data collection gaps were a maximum of 45 days long. These two
103 criteria are in line with other air quality work.³⁰ Third, sites were required to have sampled for at
104 least 40% of the time during various two-week periods that were used in two of our “common”
105 designs (described below). This sample size ensured that we could sample during these periods
106 without replacement. Fourth, sites were required to have positive readings (> 0 ppb) at least 60%
107 of the time, thus ensuring that sites had sufficient variability in their concentrations and allowing
108 us to model annual averages on the natural log scale. Finally, sites in rural and industrial settings
109 (as determined by the US EPA)³¹ were excluded since these do not represent where the majority
110 of people reside. The resulting sites were in both urban and suburban settings, in residential and
111 commercial areas.

112



113
114 *Figure 1. AQS sites included in this analysis (N=69) and their true annual average NOx measurements, as*
115 *measured by the long-term Year-Around Balanced Design Version 1 (see Methods for details).*

116

117 2.2 Sampling Designs

118

119 We conducted simulation studies to characterize the properties of three sampling designs

120 (

121 Table 1, Supplementary Information [SI] Figure S1). Each design has a long- and a short-
122 term sampling approach. Long-term approaches use all of the data that meet each design's
123 definition to estimate site annual averages and are analogous to traditional, fixed-site sampling
124 approaches where sampling at a given location occurs over an extended period of time. Short-
125 term approaches only collect 28 samples per site and are analogous to mobile monitoring
126 campaigns that collect a few repeat samples per site. (The cut-off of 28 samples reflects our
127 preliminary analyses showing that 28 hourly NO_x samples are sufficient to estimate a site's
128 annual average within about 25% error or less [SI Figure S2].) Each design has multiple versions
129 where samples are collected at slightly different times. The various design versions are intended
130 to reflect the bias produced if only certain times are included in the measurements. We simulated
131 each short-term sampling approach 30 times (Monte Carlo resampling), and hereafter refer to
132 each of these simulations as a "campaign" since each represents a potential mobile monitoring
133 study.

134

135

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

Design	Sampling Approach and Number of Samples²	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	<u>V1</u> (All Hours) <u>V2</u> (Most Hours): 5 AM – 12 AM <u>V3</u> (Truncated Hours): 6-9 AM, 1-5 PM, 8-10 PM
	Short-term (28 hourly samples per site x 30 campaign	7 samples per season	5/7 weekday samples; 2/7 weekend samples	Random hours according to V1, V2, or V3

simulations)

Two-Season Weekday “Rush Hours” Design⁴	Long-term (Max 160 hourly samples per site x 1 campaign simulation)	<u>V4-5</u> : winter & summer (2-wk period per season) <u>V6-7</u> : spring & fall (2-wk period per season)	Mon – Fri	7-10 AM, 3-6 PM
	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” Design⁴	Long-term (Max 180 hourly samples per site x 1 campaign simulation)	<u>V4-5</u> : winter & summer (2-wk period per season) <u>V6-7</u> : spring &	Mon – Fri	9 AM – 5 PM

	fall (2-wk period per season)	
Short-term	14 samples per season	Random Business Hours
(28 hourly samples per site x 30 campaign simulations)		according to V4- 5 or V6-7

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

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

145 ⁴ See SI Table S1 for each version’s exact sampling periods

146

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

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

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

163

164 2.3 Prediction Models

165

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

175 To best understand the role of design, we present results for annual average estimates,
176 predictions, and model performance statistics. In descriptive analyses, we compare design-
177 specific annual average estimates and predictions to the gold standard (long-term Balanced
178 Design Version 1). We compare predicted site concentrations against predictions from the gold
179 standard since epidemiologic air pollution studies often rely on predicted exposure, and the gold
180 standard prediction represents the best possible prediction of annual-average concentrations that
181 a study could hope to achieve. We complement this approach with model assessment evaluations
182 of design-specific site predictions against two different references: an assessment against the true
183 averages, and a traditional model assessment evaluation against the respective design-specific
184 annual average estimates. The traditional assessment compares the predicted exposures to the
185 observed site measurements from which they were derived. This allows us to document the
186 quantities that would normally be available from modeling the data measured from any specific
187 campaign. We summarize the model performance in terms of cross-validated mean squared error
188 (MSE)-based R^2 (R^2_{MSE}), regression-based R^2 (R^2_{reg}), and root mean squared error (RMSE).
189 R^2_{MSE} assesses whether two sets of measurements such as estimates and predictions are the same
190 (along the 1-1 line), and thus reflects both bias and variation around the one-to-one line (see SI
191 Equations 1-3 for definitions). R^2_{reg} , on the other hand, assesses whether observations are
192 linearly associated (based on the best fit line though not necessarily the 1-1 line) and thus adjusts
193 for bias and slopes different than one. R^2_{reg} is defined as the squared correlation between two sets
194 of measurements.

195 In sensitivity analyses, we repeated these simulations for nitrogen dioxide (NO₂) and
196 nitrogen monoxide (NO), adding a two ppb constant to all of the hourly NO readings before log-
197 transforming to eliminate negative and zero concentration readings. Furthermore, we conducted

198 NOx simulations for a subset of sites (N=17) within the Los Angeles (LA) and San Diego
199 Counties, refitting PLS models to these sites alone. This region was meant to represent a
200 potential area of interest for epidemiologic exposure assessment and one that could be more
201 feasibly covered by a mobile monitoring campaign, though it had a reduced sample size.

202

203 *All analyses were conducted in R (v 3.6.2, using RStudio v 1.2.5033).³² SI Note S1 lists the R packages*
204 *used. All map tiles were created by Stamen Design³³ under CC BY 3.0,³⁴ using data by OpenStreetMap*
205 *under ODbL.³⁵*

206

207 3 Results

208

209 3.1 Hourly Readings

210

211 Sites (N=69) had on average (SD) of 8,090 (361) hourly readings, the equivalent of 337
212 (15) days of full sampling (See SI Table S3). Average (SD) hourly NOx concentrations were 16
213 (21) ppb (See SI Table S4). Sites had seasonal, daily, and hourly concentration patterns, with
214 trends being more pronounced at some sites than others (See SI Figure S4-S6).

