Particulate Matter Air Quality from Space – Advanced Statistical Modeling

Yang Liu, PhD

October 15, 2015

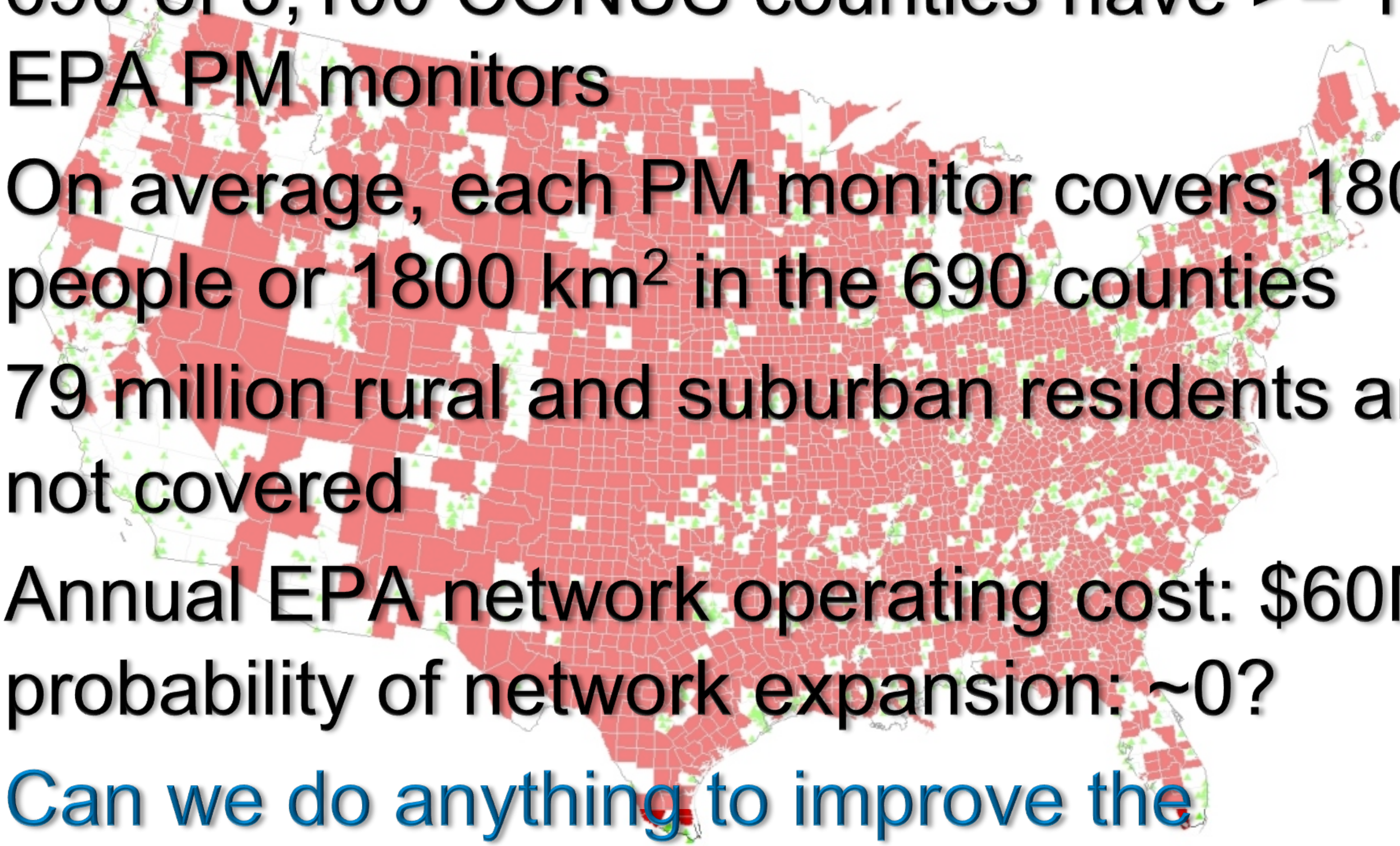
ROLLINS SCHOOL OF PUBLIC HEALTH
EMORY

ARSET

Applied Remote SEnsing Training

A project of NASA Applied Sciences

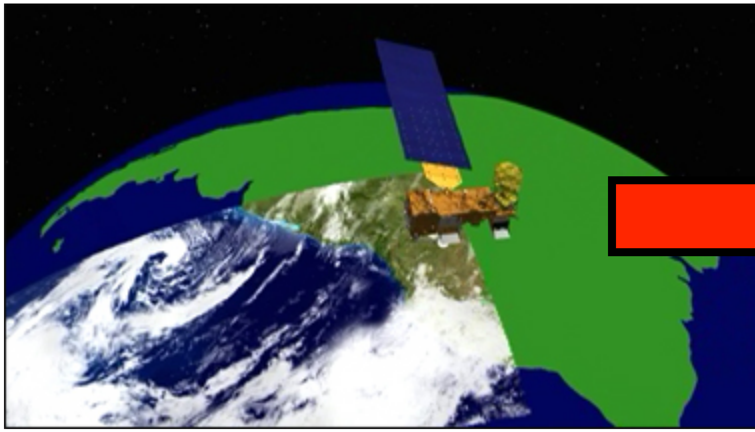
With > 1,000 $PM_{2.5}$ monitors, why bother?



- 690 of 3,100 CONUS counties have >= 1 EPA PM monitors
- On average, each PM monitor covers 180K people or 1800 km^2 in the 690 counties
- 79 million rural and suburban residents are not covered
- Annual EPA network operating cost: $60M, probability of network expansion: ~0?
- Can we do anything to improve the situation?

AOD and $PM_{2.5}$ are different

ROLLINS SCHOOL OF PUBLIC HEALTH
EMORY



AOD – Column integrated value (TOA to surface) - Optical measurement of ambient particle loading.

Relative accuracy: ~15%

$PM_{2.5}$ – dry mass concentration for particles less than 2.5 µm in aerodynamic diameter at ground level

Relative accuracy: < 5%

AOD – PM Relation

ROLLINS SCHOOL OF PUBLIC HEALTH
EMORY

$$AOD(\lambda) = \int_{\text{surface}}^{\text{Top-of-Atmosphere}} \beta_{ext,p}(\lambda, z)\,dz$$

$$C = \frac{4\rho r_e}{3Q} \times \frac{f_{PBL}}{H_{PBL}} \times AOD$$

- ρ – particle density
- Q – extinction coefficient
- r_e – effective radius
- f_{PBL} – % AOD in PBL
- H_{PBL} – mixing height

Composition

Size distribution

Vertical profile

Underlying Assumption for the AOD-PM Relation on Last Slide

ROLLINS SCHOOL OF PUBLIC HEALTH
EMORY

When most particles are concentrated and well mixed in the boundary layer, satellite AOD contains a strong signal of ground-level $PM_{2.5}$ concentrations. **In other words, they must be correlated to begin with.**

Long-range transport events, though rare, tend to break down this assumption. Ideally we manage this in the model. Otherwise, there might be a small amount of outliers.

