Spatio-temporal Modeling of Environmental Data for Epidemiologic Health Effects Analyses

Paul D. Sampson
University of Washington

**Air Quality and Health:
a global issue with local challenges**

CCAR
Center for Clean Air Research

University of Bath · UNAM · CIMAT

Research Network for Statistical Methods for Atmospheric & Oceanic Sciences
STATMOS

8 Aug 2017 -- Mexico City

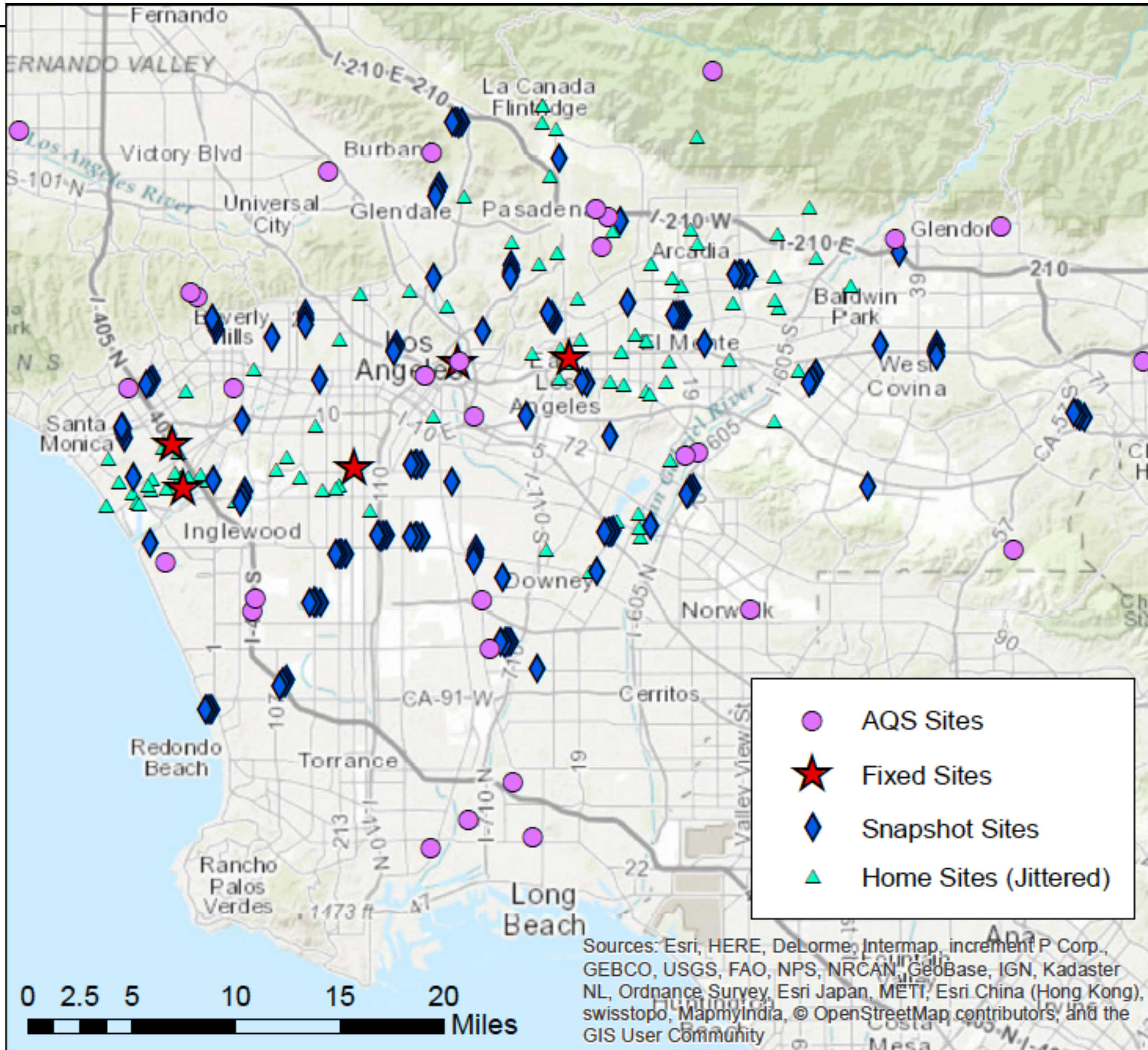
The MESA Air study

- The Multi-Ethnic Study of Atherosclerosis (MESA) is a large study of cardiovascular diseases.
- It follows more than 6 000 people from six communities.
  - Baltimore
  - Chicago
  - Los Angeles
  - Minneapolis – Saint Paul
  - New York
  - Winston–Salem

Basic problem: People don't live where we monitor.

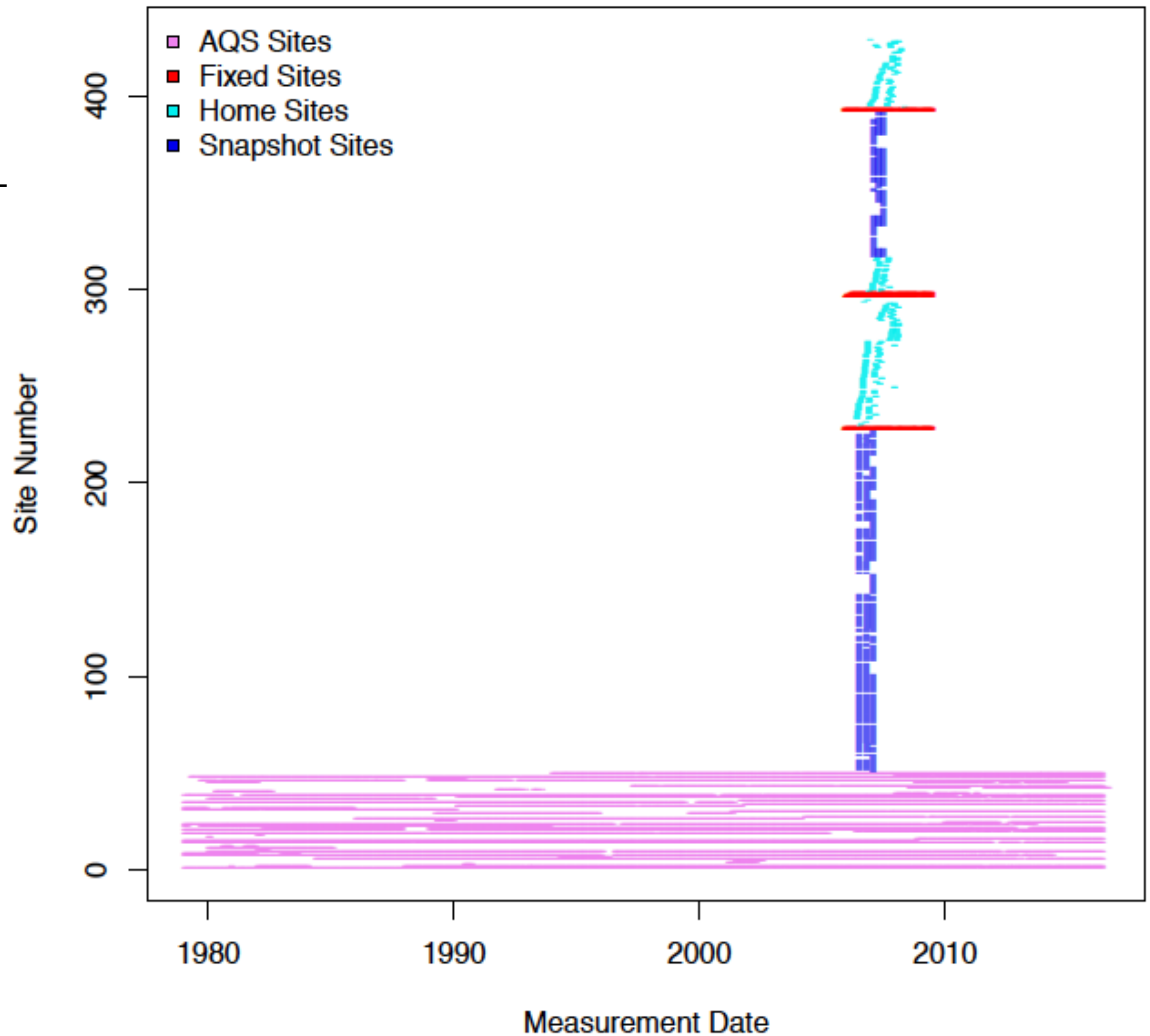
(Szpiro et al., 2010; Sampson et al., 2011; Lindström et al., 2013)

Multi-Ethnic Study of Atherosclerosis – Air Quality Monitoring Sites (MESA-Air)



Spatio-temporal modeling for MESA Air

- 2005 – 2009 …
- $PM_{2.5}$, NO_2, NO_X, BC, O_3
- 2-week measurements
  - **Fixed** Sites
    - 3 – 7 in each of 5 cities, 1 collocated with AQS site
  - **Home** Outdoor Sites
    - 1 – 3 measurements (each a 2-wk ave) from ~100 participant residence locations in each city
  - **Snapshot** Sites (NO_X and NO_2 only)
    - Clusters around roadways



Schematic of data. Each measurement is a point in space x time. **AQS**: temporally rich at multiple locations; **MESA Air** 2005-2009: 5 fixed sites, 177 "snapshot" sites, Home sites (4 monitors moving from one 2-week period to another)

Outline:

1. Background, spatio-temporal monitoring data and model structure (EOFs and land use regressions)
2. UW spatio-temporal model and `SpatioTemporal` R package
3. Model fitting procedures
4. NOx example application
5. Ozone application with CMAQ spatio-temporal covariate
6. Summary and development wish list.