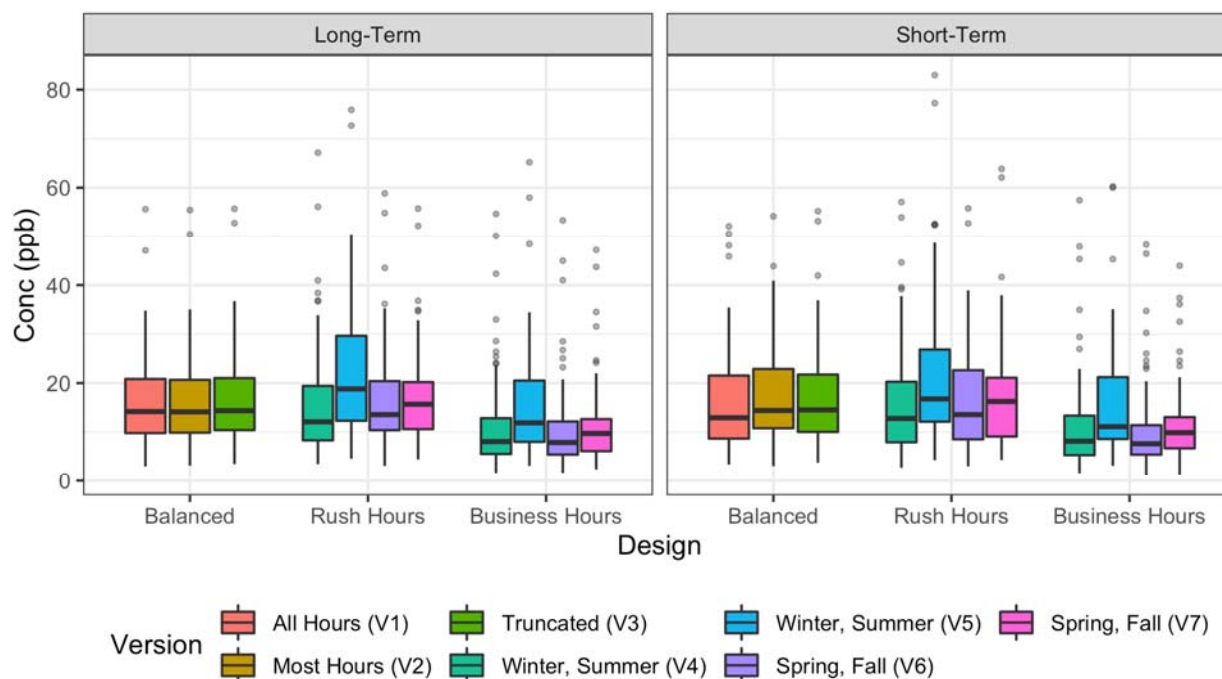
215

216 3.2 Annual Average Estimates

217

218 Across the 69 monitor locations, measured annual average concentrations (long-term
219 Balanced Design Version 1), had a median (IQR) of 14 (10 - 21) ppb and ranged from 3-56 ppb.

220 The short-term and long-term sampling approaches resulted in similar distributions of annual
 221 averages for different design versions. Figure 2 shows the long-term and a single short-term
 222 approach for each design. Overall, the long-term and short-term approach for each design
 223 version had very similar distributions. All of the Balanced Design versions resulted in only slight
 224 differences in their medians and IQRs. The Rush Hours Design versions generally resulted in
 225 slightly higher annual averages than the true averages, with some versions being more variable
 226 and having somewhat different distributions. The Business Hours Design versions resulted in
 227 annual averages that were generally lower than the true averages and less variable than the Rush
 228 Hours Design versions. See SI Table S5 for summary statistics. SI Figure S7 shows annual
 229 average estimates for all campaigns and pollutants.
 230

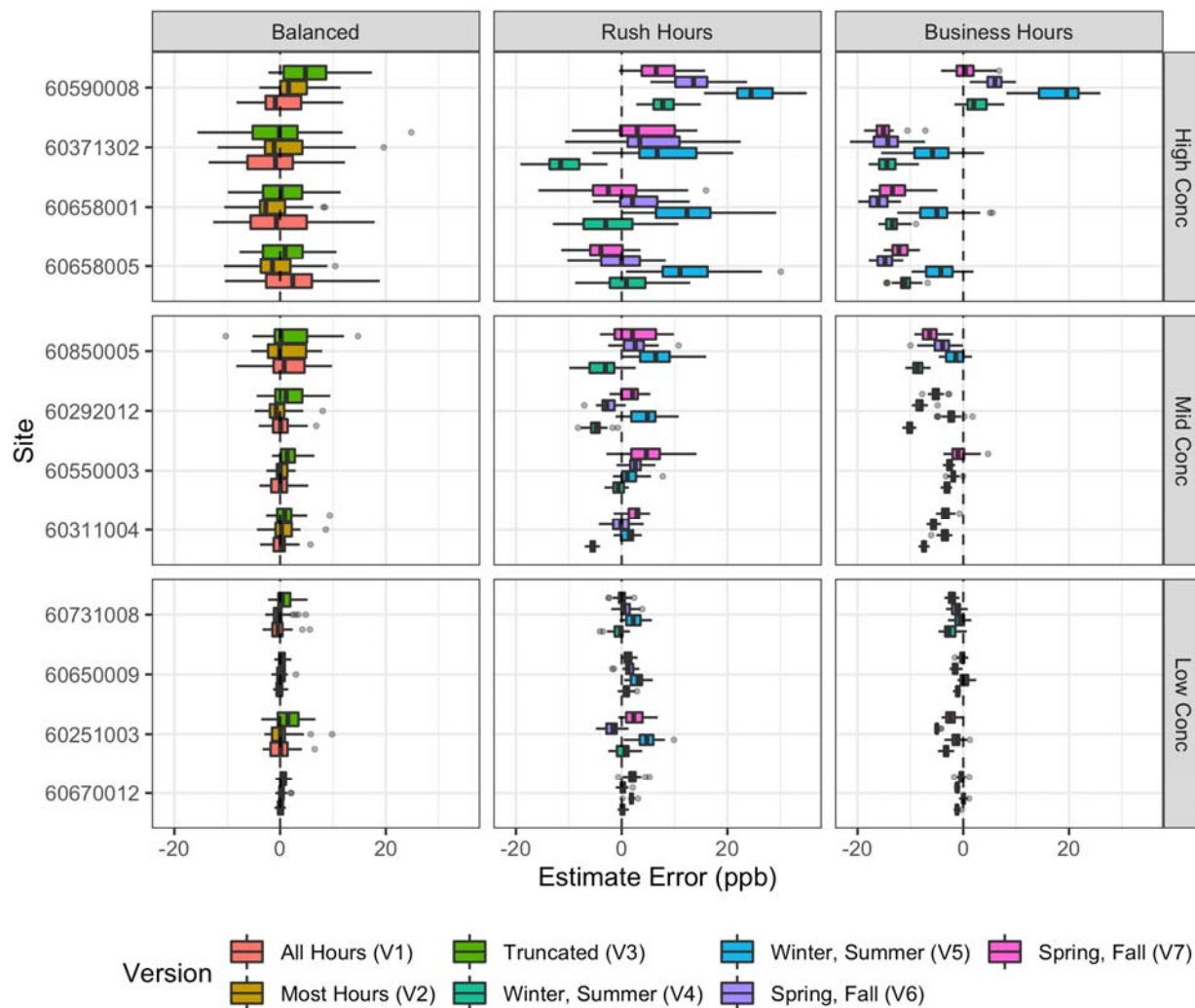


231
 232 *Figure 2. Distribution of NO_x annual averages (N=69 sites) from different design versions. Showing the*
 233 *one campaign for each long-term approach and one example campaign for each short-term approach.*

234

235 Figure 3 shows the site-specific distributions of annual averages across designs for short-
236 term approaches relative to the true averages for a stratified random sample of 12 sites. Sites are
237 stratified by whether their true mean concentration was in the low (<25th percentile), middle
238 (25th-75th percentile) or high (>75th percentile) concentration category. The variation of averages
239 across campaigns increases with concentration in all designs. Site-specific averages are similar to
240 the true averages for all Balanced Design versions while there were multiple sites from the
241 Business Hours Design versions with averages systematically lower. The Rush Hours Design
242 versions also had many biased averages, although the direction of the bias varied by site and
243 design version. SI Figure S8 shows these biases for all sites.

