



Published in final edited form as:

Atmos Environ. 2013 June 1; 71: 54–63. doi:10.1016/j.atmosenv.2013.01.038.

Estimating Spatiotemporal Variability of Ambient Air Pollutant Concentrations with A Hierarchical Model

Lianfa Li^{1,2}, Jun Wu^{1*}, Jo Kay Ghosh³, and Beate Ritz⁴

¹Program in Public Health, College of Health Sciences, University of California, Irvine, USA

²State Key Lab of Resources and Environmental Information Systems, Institute of Geographical Sciences and Natural Resources Research, Chinese Academy of Sciences, China

³Department of Preventive Medicine, Keck School of Medicine, University of Southern California, Los Angeles, USA

⁴Department of Epidemiology, School of Public Health, University of California, Los Angeles, USA

Abstract

Studies have linked exposure to air pollutants to short-term and sub-chronic health outcomes. However, individual-level air pollution exposure is difficult to measure at a high spatial and temporal resolution and for larger populations due to limitations in sampling techniques. We presented a hierarchical model to capture spatiotemporal variability of nitrogen dioxide (NO₂) and nitrogen oxides (NO_x) concentrations in Southern California by combining high temporal resolution data from routine monitoring stations with high spatial resolution data from investigator-initiated episodic measurements. In this model, the spatiotemporal field of concentrations was first decomposed into a mean and residual and the mean representing the seasonal trend was further decomposed into a constant and varying temporal basis functions. The mean of the spatially varying coefficients of temporal basis functions were modeled by local covariates using non-linear generalized additive model and least square fitting using measurements from both routine monitoring and additional episodic sampling locations, while the spatially-correlated residuals of the coefficients were co-kriged. We found traffic, land-use and wind accounted for a large portion of the variance (beyond 35%) for the long-term average trend of concentrations. Spatial residuals accounted for a large portion of the variance of the temporal components (about 30% for NO₂ and 20% for NO_x). Leave-one-out cross validation produced an R² of 0.84 for NO₂ and 0.81 for NO_x when comparing the modeled weekly concentration with the observed trends at all routine monitoring stations.

Keywords

spatiotemporal variability; temporal trend; air pollution; generalized additive model; nitrogen dioxide; nitrogen oxides

© 2013 Elsevier Ltd. All rights reserved.

*Corresponding author: Program in Public Health & Department of Epidemiology, Anteater Instruction & Research Bldg (AIRB) # 2034, University of California, Irvine CA 92697-3957. Tel: 949-824-0548, Fax: 949-824-0529, junwu@uci.edu.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

1. Introduction

Exposures to air pollution have been linked to short-term and sub-chronic adverse health outcomes, including daily respiratory and cardiovascular morbidity (Mahiyuddin et al., 2013; Zanobetti et al., 2003), and adverse pregnancy outcomes (Ritz and Wilhelm, 2008). Many time-series health effect studies relied solely on area-wide exposure measures that failed to account for spatial exposure variability among individual subjects, and thus may have underestimated the magnitude of estimated effects (Lindstrom et al., 2011). For urban areas with spatially heterogeneous emission sources (e.g. Los Angeles, California), the within-area variation of air pollutant concentrations was possibly larger than the between-area variations (Jerrett et al., 2005). In these areas, it is essential to account for individual-level exposure variation in air pollution epidemiological studies.

Spatial proximity, spatial interpolation, and land use regression (LUR) models have been extensively used to assign exposure estimates to individual subjects that are more spatially refined but generally ignore temporal patterns of variation. Some studies assign measurements taken at the nearest monitoring station to a target location. This approach usually allows one to generate exposure estimates at a moderate or high temporal resolution while the spatial resolution is relatively low due to the limited coverage of the routine monitoring stations. Spatial interpolation (e.g. kriging) derives unbiased exposure estimation based on the existing measurements (Whitworth et al., 2011) with no additional spatial information used; spatial resolution is thus still limited by the number of routine sampling sites. The LUR models incorporate local covariates, including land-use/cover, population, and traffic information, to predict intra-urban variation of air pollution based on measurements at a high spatial resolution (Jerrett et al., 2005). Most previous LUR models focused on characterizing long-term spatial (Hart et al., 2009) or seasonal exposures (Li et al., 2012) and had no or low temporal resolution (Hoek et al., 2008). Some studies superimposed temporal profiles from routine monitoring stations onto the long-term LUR-estimated concentrations (Ghosh et al., 2012). These approaches may be inadequate because they typically rely on measurements at fewer locations than used to develop the models themselves or/and assume that temporal changes are constant across the study area (Wu et al., 2011). Jerrett et al. (2012) applied LUR models repetitively on monthly measurements of particulate matter $<2.5 \mu\text{m}$; Bayesian maximum entropy (BME) was further used to estimate spatiotemporal variability of the residuals of LUR-estimated monthly means. BME is a local spatiotemporal modeling approach that incorporates measurement data, prior knowledge of neighbor information, and local spatiotemporal covariates. This study, however, did not systematically integrate temporal variability in the models.

Only a few studies have systematically modeled spatiotemporal variability of air pollutants at a high spatial and temporal resolution. BME was developed to estimate the spatiotemporal concentrations of particulate matter $<10 \mu\text{m}$ and ozone (Yu et al., 2009). Two-stage models were developed for particulate matter with a second-order stationary and isotropic assumption of spatiotemporal covariance and linear regression to correlate the covariates to spatiotemporal trend of concentrations (Cameletti and Ignaccolo, 2010). Szpiro et al. (2010) constructed a hierarchical model for nitrogen oxides (NO_x) with a flexible correlation structure and spatial stationary assumption. Lindstrom et al. (2011) adapted the hierarchical model used in Szpiro et al. (2010) by adding a spatiotemporal covariate (i.e. estimated concentration from a line-source dispersion model).

Most spatiotemporal models were trained using the routine monitoring data. In this paper, we presented a hierarchical model that fully utilizes all available measurements (both long-term data at routine monitoring stations and additional short-term data from the investigator-initiated field campaigns) and spatial covariates to reliably estimate spatiotemporal

variability of pollutant concentrations. The U.S. Environmental Protection Agency (EPA) monitoring network provides long-term measurements for the criteria air pollutants at a high temporal resolution throughout the country, although the sampling stations are usually sparsely located. Investigator-initiated field campaigns, on the other hand, usually have a relatively dense spatial coverage, but are often short-term and passive sampling only for one or two types of pollutants. In this study, we developed hierarchical two-stage models based on a couple of previous studies (Finkenstadt et al., 2007; Szpiro et al., 2010) that combined dominant temporal trends at regional scale with spatial variability at local scale to improve estimation of spatiotemporal variability of pollutant concentrations. The models were developed for nitrogen dioxide (NO₂) and NO_x, two criteria air pollutants widely used in air pollution health studies (Jerrett et al., 2012).