Modeling the Relation of AOD with $PM_{2.5}$

ROLLINS SCHOOL OF PUBLIC HEALTH
EMORY

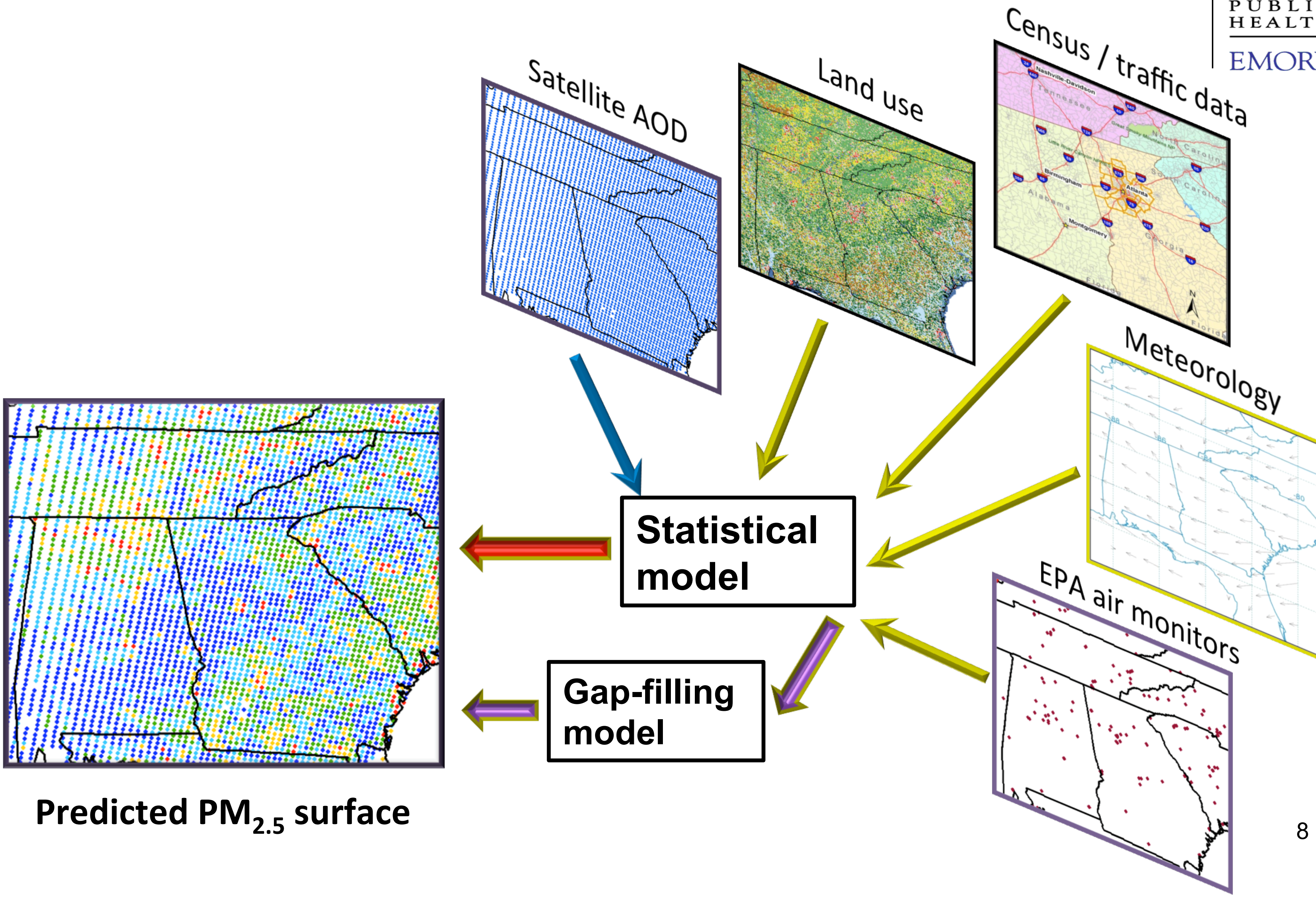
- The AOD-$PM_{2.5}$ relation depends on parameters hard to measure:
  - Vertical profile
  - Size distribution and composition
  - Diurnal variability
- We develop statistical models with variables to represent these parameters
  - Model simulated vertical profile
  - Meteorological & other surrogates
  - Average of multiple AOD measurements

Caveats

ROLLINS SCHOOL OF PUBLIC HEALTH
EMORY

- Given the complex relation between AOD and $PM_{2.5}$ and errors in all the input parameters, uncertainties in satellite $PM_{2.5}$ estimates are inevitable.
- Most high-performance models nowadays can estimate daily $PM_{2.5}$ levels with 15-20% random error and <10% systematic error.
- No one model works everywhere. Model needs to be custom built for performance.
- Regional models usually work better than national models.