1. Background: Methods in recent texts have not met our space-time modeling needs.

- Nhu D. Le, and James V. Zidek, *Statistical analysis of environmental space-time processes*, 2010
- Noel Cressie and Chris Wikle, *Statistics for Spatio-temporal data.* 2011
- Sudipto Banerjee, Brad Carlin and Alan Gelfand, *Hierarchical Modeling and Analysis for Spatial Data*, 2nd ed. 2015

What we needed when we began 17 years ago was:

- Flexible statistical modeling framework accommodating irregular space-time monitoring data from multiple networks
- Ability to incorporate effects of spatial and/or spatio-temporal covariates explaining heterogeneous spatio-temporal trend at fine spatial scales (10s of meters)
- A correspondingly convenient/flexible software system

Spatio-temporal models

A general spatio-temporal process:

$$Y(t,\mathbf{x}) = \mu(t,\mathbf{x}) + Z(t,\mathbf{x}) + \varepsilon(t,\mathbf{x})$$

where:

$\mu(t,\mathbf{x})$ is the spatio-temporal mean field or trend,

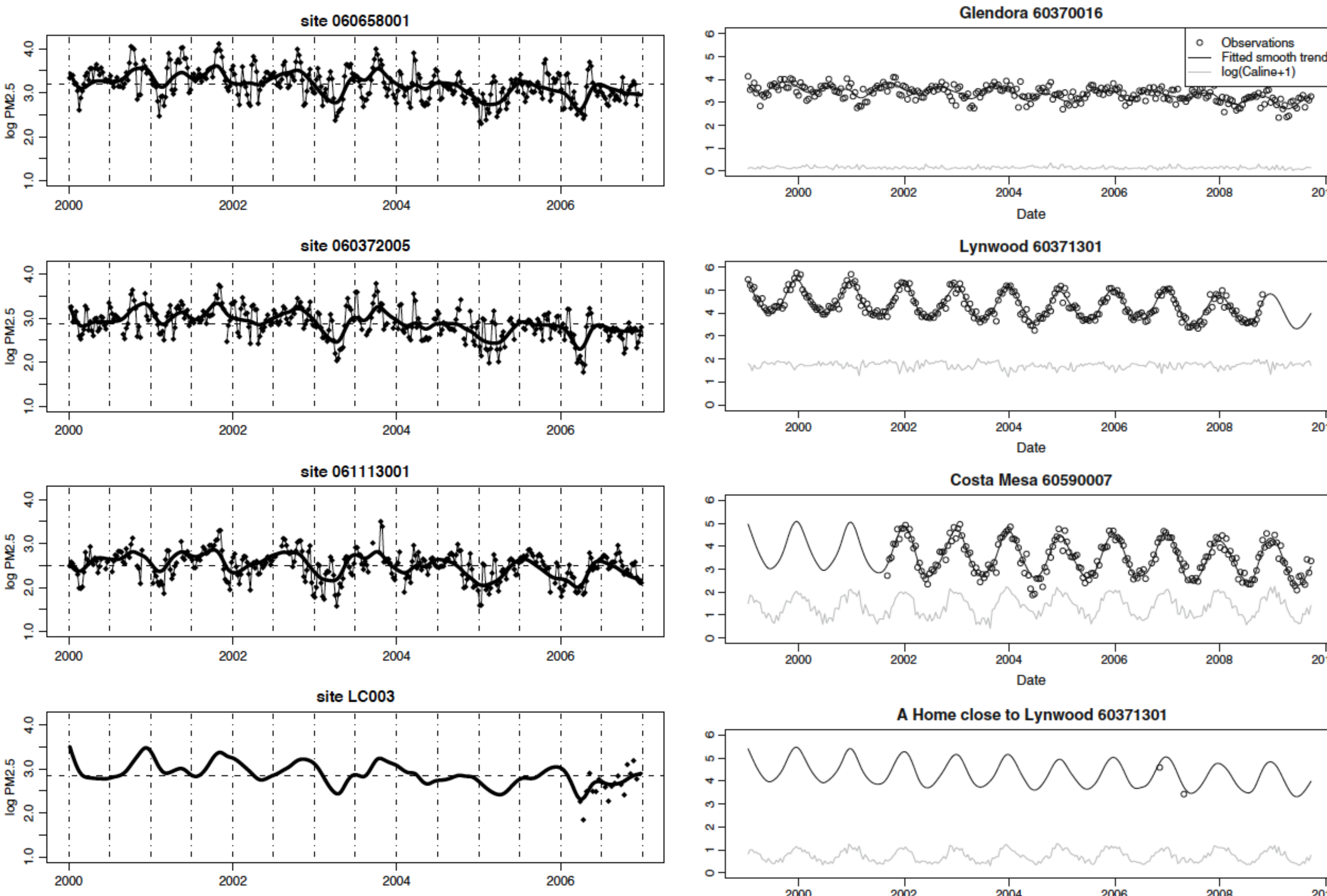
$Z(t,\mathbf{x})$ denotes a smooth spatio-temporal underlying process,

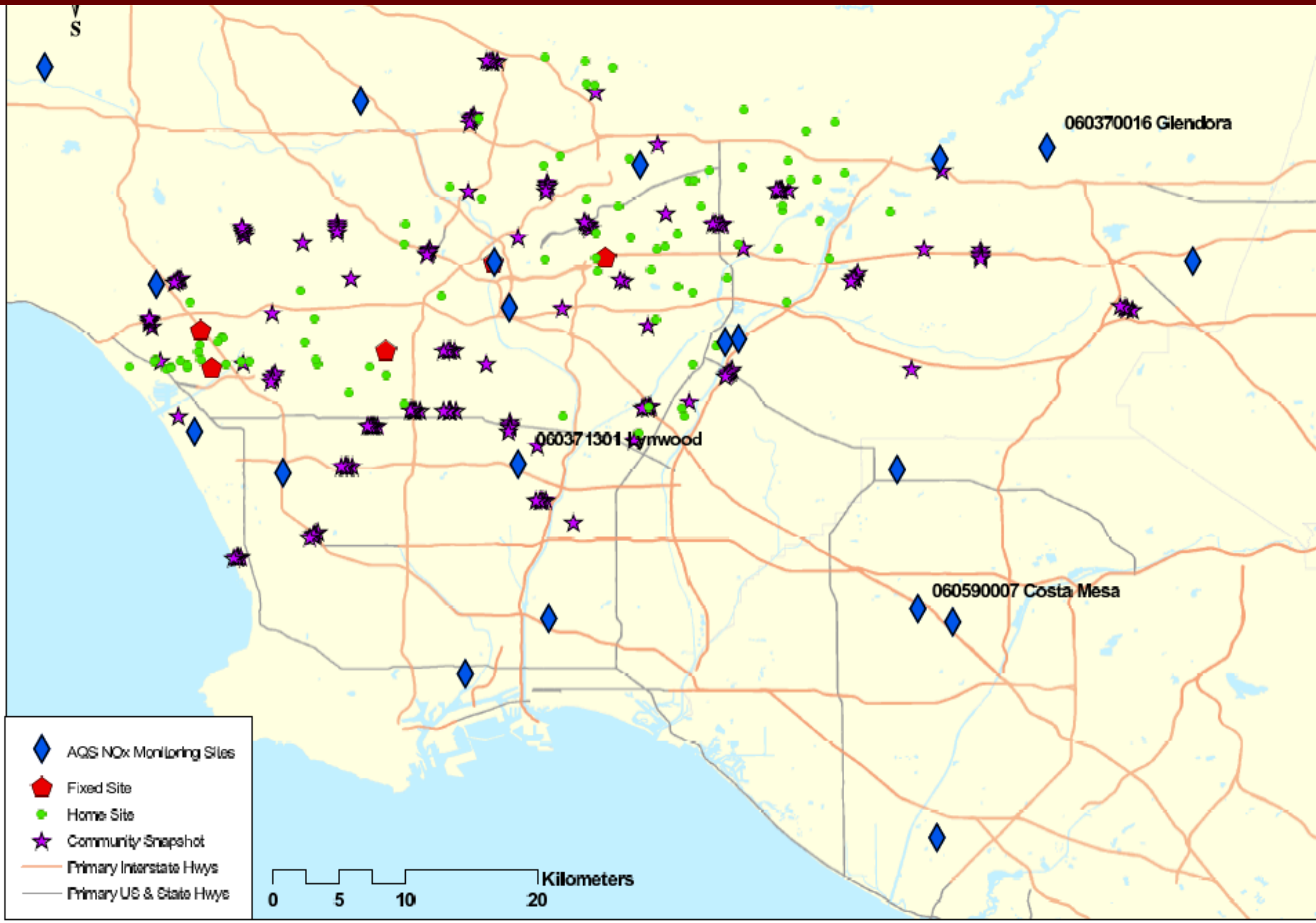
$\varepsilon(t,\mathbf{x})$ is a gaussian noise process, that represents the spatio-temporal measurement error (and small scale spatial variability).

Strategies for spatio-temporal modeling:

- *Spatial models varying (or evolving) in time*, including "dynamic" models
- *Temporal models* (at monitoring sites and arbitrary locations) *varying in space*.
- UW Spatio-temporal model and R package `SpatioTemporal` adopt this second strategy, using a notion of *temporal basis functions* or *empirical orthogonal functions*.

