Supporting Information

Authors
Mei W. Tessum[a,*], Lianne Sheppard[b,c], Timothy V. Larson[b,d], Timothy R. Gould[d], Joel D. Kaufman[b], Sverre Vedal[b].

[a] University of Illinois at Urbana-Champaign, Department of Agricultural and Biological Engineering, Urbana, IL 61801, USA

[b] University of Washington, Department of Environmental and Occupational Health Sciences, Box 357234, Seattle, WA 98195, USA

[c] University of Washington, Department of Biostatistics, Box 357232, Seattle, WA 98195, USA

[d] University of Washington, Department of Civil & Environmental Engineering, Box 352700, Seattle, WA 98195, USA

Title
Improving long-term air pollution modeling using short-term mobile monitoring in two US metropolitan regions

Number of pages: 18

Number of tables: 2

Contents

- Figure S5. Scatter plots of 2-week NO_x predictions by different model scenarios vs. observations at home sites in Los Angeles
- Figure S6. Scatter plots of long-term average NO_x predictions by different model scenarios vs. observations at AQS and fixed sites in Los Angeles
- Figure S7. Scatter plots of 2-week NOx predictions by different model scenarios vs. observations at AQS and fixed sites in Los Angeles
- Figure S8. Scatter plots of 2-week NOx predictions by different model scenarios vs. observations at home sites in Baltimore
- Figure S9. Scatter plots of long-term average NOx predictions by different model scenarios vs. observations at AQS and fixed sites in Baltimore
- Figure S10. Scatter plots of 2-week NOx predictions by different model scenarios vs. observations at AQS and fixed sites in Baltimore
- Figure S11. Site distribution of distance to roadways for AQS, MESA fixed and home sites by region. Vertical lines represent median distance to roadway for each type of monitoring site
- Figure S12. Time trends for the NOx model using AQS+Fixed Site+Mobile data in Los Angeles
- CSV files: PLS components for the long-term average ($\boldsymbol{\beta 0}$) in best models characterized by the loadings of all geographic covariates on the component scores

Spatio-temporal model description

- Spatiotemporal Model

The spatio-temporal model consists of a spatio-temporal trend model and a model to fit spatio-temporal residuals that can accommodate highly unbalanced data in time and space. This model can be written as:

$$C(s, t) = \mu(s, t) + v(s, t) \quad (S1)$$

where *C(s,t)* represents the two-week average concentration of NO_2 or NO_x at location *s* and time t. μ(s, t) represents the spatio-temporal mean surface. The term *v(s, t)* represents spatio-temporal residual variation. The spatiotemporal mean $\mu(s, t)$ can be further broken down into a long-term average $\beta 0(s)$ at location s and smooth time trends *fi(t)* with spatially-varying coefficient fields $\beta i(s)$, as written in Equation 1 in the manuscript.

With this model framework as implemented in the SpatioTemporal R package, we first derive the time trends $f_i(t)$ from AQS and MESA fixed sites that contain multiple years of measurements and therefore contain enough of the temporal structure. The time trends are estimated from a singular value decomposition (SVD) of the space-time data matrix for the AQS and MESA fixed sites, missing or incomplete monitoring data were filled using an iterative expectation maximization procedure. The number of time trends *m* was determined by the lowest mean square error and Akaike information criterion (AIC) using leave-one-site-out cross validation of the built-in SVD function in SpatioTemporal R package. The selected time trends are then smoothed using splines with degree of freedom (*df*) parameters controlling the smoothness. The smoothed time trends can be later applied with different spatial and spatial-temporal weights to short-term measurement (mobile or PSD measurements, as shown in Figure S12). The spatially-varying long-term average $\beta_0(s)$ and time trend coefficients $\beta_i(s)$ are estimated with a spatial mean and either an exponential or independent covariance structure.