Basic Ideas of Model Development



Examples of Advanced Statistical Models

- Multiple linear regression with effect modifiers (e.g., Liu et al. 2005)
- Linear mixed effects (LME) models (e.g., Lee et al. 2011)
- Geographically weighted (GWR) regression (e.g., Hu et al. 2013)
- Generalized additive models (GAM) (e.g., Liu et al. 2009, Strawa et al. 2014)
- Hierarchical models (e.g., Kloog et al. 2012, Hu et al. 2014, Ma et al. 2015)
- Bayesian models (e.g., Chang et al. 2013)
- Artificial neural network (e.g., Gupta et al. 2009)

Requirements for this job

ROLLINS SCHOOL OF PUBLIC HEALTH
EMORY

- A decent computer with large hard drives and good graphics card
- Internet to access to grab satellite & other data
- Statistical software (SAS, R, Matlab, etc.)
- Programming skill
- Knowledge of regional air pollution patterns
- Ideally, GIS software and working knowledge

Model Development Example: Estimating $PM_{2.5}$ in MA with GOES AOD, Meteorology, and GIS Information (Liu et al. 2009. EHP)

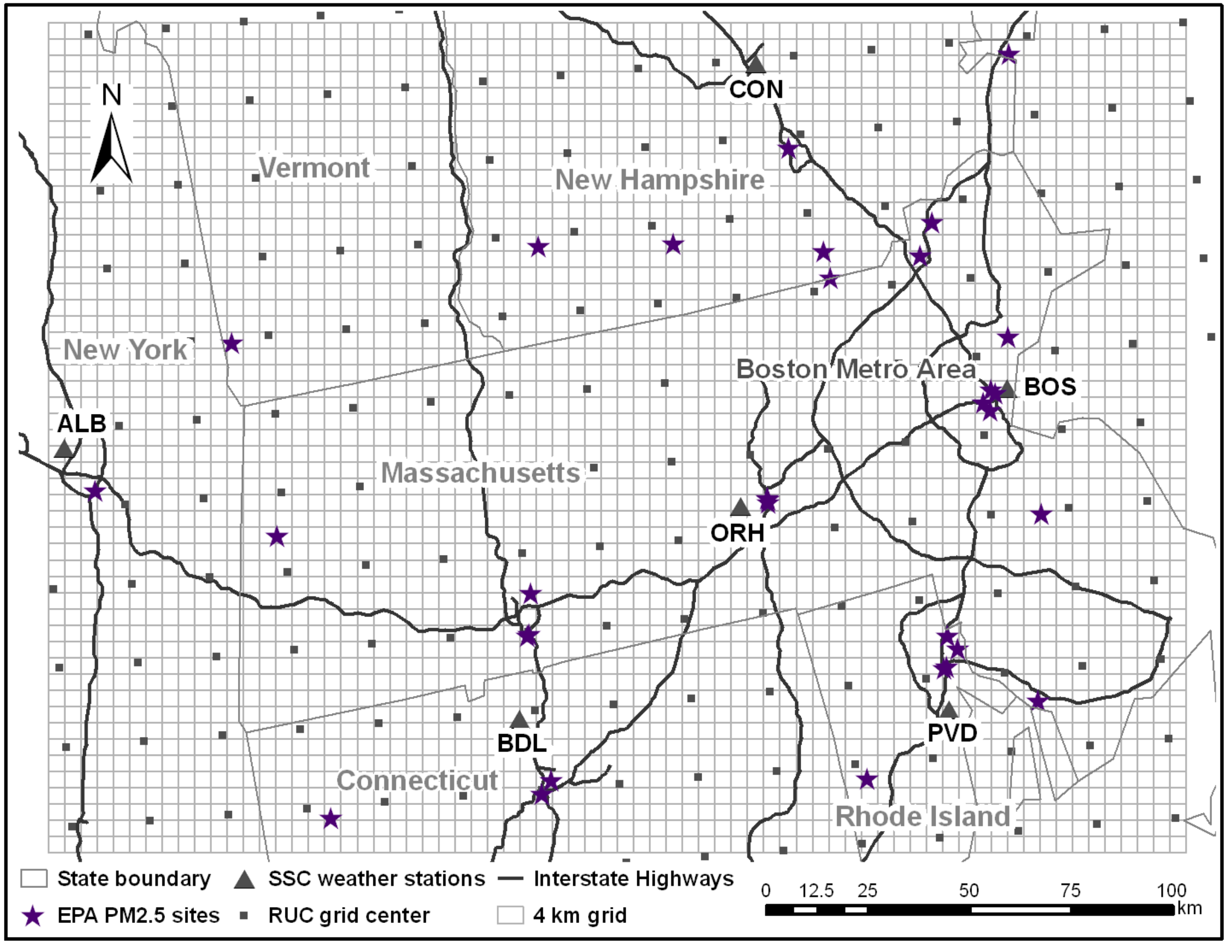
Study Objectives

- Develop a spatial model using GOES AOD, meteorology, and land use information to estimate daily $PM_{2.5}$ concentrations measured by EPA monitors in MA and nearby states

- Predict daily $PM_{2.5}$ concentrations in the modeling domain, where there are no ground measurements, for health effect studies

Modeling Domain

ROLLINS SCHOOL OF PUBLIC HEALTH
EMORY



2003/04 – 2005/06, 32 EPA sites, 4 km grid for prediction

Predictor variables

ROLLINS SCHOOL OF PUBLIC HEALTH