244



245
 246 *Figure 3. Site-specific NO_x measurement error for short-term designs (N = 30 campaigns) as compared to*
 247 *the true annual average at that site (long-term Balanced Design Version 1). Showing a stratified random*
 248 *sample of 12 sites, stratified by whether their true concentration was in the low (<25th percentile), middle*
 249 *(25th-75th percentile) or high (>75th percentile) concentration category and arranged within each stratum*
 250 *with lower concentration sites being closer to the bottom.*

251

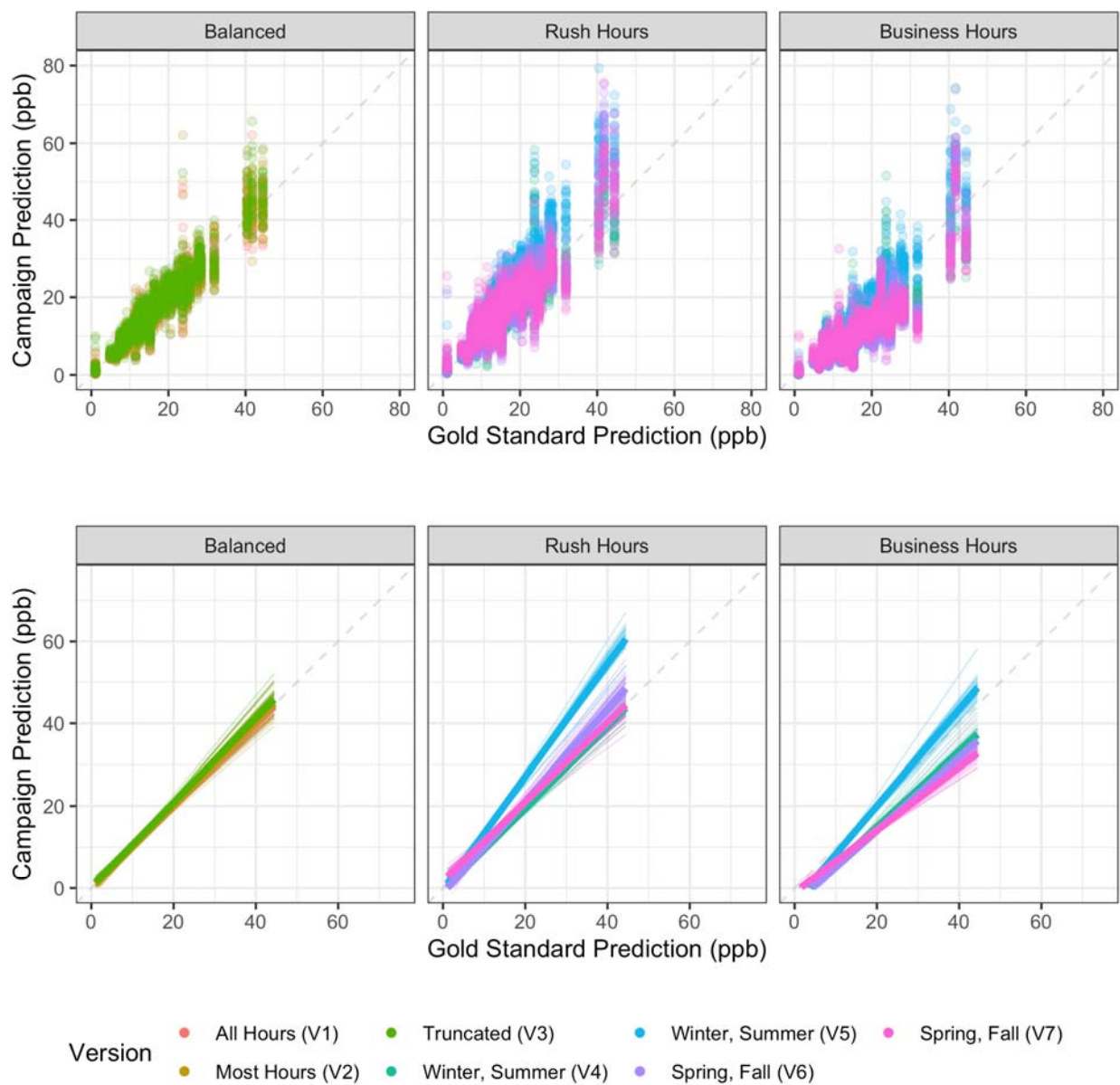
252 3.3 Model Predictions

253

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

256 We compared PLS model predictions from each short-term design to the gold standard
257 model predictions. SI Figure S9 shows the relative standard deviations of predictions by design
258 version, with 1 indicating that design predictions have the same standard deviation as the gold
259 standard model predictions. Overall, the Balanced Design predictions have similar variability to
260 those of the gold standard (range: 0.87-1.28), the Rush Hours Design predictions are more
261 variable (range: 0.90-1.74), and the Business Hours Design predictions are mixed: some less and
262 some more variable (range: 0.73-1.54). Figure 4 displays these comparisons as scatterplots and
263 best fit lines. The scatterplots show that there are a few sites, some of which have high leverage,
264 that have variable predictions in all designs. From the best fit lines, we observe that all short-
265 term Balanced Design versions resulted in the most accurate predictions on average, as indicated
266 by their overlapping general trends along the one-to-one line. The Rush Hours Design versions
267 were more likely to have a positive general trend, while the Business Hours Design versions
268 were more likely to have a negative general trend, indicating, for example, that higher
269 concentrations were more likely to be over- or under-estimated, respectively. However, there
270 was heterogeneity in this overall pattern across the Rush and Business Hours Design versions.
271 Furthermore, there was additional heterogeneity across individual campaigns. The SI contains
272 comparable figures comparing design predictions to the gold standard and additional figures for
273 NO and NO₂ (SI Figures S10-S13).

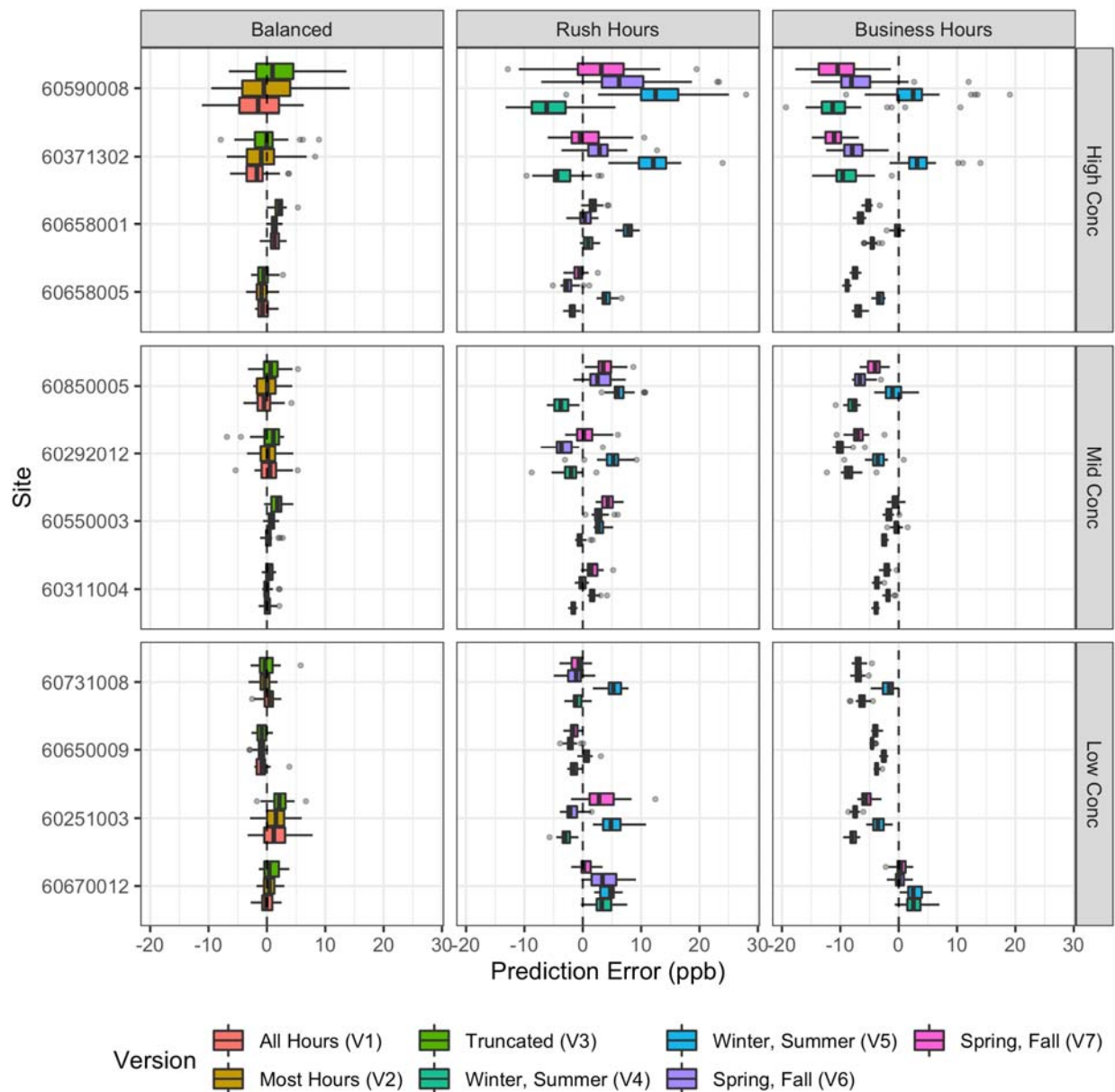
274



275
 276 *Figure 4. Scatterplots and best fit lines of cross-validated short-term predictions for 30 campaigns vs the*
 277 *gold standard predictions for NOx. Thin transparent lines are individual campaigns, colored by design*
 278 *version; thicker lines are the overall version trend. (One prediction is excluded for clarity from the Rush*
 279 *Hours Version 4 scatterplot at x=24 ppb, y=109 ppb [site 60731016] but included in the line plots.)*