2. Materials

2.1. Study domain

This study covers the metropolitan Los Angeles area (160 × 161 km²; over 12 million in population in 2010) including both Los Angeles and Orange counties in Southern California. This region has serious air pollution problems due to high emissions (e.g. traffic, industrial and port-related activities), unfavorable meteorology, and high population density. There are six major commuter and truck transport freeways within this densely populated region.

2.2. NO₂ and NO_x measurements

2.2.1. Long-term routine measurements from government-operated monitoring stations—We obtained routine measurements of hourly NO₂ and NO_x concentrations (unit: ppb) from 2006 to 2009 from the South Coast Air Quality Management District (SCAQMD) monitoring network at 32 stations in Southern California. The measurements were conducted by federally designated automated chemiluminescence methods with actively sampling instruments. Weekly average concentrations were calculated based on hourly data using a 75% completeness criterion; a total of 209 weekly concentrations were calculated from January 2006 to December 2009. Figure 1 shows the SCAQMD monitoring stations in the study domain. Among the 32 SCAQMD stations, one station was excluded because of the extreme values observed at this station based on the criteria of outer fences (Iglewicz and Hoaglin, 1993). Weekly concentrations from the 25 routine stations with complete (<3% missing values) time series data were used in the first stage of the hierarchical model to estimate the trend of constant and varying temporal basis functions. The measurements from the rest 6 stations (excluding the one having outliers) were used as episodic measurements due to more missing values (>3%).

2.2.2. Episodic measurements—Besides the measurements with >3% missing values from the 6 SCAQMD sites, episodic measurements include 161 valid NO₂ and NO_x samples from University of California, Los Angeles (UCLA) (collected in two continuous weeks, i.e. September 16–October 1, 2006 and February 10–25, 2007) and 32 valid samples of measurements for NO₂ and NO_x from University of California, Irvine (UCI) (collected at outdoor home locations of subjects in south Los Angeles and Orange counties for four weeks, i.e. July 10–18, July 24–August 1, November 13–21 and December 4–12 in 2009). For details of episodic measurements from UCI and UCLA, and their adjustment for use in combination with SCAQMD routine data, please refer to Supplemental Materials Section 2.1 and Table S1.

2.3. Spatial covariates

2.3.1. Covariates related to emission—Emission sources contribute significantly to the variability of pollutant concentrations. Traffic was one of the major sources of air

pollution in the Los Angeles area. Land-use data were used as surrogates of traffic and other emissions.

1. Roadway covariates include total roadway length (meter) inversely weighted by the distance to the sampling location within an optimal buffer size as well as the shortest distances (meter) from the sampling location to each of four roadway types, and all the four type of roadways (highway as A1, primary roads as A2, secondary or connecting roads as A3 and local roads as A4, see Supplemental Materials Section 2.2 for details). To estimate the optimal buffering distance, Pearson's correlations between the distance-weighted roadway length and the concentrations were calculated within the buffers of increasing radii (from 50 m to 5 km with interval of 50 m or 100 m); an optimal buffering distance had the highest correlation. We used 5 km as a threshold since a wider buffering distance might include influence from background and regional sources. This method was also used for the other spatial covariates that required buffer statistics. The ESRI StreetMap™ North America 9.3 (<http://www.esri.com>) including 2003 TeleAtlas® street polylines was used as the roadway data.
2. Traffic flow. We obtained annual average daily traffic (AADT) counts in 2005 for all the freeways, highways, and major surface streets from the California Department of Transportation (Caltrans). AADT was weighted by the roadway length within an optimal buffer.
3. Land-use covariates. Land-use data were obtained from the Southern California Association of Government (SCAG). The original 108 land-use types were classified into five major categories: transportation, industry, agriculture including open space and vacant, commercial area, and residential area. We calculated the percentage of area for each land-use category within an optimal buffer size.

2.3.2. Meteorological covariates—Long-term meteorological covariates were derived by averaging multiple-year (2006–2009) measurements of temperature (°C), wind speeds (meters/second abbreviated as m/s), wind direction, relative humidity (%), and precipitation (inch) from the Air Quality and Meteorological Information System managed by the California Air Resources Board (<http://www.arb.ca.gov/aqmis2/aqmis2.php>). Wind data were decomposed into two vectors by respectively multiplying wind speed with sine and cosine of wind direction, with the sine value positive from the east and the cosine value positive from north (Carslaw et al., 2007). There were 67 sites for temperature, 77 sites for wind speed and wind direction, 37 sites for relative humidity, and 55 sites for precipitation. Measurements at the nearest meteorological site were assigned to the corresponding sampling location.

2.3.3. Other location-related covariates—We obtained elevation data at 10m×10m resolution from US National Elevation Dataset (NED) (<http://nationalmap.gov/>) and assigned the elevation (meter) to each sampling location. We also calculated the shortest distance (meter) to shoreline for each sampling location.

3. Modeling approach

3.1. Stage one: decomposition of spatiotemporal field

A normally-distributed log-transformed spatiotemporal field y_{ut} ($u=1, \dots, n$ site; $t=1, \dots, m$ time slice) was decomposed into the sum of a systematic mean component (μ_{ut}) and residual ($\hat{\epsilon}_{ut}$) (Finkenstadt et al., 2007):

$$y_{ut} = \mu_{ut} + \widehat{\varepsilon}_{ut} \quad [1]$$

The mean spatiotemporal field typically represents the dominant seasonal or long-term trend of concentrations for the study domain (Szpiro et al., 2010). At the first stage, the mean component was divided into one constant and two varying temporal basis functions using the empirical orthogonal functions (EOFs) (a.k.a. independent temporal basis functions) and the long-term weekly concentrations from the 25 SCAQMD routine monitoring stations. EOFs were often used to present leading modes of variability in space-time processes in meteorology; their smoothed curves are often used to reduce noise (Finkenstadt et al., 2007). In [1], the mean μ_{ut} is essentially a projection of time series at the location, u onto the space spanned by the constant temporal basis function (represented by β_{0u}) and varying ones $f_i(t)$ ($i=1, \dots, m$) and corresponding spatially varying coefficient β_{iu} at u for $f_i(t)$:

$$\mu_{ut} = \beta_{0u} + \beta_{1u}f_1(t) + \beta_{2u}f_2(t) + \dots + \beta_{mu}f_m(t) \quad (u=1, \dots, n; t=1, \dots, m) \quad [2]$$

where β_{0u} is an intercept ($f_0(t) \equiv 1$) representing spatial variability of the constant temporal basis function. Similarly, the coefficient, β_{1u} (β_{2u}) representing spatial variability of the first (second) temporal basis function. In [1], $\varepsilon_{ut} \sim \mathcal{N}(0, \sigma)$, independent of μ_{ub} is assumed to be spatially correlated after removal of temporal auto-correlation and it can be modeled using kriging. Autocorrelation function (ACF) and partial autocorrelation function (PACF) were used to test the temporal autocorrelation of the residuals. We found that the ACF and PACF values were generally below 0.3, indicating very weak temporal autocorrelation of the residuals.

In [2], singular value decomposition (SVD) was used to generate the independent temporal basis functions. Iterative SVD (Finkenstadt et al., 2007) was also used to fill 1–5 temporal missing values (<3%) for 13 routine SCAQMD stations. The derivative-based thin plate spline penalty was applied to select a good degree of freedom to fit the temporal basis functions. With a good degree of freedom, the smoothed curve will properly represent the true temporal trend and meanwhile minimize over-fitting (Szpiro et al., 2010). We used Wood (2002)'s integrated approach of model selection and automatic smoothing parameter selection to select the degree of freedom with generalized cross-validation (GCV) criterion to determine the smoothing parameters for the temporal basis functions. More details about the temporal basis functions can be found in Supplemental Materials Section 2.3.

3.2. Stage Two: estimation of spatially varying coefficients for independent temporal basis functions

At this stage, spatially varying coefficients at the routine stations and episodic sampling sites were estimated using ordinary least square technique. We then applied a non-linear generalized additive model (GAM) plus cokriging to model these coefficients using local spatial covariates.

The product of the first three coefficients in [2], β_{iu} ($i=0, 1, 2$) and their temporal basis functions can usually capture the majority of spatiotemporal variability of concentrations (Lindstrom et al., 2011; Szpiro et al., 2010). In practice, we only need to solve the following linear equations to estimate the three coefficients at u , assuming observed values of L time slices for u :

$$y_{ut} = \mu_{ut} + \varepsilon_{ut} = \beta_{0u} + \beta_{1u}f_1(t) + \beta_{2u}f_2(t) + \varepsilon_{ut} \quad (t=1, \dots, L) \quad [3]$$

Since ε_{it} is independent of μ_{ub} ($\varepsilon_{it} \sim N(0,1)$), we used the least square fit to solve β_{iu} ($i=0, 1, 2$). A unique solution for β_{iu} ($i=0, 1, 2$) is possible in equation [3] if we have a minimum of three time slices of measurements at each sampling location according to the Rouché–Capelli theorem (David, 2006).

The UCI and SCAQMD data had at least four weekly measurements and were directly used to solve the β_i coefficients in [3]. But the UCLA data had only two bi-weekly measurements, fewer than a minimum of three required to solve β_{iu} ($i=0, 1, 2$). Thus, the initial ratios between the two continuous weekly averages of concentration and the ratios were estimated using the linear spatiotemporal model and then used to interpolate four weekly concentration averages from two bi-weekly measures of concentration averages. In this spatiotemporal model, we derived temporal basis functions and spatial coefficients using the 25 routine measurements of time series (no episodic data were used) and estimated the spatial coefficients based on two or three significant spatial covariates (roadway length, AADT or traffic land-use, and wind speed). We then estimated two continuous weekly concentrations at the UCLA sampling locations and the initial ratios respectively for 2006 and 2007. The procedure was based on maximum likelihood without the incorporation of spatial autocorrelation of residuals (Supplemental Materials Section 2.4).

Sensitivity analysis was conducted to examine the influence of the interpolation methods on model performance. We compared the interpolation method above (i.e. linear spatiotemporal model) with two other methods. One method assigned the two-week average concentration to each of the two individual weeks, assuming no variation in the concentrations over the two continuous weeks (i.e. initial ratio =1). The other one used the ratio of two weekly measurements at the nearest SCAQMD station to split the data, assuming that the ratio of concentrations at the nearest station properly reflected the variation of concentrations over the two continuous weeks.

3.2.1 Removal of outliers in episodic measurements—Outliers from episodic samples may affect the estimation of β_{iu} ($i=0, 1, 2$). We estimated β_i of the 25 routine stations, which were then used as the reference values to exclude the outliers in episodic measurements. The outer fences (Iglewicz and Hoaglin, 1993) of β_i were used to filter the outliers. A data point was treated as an outlier if the value of β_i was outside the outer fences ($Q1-3*IQR$, $Q3+3*IQR$; $Q1$ and $Q3$ are respectively the first and third quartiles, and IQR is inter-quartile range). In total, we excluded 8 outliers (4 from UCLA and 4 from UCI) for NO_2 and 6 outliers from UCI for NO_x .

The removal of too few outliers may cause poor model performance, while the removal of too many outliers may over-fit the model. Sensitivity analysis was conducted for three different definitions of outliers for β_i : 1) No outliers removed from the UCI and UCLA samples (reference); 2) Inner fences used to remove outliers ($Q1-1.5*IQR$, $Q3+1.5*IQR$); 3) Outer fences as less conservative criteria to remove outliers, introduced above and used in this study (Supplemental Materials Table S6). Compared to the models with no outlier removal, the removal of outliers improved the model performance by 5% to approximately 20% (Supplemental Materials Table S6).

3.2.2. GAM: modeling local mean of spatially varying coefficients—The GAM equation for prediction of spatially varying coefficients for temporal basis functions is:

$$\widehat{\beta}_{iu} = \widehat{\beta}_{iu}(X) + \widehat{\varepsilon}_{iis}(Z) + \widehat{\varepsilon}_{iim} \quad [4]$$

where $\hat{\beta}_{iu}$ ($i=0, 1, 2$) is estimate of spatially varying coefficient for the i^{th} basis function at location u , $\hat{\beta}_{iu}$ is the estimate of local mean at u for $\hat{\beta}_{iu}$, dependent on the set of p spatial covariates, $X=\{x_u^j\}$ ($j=1 \dots p$), $\hat{\epsilon}_{ius}$ is the estimate of spatial residual for $\hat{\beta}_{iu}$, determined by spatial residuals of samples that belong to the neighborhood of u , $Z \in Nb(u)$, $\hat{\epsilon}_{iun}$ is a random residual at u , with normal distribution, $\hat{\epsilon}_{iun} \sim \mathcal{N}(0,1)$.