EMORY

- Satellite Data
  - GOES AOD : daily average
- Meteorology
  - RUC20 : assimilated mixing height, T, RH, and wind
  - SSC : weather types
- Land use at 4 km resolution
  - Population density based on census data
  - Road lengths (Class 1, 2, 3, and total)
- Total raw data volume: ~ 1TB

Study Methodology

ROLLINS SCHOOL OF PUBLIC HEALTH
EMORY

- Modeling idea: AOD-$PM_{2.5}$ relation is non-linear in our study domain. The nonlinearity may arise from both temporal and spatial variability.
- Modeling strategy: develop a two-level model
  - Level 1: impact of temporally varying predictors on $PM_{2.5}$ at all sites
  - Level 2: impact of site-specific spatial characteristics on $PM_{2.5}$ at each site

Generalized Additive Model

ROLLINS SCHOOL OF PUBLIC HEALTH
EMORY

The purpose of GAMs is to maximize the quality of prediction of a dependent variable Y from various distributions, by estimating non-parametric functions of the predictor variables which are "connected" to the dependent variable via a link function.

- "Generalized" means we can include categorical variables in the model.
- "Additive" means instead of a single coefficient for each variable (additive term), an non-parametric function is estimated for each predictor.

Two-Level GAM Model
Fitted in R with mgcv package

Level 1: RHS only has temporally varying variables (N = 2,570)

$$Y_{(t)} \sim \mu_1 + f_t(t) + f_{AOD}(t\_AOD) + f_{PBL}(t\_PBL) + f_{RH}(t\_RH)$$
$$+ f_{TEMP}(t\_TEMP) + f_{U,V}(t\_U, t\_V) + \beta_{SSC} SSC$$

