Supporting Information

Constrained Mixed-Effect Models with Ensemble Learning for Prediction of Nitrogen Oxide Concentrations at a High Spatiotemporal Resolution

Lianfa Li[*,1,2], Fred Lurmann[3], Rima Habre[*,1], Robert Urman[1], Edward Rappaport[1], Beate Ritz[4], Jiu-Chiuan Chen[1], Frank D. Gilliland[1], Jun Wu[*, 5]

1. Department of Preventive Medicine, University of Southern California, Los Angeles, CA, USA

2. State Key Laboratory of Resources and Environmental Information System, Institute of Geographical Sciences and Natural Resources, Chinese Academy of Sciences, Beijing, China

3. Sonoma Technology, Inc., Petaluma, CA, USA

4. Department of Epidemiology, University of California, Los Angeles, CA, USA

5. Program in Public Health, College of Health Sciences, University of California, Irvine, CA, USA

*Corresponding authors' contact information: Lanfa Li: Phone, 650-388-0268; Email, lianfali@usc.edu or lspatial@gmail.com. Rima Habre: Phone, 323-442-8283; Email, habre@usc.edu. Jun Wu: Phone, 949-824-0548; Email, junwu@uci.edu.

Number of pages: 21

Number of figures: 8

Number of tables: 4

Contents

1. Overview

This paper proposed a three-stage spatiotemporal model that can reliably predict nitrogen oxide concentrations with a high spatiotemporal resolution over a long time span (>20 years). The spatially extensive highly-clustered exposure data from short-term measurement campaigns across 1-2 years and long-term central site monitoring in 1992-2013 were leveraged to develop the first stage mixed-effect models and the second stage ensemble learning with uncertainty estimates. Then at the third stage, constrained optimization was designed and implemented based on the point estimates from the first and second stages to simulate the long-term series of pollutant concentrations for any target location in the study region. The following sections provide the supplemental information about concentration measurements, the covariates selected, modeling method, and results to support the formal paper.

2. Measurements of NO_2 and NO_x concentrations

Besides the measurements of routine monitoring stations, additional data were generated in intensive field measurement campaigns conducted by the University of Southern California (USC), University of California Los Angeles (UCLA), and University of California Irvine (UCI), respectively. Passive diffusion-based Ogawa samplers [1] were used to measure NO_2 and NO_x at different time periods and at different locations (e.g. outside homes, schools, strategic outdoor sampling locations, and central monitoring sites). Our previous papers respectively provide more details about the measurement methods that generated the USC [2], UCLA [3] and UCI [4] samples. Figure S1 also shows the locations for the routine and USC sampling sites [the UCI and UCLA sampling locations are concealed to comply with specific requirements by their Institutional Review Boards].

To minimize systematic bias in the field campaign data, we compared the passive data with the active data from the routine government monitors at the co-located sites and made small adjustments to standardize the passive measurements to Federal Reference Method equivalent values. For details, please refer to Table S1.

3. The Covariates Selected

3.1 CALINE4-estimated concentrations from local traffic emissions

CALINE4 is a line source dispersion model that was used to assess the contribution of local motor vehicle emissions to ambient concentrations [5, 6]. CALINE4 was used to compute mean NO_x concentration from emissions on freeways [coded using Feature Class Codes (FCC) as FCC1] and non-freeways (FCC2, FCC3 and FCC4). Traffic count data were obtained from Caltrans and TeleAtlas/GDT and assigned to ERSI Premium Street Map roadway geometry. Emission strength was estimated using quarterly average daily traffic volumes and EMFAC2011 (for 1992-2012) [7] and EMFAC2014 (for 2013) [8], which generated air basin emission factors that were based on average vehicle speed and heavy-duty truck fraction (Caltrans post-mile truck count data by year). Wind speeds and directions were based on hourly observations of these surface meteorological variables from 72 monitoring stations of California [8].

3.2 Traffic density

Traffic density represents distance-decayed annual average daily traffic (AADT) volume in both directions from all roads (FCC1-FCC4) within a circular buffer. Traffic density is symmetric on both sides of each roadway, computed as if the wind directions were uniformly distributed around the compass. The values of traffic density were computed by the ESRI ArcGIS density function using a kernel with a 300 m search radius and 5 m grid resolution. Annual traffic density estimates were provided for the regulatory monitoring sites and local sampling locations. In addition, because these cover a long time period, the traffic densities were scaled by the South Coast Air Basin (SoCAB) EMFAC2011 vehicle fleet average NO_x emission factor for 50 mph and 6% heavy-duty vehicle fraction (normalized to 1.00 in 2002 since we used the 2002 AADT as the baseline data) to reflect the composite trend in traffic volumes and emissions over time.

4. Modeling Approach

4.1 Non-parametric additive methods

For spatiotemporal factors, we adopted non-parametric additive methods to model non-linear effects. Specifically, we used non-parametric trend functions to quantify the association, $s(\ldots)$, i.e. approximated by the weighted sum of polynomial spline (B-spline basis) functions

$$x_{min} = \zeta_0 < \zeta_1 < \zeta_2 \ldots < \zeta_{m-1} < \zeta_m = x_{max} \quad \text{(S1)}$$

$$f(x_i) = \sum_j \beta_{ij} B_{ij}(x_i) \quad \text{(S2)}$$

where ζ_i is the split for an interval for the covariate x_i (ζ_0 and ζ_m are respectively the minimum and maximum values of x_i), B_{ij} and β_{ij} respectively represent the basis function and parameter for the interval j for the covariate x_i. Penalized maximum likelihood was used to solve (eq. S2) to estimate β_{ij}.

4.2 Modeling of spatial random effects

By estimating a structured component and an unstructured component, we can distinguish between the two sources of spatial autocorrelations [9]. The spatial effects were modeled using the following formulas for both structured (S3) and unstructured spatial effects (S4).

$$f_s(r_s)|f_s(r'), r' \neq r_s, \tau^2_{str} \sim N(\frac{1}{N_{r_s}} \sum_{r' \in \delta_{r_s}} f_s(r'), \frac{\tau^2_{str}}{N_{r_s}}) \quad \text{(S3)}$$

$$f_{re}(r_s)|\tau^2_{unstr} \sim N(0, \tau^2_{unstr}) \quad \text{(S4)}$$

where r_s is the region where the observation $y(s,t)$ is located, δ_{r_s} represents a set of neighbors (r') of the polygon r_s, N_{r_s} is the number of neighboring polygons for r_s, τ^2_{str} is the total variance for the structured component, τ^2_{str} ~IG(a,b). $f_s(r')$ in (eq. S3) represents the spatial influence from neighboring polygons (r') on r_s; $f_{re}(r_s)$ in (eq. S4) represents the unstructured spatial effect with zero mean and standard deviation (τ^2_{unstr}) for r_s.

Thiessen polygons were constructed around the central points (derived by averaging the coordinates of all the routine or/and campaigns sampling locations within a certain distance) to simulate spatial effects. Rook adjacency was used for spatial adjacency: two polygons were assumed to be neighbors if they share a common border. We conducted sensitivity tests for a series of aggregation distances (100 m, 300 m, 500 m and 3 km) and finally selected an optimal aggregate distance (500 m) that provided a good balance between model accuracy and computing efficiency. We used the packages of rgdal and spdep in the statistics software R (Version 3.3) for generation of the Thiessen polygons with their spatial weight matrix.