280

281 Figure 5 shows site-specific comparisons of predictions across 30 short-term campaigns
282 relative to the gold standard predictions for a stratified random sample of 12 sites in order to
283 characterize relative bias (see SI Figure S14 for all sites). Overall, the short-term Balanced
284 Design predictions had a median (IQR) bias of 0.2 (-1 – 1.4) ppb relative to the gold standard
285 predictions (see SI Table S7 for details). All Balanced Design predictions were very similar to
286 the gold standard predictions, though some sites frequently had larger biases. The Rush Hours
287 and Business Hours Design versions were more likely to consistently produce biased site
288 predictions, with a median (IQR) bias of 1.2 (-1.2 – 4) ppb and -3.8 (-6.6 – -1.4) ppb,
289 respectively. While the Rush Hours Design versions generally resulted in higher predictions
290 across sites (with some inconsistency across versions for a few sites), the Business Hours Design
291 versions resulted in predictions that were both lower and higher than the gold standard
292 predictions across sites. There were also a few sites that tended to have more biased and/or more
293 variable predictions relative to the gold standard across all designs. We observed similar patterns
294 when looking at estimate (rather than prediction) biases (See Figure 3, SI Figure S8).
295



296

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

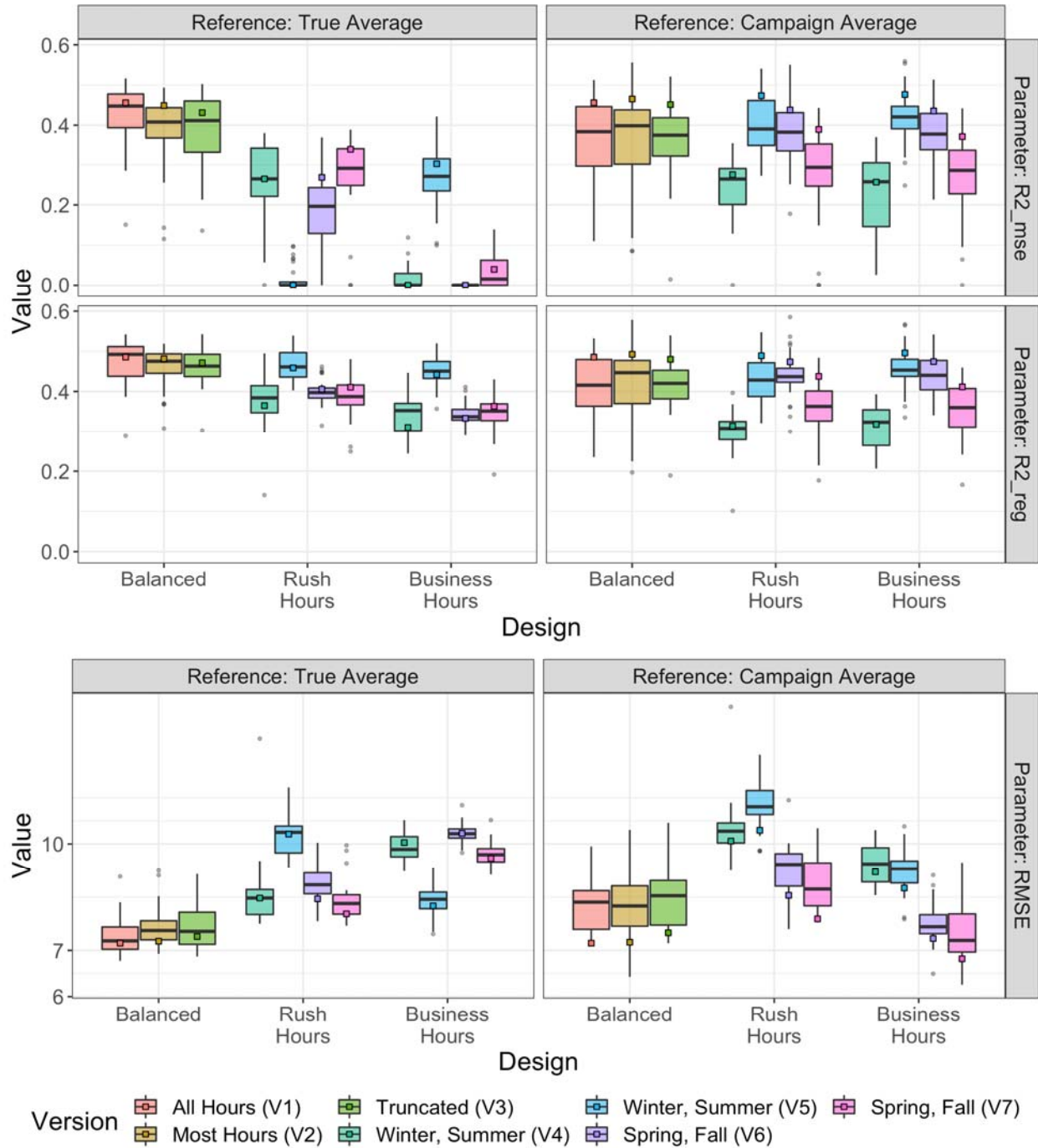
302

303 3.4 Model Assessment

304

305 Figure 6 shows the out-of-sample prediction performances relative to the observations
306 from the true averages (left column) and the specific design (right column), for both the long-
307 term and short-term approaches. The boxplots quantify the distribution of performance statistics
308 across all 30 short-term campaigns while the squares show the performance for the long-term
309 approach of the same design version. When assessed against the true averages, all the Balanced
310 Design versions generally perform better than either the Rush Hours or Business Hours Design
311 versions with higher $CV R^2_{MSE}$ and $CV R^2_{reg}$, and lower $CV RMSE$ estimates. This is particularly
312 apparent for the long-term approach. Furthermore, within design the performance for the long-
313 term approach is better than the majority of the short-term campaigns. There is considerable
314 heterogeneity in performance across the Rush Hours and Business Hours Design versions. In
315 contrast, when assessed against observations from the same design, as would typically be done in
316 practice, the role of sampling design on prediction performance is not as evident. The superior
317 performance of the Balanced Design is not as apparent, and some of the Rush Hours and
318 Business Hours Design versions appear to perform better. There are also a few campaigns that
319 show poor performance, even under the Balanced Design. SI Figure S15-S16 show similar
320 results for NO_2 and NO , with NO showing more variability and some lower performing statistics.
321 Stratifying by whether sites were considered to have high or low variability (based on hourly
322 standard deviation estimates) showed similar R^2 and $RMSE$ patterns (data not shown).