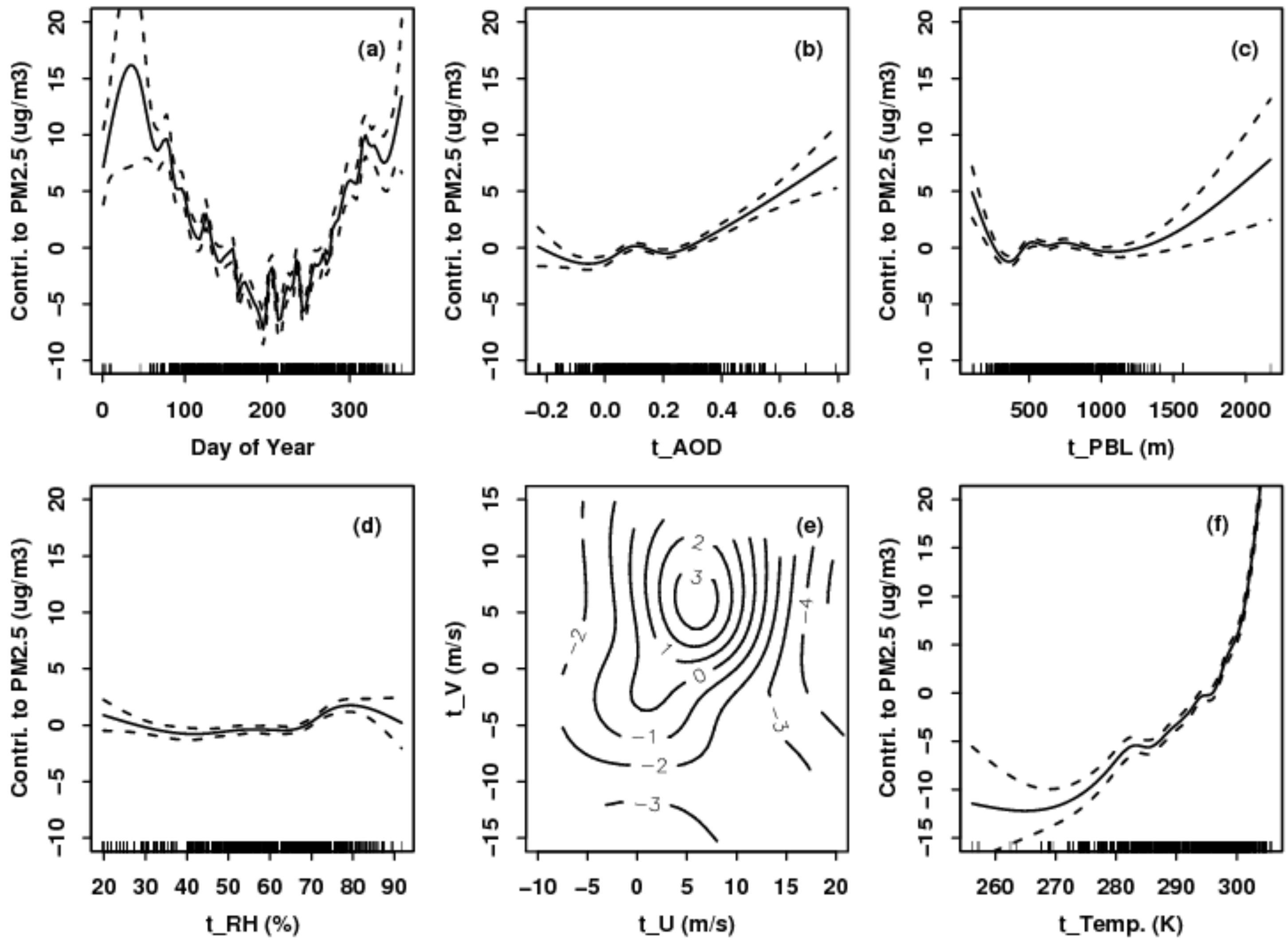
Level 2: RHS only has spatially varying variables (N = 32)

$$Y_{(site)} = \overline{Y_{(t,site)} - \hat{Y}_{(t)}} \sim \mu_2 + \beta_{AOD} AOD_{site} + \beta_{POP} POP$$
$$+ f_{x,y}(x, y) + f_{CLASS\_3}(CLASS\_3)$$