4.3 Aggregated predictions by ensemble learning

The aggregated predictions (mean and standard deviation) are the weighted summary of all trained models, where the weighting is the square of each model's R^2.

$$m_f(s,t) = \sum_i h_i(d_b, f_r) w_i \quad \text{(S5)}$$

$$\sigma_f(s,t) = \sqrt{\frac{\sum_i w_i (h_i(d_b,f_r) - m_f(s,t))^2}{\frac{M-1}{M} \sum_i w_i}} \quad \text{(S6)}$$

$$w_i = {R_i^2}^2 / \sum_i {R_i^2}^2 \quad \text{(S7)}$$

where $m_f(s,t)$ is the aggregated prediction (the weighted mean), $h_i(d_b,f_r)$ is the prediction by the i^{th} spatiotemporal model (eq. 1) trained using the bootstrap sample (d_b) and selected set of predictors (f_r); w_i is the normalized weight derived from the i^{th}

model's performance measure; $\sigma_f(s,t)$ is the standard deviation from the output of multiple models, M is the number of nonzero weights.

Table S1. Correlation between average biweekly passive measurements at USC, UCLA, and UCI sites and collocated routine monitoring sites, and their linear regression coefficients used for consistent adjustment

Cover		Pollutant	Number of collocated locations	Sampling period	Correlation coefficient	Parameters	
						Slope	Intercepts
USC	ICV1[a]	NO_2		Mixed dates for 2009-2013	0.94	0.89	3.57
		NO_x		Mixed dates for 2009-2013	0.93	0.83	6.23
	ICV2[a]	NO_2		Mixed dates for 2009-2013	0.92	0.84	3.98
		NO_x		Mixed dates for 2009-2013	0.96	0.82	5.40
UCLA		NO_2	14	Sept. 9- Sept. 22, 2006	0.94	0.68	4.43
				Feb. 10- Feb. 23, 2007	0.95	1.00	0.29
		NO_x	14	Sept. 9- Sept. 22, 2006	0.98	0.80	2.38
				Feb. 10- Feb. 23, 2007	0.97	0.81	12.05
UCI		NO_2	11	Jul. 10- Jul. 18, 2009	0.98	0.88	5.20
				Jul. 24-Aug. 1, 2009	0.99	0.94	3.56
				Nov. 13-Nov. 21, 2009	0.996	0.58	12.91
				Dec. 4- Dec. 12, 2009	0.95	0.65	7.74
		NO_x	11	Jul. 10- Jul. 18, 2009	0.96	0.69	8.53
				Jul. 24-Aug. 1, 2009	0.95	0.69	5.46
				Nov. 13-Nov .21, 2009	0.96	1.22	-13.41
				Dec. 4- Dec.12, 2009	0.97	0.72	14.38

[a]. ICV: The Intra-Community Variability study

Table S2. Variance explained by the predictors included in the Stage 1 mixed effects model

Covariate	Unit	Source (buffer distance)	Threshold[a]	Variances explained	
				NO_2	NO_x
Wind speed	Meter /second	Gridded Surface Meteorological Data	-	3%	3%
Minimum air temperature	Celsius (˚C)	Gridded Surface Meteorological Data	-	5%	4%
Spatiotemporal basis 1[b]	Log ppb	Singular value decomposition by temporal basis function	-	19%	13%
Spatiotemporal basis 2	Log ppb		-	3%	1%
CALINE4 on freeways	ppb	CALINE4 Dispersion model NO_x from freeway	>180	9%	13%
Caline4 on non-freeways	ppb	CALINE4 Dispersion model NO_x from non-freeway	-	3%	5%
Traffic density (300m-5km)	Vehicles/day	Distance-decayed annual traffic volume in a scaled by vehicle emission factors (in a donut radii =300 m, 5 km).	-	11%	9%
Distance to FCC1[c]	Meter	Distance to FCC1	>15 km	7%	8%
Population Density			>21830	5%	11%
Region-level yearly mean	ppb	Annual mean concentration for the sub region determined from routine monitoring data	-	12%	13%
Spatial autocorrelation		Simulated using Thiessen polygons	-	13%	11%
Total				90%	91%

[a]: threshold defined to remove the outliers for the covariate;

[b]: bold font highlights the variance explained ≥10% by the variable;

[c]: Feature Class Codes for freeways and highways

Table S3. NO_2 model performance by CHS community in cross validation

ICV 1[a]	Samples	Mean (ppb)	Correlation	RMSE[b]	NRMSE[c]	CVRMSE[d]
Alpine	156	8.71	0.89	2.51	0.1	0.29
Anaheim	128	30.07	0.69	4.02	0.13	0.13
Glendora	221	20.84	0.88	3.05	0.09	0.15
Lake Arrowhead	121	8.93	0.61	2.36	0.15	0.28
Lake Elsinore	165	10.93	0.92	1.46	0.08	0.13
Long Beach	162	20.2	0.98	2.59	0.07	0.13
Mira Loma	189	13.66	0.98	1.46	0.06	0.11
Riverside	215	16.55	0.93	2.84	0.1	0.17
San Bernardino	138	15.17	0.97	2.13	0.08	0.14
San Dimas	174	25.17	0.9	2.7	0.08	0.11
Santa Barbara	147	11.29	0.93	2.34	0.1	0.21
Santa Maria	153	8.39	0.64	1.02	0.14	0.12
Upland	232	20.14	0.95	2.73	0.08	0.14

ICV2[a]	Samples	Mean (ppb)	Correlation	RMSE[b]	NRMSE[c]	CVRMSE[d]
Anaheim	26	19.56	0.77	3.32	0.15	0.17
Glendora	30	17.03	0.75	2.45	0.19	0.14
Long Beach	27	20.42	0.92	3.15	0.13	0.15
Mira Loma	26	18.24	0.98	1.68	0.08	0.09
Riverside	28	13.82	0.51	2.35	0.28	0.17
San Bernardino	40	11.07	0.71	2.41	0.16	0.22
San Dimas	28	20.16	0.83	2.66	0.19	0.13
Upland	28	15.79	0.75	2.56	0.18	0.16

[a]: ICV, The Intra-Community Variability study; [b]: RMSE, root mean square error; [c]: RMSE, root mean square error; NRMSE, normalized RMSE; [d]:CV RMSE, coefficient of variation of the RMSE.

Table S4. NO_x model performance by CHS community in cross validation

ICV 1[a]	Samples	Mean (ppb)	Correlation	RMSE[b]	NRMSE[c]	CVRMSE[d]
Alpine	156	18.8	0.9	4.29	0.08	0.23
Anaheim	128	67.25	0.95	6.59	0.08	0.1
Glendora	221	41.82	0.9	5.34	0.07	0.13
Lake Arrowhead	121	16.12	0.68	3.7	0.13	0.23
Lake Elsinore	165	20.34	0.91	2.5	0.08	0.12
Long Beach	162	63.01	0.98	8.78	0.07	0.14
Mira Loma	189	31.64	0.91	3.17	0.07	0.1
Riverside	215	35.86	0.91	5.99	0.08	0.17
San Bernardino	138	39.42	0.91	4.74	0.07	0.12
San Dimas	174	52.64	0.94	5.73	0.06	0.11
Santa Barbara	147	26.98	0.94	6.65	0.09	0.25
Santa Maria	153	14.21	0.67	1.43	0.13	0.1
Upland	232	41.37	0.93	5.83	0.06	0.14

ICV2[a]	Samples	Mean (ppb)	Correlation	RMSE[b]	NRMSE[c]	CVRMSE[d]
Anaheim	26	33.4	0.76	8.86	0.17	0.27
Glendora	30	24.47	0.85	3.05	0.14	0.12
Long Beach	27	48.92	0.97	8.09	0.08	0.17
Mira Loma	26	33.9	0.93	6.98	0.12	0.21
Riverside	28	19.29	0.83	2.78	0.14	0.14
San Bernardino	40	17.58	0.83	4.21	0.11	0.24
San Dimas	28	29.51	0.83	4.43	0.18	0.15
Upland	28	22.59	0.89	3.05	0.12	0.14

[a]: ICV, The Intra-Community Variability study; [b]: RMSE, root mean square error; [c]: RMSE, root mean square error; NRMSE, normalized RMSE; [d]:CV RMSE, coefficient of variation of the RMSE.