We used the GAM package in R statistical software (Version 2.11.1) to develop and validate the model. The following is to correlate local spatial covariates to the mean of $\hat{\beta}_{iun}$:

$$g(E(\hat{\beta}_{iu})) = \mu_{i0} + \sum_{j=1}^q s_j(x_u^j) + s_w(w_{-s_u}, w_{-c_u}) + \sum_{k=q+1}^p \gamma_k x_u^k \quad [5]$$

where μ_{i0} is the model intercept, w_{-s_u} ([wind speed] · sine([wind direction]) and w_{-c_u} ([wind speed] · cosine([wind direction]) respectively represent two wind vectors, at u , x_u^j or $x_u^k \in X$ are other local covariates, $s_j(\dots)$ is the smooth function to model the non-linear relationship between x_u^j and $g(\hat{\beta}_{iu})$, $s_w(w_{-s_u}, w_{-c_u})$ is a bivariate smooth function to model the interaction between wind speed and wind direction, γ_k are the linear parameters used to construct the linear relationship between x_u^k and $g(E(\hat{\beta}_{iu}))$, and p is the number of covariates. For normally distributed $\hat{\beta}_{iu}$ ($i=0, 1, 2$), the link function is $g(E(\hat{\beta}_{iu}))$. If $\hat{\beta}_{iu}$ does not follow the normal distribution, a power transformer can be used to make them normal (Handelman, 2002). For the degree of freedom in non-linear fitting of $\hat{\beta}_p$ we applied the same approach that was used to fit the temporal basis functions.

For selection of covariates, we first used Pearson's correlation of 0.1 as a threshold to filter out the irrelevant ones and then selected optimal covariates by examination of multicollinearity and combinational tests. See Supplemental Materials Section 2.5 for details.

3.2.3 Kriging of residuals to minimize error variance

Spatial residuals $\hat{\epsilon}_{ius}$ in [4] were modeled by cokriging them with regional residuals at nearby samples with the assumption of a stable spatial domain after removal of local means (Christakos, 1990). Assuming $\mathbf{e} = [e_{us/rs}(u_i)] \sim \mathcal{N}(0, V(\theta))$ ($i=0, 1, 2, \dots, f$), u_0 is the location to be estimated, θ is the vector of variogram parameters (characterized by range ϕ , partial sill, σ^2 and optional model efficient, τ , nugget supposed to be 0 given the assumption of highly spatial correlation), $Z = \{u_1, u_2, \dots, u_f\}$ is the set of spatial samples belonging to the neighborhood of u_0 ; $\hat{\epsilon}_{us}(u_i)$ is the estimate of spatial residual at the neighboring location, u_i , and is derived by subtracting the GAM local mean from β_{ku_j} at a sampling location; $\hat{\epsilon}_{rs}(u_i)$ is the estimate of the regional residual or total variation of β_{ku_j} at the regional scale and derived by subtracting the population's mean from β_{ku_j} at each sampling location. Cokriging used maximum likelihood based on estimates of variogram, θ and we can get the estimate of the residual for u_0 :

$$\hat{\epsilon}_{us}(u_0) = \sum_{i=1}^{n_u} \lambda_{u_i}^u \hat{\epsilon}_{us}(u_i) + \sum_{i=1}^{n_u} \lambda_{u_i}^r \epsilon_{rs}(u_i) \quad [8]$$

where $\lambda_{u_i}^u$ and $\lambda_{u_i}^r$ are the optimal cokriging weights estimated according to the minimum unbiased optimal principle of Kriging.

3.3. Cross validation

For model evaluation, we used leave-one-out cross validation (LOOCV), which used an observation as the validation data and the remaining observations as the training data. This is repeated such that each observation in the sample is used once as the data point for which

the validation is performed. In LOOCV, we validated how the estimated trend of concentrations captured spatiotemporal variability of the observed time series for the 25 routine SCAQMD stations, by measuring the 95% confidence intervals of Pearson's correlations of time series for the 25 stations and R^2 for all the temporal trends of all the stations. We compared our approach with the spatiotemporal model based on linear regression (Finkenstadt et al., 2007). Further, we calculated R^2 of the estimated versus observed averages of the concentrations over 4 years (2006–2009) and the corresponding square root of the mean of the squared errors (RMSE).

4. RESULTS

4.1. Temporal basis functions and their spatially varying coefficients

Figure 2 depicts the first two varying temporal basis functions as trend curves smoothed by GAM for the study domain. The first component (product of temporal basis function and its spatially varying coefficient; the temporal trend alone unable to reflecting its seasonal trend) of the temporal basis trends showed clear seasonal changes, i.e. high in the winter and low in the summer. For NO_2 , the first component of temporal basis trend accounted for 59% of variance. For NO_x , the first component accounted for 56% (except outliers, its spatially varying coefficient, β_1 generally smaller than 0, Supplemental Materials Table S3). The trend of the second temporal basis function explaining a much lower percentage of variance (about 9% for NO_2 and 9% for NO_x), was inconsistent with the seasonal trend, and seemed more complicated. Supplemental Materials Table S3 lists the descriptive statistics of β_j generated by least square fitting based on all routine and episodic measurement data. Supplemental Materials Figure S1 presents the box plots.

4.2. Step one of spatial modeling: correlate local covariates to spatially varying coefficients

Table 1 lists the variances explained in total and by local covariate and spatial autocorrelation of the residuals. Figure 3 shows the non-linear relationship of change trend between smoothed β_0 and traffic factors, land-use and long-term average wind covariate with the 95% confidence intervals (the gray area). Supplemental Materials Figure S2–S7 show the changing trends between β_j and the covariates used as regressors.