Examples of data series for log PM2.5 (left) and log NO2 (right)




S
060370016 Glendora
060371301 Lynwood
060590007 Costa Mesa
AQS NOx Monitoring Sites
Fixed Site
Home Site
Community Snapshot
Primary Interstate Hwys
Primary US & State Hwys
Kilometers
0
5
10
20

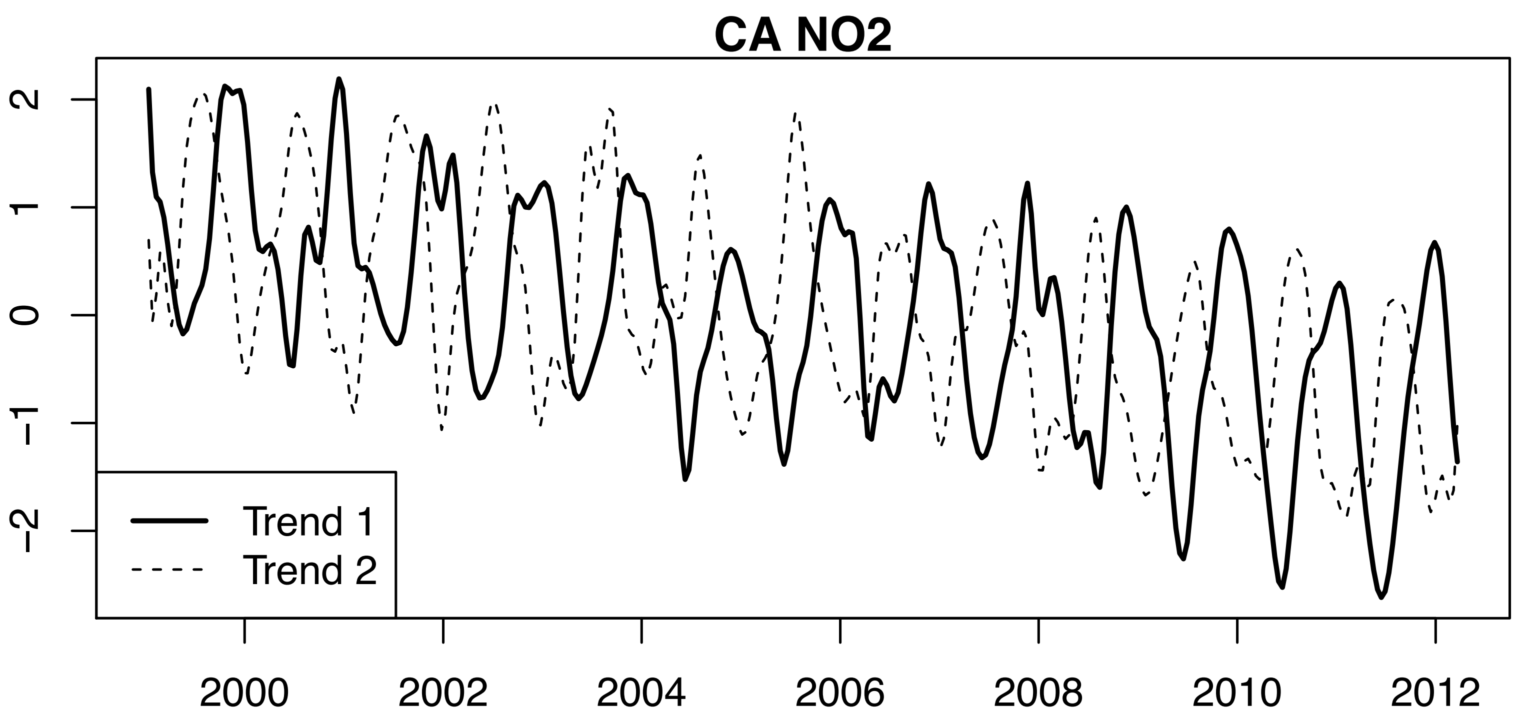
A spatio-temporal model based on exploratory EOF

- Cressie & Wikle's chapter 5 on Exploratory Methods for Spatio-Temporal Data includes a section on *Empirical Orthogonal Function* (EOF) Analysis.
- Our modeling strategy can be viewed as a hierarchical statistical model built on the foundation of an EOF analysis, with spatio-temporal variation explained by spatial (geographic or *land use*) and spatio-temporal covariates.

NO_2 Time Trends in CA

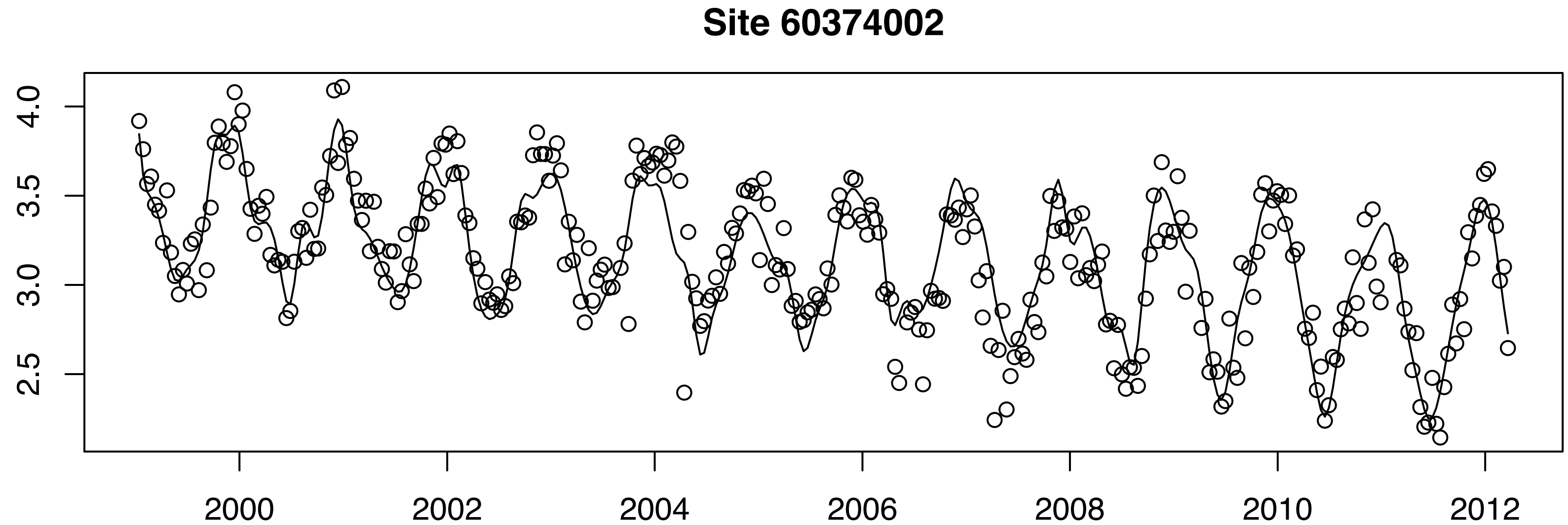
$$\beta_0(s) + \sum_{i=1}^{n} \beta_i(s) f_i(t) + \nu(s,t)$$

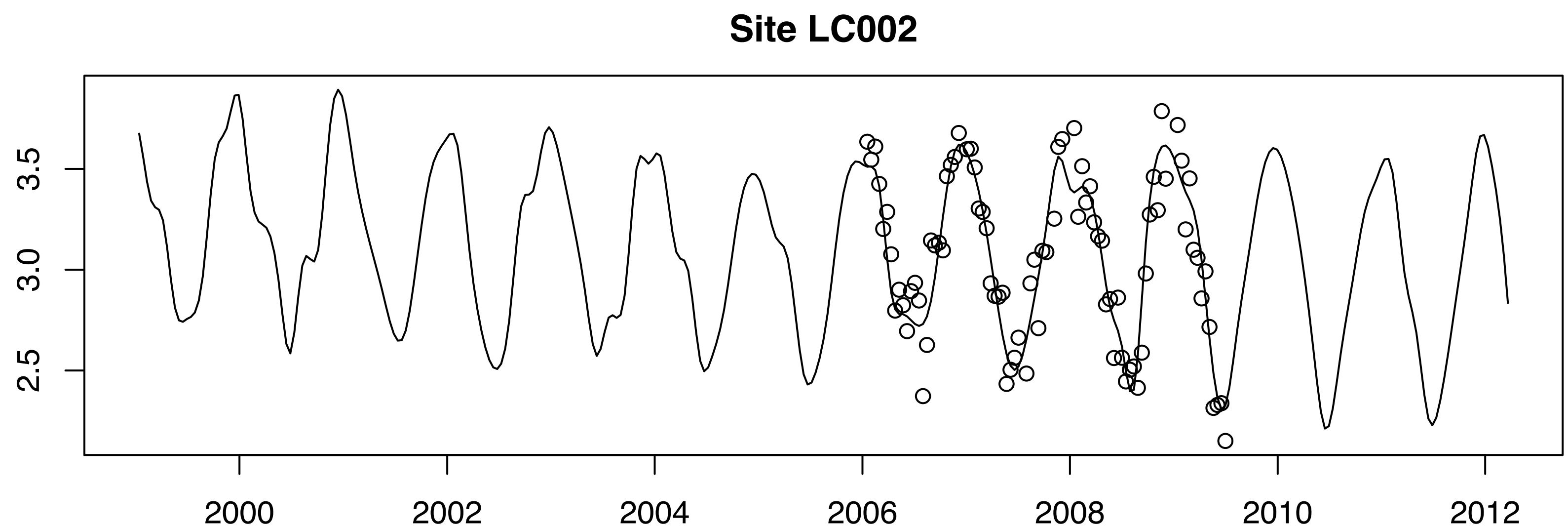
- Represent time trends as combinations of *smooth*, temporal basis functions derived from the right singular vectors of an SVD of the AQS space x time data matrix
- Because these derive from singular vectors, they represent ***temporal structure shared across monitoring sites***.



NO_2 Time Trends in CA

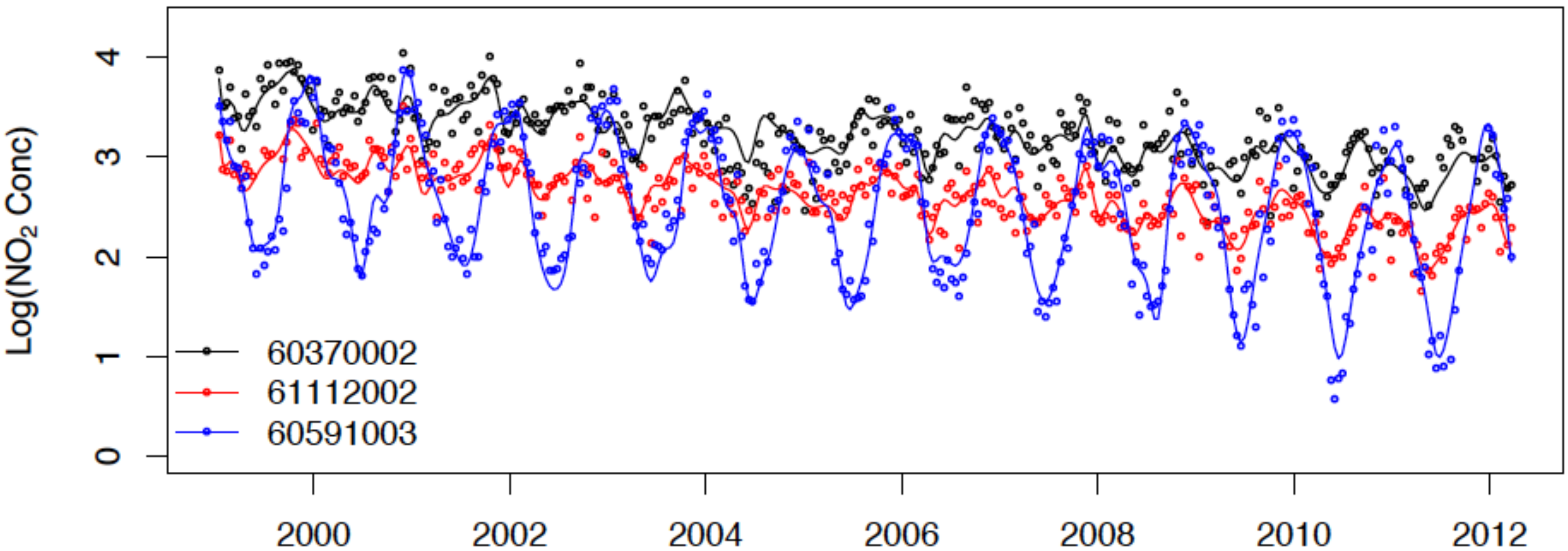
$$\beta_0(s) + \sum_{i=1}^{n} \beta_i(s) f_i(t) + \nu(s,t)$$