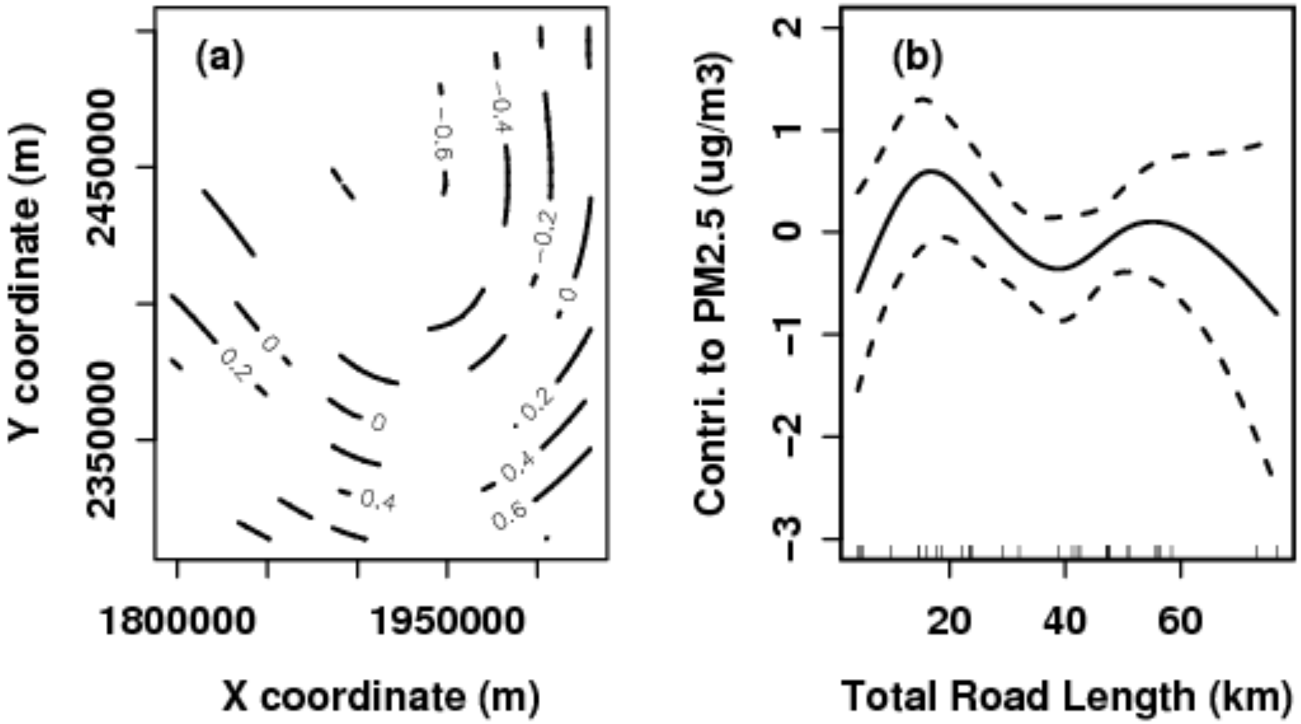
Final prediction (N = 2,570)

$$\hat{Y}_{(t,site)} = \hat{Y}_{(t)} + \hat{Y}_{(site)}$$

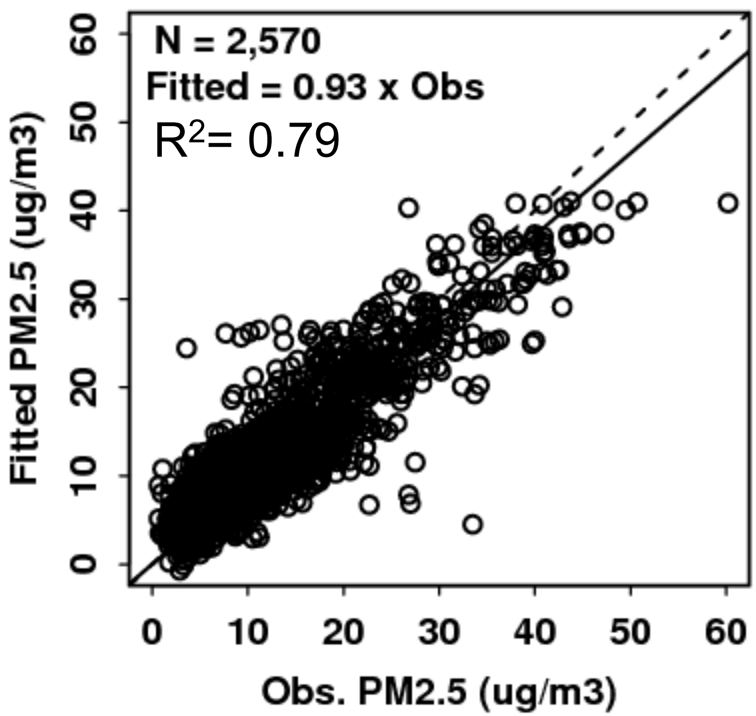
L1 Temporal Model Fitting Results
standard mgcv outputs



L2 Spatial Model Fitting Results



Combined Model Fitting Results



- Mean EPA $PM_{2.5}$ = 10.7 $\mu g/m^3$
- Mean fitted $PM_{2.5}$ = 10.7 $\mu g/m^3$
- Mean abs. diff. = 2.4 $\mu g/m^3$