For β_0 , significant local covariates included traffic-related factors (AADT, distance-weighted roadway length, traffic land-use or distance to roadways), meteorological covariate (long-term ambient wind covariate), shortest distances to the shorelines, land-use area proportions (residential, commercial or agricultural/open field) or elevation. Traffic factors together accounted for an important part of variance (32.9% for NO_2 and 19% for NO_x). Except for traffic land-use, other land-use factors together accounted for about 17% of variance for NO_2 or 10% for NO_x . Major traffic-related factors, residence land-use and shortest distance to shorelines present positive correlation with β_0 . Wind speed and wind direction were an important predictor, accounting for about 20.2–34.5% of variance. Lower wind speed was associated with higher β_0 , while wind from the northwest was associated with lower β_0 (Figure 3e and 3f). Elevation and shortest distance to the shorelines together explained 13.0% of the variance for NO_2 and 19.9% for NO_x .

For β_1 , the NO_2 variance was more influenced by long-term average wind (25.2%), land-use covariates (22.2%) and traffic-related factors (8.0%), while the variance of NO_x was more explained by long-term average wind (22.1%), weighted road length (15.4%) and land-use (24.6%). Wind from the north had stronger influence on NO_2 , while wind from the west had more influence on NO_x .

For β_2 , the variance of NO_2 was more influenced by long-term average wind (21.1%), commercial land-use (14.3%) and distance-weighted road length (6.8%), while the variance of NO_x was more explained by long-term average wind (33.5%), elevation (14.3%) and traffic-related factors (16.8%; including traffic land-use). The south wind had major contribution for β_2 of NO_2 .

4.3. Step two of spatial modeling: variogram of spatial residuals

Although only explaining a small-modest part of variance for β_0 (4.2–8%), spatial autocorrelation of the residuals for β_1 and β_2 accounted for a considerable proportion of the variance explained 19.4–31.8% (Table 1). Supplemental Materials Table S4 shows the variogram models and relevant coefficients for spatial residuals of β_i . We found that β_0 for the constant basis function had a longer range and smaller partial sills, indicating a more gradual spatial variability. β_1 and β_2 of the varying temporal basis functions had a shorter range and higher partial sills, indicating higher spatial variability, probably more affected by wind.

4.4. Comparison of modeled vs. observed temporal trend

Figure 4 shows the typical curves of two SCAQMD stations (one in Orange County and the other in Los Angeles County) simulated with our method based on the time series data at the other 24 SCAQMD stations and episodic measurement data vs. the practical smoothed curves derived from the observed time series. The simulated curves and the observed values showed a strong association. For the temporal trends of the 25 routine SCAQMD stations, the median of the correlation between the LOOCV predicted values and observed time series was 0.90 (mean: 0.87) with the 95% confidence interval (CI) of [0.83, 0.91] for NO_2 and 0.96 (0.92) with 95% CI of [0.88, 0.96] for NO_x .

For all 5225 (25×209) weekly samples, our model had a CV R^2 of 0.67 for NO_2 and 0.66 for NO_x between the predicted time series and observed values, but the model had a better CV R^2 (0.84 for NO_2 ; 0.81 for NO_x) between our predicted time series and the GAM-smoothed trend curves of weekly measurement data. Our model outperformed the linear spatiotemporal model in predicting the smoothed trend curve (0.84 vs. 0.56 for NO_2 ; 0.81 vs. 0.65 for NO_x in Table 2).

4.5. Comparison of modeled vs. observed long-term means

For long-term average concentrations, CV R^2 was 0.89 for NO_2 (RMSE =2.28) and 0.77 for NO_x (RMSE =6.8). Supplemental Materials Figure S8 shows the plots between predicted and observed long-term means for NO_2 and NO_x . Similar to our time-series predictions, we found that GAM plus cokriging improved R^2 by over 19% in variance (Table 2) for NO_2 and NO_x over linear regression to model spatially varying coefficients.

5. DISCUSSION

Our hierarchical two-stage models decomposed the spatiotemporal field into temporal components (temporal functions and their spatially-varying coefficients) at the first stage and then modeled the spatial fields of temporal components using a flexible spatial correlation structure at the second stage. Compared to a similar hierarchical modeling approach (Szpiro et al., (2010); Lindstrom et al., (2011)), our approach used the non-linear GAM to estimate spatial coefficients (β_i) of the temporal basis functions, incorporated regional residuals using cokriging, and examined additional spatial covariates including long-term meteorology. The GAM provided a more flexible modeling framework than a linear model. For example, in our study, long-term average wind covariate contributed 21.2–

34.5% of the β_j variability in GAM, but in the linear model it only accounted for about 8.0–10.0%.

Our model's R^2 of 0.77 for the long-term average estimates of NO_x was slightly better than the reported R^2 (0.58–0.67) from the spatiotemporal models of NO_x for the similar Los Angeles region (Lindstrom et al., 2011; Szpiro et al., 2010). Given different sources of measurements, direct comparison is inappropriate although our approach outperformed the linear model by over 15% in variance explained based on our data. Our estimation of long-term NO_2 had a R^2 of 0.89, better than that for NO_x .

In this study, traffic-related variables (such as traffic flow or roadway length), land-use, meteorological or elevation factors had more or less influence on spatiotemporal variability of NO_2 and NO_x (Table 1, Figure 3 and Supplemental Materials Figure S2–S7). As expected, the shortest distance to roadways was negatively associated with the long-term averages (β_0), while the other traffic-related factors were positively associated. They together accounted for a large portion of variance. β_0 tended to be higher further away from the shorelines since the predominant wind in Los Angeles Basin blows from ocean to inland and wind carries emissions generated in the coastal and upwind regions to inland areas. These results were consistent with the conclusion in Szpiro et al. (2010) and Lindstrom et al. (2011). Among the meteorological factors, wind explained a large portion of variance on spatial variability of the temporal basis functions. Lower wind speed limited the dispersion of air pollutants, thus resulting higher long-term average β_0 . Elevation was negatively associated with pollutant concentrations likely because of lower local emissions (e.g. a sparse roadway network) and stronger pollutant dispersion in hilly or mountainous areas. Since wind and elevation data are widely available in the U.S. (e.g. wind data from airport measurements, certain air monitoring station and other sources and elevation data from USGS), the exploration of such data in spatiotemporal modeling of air pollutant concentrations is recommended for epidemiological studies about acute or sub-chronic effects of air pollution. Further, since spatial autocorrelation contributed substantially to the estimation of spatial varying coefficients, spatial autocorrelation should not be neglected in similar modeling studies.