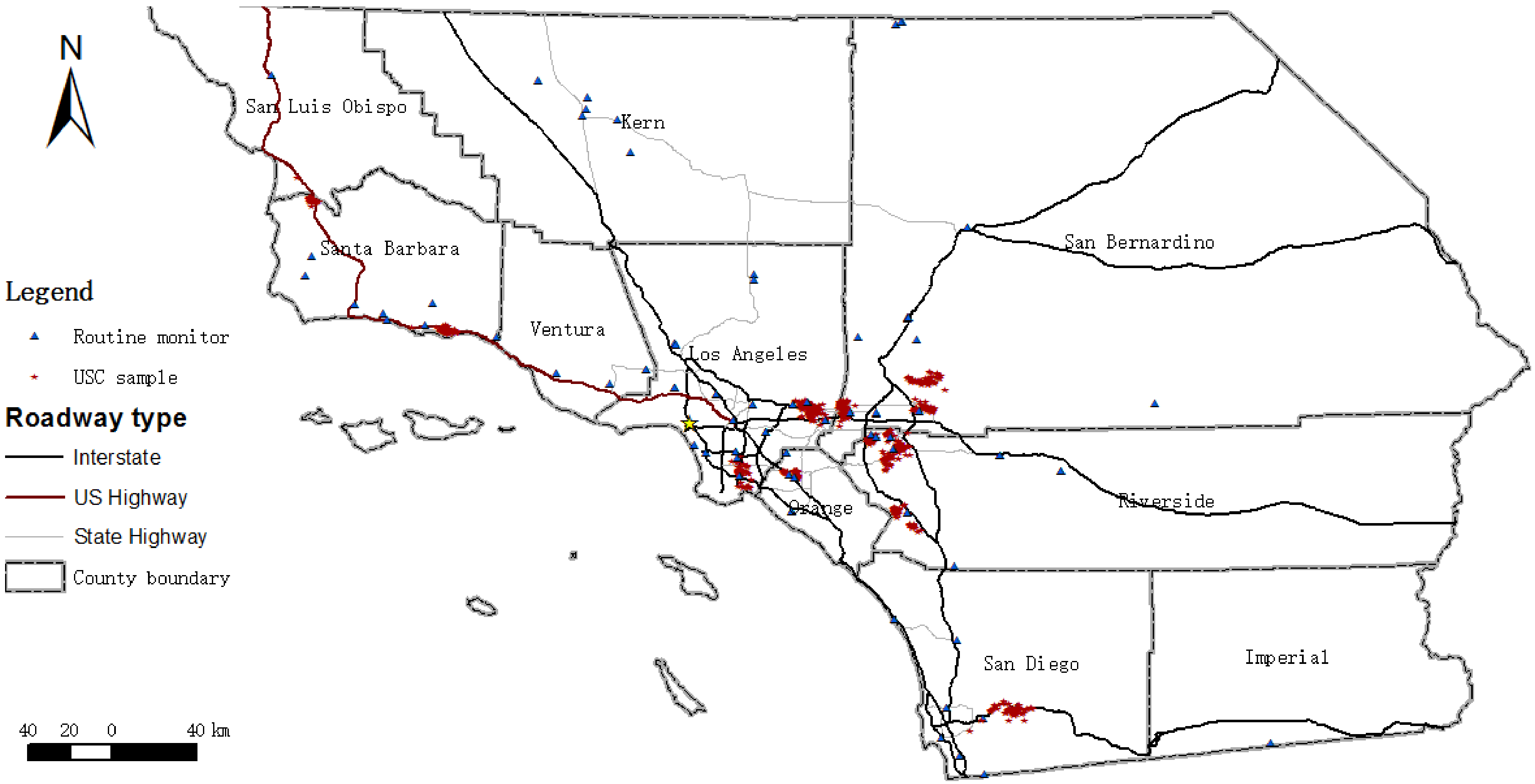
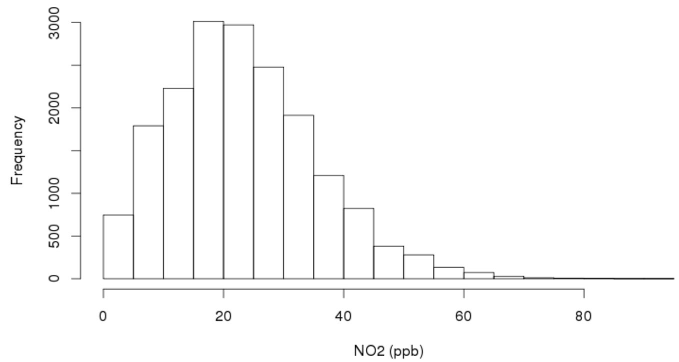
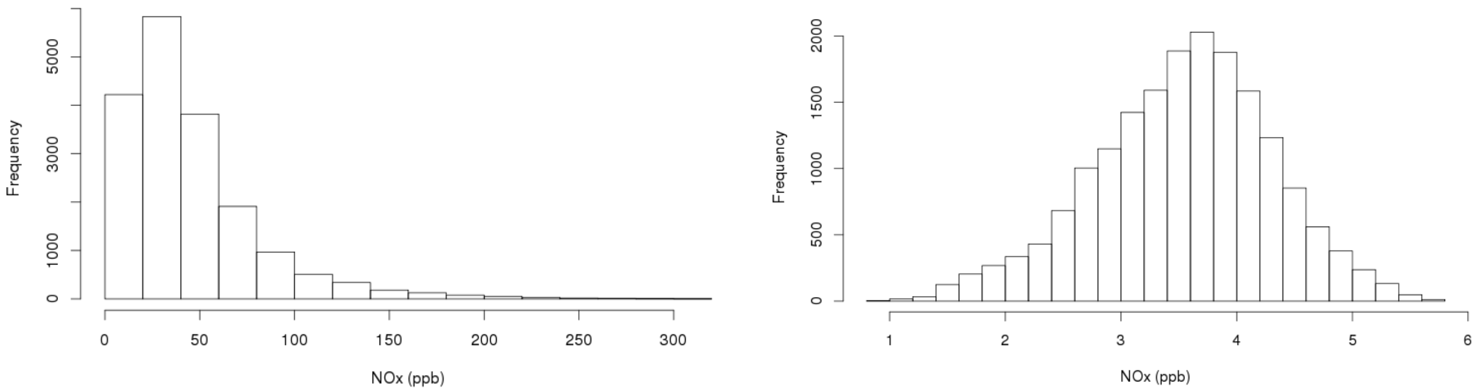


Figure S1. Study region with routine monitoring locations and USC sampling locations (the UCLA and UCI data are not shown due to IRB restrictions)