- Geographic Covariates

More than 300 geographic variables were compiled for use in our model, including proximity variables (e.g., distance to nearest major road) and buffer variables (e.g., total road length within a buffer area of various radii). Variables were selected by removing the ones with highly influential values or limited variability with the following criteria: 1) > 80% of monitoring locations had the same value, 2) > 2% of observations were > 5 standard deviations away from mean, 3) the standard deviation of the distribution of values at participant homes was > 5 standard deviations of the distribution of values at monitoring locations, or 4) for land-use variables, the maximum value was 10% among all monitoring locations. Approximately 200 geographic variables remained after applying the selection criteria.

- Partial Least Squares (PLS) Scores

At each AQS and fixed site, ordinary least squares regression with mean function was used to regress the time series of concentration observations on the smoothed time trends. At each mobile/PSD monitoring and MESA home site, PLS regression was used to compute the matrix of geographic covariates (*X(s)*) as the predictors. In general, PLS provides a decomposition of the large geographic covariate matrix *X(s)* into a sequence of orthogonal PLS scores computed to maximize the covariance between long-term mean concentrations at *s* monitoring sites and their prediction by these score vectors. The number of PLS scores was later chosen by cross-validation to give the best predictions. Typically, three or fewer PLS scores are needed to capture the multidimensional information and avoid overfitting.

- Parameter Estimation

Once the PLS scores *X(s)* and time trends *fi(t)* were estimated, the remaining regression coefficients were obtained using maximum likelihood which is a built-in function in the SpatioTemporal package. Additionally, model hyperparameters were determined independently for each pollutant in each metropolitan region based on best-fitting model to represent different model structure settings, and were manually chosen from the following options: the number of time trends (either 1 or 2); the degrees of freedom for smoothing time trends (either 4 or 8 per year); the number of PLS scores per time trend (2 or 3); and the presence of a covariance structure of the βi fields (spatial smoothing or no spatial smoothing). Those possible options are empirically determined by previous studies in model development (Sampson et al. 2011, 2013, Lindström et al. 2014).

PSD and Mobile Monitoring Campaign Instrumentation

In the PSD monitoring campaign, Ogawa samplers (Ogawa & Company, USA, Inc., Pompano Beach, FL) were used to measure NO_2 and NO_x. The Ogawa sampler contained two filters in separate chambers for measuring NO_2 (triethanolamine coating) and NO_x (2-Phenyl-4,4,5,5- tetramethylimidazoline-3-oxide-1-oxyl coating), respectively. During the two-week measurement period, PSDs were attached to custom built, aluminum/stainless steel sampler shelters and were mounted on utility poles approximately 3 meters above the ground. All collected samplers were sent to the Environmental Health Laboratory at the University of Washington for analysis.

In the mobile monitoring campaign, a CAPS NO_2 monitor (Aerodyne Research, Inc., Billerica, MA) with a measurement range of 0-3000 ppb and a NO Model 410 with a range of 0-2000 ppb (Converter #401, 2B Technologies, Boulder, CO) were used to measure NO_2 and NOx respectively. Both instruments were calibrated for flow, zero, and span at the University of Washington before monitoring campaign. During monitoring campaign, we have performed in-field calibration to check for zero and span periodically between runs.

Table S1. Geographic variables used in the spatiotemporal model for predicting NO_x concentration in Los Angeles

Variable	**Distance or buffer size (m)**
Elevation	from sea level
Count of points of same elevation within or more than 20 or 50 m	above, at, below 1 km and 5 km
Average impervious surface value	50, 100, 150, 300, 400, 500, 750, 1000, 3000, 5000
Distance to major intersection: A1-A1, A1-A3, A2-A2, A2-A3, A3-A3	500, 1000, 3000
Lengths of major road: A1, A2, A3	100, 150, 300, 400, 500, 750, 1000, 1500, 3000, 5000
Lengths of truck route	500, 750, 1000, 1500, 3000, 5000, 10000, 15000
Distance to emission sources: SO_2, NO_x, $PM_{2.5}$, PM_{10}, CO (log scale)	3000, 15000, 30000
Land use percentage: shrub land, grassland, water, high/ median/ low intensity development, open development	50, 100, 150, 300, 400, 500, 750, 1000, 3000, 5000
Total population	500, 1000, 1500, 2000, 2500, 3000, 5000, 10000, 15000
Vegetative index (NDVI): 25th, 50th, 75th quartiles; winter; summer	250, 500, 1000, 2500, 5000, 7500, 10000
Distance to nearest major road (A1, A2, A3), intersection (A1-A1, A1-A2, A1-A3, A2-A2, A2-A3, A3-A3), airport, coast, commercial zone, railyard, railroad, truck route (log scale)	--