This study has several limitations. First, due to the unavailability of year-specific data, we used the 2003 roadway and 2005 AADT data to derive traffic-related covariates for 2006–2010. Although roadway and traffic data unlikely changed dramatically over the study region, these temporally-mismatched data may influence the performance of the models. Second, although we have examined the contribution of diesel truck emissions using Caltrans weigh-in-motion (WIM) data on freeways and highways, the fraction of diesel truck counts was not a statistically significant variable in any of the models. This is likely due to large uncertainties in the estimated truck count data since there were only five WIM stations in the study region (one in Orange County and four in Los Angeles). Third, the approach of splitting two continuous bi-week measurements to the four weekly estimates for the UCLA data may have introduced bias. The test using three different splitting methods (Supplemental Materials Table S5) showed limited influence of the interpolation methods on the prediction likely because of the small variation of concentration between two continuous weeks in our datasets. Ideally more than two time slices should be included whenever possible though. Fourth, the use of non-linear regression and residual spatial autocorrelation might cause an over-fitting problem. We tried to avoid overfitting by using cross validation, restricting the degree of freedom in the GAM and only incorporated non-linear covariates when necessary. Furthermore, the mgcv package in R used in our study controls the complexity of the splines by imposing the penalty on the parameters, lowering the over-fitting risk (Gill, 2001).

6. CONCLUSION

We used a hierarchical spatiotemporal model to combine routine time series with episodic measurements at a large number of locations to improve the estimation of spatiotemporal variability of NO₂ and NO_x concentrations in a dense urban area of southern California. Our model decomposed the spatiotemporal field of concentrations into the mean and residual. The mean was further decomposed into independent constant and varying temporal basis functions and local covariates were used to model spatially varying coefficients of the temporal basis functions with GAM plus cokriging. We found traffic, land-use and wind accounted for a large portion of the variance (beyond 35%) for the long-term average trend of concentrations. Spatial residuals accounted for a large portion of the variance of the temporal components (about 30% for NO₂ and 20% for NO_x). Leave-one-out cross validation produced an R² of 0.84 for NO₂ and 0.81 for NO_x when comparing the modeled weekly concentration with the observed trends at all routine monitoring stations. The approach has practical implication in studies of sub-chronic or short-term health effect that require high spatial and temporal resolutions of exposure.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

References

- Cameletti, M.; Ignaccolo, R. Comparing spatio-temporal hierarchical models for air quality data, Scientific Meeting of SIS. 45th Scientific meeting of Italian Statistical Society University of Padua; 2010.
- Carlaw CD, Beevers DS, Tate EJ. Modeling and assessing trends in traffic-related emissions using a generalized additive model. *Atmospheric Environment*. 2007; 41:5289–5299.
- Christakos G. A Bayesian maximum entropy view to the spatial estimation problem. *Mathematical Geology*. 1990; 22:763–777.
- David, P. *Linear Algebra: A Modern Introduction*. 2. Brooks/Cole; 2006.
- Finkelstadt, B.; Held, L.; Isham, V. *Statistical Methods for Spatio-Temporal Systems*. Chapman & Hall/CRC; New York; 2007.
- Ghosh JK, Wilhelm M, Su J, Goldberg D, Cockburn M, Jerrett M, Ritz B. Assessing the influence of traffic related pollution on term LBW based on land use regression models and air toxics measures. *American Journal of Epidemiology*. 2012; 175:1262–1274. [PubMed: 22586068]
- Gill, J. *Generalized Linear Models: A Unified Approach*. Sage; Thousand Oaks, CA; 2001.
- Handelman D. Optimal Power Transformations for Analysis of Sperm Concentration and Other Semen Variables. *Journal of Andrology*. 2002; 23
- Hart EJ, Yanosky DJ, Puett R, Ryan J, Dockery WD, Smith JT, Garshick E, Laden F. Spatial Modeling of PM₁₀ and NO₂ in the Continental United States, 1985–2000. *Environmental Health Perspectives*. 2009; 117:1690–1696. [PubMed: 20049118]
- Hoek G, Beelen R, Hoogh K, Vienneau D, Gulliver J, Fischer P, Briggs D. A review of land-use regression models to assess spatial variation of outdoor air pollution. *Atmospheric Environment*. 2008; 42:7561–7578.
- Iglewicz, B.; Hoaglin, CD. How to Detect and Handle Outliers. In: Mykytka, FE., editor. *The ASQ Basic References in Quality Control: Statistical Techniques*. American Society for Quality; Milwaukee; 1993.
- Jerrett M, Arain A, Kanaroglou P, Beckerman B, Potoglou D, Sahsuvaroglu T, Morrison J, Giovis C. A review and evaluation of intraurban air pollution exposure models. *Journal of Exposure Analysis and Environmental Epidemiology*. 2005; 15:185–204. [PubMed: 15292906]
- Jerrett M, Burnett R, Pope A, Krewski D, Thurston G, Christakos G, Hughes E, Ross Z, Hi SY, Thun M. *Spatiotemporal Analysis of Air Pollution and Mortality in California Based on the American Cancer Society Cohort: Final Report*. State of California Air Resources Board. 2012

- Li L, Wu J, Wilhelm 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. [PubMed: 23439926]
- Lindstrom J, Szpiro AA, Sampson DP, Sheppard L, Oron A, Richards M, Larson T. A flexible spatio-temporal model for air pollution: allowing for spatio-temporal covariates. *UW Biostatistics Working Paper Series*. 2011
- Mahiyuddin W, Sahani M, Aripn R, Latif M, Thach T, Wong C. Short-term effects of daily air pollution on mortality. *Atmospheric Environment*. 2013; 65:69–79.
- Ritz B, Wilhelm M. Ambient air pollution and adverse birth outcomes: methodologic issues in an emerging field. *Basic & Clinical Pharmacology & Toxicology*. 2008; 102:182–190. [PubMed: 18226073]
- Szpiro AA, Sampson DP, Sheppard L, Lumley T, Adar DS, Kaufman DJ. Predicting intra-urban variation in air pollution concentrations with complex spatio-temporal dependencies. *Environmetrics*. 2010; 21:606–631.
- Whitworth KW, Symanski E, Lai D, Coker LA. Kriged and modeled ambient air levels of benzene in an urban environment: an exposure assessment study. *Environmental Health*. 2011; 10:21. [PubMed: 21418645]
- Wood SN, Augustin NH. GAMs with integrated model selection using penalized regression splines and applications to environmental modelling. *Ecological Modelling*. 2002; 157:157–177.
- Wu J, Wilhelm M, Chung J, Ritz B. Comparing exposure assessment methods for traffic-related air pollution in an adverse pregnancy outcome study. *Environmental Research*. 2011; 111:685–692. [PubMed: 21453913]
- Yu H, Chen J, Christakos G, Jerrett M. BME estimation of residential exposure to ambient PM10 and ozone at multiple time scales. *Environmental Health Perspectives*. 2009; 117:537–544. [PubMed: 19440491]
- Zanobetti A, Schwartz J, Samoli E, Gryparis A, Touloumi G, Peacock J, Anderson HR, Tertre LA, Bobros J, Celko M, Goren A, Forsberg B, Michelozzi P, Rabczenko D, Hoyos PS, Wichmann EH, Katsouyanni K. The temporal pattern of respiratory and heart disease mortality in response to air pollution. *Environmental Health Perspective*. 2003; 111:1188–1193.