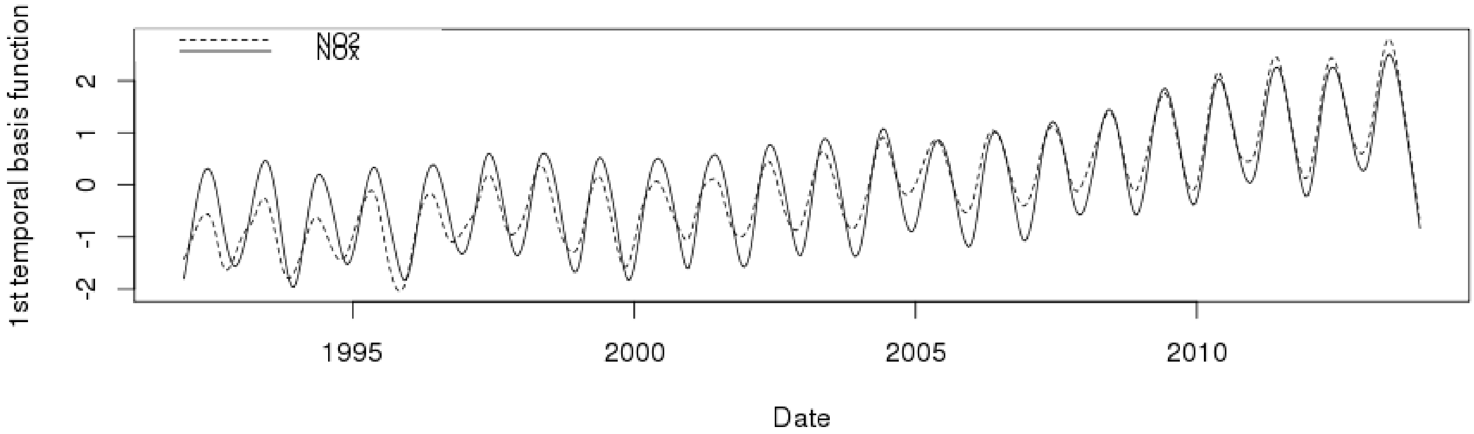


a. Histogram of NO_2 (with a small right skewness of 0.6)

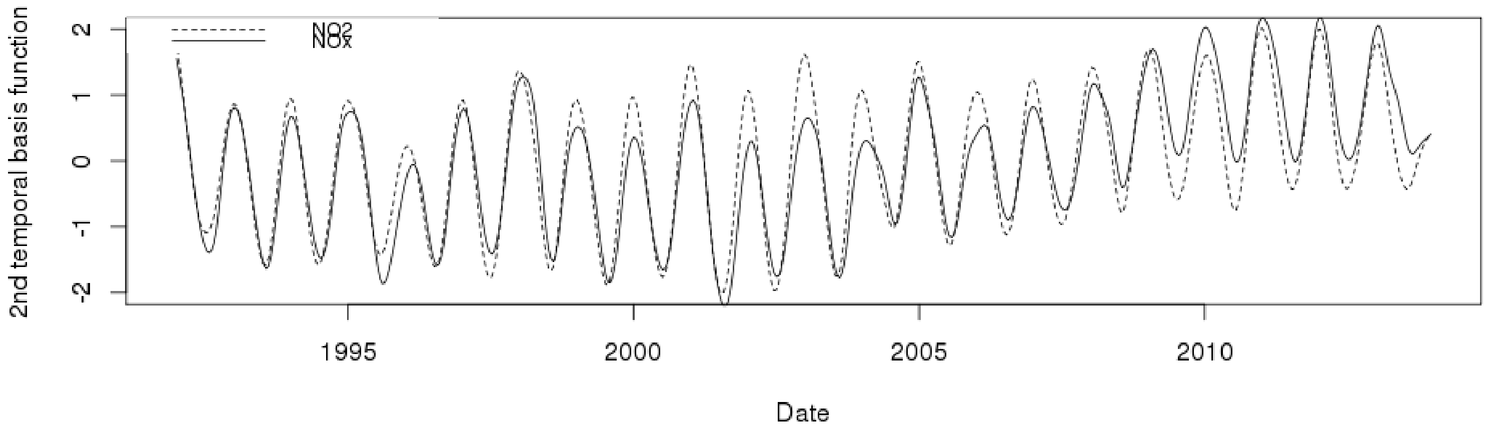


b. Histogram of NO_x (left: the original NO_x measurements with a big right skewness of 2.0; right: log-transformation with a small skewness of -0.25)

Figure S2. Histograms for NO_2 and NO_x to determine log-transformation (no log transformation for NO_2; log transformation for NO_x)

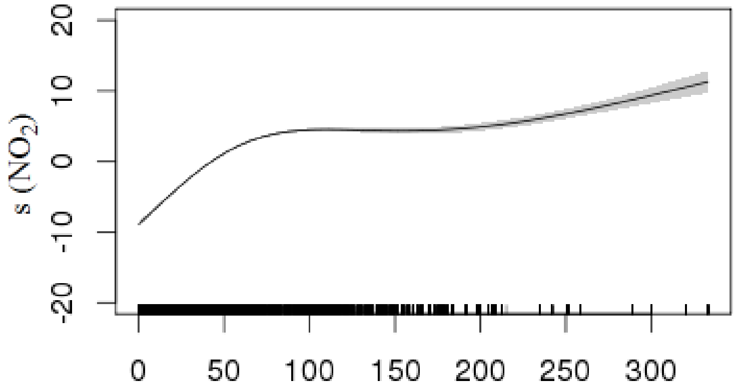


a. the first temporal basis function

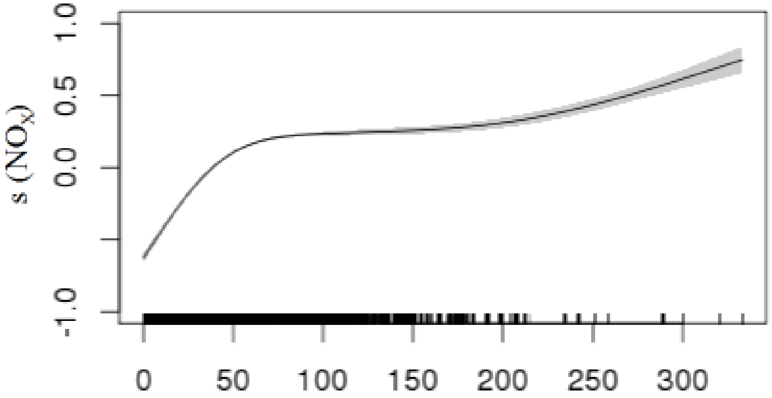


b. the second temporal basis function

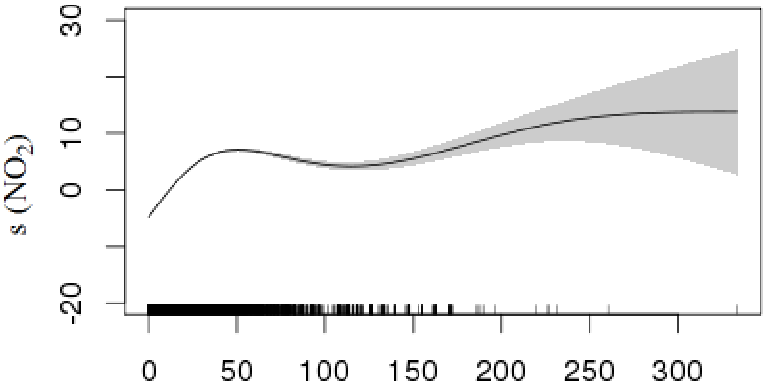
Figure S3. The first and second temporal basis functions to reflect the seasonal variability for the study region