Table S2. Summary of monitoring data at AQS and MESA fixed and home site by pollutant and region

Pollutant	Site type	Number of sites	Mean ± SD
NO_2 (ppb)	Los Angeles, CA		
	AQS	21	18.4±7.92
	MESA fixed	7	22.5±9.68
	MESA home	120	20.7±8.26
	Mobile	43	12.3±3.22
	PSD	43	14.9±3.37
NO_x (ppb)	Los Angeles, CA		
	AQS	21	31.2±21.5
	MESA fixed	7	50.1±35.2
	MESA home	120	42.1±25.4
	Mobile	43	31.6±8.46
	PSD	43	36.9±13.1
	Baltimore, MD		
	AQS	11	20.7±13.8
	MESA fixed	5	29.6±20.6
	MESA home	85	20.6±15.4
	Mobile	42	29.9±9.91
	PSD	84	26.2±16.6

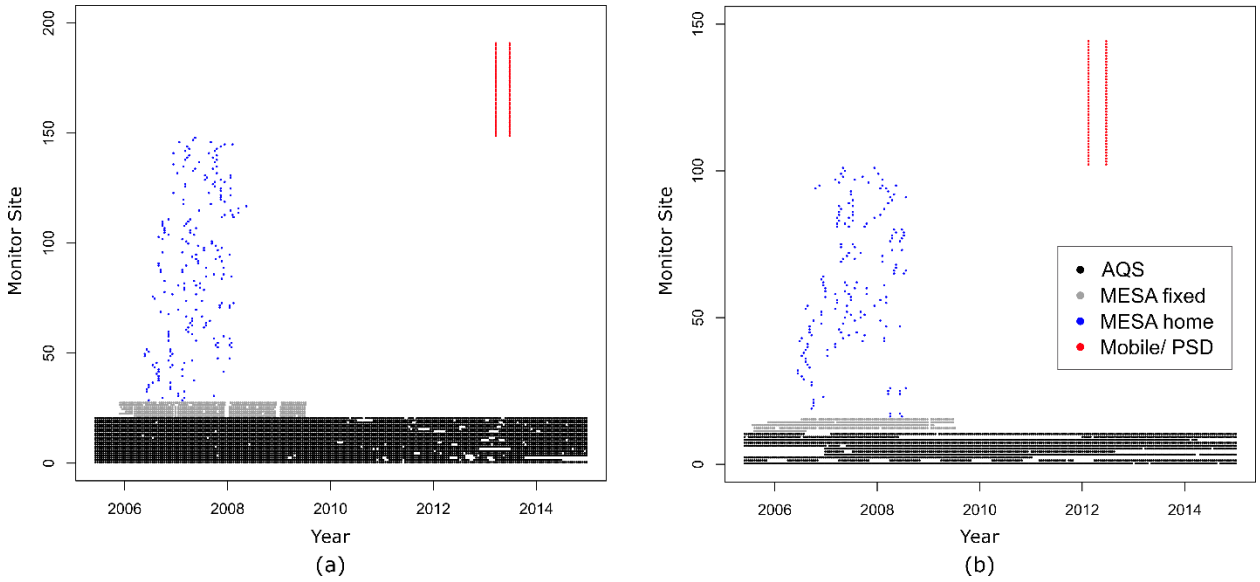


Figure S1. Density plots of NO_x observations from all five types of monitoring site in Los Angeles (a) and Baltimore (b).