NO_2 Time Trends in CA

$$\beta_0(s) + \sum_{i=1}^{n} \beta_i(s) f_i(t) + \nu(s,t)$$



060370002: Azusa, Los Angeles county
061112002: Simi Valley, Ventura county
060591003: Costa Mesa, Orange county

Smooth temporal basis functions

- The $f_i(t)$ represent the shared temporal variability in the monitoring observations. These could be specified a priori as deterministic (e.g. sinusoids), or as done here, obtained as smoothed right singular vectors of the *n* x T data matrix on *n* sites. But monitoring data *always* have missing observations.
- Guttorp, Fuentes & Sampson, 2006, introduced an EM-like algorithm for imputing missing data and computing the SVD.
- Then use `smooth.spline` in R to smooth the temporal (left) singular vectors
- Cross-validation to determine how many singular vectors/temporal basis functions to retain.

(Tools for these computations are provided in the *SpatioTemporal* package.)

2. UW Spatio-temporal Statistical Model

$$C(s,t) = \beta_0(s) + \sum_{i=1}^{m} \beta_i(s) f_i(t) + \nu(s,t)$$

- $C(s, t)$ – (log) conc. at location *s* at time *t*
- $\beta_i(s)$ – spatial random fields
  - $\beta_i(s) \sim N\left(\mathbf{X}_i(s)\alpha_i, \Sigma(\tau_i, \sigma_i, \varphi_i)\right)$, $i = 0,\ldots,m$
  - $\mathbf{X}_i(s)$ – geographic ("land use") covariates
  - $\Sigma(\tau_i, \sigma_i, \varphi_i)$ – spatial covariance function
  - $\beta_0(s)$ – long-term means at location s
- $f_i(t)$ – temporal trend basis functions
- $\nu(s,t)$ – residual space-time field

2. UW Spatio-temporal Statistical Model

We model the logarithm of each 2-week average as

$$y(s,t) = \sum_{l=1}^{L} \gamma_l \mathcal{M}_l(s,t) + \sum_{i=1}^{m} \beta_i(s) f_i(t) + \nu(s,t).$$

$\mathcal{M}_l(s,t)$ Spatio-temporal covariates with coefficients γ_l.

$f_i(t)$ Smooth temporal trends with $f_1(t) \equiv 1$ and $f_2(t), \ldots, f_m(t)$ mean zero.

$\beta_i(s)$ Spatially varying coefficients for the temporal trends.

$\nu(s,t)$ Residuals, modelled as a mean zero Gaussian field that is independent in time but has spatial structure.

The smooth temporal trends, $f_i(t)$ are compute using a singular value decomposition of the data matrix, Y (see Fuentes et al., 2006, and the computer exercise).

2. UW Spatio-temporal Statistical Model

$$\beta_i(s) \in \mathrm{N}\left(X_i\alpha_i, \Sigma_{\beta_i}(\theta_B)\right)$$

X_i Design matrices, that includes geographical covariates (different for each *i*).

α_i Regression coefficients.

Σ_{β_i} Covariance matrix describing additional spatial dependence not captured by the geographical covariates.

θ_B Parameters of the covariance matrices.

$$\nu(s,t) \in \mathrm{N}\left(0, \Sigma_\nu(\theta_\nu)\right)$$

Σ_ν Block diagonal covariance matrix for the residuals.

θ_ν Parameters of the residual covariance matrix.

Remark:

- We "assume" that the residual space-time field is *correlated in space* but *independent in time*. We have generally not tried to fit effects of meteorological parameters, which drive correlated temporal patterns (over metropolitan spatial domains). A model fit that leaves temporally uncorrelated residuals has implicitly accounted for effects of meteorological factors either in the shared temporal structure of the trend basis functions (if they are not too smooth) or the spatially correlated residuals.

The "land use" regressions in the space-time model

$$\beta_i(s) \sim N\left(\mathbf{X}_i(s)\alpha_i, \Sigma(\tau_i, \sigma_i, \varphi_i)\right)$$

where $X_i(s)$ is a vector of covariates measured at location s. At the U.W. we have coded as many as 300 geographic, GIS-based covariates, including:

Distance to:	Buffer values:	Misc:
A1, A2, A3 road	A1, A2, A3 road length	Street canyon metrics (NY & IL)
Truck Route	Intersections	
Railroad	Truck Route Length	Absolute Elevation
Railyard	Emissions	Long-term CALINE value:
Airport	Residual Oil output (NY)	• Traffic volumes
Port	% Land Use	• Roadway locations
Coast	Impervious surface	• Diurnal traffic patterns
Oil Boiler (NY)	NDVI	• Wind speed, wind direction, mixing height
	Population	

Covariate Selection

$$\beta_0(s) + \sum_{i=1}^{n} \beta_i(s) f_i(t) + \nu(s,t)$$

- Variable selection or dimension reduction for the land use regressions of $\beta_i(s)$ on $X_i(s)$.
  - with a large number of highly correlated covariates (e.g. lengths of roads in buffers of 50m, 100m, 150m, …), we usually choose dimension reduction
- Partial Least Squares (PLS)
  - Conceptually similar to PCA, but components computed as linear combinations of original covariates to *maximize covariance* between Y and X.
  - Typically retain small number of combinations, 1, 2, 3.
  - PLS scores pre-computed from empirical estimates of $\beta_i(s)$.

The `SpatioTemporal` R package

Lindström et al, 2014. A flexible spatio-temporal model for air pollution with spatial and spatio-temporal covariates. *Environmental and Ecological Statistics*.

1. Extends the model above to include spatio-temporal covariates (we'll see this below)
2. Evaluates predictions of long term averages using a cross-validation strategy accounting for the complex monitoring design and for temporal effects
3. Uses "smart" methods (taking advantage of block structure of covariance residual covariance matrices) to make MLE computationally feasible.

The `SpatioTemporal` R package: Computation

Introducing

$$B = \begin{bmatrix} X_1 & 0 & 0 \\ 0 & X_2 & 0 \\ 0 & 0 & X_3 \end{bmatrix} \qquad \Sigma_B(\theta_B) = \begin{bmatrix} \Sigma_{\beta_1}(\theta_B) & 0 & 0 \\ 0 & \Sigma_{\beta_2}(\theta_B) & 0 \\ 0 & 0 & \Sigma_{\beta_3}(\theta_B) \end{bmatrix}$$

our model becomes

$$[Y|\theta_B, \theta_\nu, \alpha] \in \mathrm{N}\left(FX\alpha, \Sigma_\nu(\theta_\nu) + F\Sigma_B(\theta_B)F^\top\right),$$

and parameters can be estimated by maximising the log-likelihood

$$l(\theta_B, \theta_\nu, \alpha|Y).$$

The `SpatioTemporal` R package: Computation

- Matrix algebra can be used to "simplify" the likelihood (Harville, 1997; Petersen and Pedersen, 2012).
- As an example we study the determinant of the log-likelihood

$$\log\left|\Sigma_\nu + F\Sigma_B F^\top\right| = \log\left|\Sigma_\nu\right| + \log\left|\Sigma_B\right| + \log\left|\Sigma_B^{-1} + F^\top\Sigma_\nu^{-1}F\right|$$

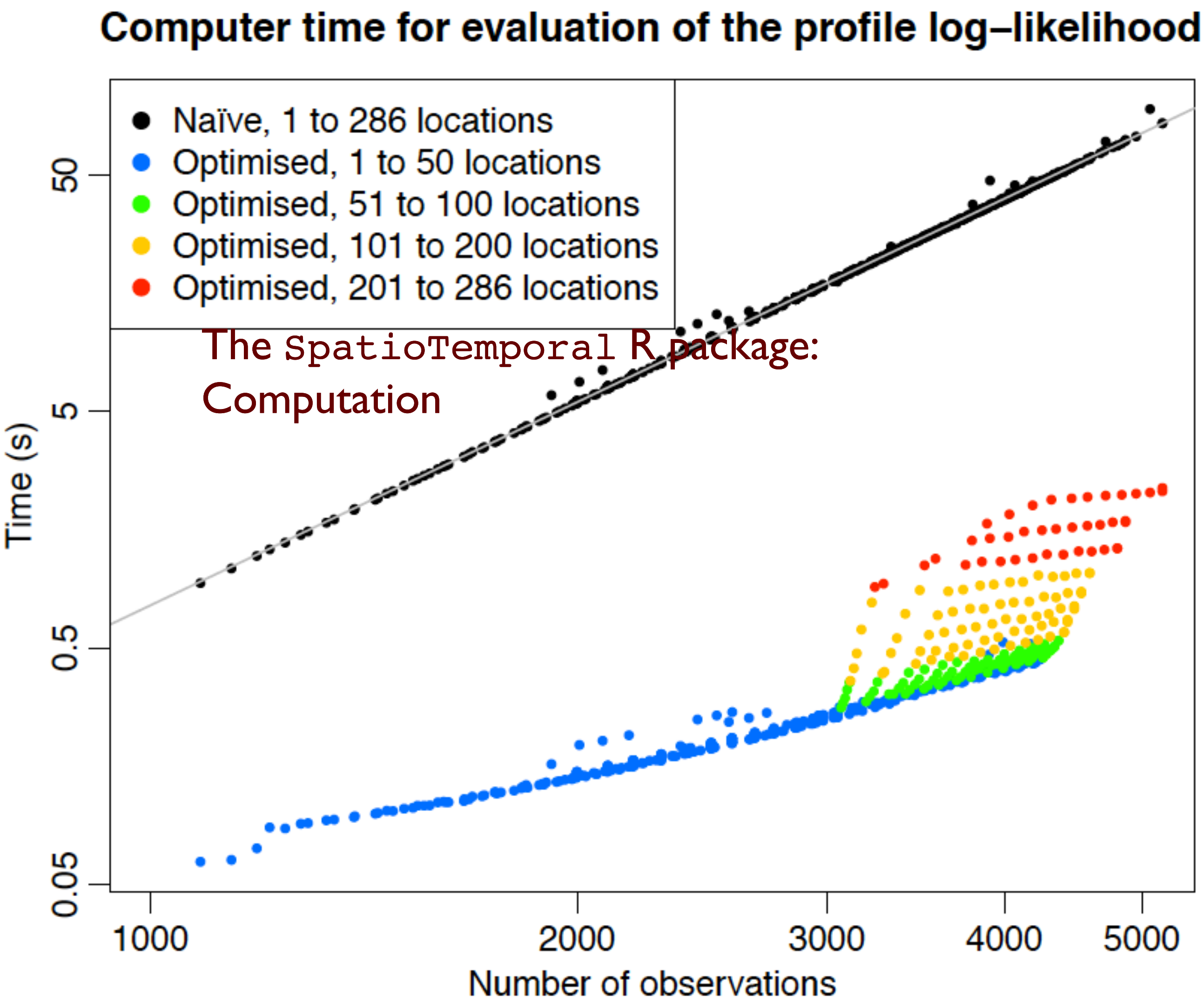
This may not seem simpler but:

1. $\Sigma_\nu + F\Sigma_B F^\top$ is dense $N \times N$-matrix, and computing the determinant requires $\mathcal{O}\left(N^3\right)$ operations.
2. Σ_ν and Σ_B are both block diagonal, with "small" blocks.
3. $\Sigma_B^{-1} + F^\top\Sigma_\nu^{-1}F$ is a dense $mn \times nm$-matrix. Computing the determinant requires $\mathcal{O}\left(m^3n^3\right)$ operations, with $mn \ll N$.

Where:

N	Total number of observations.
n	Total number of observed sites.
m	Number of temporal basis functions (incl. intercept).

The SpatioTemporal R package: Computation



The SpatioTemporal R package: Computation

SpatioTemporal R package summary:

- Maximum likelihood estimation with unbalanced spatio-temporal response data
- Incorporates spatio-temporal covariates $M(s,t)$, such as meteorology, AQ model predictions (CTM or CMAQ), or satellite predictions

$$C(s,t) = \beta_0(t) + \sum \beta_i(s) f_i(t) + \gamma M(s,t) + \nu(s,t)$$

- Predictions (w/ se's) from ML-based "universal kriging"

$$\hat{C}(s_0,t) = \hat{\beta}_0(s_0) + \sum_{i=1}^{m} \hat{\beta}_i(s_0) f_i(t) + \gamma M(s_0,t) + \hat{\nu}(s_0,t)$$

where $\hat{\beta}_i(s_0)$ derive from ML-based kriging of these spatial fields and $\hat{\nu}(s_0,t)$ similarly.

- The current program permits only the coefficients $\beta_i(s)$, multiplying the temporal basis functions, to vary spatially. There is an ad hoc approach to allowing the coefficient γ_l to vary spatially by introducing interactions of $\mathcal{M}_l(s,t)$ with a set of basis functions for spatial splines, but more is possible.

3. Model fitting procedures: Model Selection

- Vary model parameters
  - Spatial covariance structure for the $\beta_i(s)$
    - Exponential variogram (smoothing) or Independent (No Smoothing)
  - Number of Time Trends: 1 or 2
  - Number of PLS components: 2 or 3
- Best model for each city and pollutant chosen by cross-validation

Evaluation of model predictions

- Standard cross-validation: repeatedly
  1. Split data into training and test sets
  2. Estimate parameters using training data
  3. Predict at space-time locations of test data

  Compute MSE and R^2 of predicted vs. observed data
  - $R_{cv}^2 = \max\left(0, 1 - MSE_{pred} / MSE_{obs}\right)$
  - measures fit to 1-1 line, not simply correlation

For the case of the monitoring data in MESA Air

- AQS and Fixed sites
  - Leave-one-out CV
- Home sites
  - 10-fold CV
- Snapshot Sites
  - 10-fold CV by cluster

Temporally adjusted R^2

- Challenging to separate the spatial and temporal contributions to R^2_{cv} for cross-validation of temporally sparse datasets like MESA Air home sites.
- Lindström et al. (2013) introduced three temporally-adjusted adaptations of R^2_{cv} using data from neighboring AQS and fixed sites for the reference MSE instead of MSE_{obs} in order to focus on *spatial prediction accuracy*.

4. NOx example

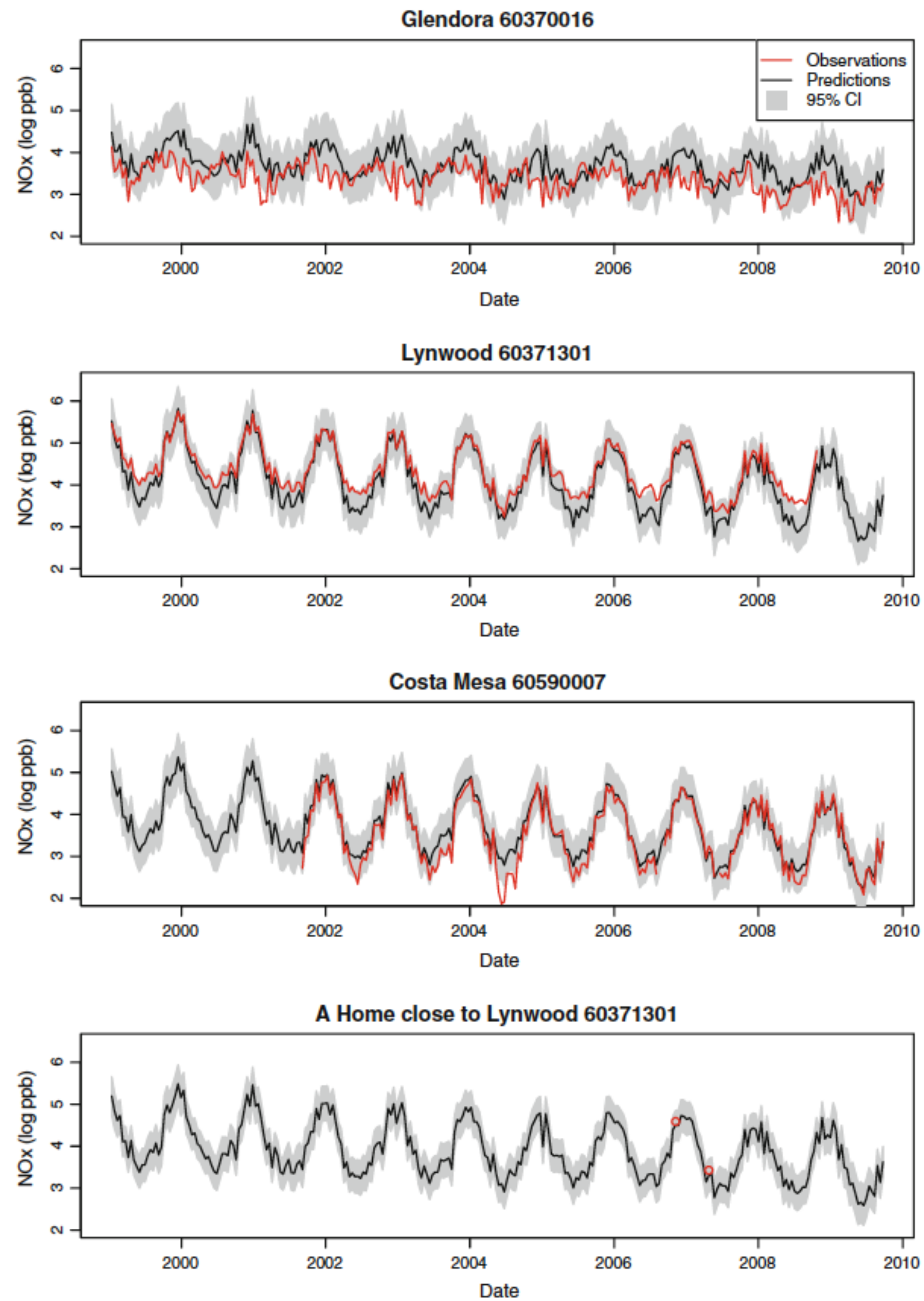
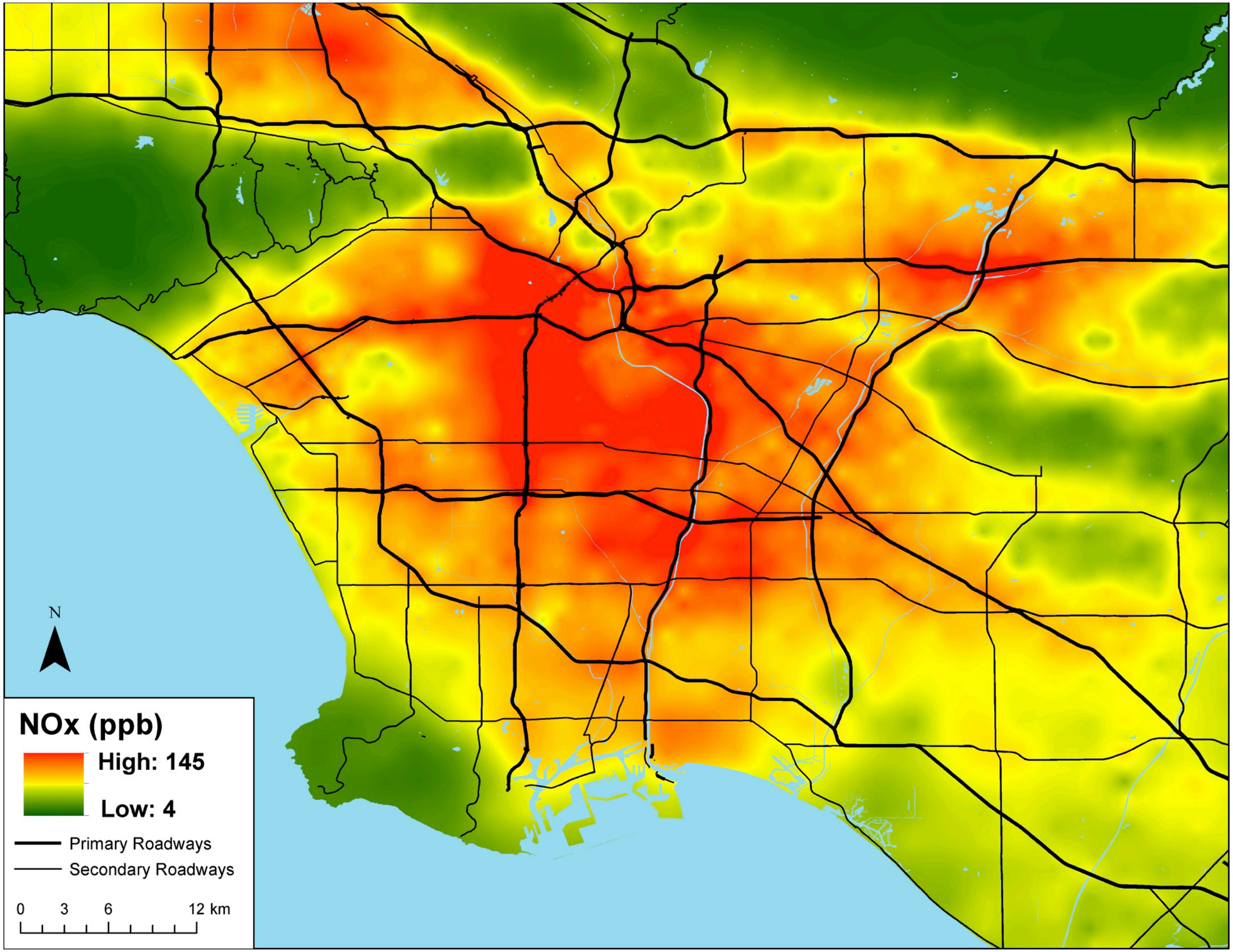