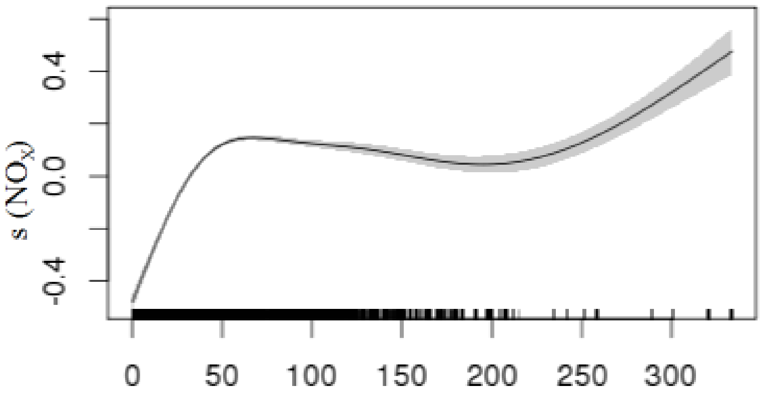
a. Traffic density (buffer distance: 0.3-5 km) for NO_2



b. Traffic density (buffer distance: 0.3-5 km) for NO_X



c. CALINE4 output (ppb) on freeways for NO_2



d. CALINE4 output (ppb) on freeways for NO_X

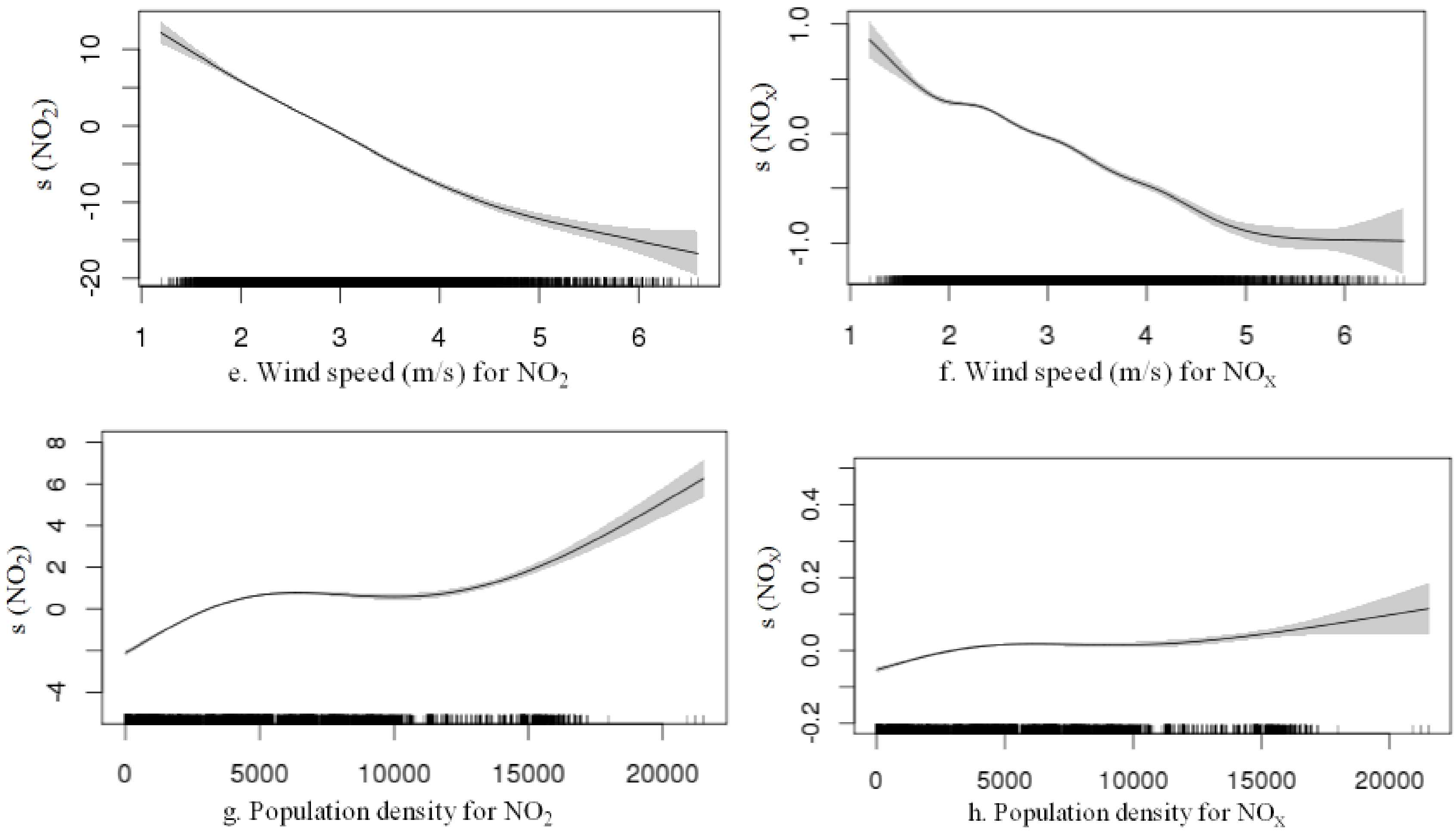


Figure S4. Non-linear association between predictive variables and concentrations by mixed models