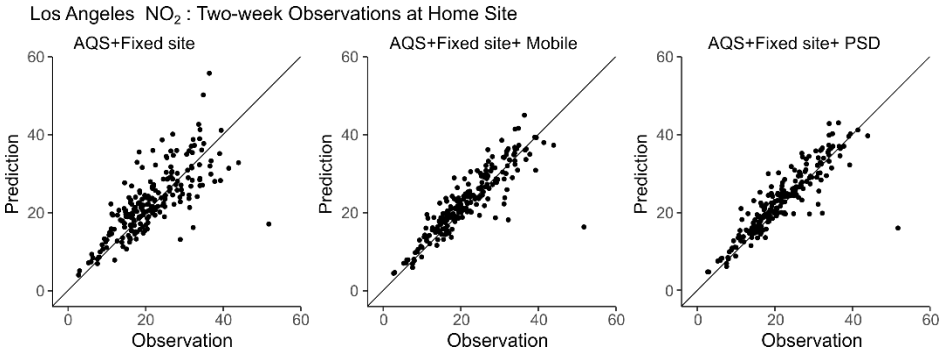


Figure S2. Scatter plots of 2-week NO_2 predictions by different model scenarios vs. observations at home sites in Los Angeles.

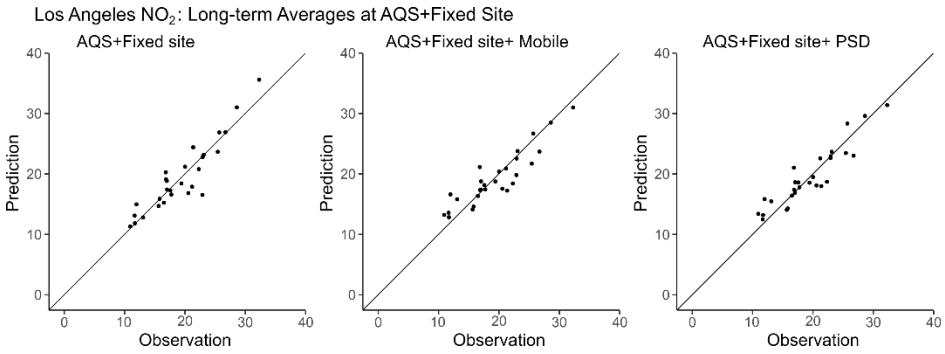


Figure S3. Scatter plots of long-term average NO_2 predictions by different model scenarios vs. observations at AQS and fixed sites in Los Angeles.

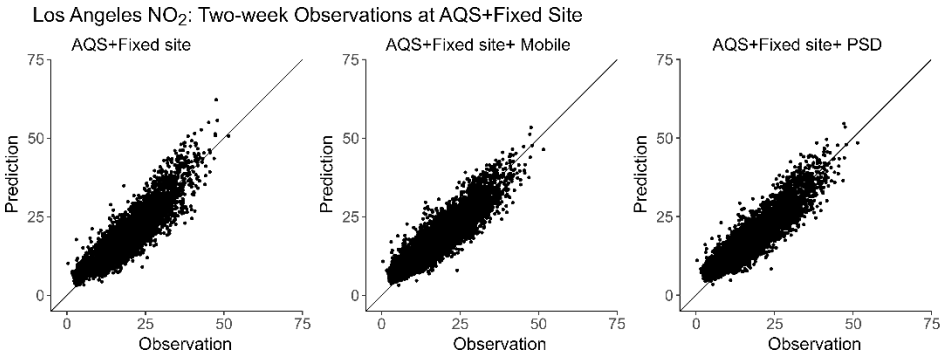


Figure S4. Scatter plots of 2-week NO_2 predictions by different model scenarios vs. observations at AQS and fixed sites in Los Angeles.

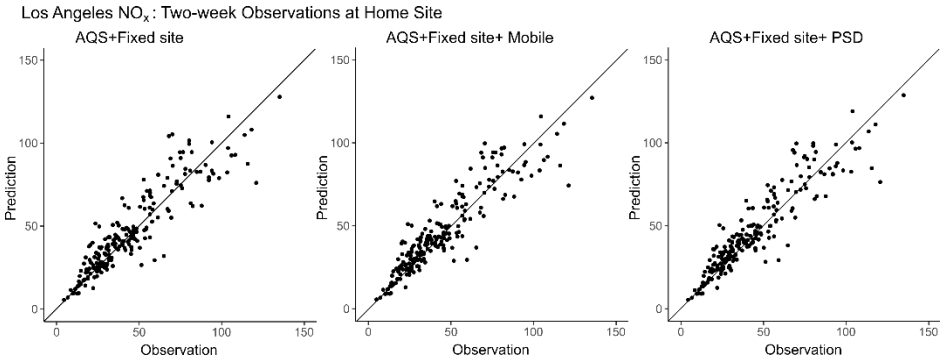


Figure S5. Scatter plots of 2-week NO_x predictions by different model scenarios vs. observations at home sites in Los Angeles.