323



324

325 *Figure 6. Model performances (MSE-based R2, Regression-based R2, and RMSE), as determined by each*

326 *campaign's cross-validated predictions relative to: a) the true averages (long-term Balanced Version 1),*

327 *and b) its respective campaign averages. Boxplots are for short-term approaches (30 campaigns), while*

328 *squares are for long-term approaches (1 campaign).*

329

330 3.5 Sensitivity Analyses

331

332 Findings were similar for sensitivity analyses (see the SI for NO and NO₂ results). Figure

333 7 and

334 Table 2 further illustrate the resulting predictions for the Los Angeles-San Diego analysis for the

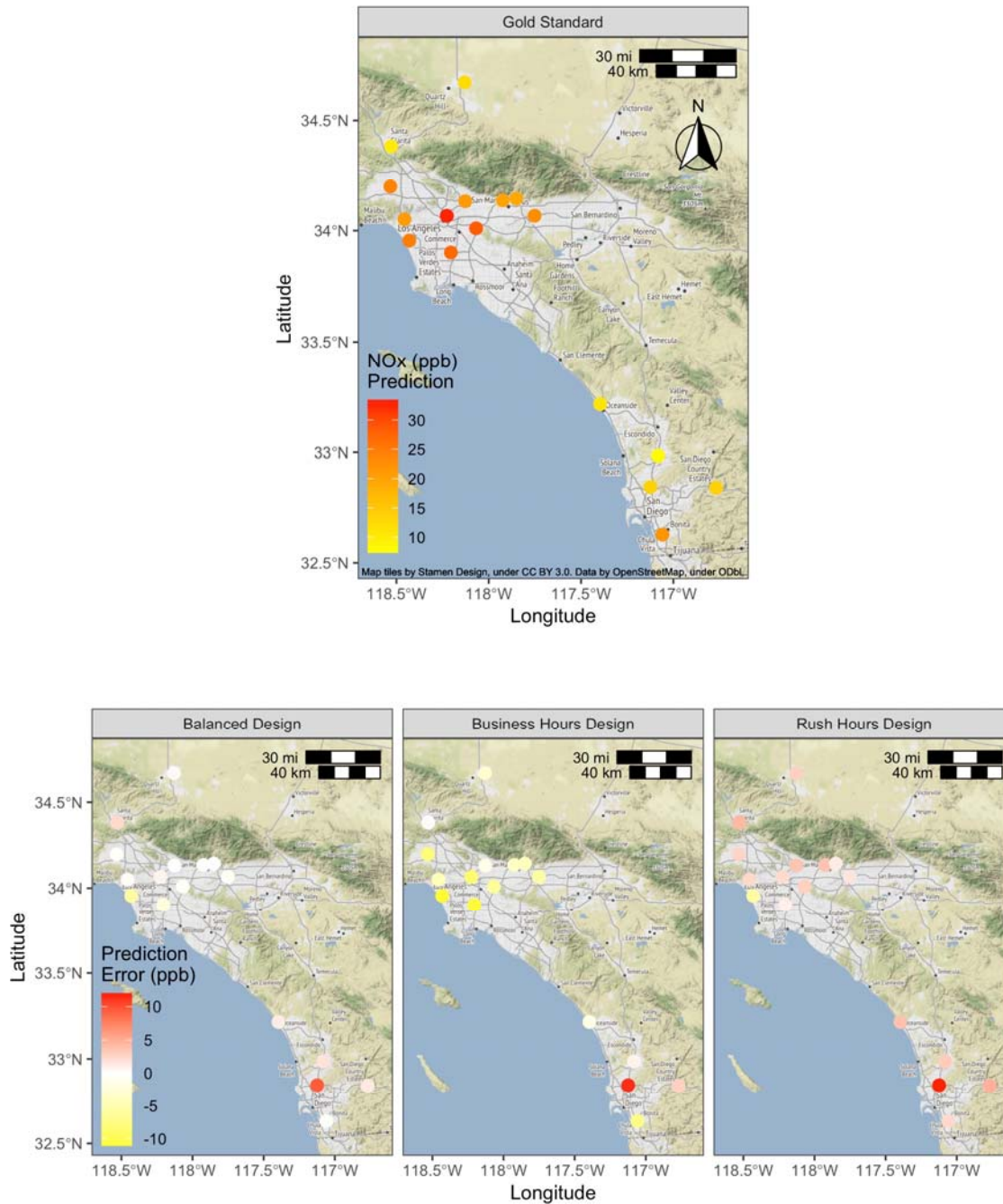
335 gold standard campaign (long-term Balanced Version) and each of the short-term designs. Short-

336 term designs estimates are for the average site prediction across all simulations and design

337 versions for simplicity. Compared to the gold standard campaign, the median prediction bias

338 (and percent error) for the Balanced, Rush Hours and Business Hours designs was about 0.0 ppb

339 (13.2%), 2.1 ppb (20.4%) and -4.0 (27.5%), respectively.



340

341 *Figure 7. Site predictions from the gold standard campaign (long-term, Balanced Design, All Hours) and*
342 *prediction errors from each short-term design, as compared to the gold standard campaign, for the Los*
343 *Angeles-San Diego sensitivity analysis (N = 17 sites).*

344

345 *Table 2. Site prediction error by design relative to the gold standard campaign predictions for the*
346 *southern California sensitivity analysis (No. Predictions = 17 sites x 30 simulations/version x 3-4*
347 *versions/design).*

Design	No. Predictions	Absolute Error (ppb)		Percent Error (%)	
		Median	IQR	Median	IQR
Balanced	1530	0.0	4.6	13.2	17.2
Rush Hours	2040	2.1	6.8	20.4	28.7
Business Hours	2040	-4.0	7.6	27.5	26.0

348

349 4 Discussion

350

351 In this paper we have used existing regulatory monitoring data to deepen our
352 understanding of the importance of short-term stationary mobile monitoring study design for
353 application to epidemiologic cohort studies. Others have shown that short-term data can be used
354 to estimate long-term averages.^{8,9} What has been missing from the literature until now, however,
355 is the impact of short-term stationary mobile monitoring study design on the accuracy and
356 precision of long-term exposure estimates and model predictions, particularly when the goal is to
357 produce predictions for an epidemiologic study. Our results indicate that for designs with a
358 sufficient number of short-term samples at each location (about 28 or more), the design rather
359 than the sampling approach (i.e., sampling duration at a given site) has the largest impact on the
360 estimated annual averages. We focus the rest of this discussion on the short-term approaches for

361 each design, which resemble mobile monitoring, though the long-term approaches produced
362 similar results.

363 In terms of specific design, we found that all of the Balanced Design versions resulted in
364 similar annual averages as the true averages (long-term Balanced Version 1), while the Rush
365 Hours and Business Hours Design versions were more likely to result in more biased and more
366 or less variable annual average estimates. Specifically, the Rush Hours Design was more likely
367 to overestimate, while the Business Hours Design was more likely to underestimate site
368 averages. This result was likely because the Balanced Design captured much of NO_x's temporal
369 variability by allowing for samples to be collected during each season, day of the week, and all
370 or most times of the day, all periods during which meteorology and traffic activity patterns
371 impact air pollution concentrations (SI Figure S4-S6). The Rush Hours Design, on the other
372 hand, was restricted to two sampling seasons and was more likely to sample during high
373 concentration times of day and days of the week. The Business Hours Design had similar
374 limitations though it was more likely to sample during low concentration times. These
375 conclusions were the same in the Los Angeles-San Diego sensitivity analysis, which is more
376 representative of a geographic area that could be realistically sampled by a mobile campaign.