Fig. 6 Example of out-of-sample predictions of the log-transformed 2-week average NO_x concentrations at three AQS monitors and one home site in the Los Angeles area. Observations, predictions, and 95 %

NO_X Predictions: 2000 Average



NO_X Predictions: 2000 Average

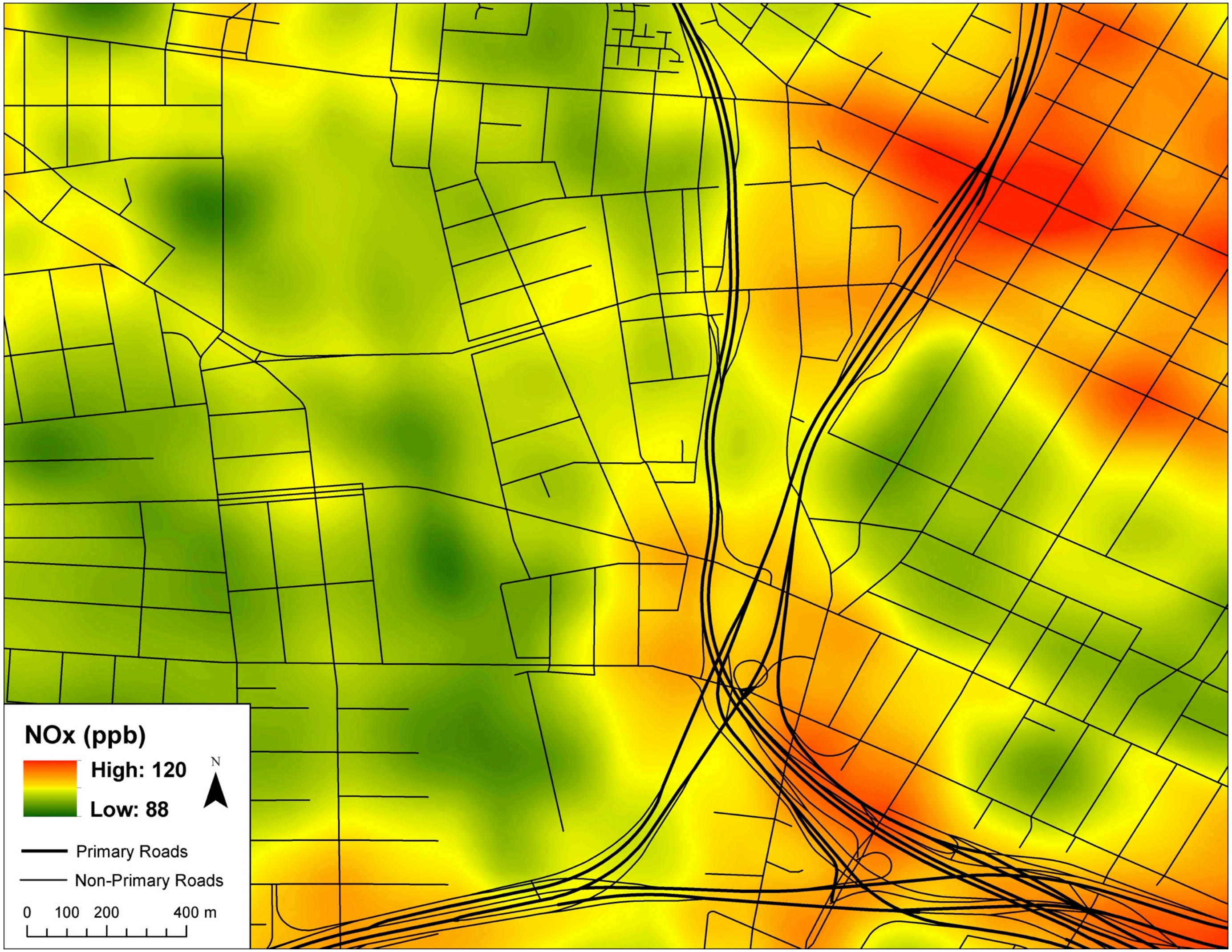


Table 4 Cross-validation results for the models with all GIS covariates, without and with Caline

	With road covariates					
	No Caline			Caline		
	RMSE	R^2	Cov.	RMSE	R^2	Cov.
AQS and MESA fixed						
2-Week	17.90	0.80	0.91	18.12	0.79	0.90
Long-term avg.	11.97	0.58		12.26	0.56	
Snapshot						
2006-07-05	7.94	0.52	0.93	7.62	0.56	0.95
2006-10-25	13.32	0.68	0.97	13.32	0.68	0.95
2007-01-31	15.69	0.66	0.99	15.77	0.66	0.98
Home sites	9.34	0.89	0.97	9.06	0.90	0.95
Average		0.67			0.69	
Closest		0.74			0.76	
Smooth		0.74			0.76	

The table gives RMSE, R^2, and coverage for 95 % predictions intervals for the out-of-sample predictions. For the Home sites the three adjusted R^2, showing improvement over simple temporal models, are also provided. All values are computed on the back transformed scale (ppb NO_x)

5. Gridded spatio-temporal covariates

- CMAQ *chemical transport model:* gridded air quality model predictions as a spatio-temporal covariate:
  - Must *map CMAQ grid predictions to point observations* and get clever to permit coefficient γ of CMAQ predictions $M(s,t)$ to vary spatially, representing *sub-grid scale spatial variability*: incorporate a LUR model for the coefficient γ.
- Satellite based aerosol optical depth (AOD) as a spatio-temporal covariate. Irregular gridded data full of missing data due to satellite path and meteorology (clouds, snow).
  - Practical strategies to fill in missing data and *calibrate* the coefficient of AOD to vary spatially due to influences of meteorology and other factors
- Models for daily time scale, with or without incorporating meteorology.

Combining Land Use Regression and Chemical Transport Modeling in a Spatio-temporal Geostatistical Model for Ozone and $PM_{2.5}$ in Los Angeles, California

Meng Wang[1], Paul D. Sampson[2], Michael Kleeman[3], Joshua P. Keller[4], Casey Olives[1], Adam A. Szpiro[4], Sverre Vedal[1], Joel D. Kaufman[1]

- CTM uses deterministic equations with data on emissions, meteorological conditions and topographical data to dynamically simulate the physico-chemical processes of transport and atmospheric chemistry to simulate air pollution concentrations.
- Relatively coarse spatial resolution of a CTM (≥ 4 km) cannot represent concentrations at very local scales (i.e. meters) for exposure assessment.

Details

- UC-Davis-California Institute of Technology (UCD-CIT) air quality model to simulate O_3 (8-hour averaged) and daily $PM_{2.5}$ concentrations over the South Coast Air Basin (SoCAB) from 2000 to 2008 (Figure below).
- Inverse distance weighing (IDW) mapping of grid cell O_3 and $PM_{2.5}$ predictions to the monitoring sites.
- Two-stage calibration of daily CTM predictions to MESA Air 2-week average observations:
  1. Regress daily average monitoring observations on CTM predictions at each of the AQS and fixed sites, allowing slope and intercept to vary from site to site.
  2. Spatial variation in these calibration slopes and intercepts explained by land use regressions using Partial Least Squares (PLS).