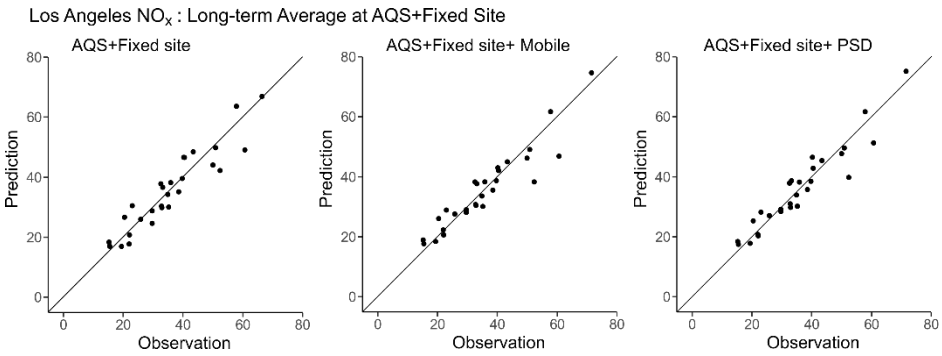


Figure S6. Scatter plots of long-term average NO_x predictions by different model scenarios vs. observations at AQS and fixed sites in Los Angeles.

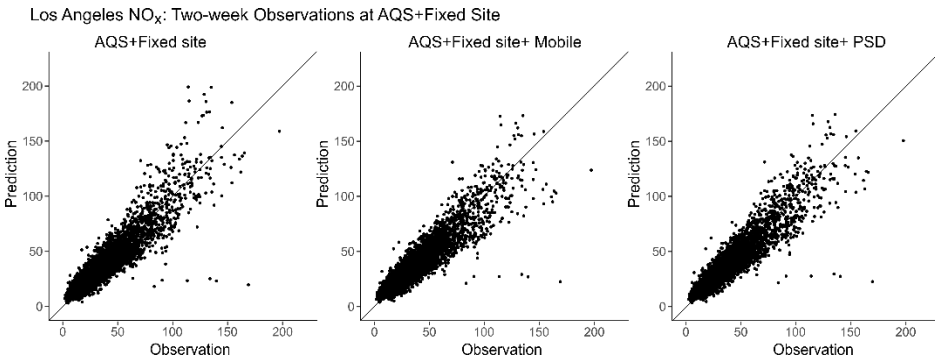


Figure S7. Scatter plots of 2-week NO_x predictions by different model scenarios vs. observations at AQS and fixed sites in Los Angeles.

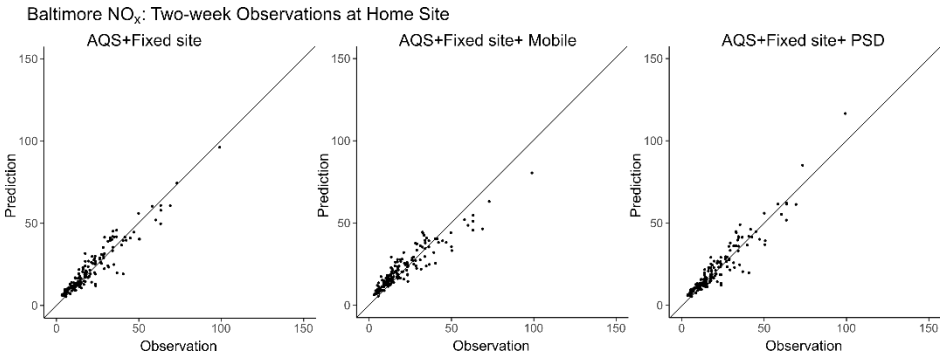


Figure S8. Scatter plots of 2-week NO_x predictions by different model scenarios vs. observations at home sites in Baltimore.