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

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

391 At the site level, we saw that while any individual study campaign had the potential to
392 produce biased estimates and predictions, the Rush Hours and Business Hours Designs were
393 more likely to do so than the Balanced Design. The direction and magnitude of bias varied by
394 site and depended upon the sampling design and the typical seasonal, day of week, and time of
395 day patterns of pollution at that site. This suggests a simple correction factor to time-adjust short-
396 term measurements based on long-term observations at a small number of reference sites, for
397 example using regulatory fixed sites, is unlikely to fully adjust for bias at the site level.²² While
398 many past campaigns have taken this approach to account for the fact that short-term stationary
399 mobile sampling inherently misses some observations, this approach makes a strong assumption
400 that all sites have the same temporal trends. SI Figure S17 – S19 illustrate the temporal trends for
401 sites included in the Los Angeles-San Diego analysis and clearly shows how lower concentration
402 “background” sites are also more likely to have less temporal variation when compared to other
403 sites. Using these “background” sites (or any other site for that matter) to adjust readings at other
404 sites would not substantially reduce the bias from an unbalanced sampling design. This may be
405 especially pertinent for mobile monitoring campaigns since their increased spatial coverage is
406 more likely to capture localized pollution hotspots that may have even more temporal variation.

407 We thus argue that sampling design should be prioritized in mobile monitoring campaigns.

408 Analytical methods such as temporal adjustment factors, on the other hand, should be further
409 investigated to establish their true value given their strong assumptions.

410 Furthermore, non-balanced designs may misrepresent some sites more than others and
411 lead to differential exposure misclassification in epidemiologic studies since higher
412 concentration sites were more likely to have greater degrees of bias and variation (Figure 4 –
413 Figure 5). Thus, while non-balanced designs may be appropriate for non-epidemiologic purposes
414 including characterizing the spatial impact of traffic related air pollutants during peak hours for
415 urban planning and policy purposes, these could be misleading in epidemiologic applications.

416 In this study we were able to evaluate prediction model performance against the true
417 annual average NO_x exposure as well as against the observations typically available for model
418 performance assessment. Performance assessment against the true averages indicates that the
419 Balanced Design is clearly the best, and that there is little degradation in performance across
420 design versions. This means it is possible to design high quality short-term stationary mobile
421 monitoring studies that accommodate some measure of logistical feasibility, for example, by not
422 requiring sampling in the middle of the night. In contrast, the performance of the Rush Hours and
423 Business Hours Designs is comparatively worse, indicating that the logistically appealing
424 approach that samples only four weeks during two seasons, during daytime hours, and only
425 during weekdays is inadequate for providing high quality estimates of annual averages. Further,
426 the performance of these designs varies considerably and unpredictably depending upon the
427 specific pair of two-week periods that are selected for sampling. Additionally, comparison of the
428 two R^2 estimates (R^2_{MSE} and R^2_{reg}) indicates that not all of their poor performance is due to the
429 inability to predict the same value as the truth (R^2_{MSE}), but due to systematic bias in the design.

430 As noted earlier, R^2_{MSE} assesses whether two measurements are the same - along the 1-1 line,
431 whereas R^2_{reg} simply assesses whether they are linearly associated. SI Figure S13, for example,
432 shows that Balanced Designs generally produce predictions that are more similar to the “true”
433 estimates from a gold standard campaign (closer to the 1-1 line), whereas the Rush Hours and, in
434 particular, the Business Hours Designs are more likely to produce predictions away from the 1-1
435 line. This results in the Balanced Designs having R^2_{MSE} estimates that are only slightly lower
436 than R^2_{reg} estimates, whereas this drop in performance is greater for the Rush Hours and Business
437 Hours Designs, as seen in Figure 6.

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

445 It is notable that the performance of short-term stationary mobile campaigns were fairly
446 consistent with, though generally slightly worse than, the performance observed in the longer-
447 term campaigns for each design version (Figure 6). However, occasionally there was an
448 “unlucky” short-term campaign with meaningfully poorer performance than the other campaigns
449 of the same design. This was more likely in the non-balanced designs, though even the Balanced
450 Design versions had 1-2 of the 30 campaigns (~3-6%) with notably worse performance as
451 quantified by R^2 . It may be possible that this result is driven by a few high-leverage outlier sites
452 that impact the prediction model performance. In practice, mobile monitoring study investigators

453 are likely to investigate high-leverage sites and address their influence in their prediction
454 modeling.

455 Our study focused on short-term stationary mobile campaigns with 28 repeat samples per
456 site. We did not consider campaigns with fewer or more visits. As evident in SI Figure S2, the
457 percent error in estimating the annual average from fewer than 25 visits skyrockets, suggesting
458 that site estimates will be considerably noisier in mobile campaigns with few repeat visits,
459 regardless of the study design. Prediction model performance is thus likely to decrease as the
460 number of visits per site decrease. Logistically, it is also difficult to achieve balance in sampling
461 over time across season, day of week, and time of day with fewer than 28 samples per site.
462 Furthermore, we note that this study focused on a few generalizable, common designs in the
463 literature, though many other approaches have been taken. We expect that the variety of mobile
464 campaign designs that have been implemented will all produce slightly different results.

465 In putting these results in context, it is important to recognize that in this simulation study
466 we are using NO_x hourly averages to approximate much shorter-term sampling durations (e.g., a
467 few minutes or less) than would be collected during a mobile monitoring campaign. Shorter
468 duration sampling will affect the noise in the data, to an amount that depends on the environment
469 (e.g., temporal patterns in the concentrations of the pollutant being measured) and the
470 instruments. For comparison, however, our additional evaluations of minute-level data suggest
471 that the decrease in percent error in going from two-minute to hour-long samples is at most a few
472 percent because of serial correlation in the data. This thus gives us confidence that the findings
473 from this work are still generalizable to more common, shorter-term stationary monitoring
474 campaigns with sampling periods closer to a few minutes.

475 Further, our study took place throughout California, a large, geographically diverse area
476 with varying climate profiles.³⁶ While such a large sampling domain would be challenging for a
477 real-world mobile monitoring campaign, the overall conclusions of this study – the importance of
478 temporally-balanced sampling, are also supported in the Los Angeles-San Diego sensitivity
479 analysis. In terms of the siting criteria for the regulatory monitoring sites where the data came
480 from, locations are generally meant to capture representative population exposures, including
481 near roadway, at various spatial scales ranging from microscales (< 100 m range) to regional
482 scales in order to inform regulatory compliance.^{37,38} This should thus have provided us with
483 decent spatial coverage and concentration variability. Most air pollution studies, in fact, rely on
484 this network of regulatory monitors.³⁹ Still, when compared to most mobile monitoring
485 campaigns, this study’s larger domain and reduced exposure variability may have produced
486 lower prediction model performances than would be expected from mobile monitoring
487 campaigns.

488 Another distinction is that while we sampled measurements within sites at random,
489 mobile campaigns typically sample from sites along a fixed route or in a designated area. The
490 actual sampling scheme will thus depend on the exact route developed and the number of
491 platforms deployed, both of which are beyond the scope of this paper. In general, sampling along
492 a route also induces some spatial correlation in the mobile monitoring data. This dependence is
493 often overlooked in mobile monitoring campaigns and was not addressed in this study.
494 Furthermore, we did not consider the importance of the distribution of sampling locations in this
495 study, which is particularly relevant when the exposure assessment goal is an epidemiologic
496 application. Selecting sites that are representative of the target cohort’s residence locations will
497 ensure the spatial compatibility assumption is met, which is an important way to reduce the role

498 of exposure measurement error in epidemiologic inference.⁴⁰ This consideration is especially
499 relevant for mobile monitoring near major sources (e.g., airports, marine activity, and
500 industry),^{8,9,41-47} which may or may not represent a study cohort of interest.