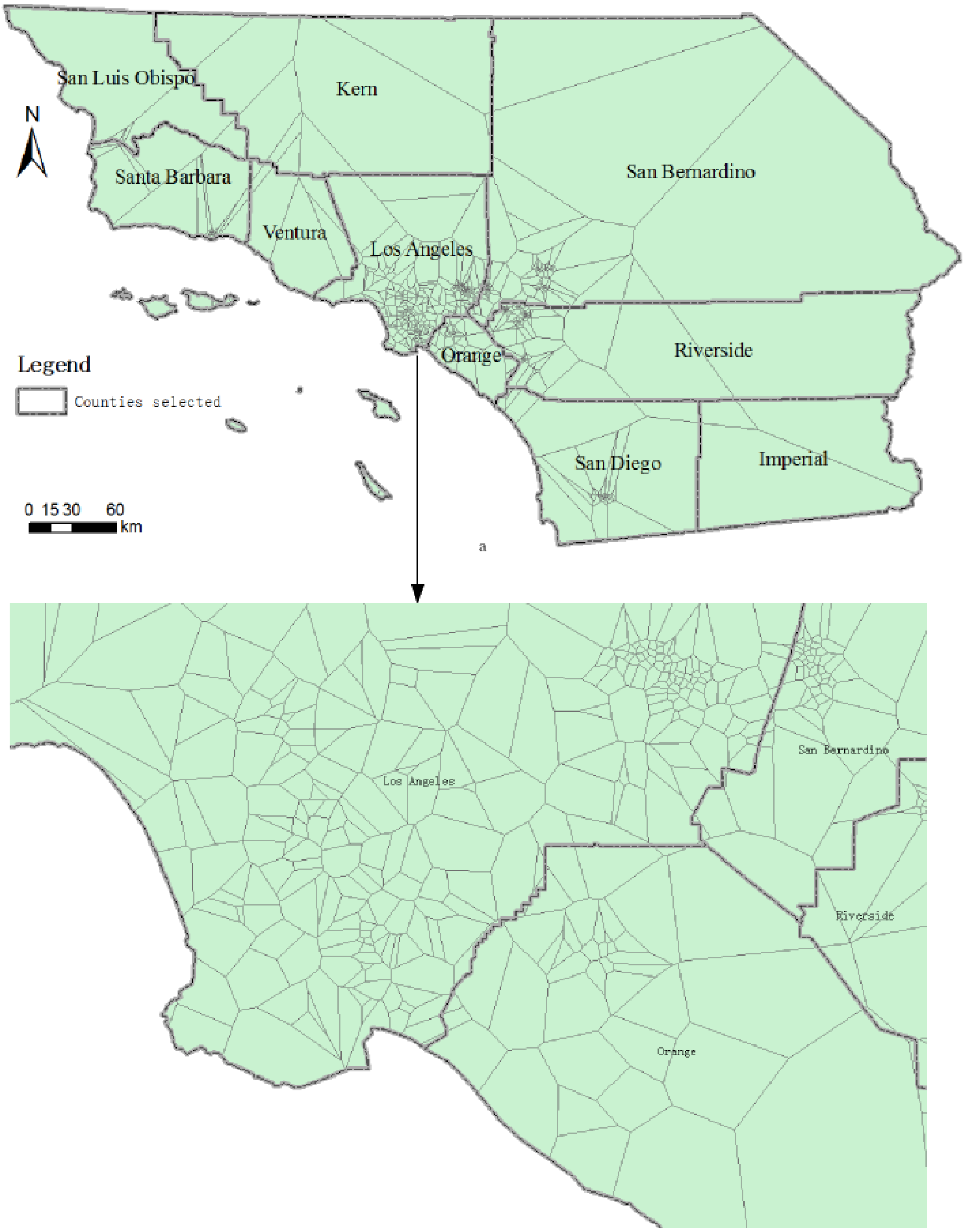
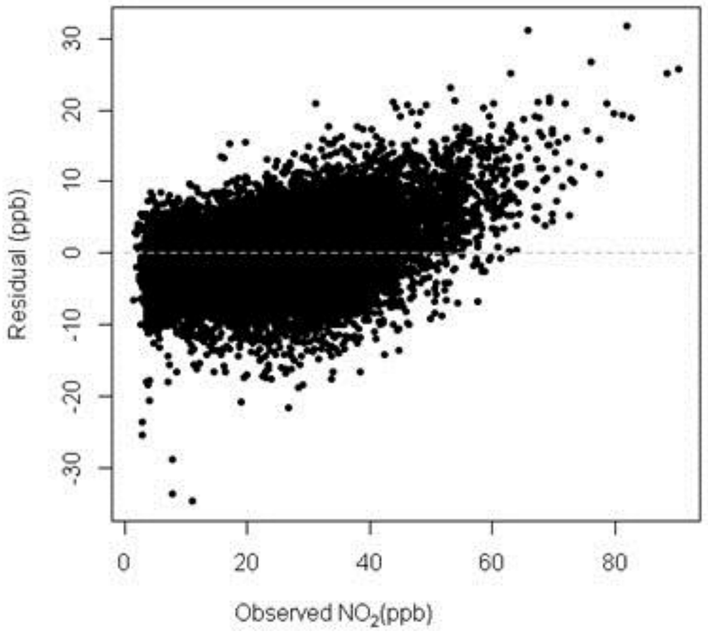
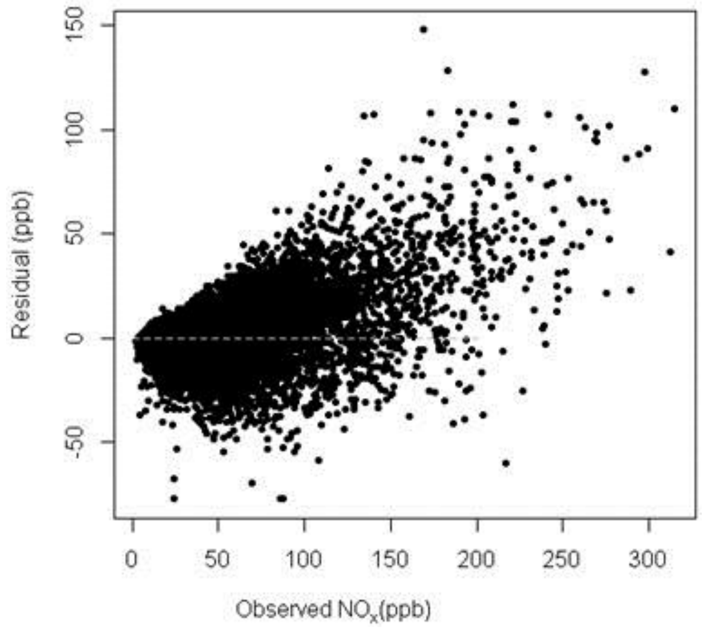


Figure S5. Spatial topology for spatial effect modeling by Thiessen polygons (aggregate distance: 500 m)

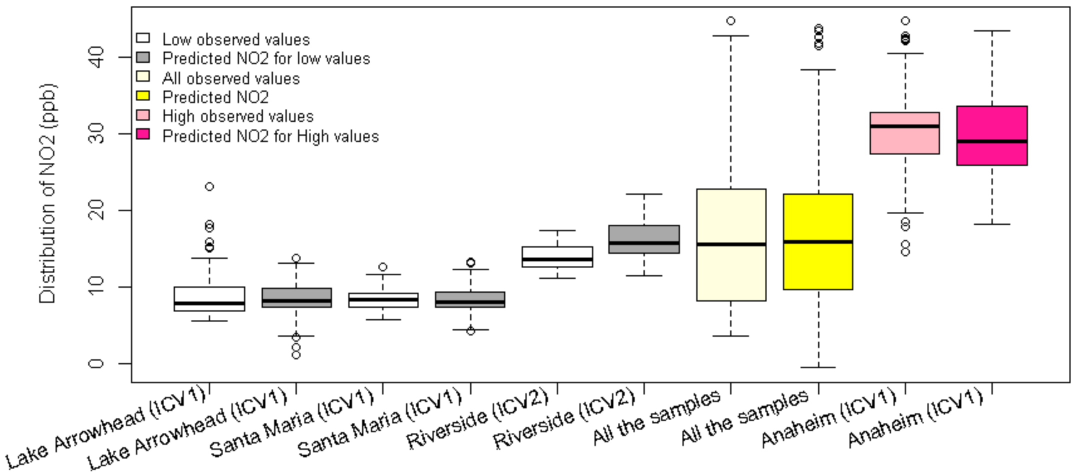


a

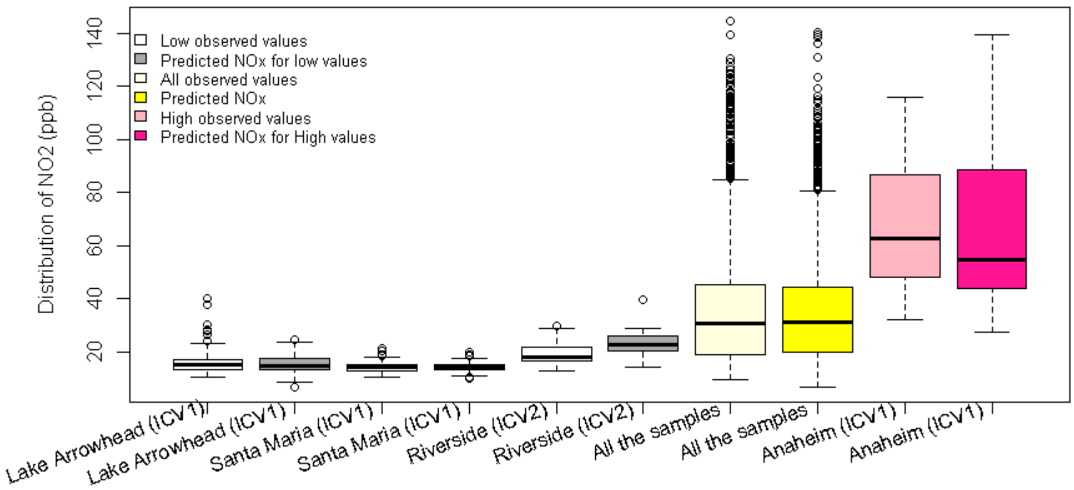


b

Figure S6. Residual plot between observed values vs. residuals (a. NO_2; b. NO_x) for the sample selected by bootstrap aggregating