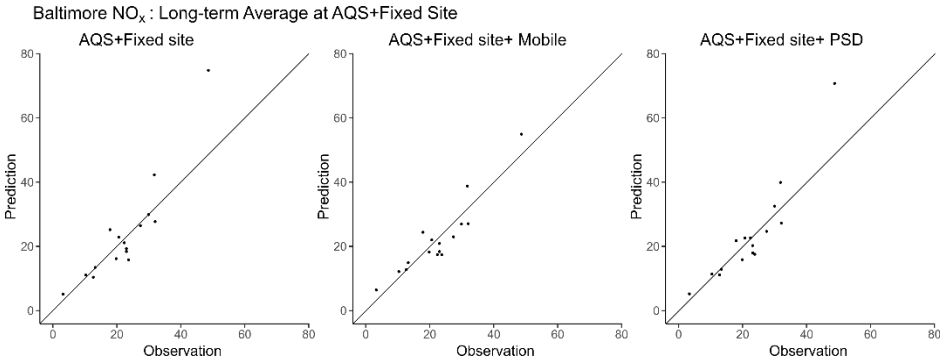


Figure S9. Scatter plots of long-term average NO_x predictions by different model scenarios vs. observations at AQS and fixed sites in Baltimore.

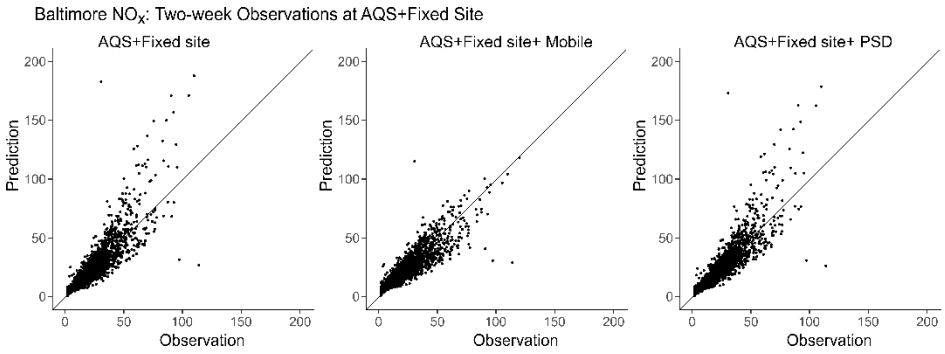


Figure S10. Scatter plots of 2-week NO_x predictions by different model scenarios vs. observations at AQS and fixed sites in Baltimore.