Highlights

- Hierarchical spatiotemporal model predicted weekly NO₂ and NO_x concentrations well
- Episodic measurements supplemented routine time-series measurements in modeling
- Land-use, meteorology, and roadway were important in concentration prediction

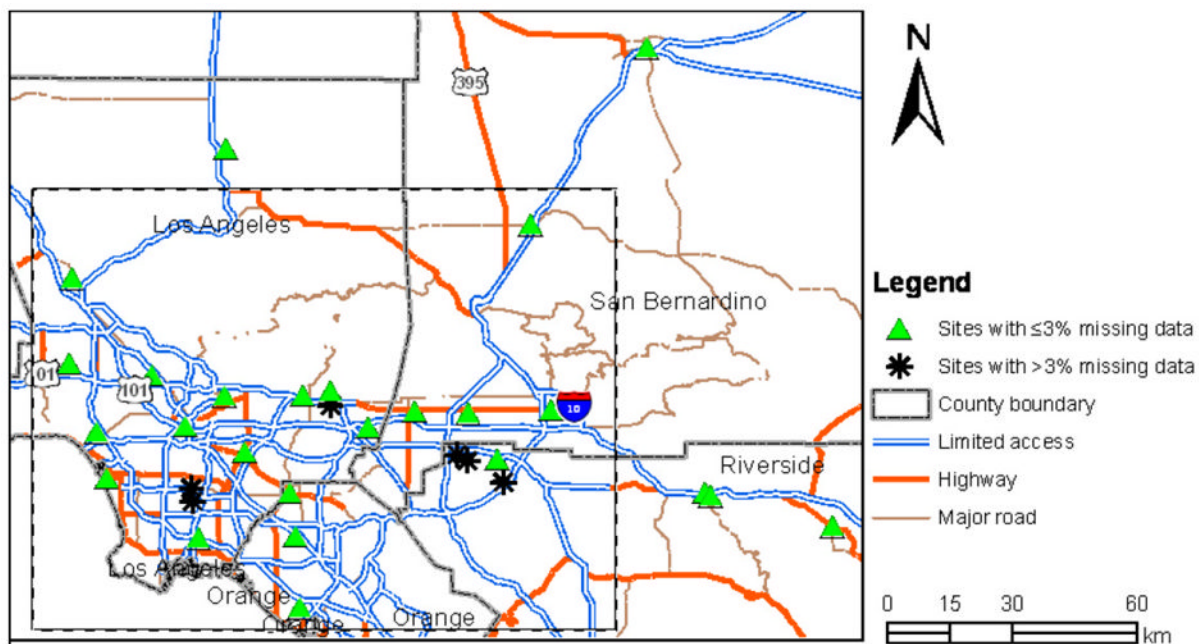
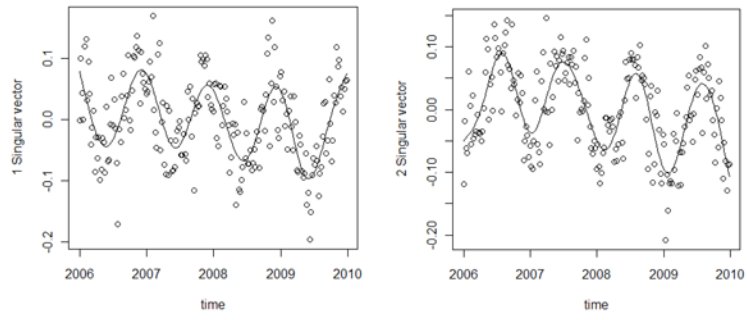
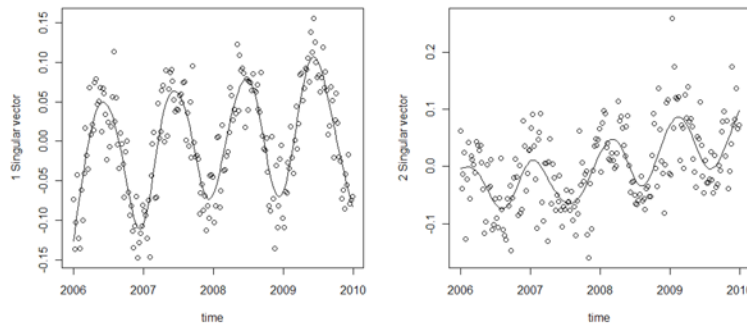


Figure 1. Study domain with 31 SCAQMD sites (25 sites with $\leq 3\%$ missing data used for generating temporal basis functions of the domain and 6 sites with $>3\%$ missing data; all SCAQMD sites used for spatial modeling with UCI and UCLA measurements)



a. First temporal basis function for NO_2 (β_1)

b. Second temporal basis function for NO_2 (β_2)

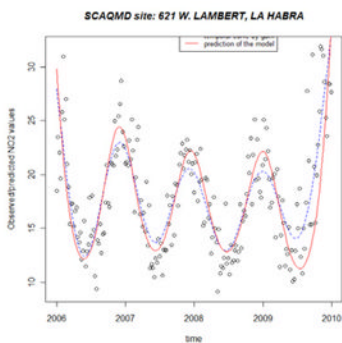


c. First temporal basis function for NO_x (β_1)

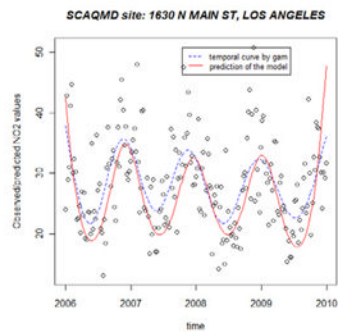
d. Second temporal basis function for NO_x (β_2)

c.