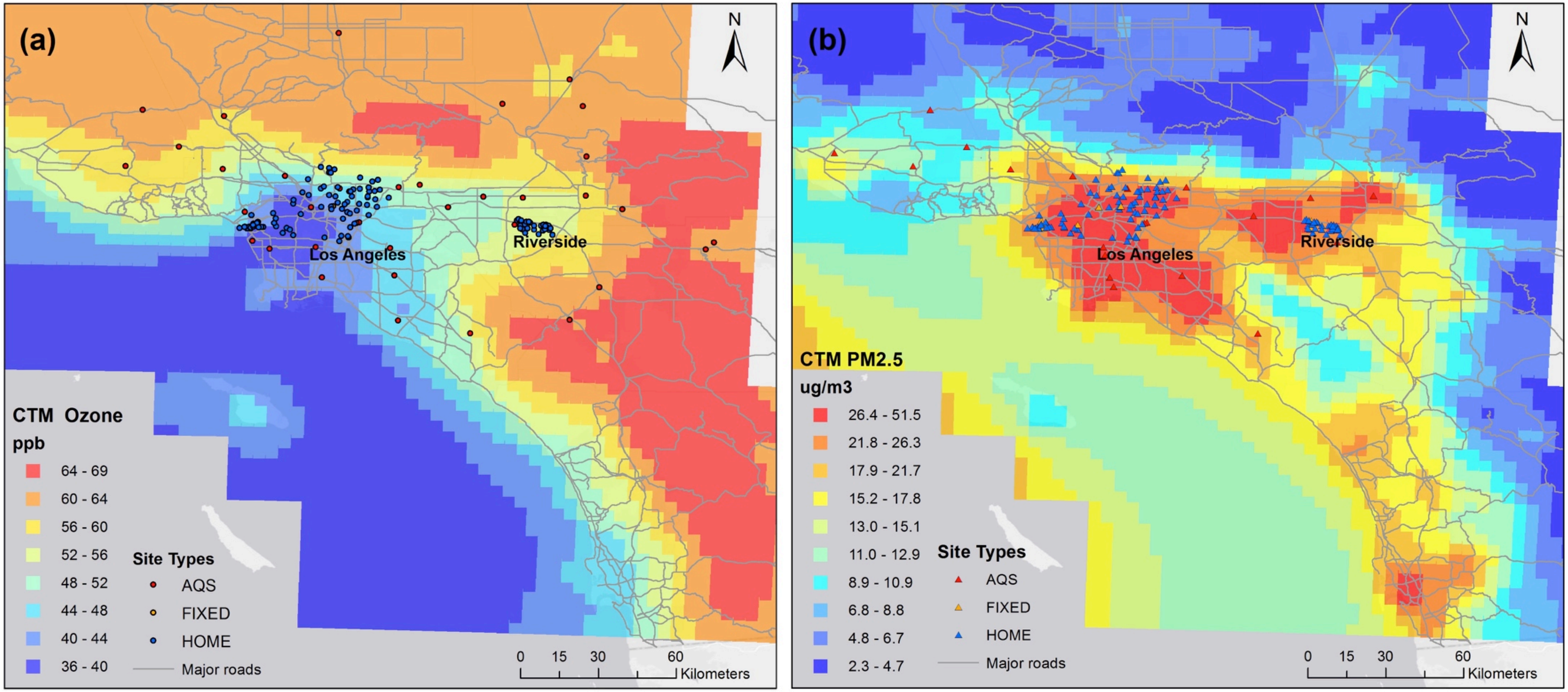


Figure 1. Long-term average of (a) O3 (daily 8 h maximum) and (b) PM2.5 (daily 24 h average) concentrations estimated by CTM with 4x4 km resolution and distribution of monitoring sites.

Tables

Table 1 Spatio-temporal model specifications for O_3 and PM2.5

	No. of AQS Sites	No. of Fixed Sites	No. of Home Sites	No. of Time Trends	[1]No. of PLS Scores	[2]Spatial Smoothing Long Term (β_0)	Time Trend (β_i)
O3	37	2	117	2	(3,1,1)	Yes	No
PM2.5	22	7	113	2	(2,1,2)	Yes	No

1. Number of PLS scores for (long-term mean, 1st time trend, 2nd time trend);

2. Yes, exponential covariance structure. No, independent covariance structure

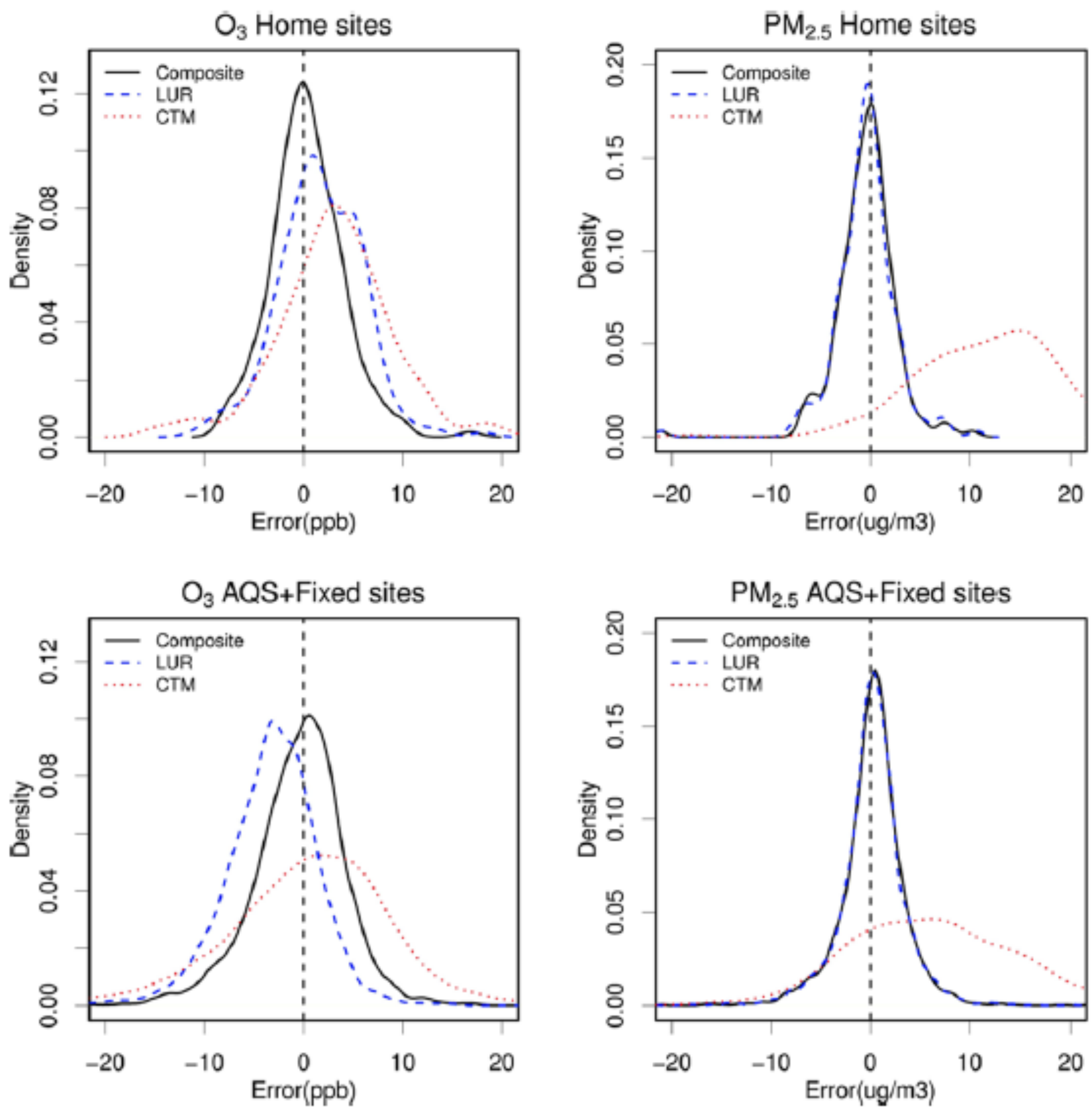


Figure 2. Distribution of prediction errors (predicted value minus observed value) at the home and routine monitoring sites for O_3 and $PM_{2.5}$ using the three modeling approaches.

Table 2 Model performances at the home and AQS+Fixed sites.

R2 (RMSE)	O3			PM2.5		
	CTM	ST-LUR	ST-LUR +CTM	CTM	ST-LUR	ST-LLUR +CTM
H: Overall[1]	0.60 (6.55)	0.78 (4.56)	0.84 (3.62)	0.32 (13.73)	0.78 (3.20)	0.80 (3.10)
A: Overall[2]	0.56 (8.83)	0.86 (5.78)	0.87 (4.64)	0.36 (10.15)	0.80 (3.42)	0.81 (3.37)
A: Spatial[3]	0.42 (7.35)	0.75 (4.90)	0.77 (3.56)	0.59 (7.11)	0.89 (1.32)	0.93 (1.18)
A: Temporal[4]	0.67 (6.78)	0.89 (4.81)	0.91 (3.81)	0.28 (9.58)	0.82 (3.26)	0.82 (3.20)