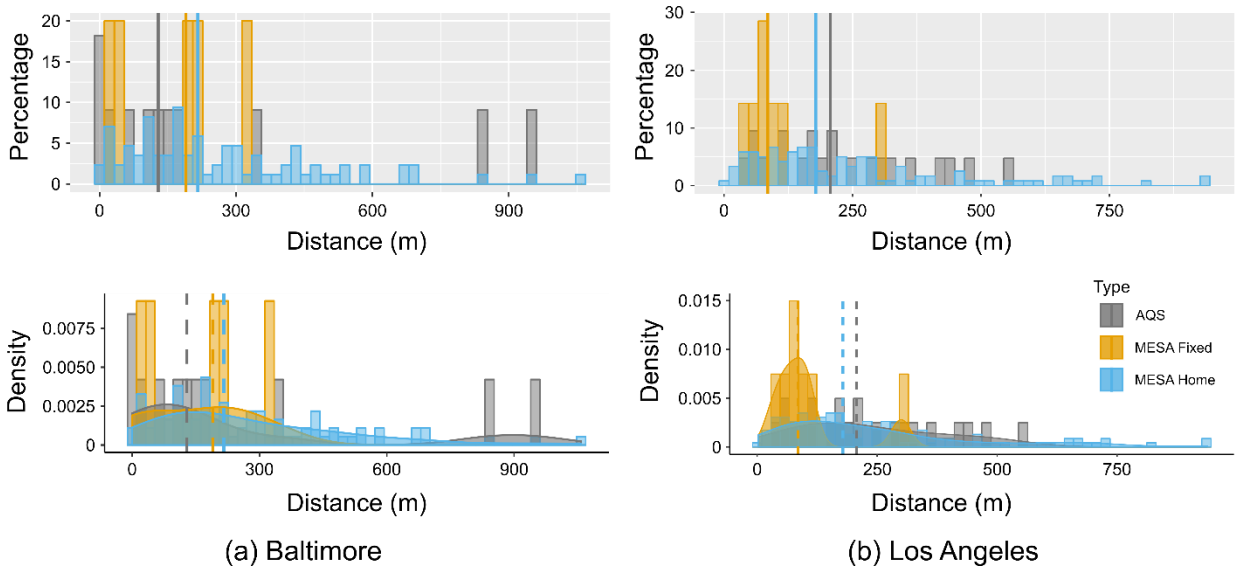
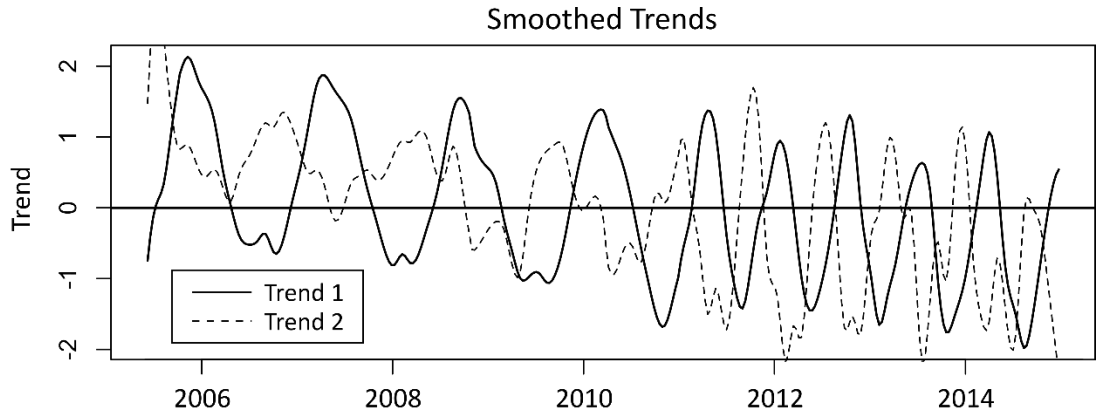
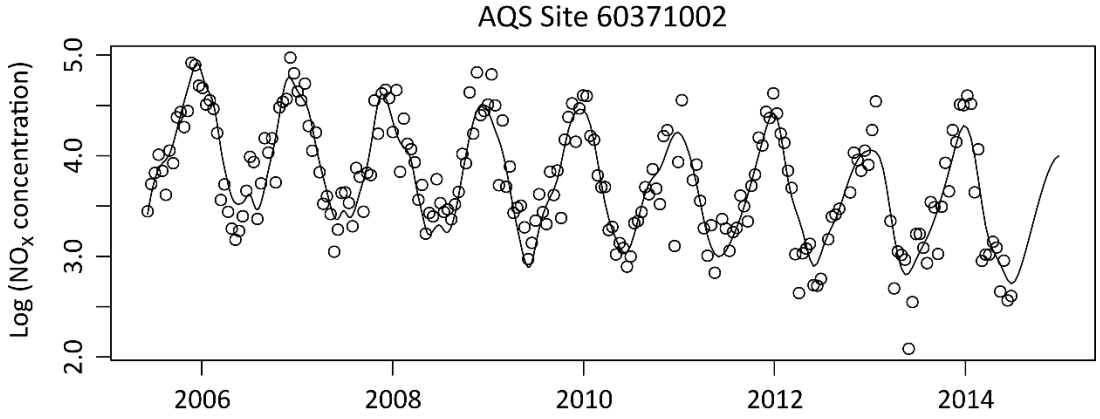
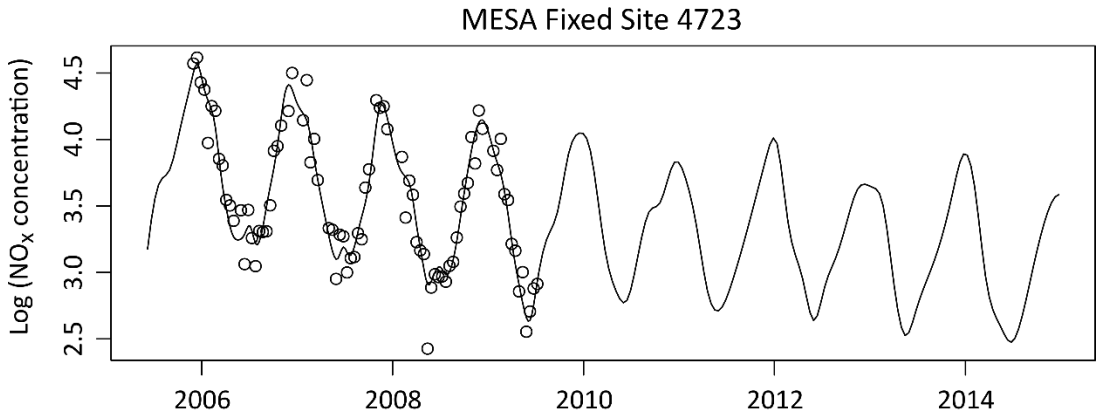