501 Our evaluation focused on NO_x, NO, and NO₂, which are quickly and moderately
502 decaying air pollutants (concentrations reach background levels approximately 400-600 m from
503 roadway sources).²¹ Campaigns that measure these pollutants may be more susceptible to
504 sampling design than campaigns that measure less spatially- and/or temporally-variable
505 pollutants such as PM_{2.5}.²⁷ We selected NO_x, NO, and NO₂ because these traffic-related
506 pollutants are often measured in short-term campaigns, and data for these pollutants are more
507 widely available. Non-criteria pollutants, for example ultrafine particulates (UFP), however,
508 have also received increasing attention in recent years given their emerging link to adverse health
509 effects.⁴⁸⁻⁵¹ Still, high-quality information about their spatial distribution is essentially absent,
510 and most studies have implemented short-term mobile sampling approaches⁴⁷ that may not be
511 temporally balanced and potentially be misleading. Finally, while other discrepancies surely
512 exist between this simulation study and realized mobile monitoring campaigns, we expect our
513 overall conclusions on the importance of temporally-balanced sampling to remain the same.

514 An important next step in this work is to understand whether the differences in exposure
515 estimates that we observed across study designs have a meaningful impact on epidemiologic
516 inferences. This is of particular interest considering that year-around, balanced designs are
517 resource-intensive and rare, while shorter, more convenient campaigns are more common in the
518 literature. More research is needed to better understand how and whether unbalanced mobile
519 monitoring campaigns may contribute high quality exposure assessments for epidemiology.
520 Regardless of design, we expect that the predictions from all of the campaigns will result in both

521 classical-like and Berkson-like error.^{40,52–54} Specifically, the predictions capture only part of the
522 true long-term exposure (Berkson-like error), while the parameters in the prediction model are
523 inherently noisy (classical-like error). However, these measurement error methods have not to
524 date considered exposure assessment study design, beyond considering the importance of spatial
525 compatibility, i.e., that distribution of monitoring locations is the same as the distribution of
526 participant locations. Our work suggests that deeper understanding of the role of exposure
527 assessment design on epidemiologic inference is an important area of research.

528

529 4.1 Conclusions and Recommendations for Mobile Monitoring Campaigns

530

531 Mobile monitoring study design should be an important consideration for campaigns
532 aiming to assess long-term exposure in an epidemiologic cohort. Given the temporal trends in air
533 pollution, campaigns should implement balanced designs that sample during all seasons of the
534 year, days of the week, and hours of the day in order to produce unbiased annual averages.
535 Nonetheless, restricting the sampling hours in balanced designs, for example due to logistical
536 considerations, will still generally produce unbiased estimates at most sites. On the other hand,
537 unbalanced sampling designs like those often seen in the literature are more likely to produce
538 biased annual estimates, with some sites being more biased than others. And while predictions
539 from these restricted designs may at times perform similarly to balanced designs (or, more
540 problematically, may erroneously *appear* to perform similarly when evaluated against
541 measurements which are themselves biased samples), this performance may strongly depend on
542 the exact sampling period chosen and may thus be difficult or impossible to anticipate prior to
543 conducting a new sampling campaign. Furthermore, the differential exposure misclassification

544 that may result from these designs may be problematic in epidemiologic investigations. Finally,
545 studies that implement unbalanced sampling designs are likely to have hidden exposure
546 misclassification given that both the observations and model predictions may be systematically
547 incorrect. By implementing a balanced sampling design, campaigns can thus increase their
548 likelihood of capturing accurate annual average exposure averages.

549

550 5 Funding

551

552 This work was funded by the Adult Changes in Thought – Air Pollution (ACT-AP) Study
553 (National Institute of Environmental Health Sciences [NIEHS], National Institute on Aging
554 [NIA], R01ES026187), and BEBTEH: Biostatistics, Epidemiologic & Bioinformatic Training in
555 Environmental Health (NIEHS, T32ES015459). Research described in this article was conducted
556 under contract to the Health Effects Institute (HEI), an organization jointly funded by the United
557 States Environmental Protection Agency (EPA) (Assistance Award No. CR-83998101) and
558 certain motor vehicle and engine manufacturers. The contents of this article do not necessarily
559 reflect the views of HEI, or its sponsors, nor do they necessarily reflect the views and policies of
560 the EPA or motor vehicle and engine manufacturers.

561 6 References

562

563 1. Hoek, G. *et al.* Long-term air pollution exposure and cardio- respiratory mortality: a review.
564 *Environ. Health* **12**, 43 (2013).

565 2. Kampa, M. & Castanas, E. Human health effects of air pollution. *Environ. Pollut.* **151**, 362–
566 367 (2007).

567 3. R ckerl, R., Schneider, A., Breitner, S., Cyrys, J. & Peters, A. Health effects of particulate
568 air pollution: a review of epidemiological evidence. *Inhal. Toxicol.* **23**, 555–592 (2011).

569 4. Schwartz, J. Air pollution and daily mortality: a review and meta analysis. *Environ. Res.* **64**,
570 36–52 (1994).

571 5. Chen, H., Goldberg, M. & Villeneuve, P. A systematic review of the relation between long-
572 term exposure to ambient air pollution and chronic diseases. *Rev. Environ. Health* **23**, 243–
573 298 (2008).

574 6. Pope, C. A., Dockery, D. W. & Schwartz, J. Review of epidemiological evidence of health
575 effects of particulate air pollution. *Inhal. Toxicol.* **7**, 1–18 (1995).

576 7. Hankey, S. & Marshall, J. D. Land Use Regression Models of On-Road Particulate Air
577 Pollution (Particle Number, Black Carbon, PM2.5, Particle Size) Using Mobile Monitoring.
578 *Environ. Sci. Technol.* **49**, 9194–9202 (2015).

579 8. Apte, J. S. *et al.* High-Resolution Air Pollution Mapping with Google Street View Cars:
580 Exploiting Big Data. *Environ. Sci. Technol.* **51**, 6999–7008 (2017).

581 9. Hatzopoulou, M. *et al.* Robustness of Land-Use Regression Models Developed from Mobile
582 Air Pollutant Measurements. *Environ. Sci. Technol.* **51**, 3938–3947 (2017).

- 583 10. Patton, A. P. *et al.* Spatial and temporal differences in traffic-related air pollution in three
584 urban neighborhoods near an interstate highway. *Atmos. Environ.* (2014)
585 doi:10.1016/j.atmosenv.2014.09.072.
- 586 11. Van den Bossche, J. *et al.* Mobile monitoring for mapping spatial variation in urban air
587 quality: Development and validation of a methodology based on an extensive dataset. *Atmos.*
588 *Environ.* (2015) doi:10.1016/j.atmosenv.2015.01.017.
- 589 12. Kerckhoffs, J. *et al.* Comparison of ultrafine particle and black carbon concentration
590 predictions from a mobile and short-term stationary land-use regression model. *Environ. Sci.*
591 *Technol.* **50**, 12894–12902 (2016).
- 592 13. Xie, X. *et al.* A Review of Urban Air Pollution Monitoring and Exposure Assessment
593 Methods. *ISPRS International Journal of Geo-Information* vol. 6 (2017).
- 594 14. Hatzopoulou, M. *et al.* Robustness of Land-Use Regression Models Developed from Mobile
595 Air Pollutant Measurements. *Environ. Sci. Technol.* **51**, 3938–3947 (2017).
- 596 15. Messier, K. P. *et al.* Mapping Air Pollution with Google Street View Cars: Efficient
597 Approaches with Mobile Monitoring and Land Use Regression. *Environ. Sci. Technol.* **52**,
598 12563–12572 (2018).
- 599 16. Yu, C. H. *et al.* A novel mobile monitoring approach to characterize spatial and temporal
600 variation in traffic-related air pollutants in an urban community. *Atmos. Environ.* **141**, 161–
601 173 (2016).
- 602 17. Batterman, S., Cook, R. & Justin, T. Temporal variation of traffic on highways and the
603 development of accurate temporal allocation factors for air pollution analyses. *Atmos.*
604 *Environ.* **107**, 351–363 (2015).