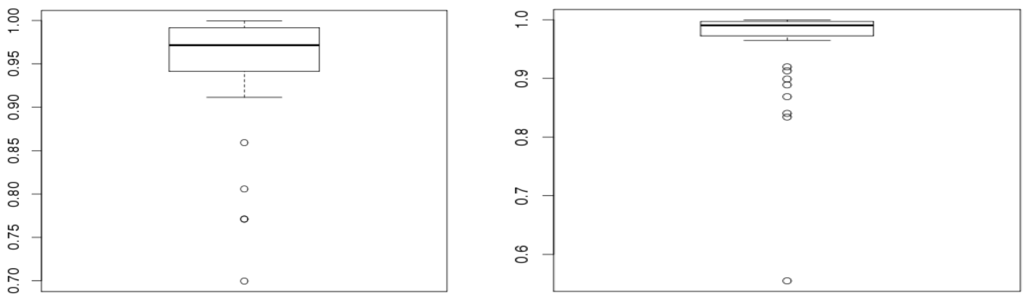


a. NO_2



b. NO_x

Figure S7. Spatial distribution of the predicted and observed NO_2 and NO_x concentration means across the CHS communities



a. NO_2 b. NO_x

Figure S8. Boxplot for correlation between constrained prediction and observed values for the time series of 51 routine monitoring stations

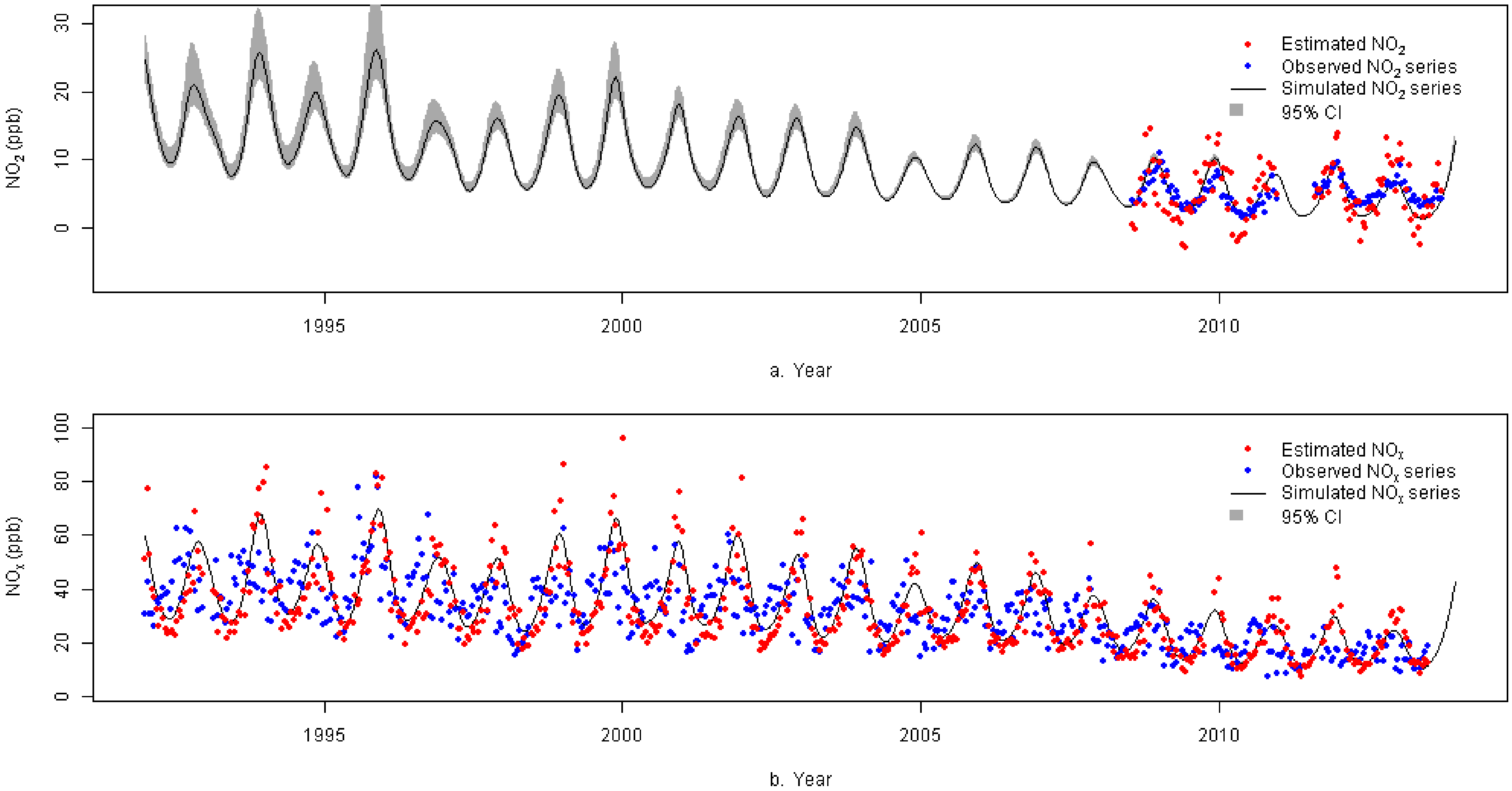
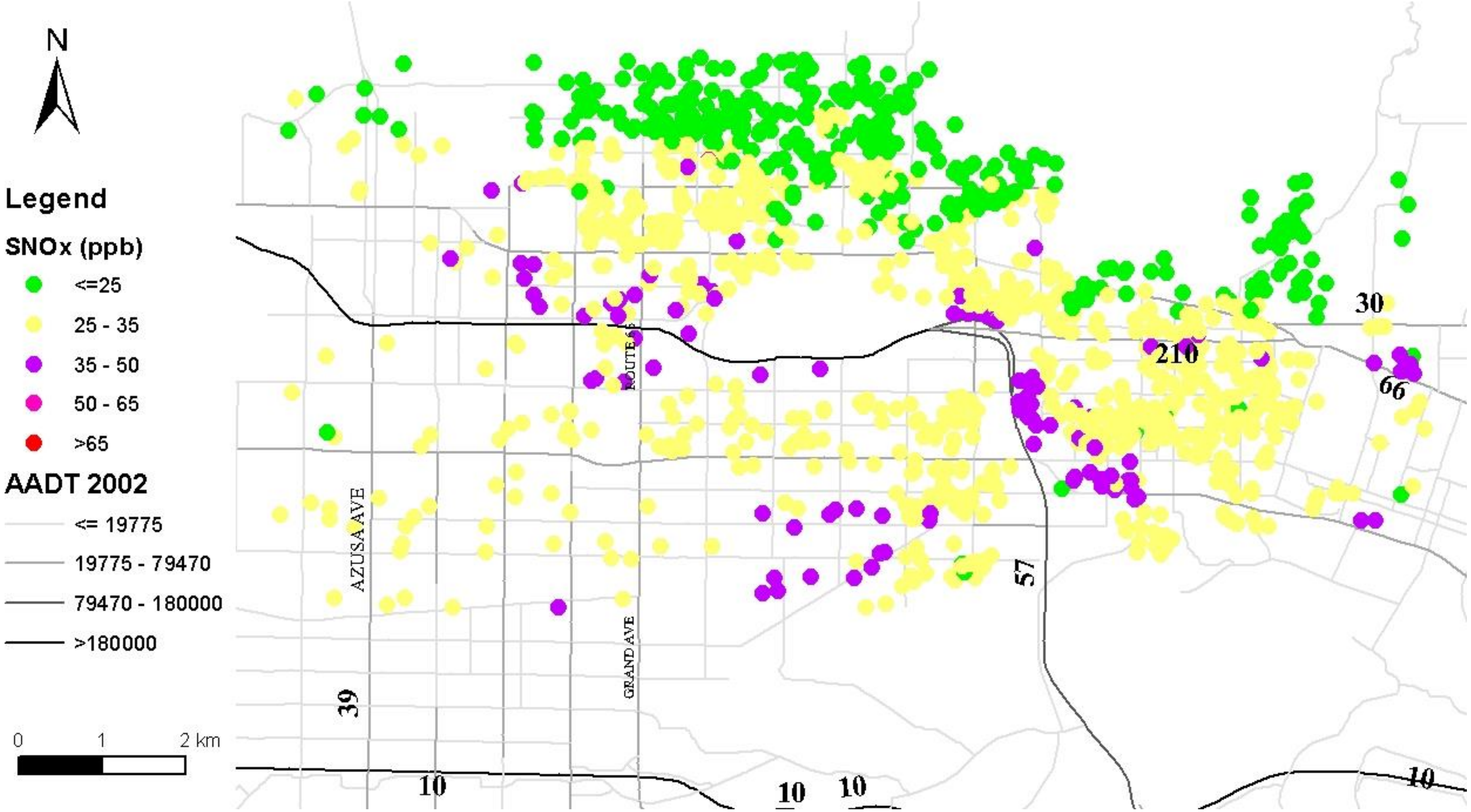
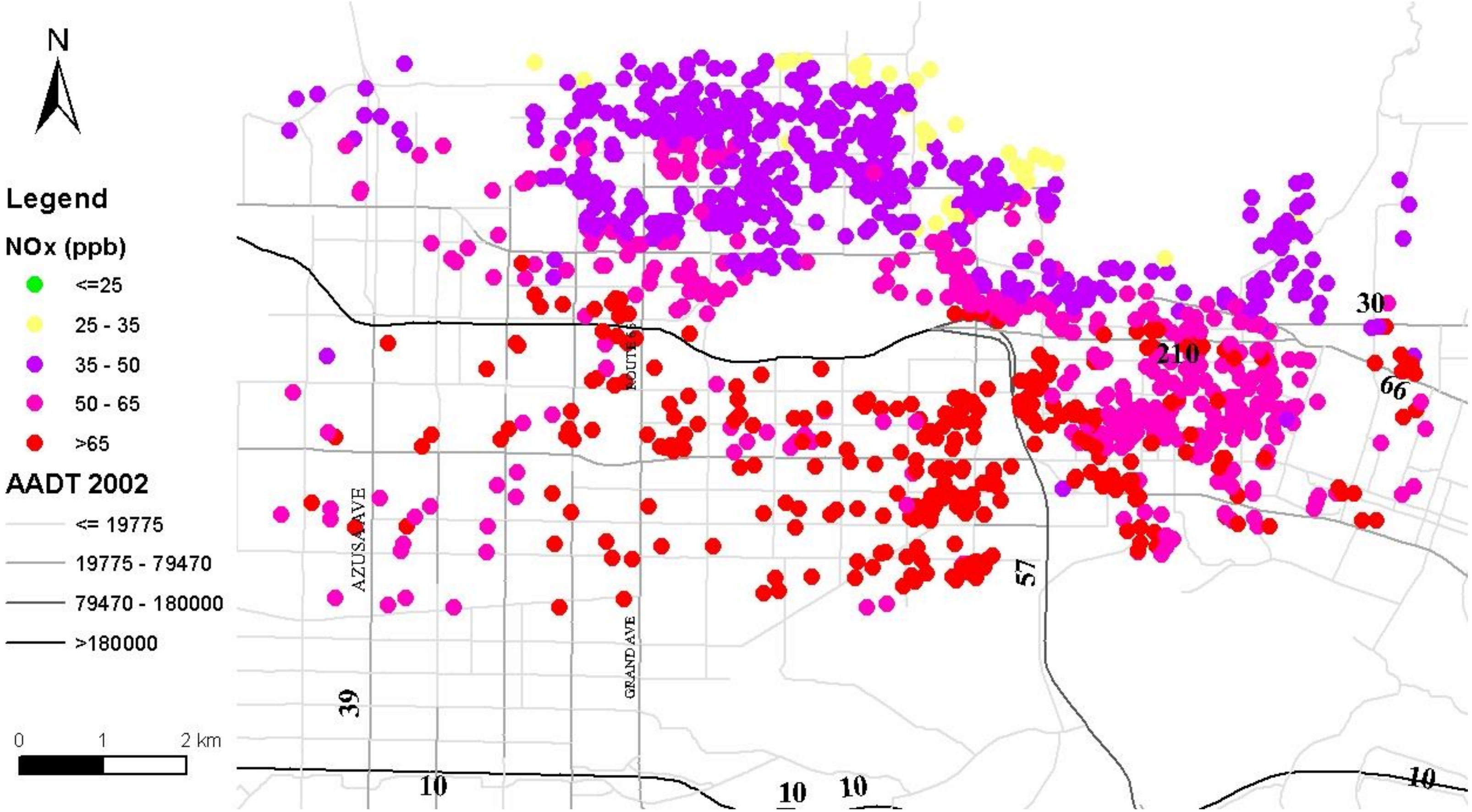