Figure S11. Site distribution of distance to roadways for AQS, MESA fixed and home sites by region. Vertical lines represent median distance to roadway for each type of monitoring site.


Smoothed Trends
Trend
2
1
0
-1
-2
Trend 1
Trend 2
2006
2008
2010
2012
2014


AQS Site 60371002
Log (NO_X concentration)
5.0
4.0
3.0
2.0
2006
2008
2010
2012
2014


MESA Fixed Site 4723
Log (NO_X concentration)
4.5
4.0
3.5
3.0
2.5
2006
2008
2010
2012
2014

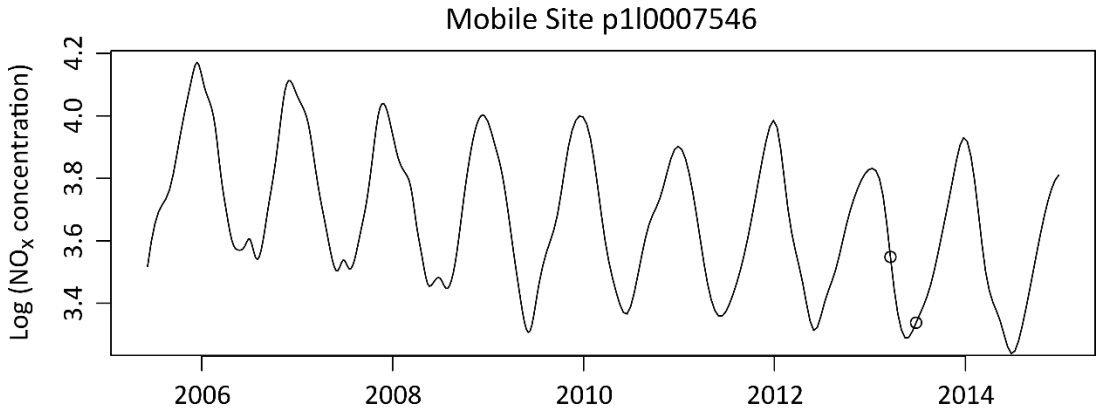

Mobile Site p1l0007546
Log (NO_X concentration)
4.2
4.0
3.8
3.6
3.4
2006
2008
2010
2012
2014

Figure S12. Time trends for the NO_x model using AQS+Fixed Site+Mobile data in Los Angeles. The top panel shows the smoothed time trends calculated from AQS and fixed sites. The second to fourth panels show the observed data and fitted trends at an AQS site, a fixed site and a Mobile site, respectively