- 605 18. Weichenthal, S. *et al.* A land use regression model for ambient ultrafine particles in
606 Montreal, Canada: A comparison of linear regression and a machine learning approach.
607 *Environ. Res.* **146**, 65–72 (2016).
- 608 19. Minet, L., Gehr, R. & Hatzopoulou, M. Capturing the sensitivity of land-use regression
609 models to short-term mobile monitoring campaigns using air pollution micro-sensors.
610 *Environ. Pollut.* **230**, 280–290 (2017).
- 611 20. Saha, P. K., Li, H. Z., Apte, J. S., Robinson, A. L. & Presto, A. A. Urban Ultrafine Particle
612 Exposure Assessment with Land-Use Regression: Influence of Sampling Strategy. *Environ.*
613 *Sci. Technol.* **53**, 7326–7336 (2019).
- 614 21. Saha, P. K. *et al.* Quantifying high-resolution spatial variations and local source impacts of
615 urban ultrafine particle concentrations. *Sci. Total Environ.* **655**, 473–481 (2019).
- 616 22. Chastko, K. & Adams, M. Assessing the accuracy of long-term air pollution estimates
617 produced with temporally adjusted short-term observations from unstructured sampling. *J.*
618 *Environ. Manage.* **240**, 249–258 (2019).
- 619 23. Kerckhoffs, J., Hoek, G., Portengen, L., Brunekreef, B. & Vermeulen, R. C. H. Performance
620 of Prediction Algorithms for Modeling Outdoor Air Pollution Spatial Surfaces. *Environ. Sci.*
621 *Technol.* **53**, 1413–1421 (2019).
- 622 24. Kerckhoffs, J., Hoek, G., Gehring, U. & Vermeulen, R. Modelling nationwide spatial
623 variation of ultrafine particles based on mobile monitoring. *Environ. Int.* **154**, 106569
624 (2021).
- 625 25. Minet, L. *et al.* Development and Comparison of Air Pollution Exposure Surfaces Derived
626 from On-Road Mobile Monitoring and Short-Term Stationary Sidewalk Measurements.
627 *Environ. Sci. Technol.* **52**, 3512–3519 (2018).

- 628 26. Simon, M. C. *et al.* Comparisons of Traffic-Related Ultrafine Particle Number
629 Concentrations Measured in Two Urban Areas by Central, Residential, and Mobile
630 Monitoring. *Atmospheric Environ. Oxf. Engl.* 1994 **169**, 113–127 (2017).
- 631 27. Karner, A. A., Eisinger, D. S. & Niemeier, D. A. Near-roadway air quality: Synthesizing the
632 findings from real-world data. *Environ. Sci. Technol.* **44**, 5334–5344 (2010).
- 633 28. Riley, E. A. *et al.* Multi-pollutant mobile platform measurements of air pollutants adjacent to
634 a major roadway. *Atmos. Environ.* **98**, 492–499 (2014).
- 635 29. US EPA. Air Quality System (AQS). *US Environmental Protection Agency*
636 <https://www.epa.gov/aqs> (2019).
- 637 30. MESA Air. *Data Organization and Operating Procedures (DOOP) for the Multi-Ethnic*
638 *Study of Atherosclerosis and Air Pollution (MESA Air) and Associated Studies.* (MESA Air,
639 2019).
- 640 31. US EPA. AirData Pre-Generated Data Files. *US Environmental Protection Agency*
641 https://aqs.epa.gov/aqsweb/airdata/download_files.html (2019).
- 642 32. R Core Team. R: A Language and Environment for Statistical Computing. *R Foundation for*
643 *Statistical Computing* <https://www.r-project.org> (2019).
- 644 33. Stamen Design. <http://maps.stamen.com/#terrain/12/37.7707/-122.3783> (2021).
- 645 34. Creative Commons. Attribution 3.0 Unprotected (CC BY 3.0).
646 <https://creativecommons.org/licenses/by/3.0/> (2021).
- 647 35. OpenStreetMap contributors. <https://www.openstreetmap.org/copyright> (2021).
- 648 36. Li, L. *et al.* Ensemble-based deep learning for estimating PM_{2.5} over California with
649 multisource big data including wildfire smoke. *Environ. Int.* **145**, 106143 (2020).

- 650 37. US EPA. Electronic Code of Federal Regulations (e-CFR), Title 40, Chapter 1, Subpart C,
651 Part 58, Appendix E to Part 58 - Probe and Monitoring Path Siting Criteria for Ambient Air
652 Quality Monitoring. (2021).
- 653 38. CARB. *Annual Network Plan - Covering Monitoring Operations in 25 California Air*
654 *Districts*. (California Air Resources Board (CARB), 2021).
- 655 39. Hoek, G. Methods for Assessing Long-Term Exposures to Outdoor Air Pollutants. *Curr.*
656 *Environ. Health Rep.* **4**, 450–462 (2017).
- 657 40. Szpiro, A. A. & Paciorek, C. J. Measurement error in two-stage analyses, with application to
658 air pollution epidemiology. *Environmetrics* (2013) doi:10.1002/env.2233.
- 659 41. Dodson, R. E., Houseman, E. A., Morin, B. & Levy, J. I. An analysis of continuous black
660 carbon concentrations in proximity to an airport and major roadways. *Atmos. Environ.* **43**,
661 3764–3773 (2009).
- 662 42. Riley, E. A. *et al.* Correlations between short-term mobile monitoring and long-term passive
663 sampler measurements of traffic-related air pollution. *Atmos. Environ.* **132**, (2016).
- 664 43. Austin, E. *et al.* *Mobile Observations of Ultrafine Particles: The MOV-UP study report*.
665 (2019).
- 666 44. Hudda, N., Gould, T., Hartin, K., Larson, T. V. & Fruin, S. A. Emissions from an
667 international airport increase particle number concentrations 4-fold at 10 km downwind.
668 *Environ. Sci. Technol.* **48**, 6628–6635 (2014).
- 669 45. Lack, D. A. & Corbett, J. J. Black carbon from ships: a review of the effects of ship speed,
670 fuel quality and exhaust gas scrubbing. *Atmospheric Chem. Phys.* **12**, (2012).

- 671 46. Kozawa, K. H., Fruin, S. A. & Winer, A. M. Near-road air pollution impacts of goods
672 movement in communities adjacent to the Ports of Los Angeles and Long Beach. *Atmos.*
673 *Environ.* **43**, 2960–2970 (2009).
- 674 47. Riffault, V. *et al.* Fine and Ultrafine Particles in the Vicinity of Industrial Activities: A
675 Review. *Crit. Rev. Environ. Sci. Technol.* **45**, 2305–2356 (2015).
- 676 48. Kilian, J. & Kitazawa, M. The emerging risk of exposure to air pollution on cognitive decline
677 and Alzheimer ' s disease e Evidence from epidemiological and animal studies. *Biomed. J.*
678 **41**, 141–162 (2018).
- 679 49. Lane, K. J. *et al.* Association of modeled long-term personal exposure to ultrafine particles
680 with inflammatory and coagulation biomarkers. *Environ. Int.* **92–93**, 173–182 (2016).
- 681 50. Weichenthal, S. *et al.* Within-city Spatial Variations in Ambient Ultrafine Particle
682 Concentrations and Incident Brain Tumors in Adults. *Epidemiology* **31**, (2020).
- 683 51. US EPA. Integrated science assessment (ISA) for particulate matter (final report, Dec 2019).
684 *US Environ. Prot. Agency* (2019).
- 685 52. Gryparis, A., Paciorek, C. J., Zeka, A., Schwartz, J. & Coull, B. A. Measurement error
686 caused by spatial misalignment in environmental epidemiology. *Biostatistics* **10**, 258–274
687 (2009).
- 688 53. Szpiro, A. A., Sheppard, L. & Lumley, T. Efficient measurement error correction with
689 spatially misaligned data. *Biostatistics* **12**, 610–623 (2011).
- 690 54. Sheppard, L. *et al.* Confounding and exposure measurement error in air pollution
691 epidemiology. *Air Qual. Atmosphere Health* **5**, 203–216 (2012).
- 692