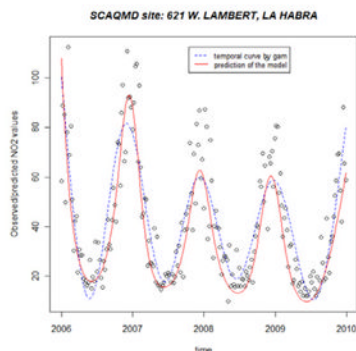
Figure 2. Plots of the first and second independent temporal basis functions for NO_2 and NO_x



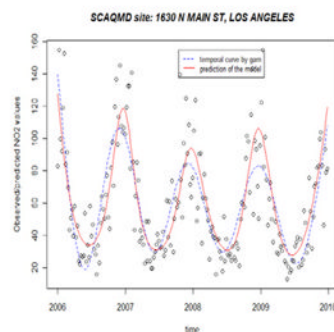
a. NO₂ at 621 W Lambert, LA Habra
(Orange county; r=0.93)



b. NO₂ at 1603 N Main ST, Los Angeles
(Los Angeles; r=0.94)



c. NO_x at 621 W Lambert, LA Habra
(Orange county; r=0.93)



b. NO_x at 1603 N Main ST, Los Angeles
(Los Angeles; r=0.92)

Figure 3. Influence of traffic, land-use and wind covariates for spatially varying coefficient, β_0 of the NO₂ and NO_x constant temporal basis functions (the variance explained in the parenthesis)

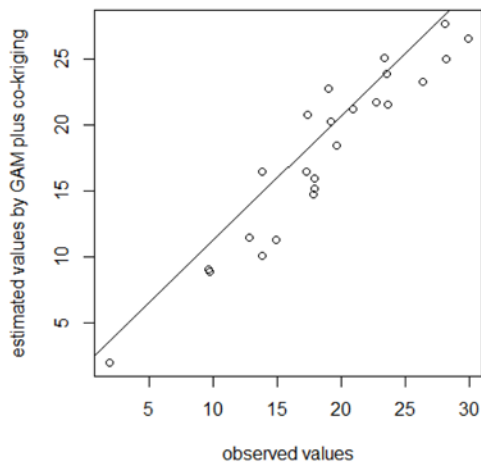
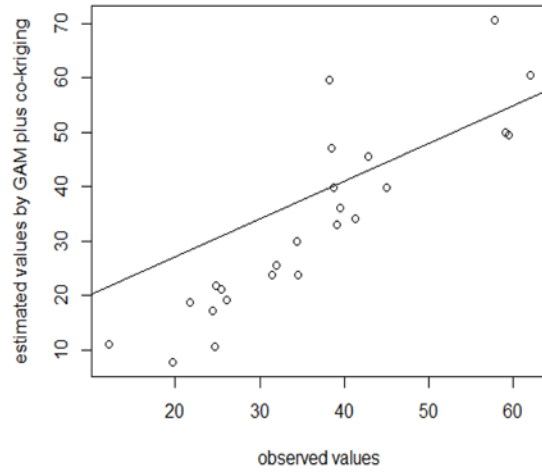
a. NO₂ ($R^2=0.84$)b. NO_x ($R^2=0.70$)

Figure 4. Good match between observed values vs. simulated values for NO₂ and NO_x (r: correlation)

Table 1

Statistics of fit parameters by the routine time series and sporadic samples

Statistics	NO ₂			NO _x		
	β_0	β_1	β_2	β_0	β_1	β_2
Min	0.59	0.02	-12.80	0.69	-16.09	-9.37
Max	3.46	20.96	9.93	4.32	-0.43	11.04
Mean	2.90	0.33	0.33	3.51	-9.24	0.86
Variance	0.13	8.08	10.16	0.20	6.90	14.11
Median	3.00	5.29	-0.16	3.62	-9.27	1.07

Table 2

Explained variance and influence of local covariates and spatial autocorrelation for estimation of parameters (p-value<0.1)

Statistics	NO ₂						NO _x					
	β_0		β_1		β_2		β_0		β_1		β_2	
	Coef	V.E.	Coef	V.E.	Coef	V.E.	Coef	V.E.	Coef	V.E.	Coef	V.E.
Total variance explained	88.9%		86.7%		86.7%		87.0%		87.4%		81.4%	
Fcc1mdis	9.6%										12.5%*	
Fcc2mdis												
Fcc3mdis												
Fcc4mdis												
Fccmdis	1.9%		-2.4e-4	0.1%	8.8%*		1.2%					
Weighted road length	18.9%			2.6%*			2.4%		21.8%			
Weight AADT											2.4%	
Long-term average temperature												19.9%*
Long-term average precipitation												
Long-term average wind speed	28.4%*				15.2%		4.9%					
Distance to the shorelines	8.5%*				9.4%		13.6%*		12.1%*			
Ratio of traffic land-use					6.8%		19.2%*					
Ratio of industry land-use	2.2%											
Ratio of commercial land-use	5.8%*				6.9%		15.9%*		-			
Ratio of residential land-use	2.5%											
Ratio of agricultural and open land-use					4.9%*		14.8%*		-			
Spatial autocorrelation of residuals (co-kriging)	11.1%*				34.8%*		15.5%*		25.4%*			
												20.7%*

* the variance explained is beyond 10% or the covariate is the only influential factor on one of β parameters (β_0 , β_1 and β_2).

Coef: linear regression coefficient given if a covariate was used in the model.

Table 3

Variogram Modeling of Spatial Residuals for β_0 , β_1 and β_2

Statistics	NO ₂			NO _x		
	β_0	β_1	β_2	β_0	β_1	β_2
Model	Stable	Stable	Stable	Stable	BesselK	Stable
Parameter	0.3	0.2	0.2	0.5	2.19	0.2
Range	4.9 km	3.8 km	3.8 km	5.5 km	4.3 km	3.3 km
Partial sill	0.013	0.87	2.22	0.012	2.73	2.96
						0.08

Table 4

Comparison of our models with the linear model at 25 SCAQMD sites

Models	NO ₂			NO _x		
	CV R ² for time series	R ² for long-term averages	RMSE	CV R ² for time series	R ² for long-term averages	RMSE
N samples	5225	25	25	5225	25	25
Our model	0.82	0.84	2.46	0.75	0.70	6.4
Linear regression model	0.53	0.63	4.61	0.48	0.51	8.2