Figure S9. Time series simulated for the routine monitoring sites with the minimum correlation between constrained prediction and observed values (a.0.55 for NO_2; 0.70 for NO_x)


N
Legend
SNOx (ppb)
<=25
25 - 35
35 - 50
50 - 65
>65
AADT 2002
<= 19775
19775 - 79470
79470 - 180000
>180000
0
1
2 km
AZUSA AVE
ROUTE 39
GRAND AVE
39
10
10
10
10
57
210
30
66

(a)



(b)

Figure S10. Summer (a) and winter (b) averages of the 2005-2006 biweekly NO_x at USC ICV1 sampling locations for San Dimas

References

1. Koutrakis, P.; Wolfson, J. M.; Bunyaviroch, A.; Froehlich, S. A. *Passive Ozone Sampler Based On A Reaction with Nitrite*; Health Effects Institute: Boston, MA, 1994; pp 19-47.

2. Franklin, M.; Vora, H.; Avol, E.; McConnell, R.; Lurmann, F.; Liu, F.; Penfold, B.; Berhane, K.; Gilliland, F.; Gauderman, W. J., Predictors of intra-community variation in air quality. *Journal of exposure science & environmental epidemiology* **2012,** *22*, (2), 135-47.

3. Su, G. J.; Jerrett, M.; Beckerman, B.; Wilhelm, M.; Ghosh, K. J.; Ritz, B., Predicting traffic-related air pollution in Los Angeles using a distance decay regression selection strategy *Environmental Research* **2009,** *109*, (2009), 657-670.

4. Li, L.; Wu, J.; Wilhelmc, M.; Ritz, B., Use of generalized additive models and cokriging of spatial residuals to improve land-use regression estimates of nitrogen oxides in Southern California. *Atmospheric Environment* **2012,** *55*, 220-228.

5. Benson, P. *CALINE4: A Dispersion Model for Predicting Air Pollutant Concentrations near Roadways*; California Department of Transportation: Sacramento, CA, 1989.

6. Wu, J.; Houston, D.; Lurmann, F.; Ong, P.; Winer, A., Exposure of PM2.5 and EC from diesel and gasoline vehicles in communities near the Ports of Los Angeles and Long Beach, California. *Atmos Environ* **2009,** *43*, (12), 1962-1971.

7. California Air Resources Board *EMFAC2011 Technical Document*; 2011.

8. California Air Resources Board *EMFAC2014 Volume III -Technical Documentation*; 2015.

9. Besag, J.; York, J.; Molli¶e, A., Bayesian image restoration with two applications in spatial statistics *Annals of the Institute of Statistical Mathematics* **1991,** *43*, 1-59.