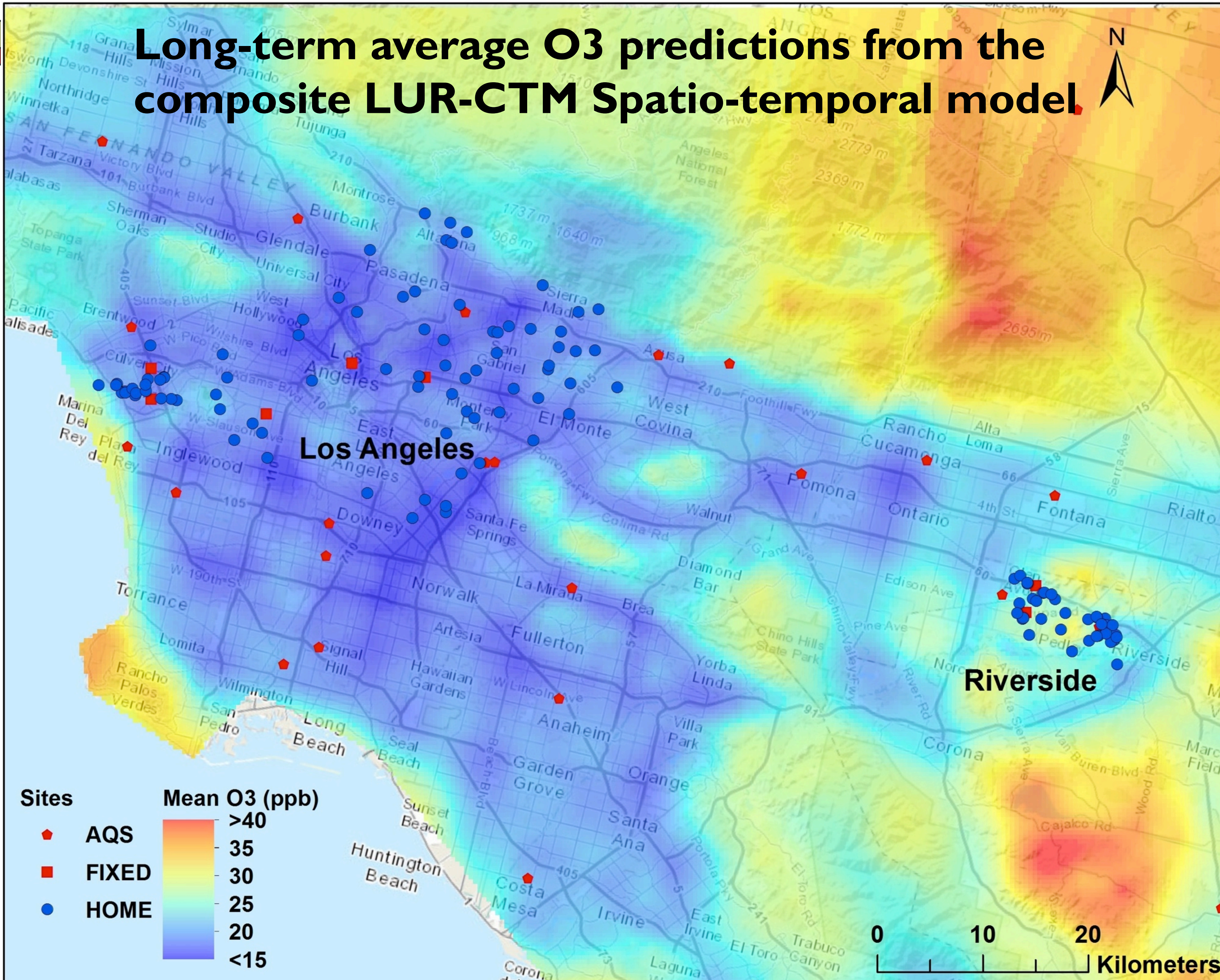
1. H: Home sites. Ten-fold CV (Note: CTM predictions are not cross-validated);

2. A: AQS and fixed sites. LOOCV (Note: O3 predictions are cross-validated in the calibration of the PLS regression; PM2.5 CTM predictions are not cross-validated);

3. Median RMSE and R^2 based on annual averages at each AQS and fixed site across years, which reflects spatial prediction ability of the models;

4. Median RMSE and R^2 between predictions and observations at two-week time points across the entire study period for individual sites.

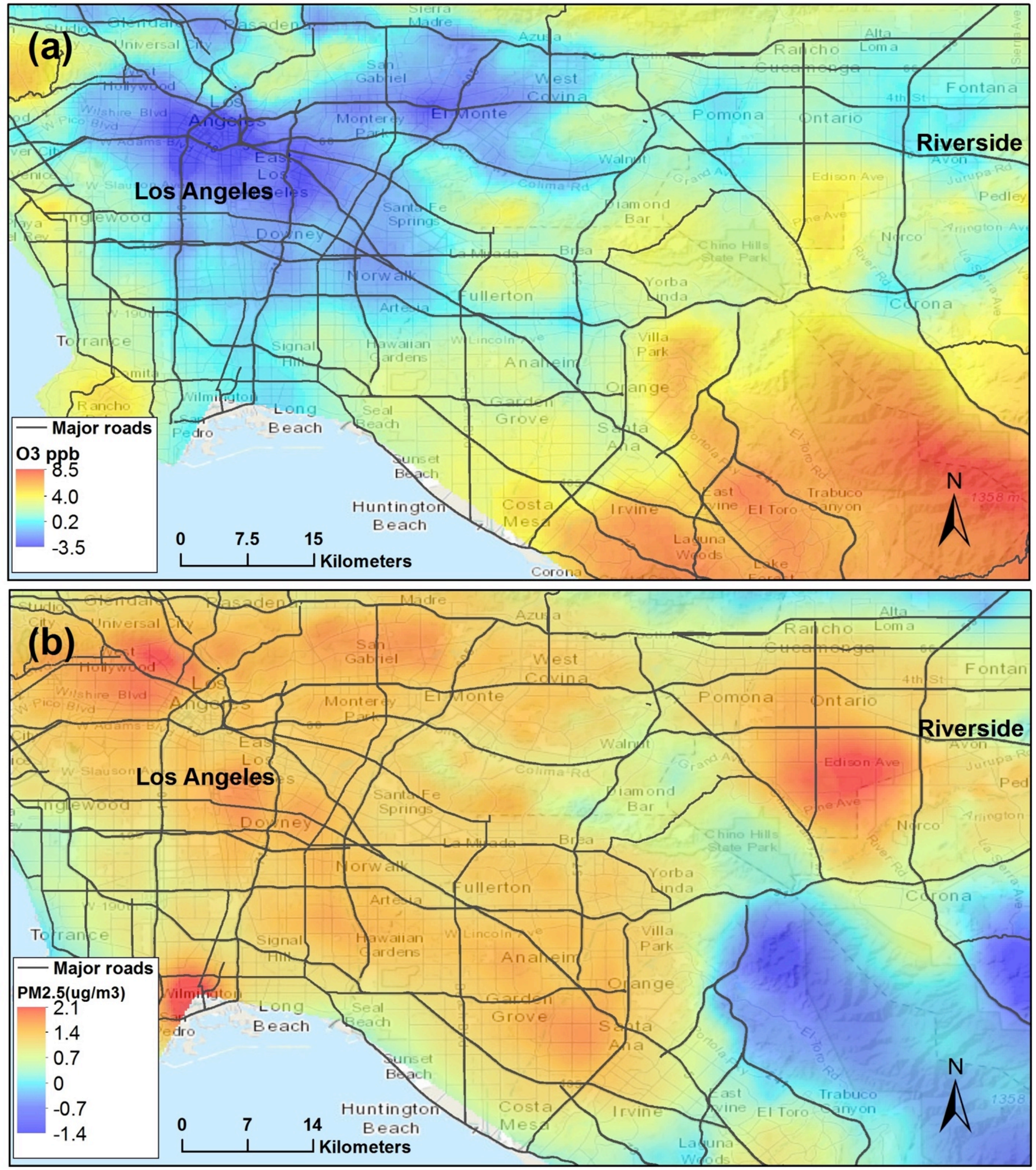
Long-term average O3 predictions from the composite LUR-CTM Spatio-temporal model



Differences between composite LUR-CTM ST model and LUR ST model alone.

Levels of O3 estimates were substantially higher in the rural areas and lower in urban areas by hybrid model than those by LUR model alone.

Levels of PM2.5 estimates were slightly higher in a large region across the Los Angeles area, except the mountain area.



6. Summary and wish list:

- Spatiotemporal model summary
  - Incorporates cohort-specific monitoring data
  - Allows for unbalanced monitoring design
  - Provides predictions at flexible time scales and with fine-scale spatial resolution
  - Can use both spatial and spatio-temporal covariates, including deterministic CTM model or remote sensing (satellite) predictions.

6. Summary and wish list:

- Topics for future development
  - Spatially varying coefficients of spatio-temporal covariates, notably for satellite and CTM predictions as covariates
  - Investigating strategies where there are too few sites (and perhaps too few temporal observations) for confidence in the SVD approach to temporal basis functions.
  - Examining contribution of spatiotemporal meteorological data in addition to temporal basis functions (which can pick up most of the temporal variation attributable to meteorology).

6. Summary and wish list (cont.):

- New, more flexible strategies for covariate data, including
  - imputation of missing covariate data (critical for satellite data); using CMAQ for imputation of missing satellite data
  - more flexible strategies for covariate selection, such as penalized methods (alternative to PLS which requires multiple fixed sites)
- Incorporate nonstationary spatial covariance

7. Some References:

2010 Szpiro AA, Sampson PD, Sheppard L, Lumley T, Adar SD, Kaufman JD (2010). Predicting intra-urban variation in air pollution concentrations with complex spatio-temporal dependencies. *Environmetrics*, 21(6), 606-631.

2011 Sampson PD, Szpiro AA, Sheppard L, Lindstrom J, and Kaufman JD. (2011) Pragmatic estimation of a spatio-temporal air quality model with irregular monitoring data. *Atmospheric Environment*, 45, 6593-6606.

2013 Johan Lindstrom, Adam Szpiro, Paul D. Sampson, Silas Bergen and Assaf P. Oron (2013). SpatioTemporal: Spatio-Temporal Model Estimation. R package version 1.1.7. https://CRAN.R-project.org/package=SpatioTemporal

2014 Lindström, J. Szpiro, A.A., Sampson, P.D., Oron, A.P., Richards, M., Larson, T.V., Sheppard, L. (2014). A flexible spatio-temporal model for air pollution allowing for spatio-temporal covariates. *Environmental and Ecological Statistics*, 21, 411-433.

2015 Keller JP, Olives Kim SY, Sheppard L, Sampson PD, Szpiro AA, Oron AP, Lindstrom J, Vedal S, Kaufman JD. (2015). A Uniform Spatiotemporal Modeling Approach for Prediction of Multiple Air Pollutants in the Multi-Ethnic Study of Atherosclerosis and Air Pollution. *Environmental Health Perspectives*.

2015 Wang, M., Keller, J.P, Adar, S.D., Kim, S-Y., Larson, T.V., Olives, C., Sampson, P.D., Sheppard, L., Szpiro, A.A., Vedal, S., Kaufman, J.D. (2015). Development of Spatiotemporal Models for Ozone in Six Metropolitan regions of the United States: The MESA Air Study. *Atmospheric Environment*, 123, 79-87. 2011.

2016 Wang, M., Sampson, P.D., Hu, J., Kleeman, M., Keller, J.P., Olives, C., Szpiro, A.A., Vedal, S., Kaufman, J.D. (2016) Combining Land-Use Regression and Chemical Transport Modeling in a Spatio-temporal Geostatistical Model for Ozone and PM2.5. *Environmental Science & Technology*, 50(10), 5111-5118.