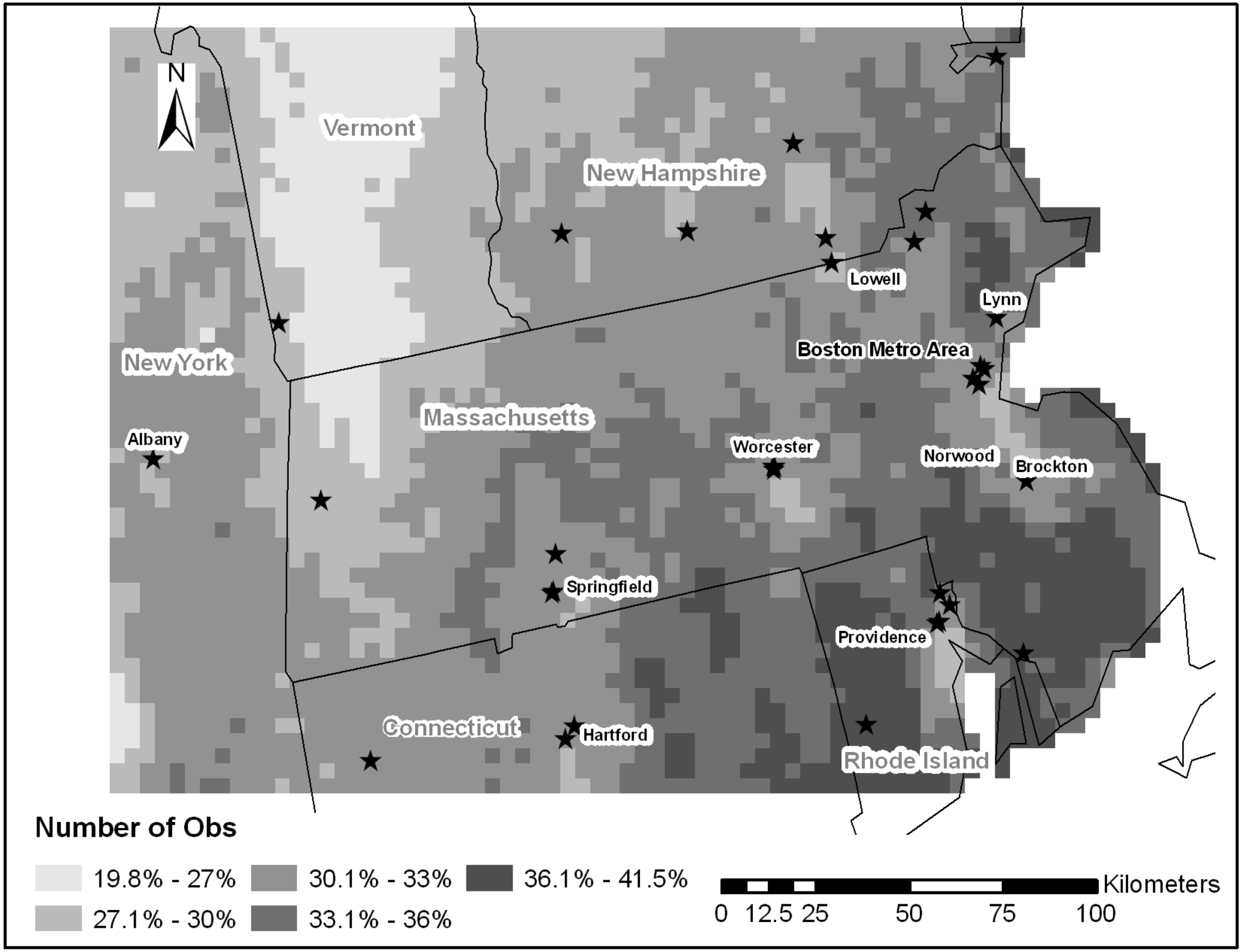
$$\frac{\sum abs(\, fitted - observed\,)}{N}$$

- Mean relative error = 30%

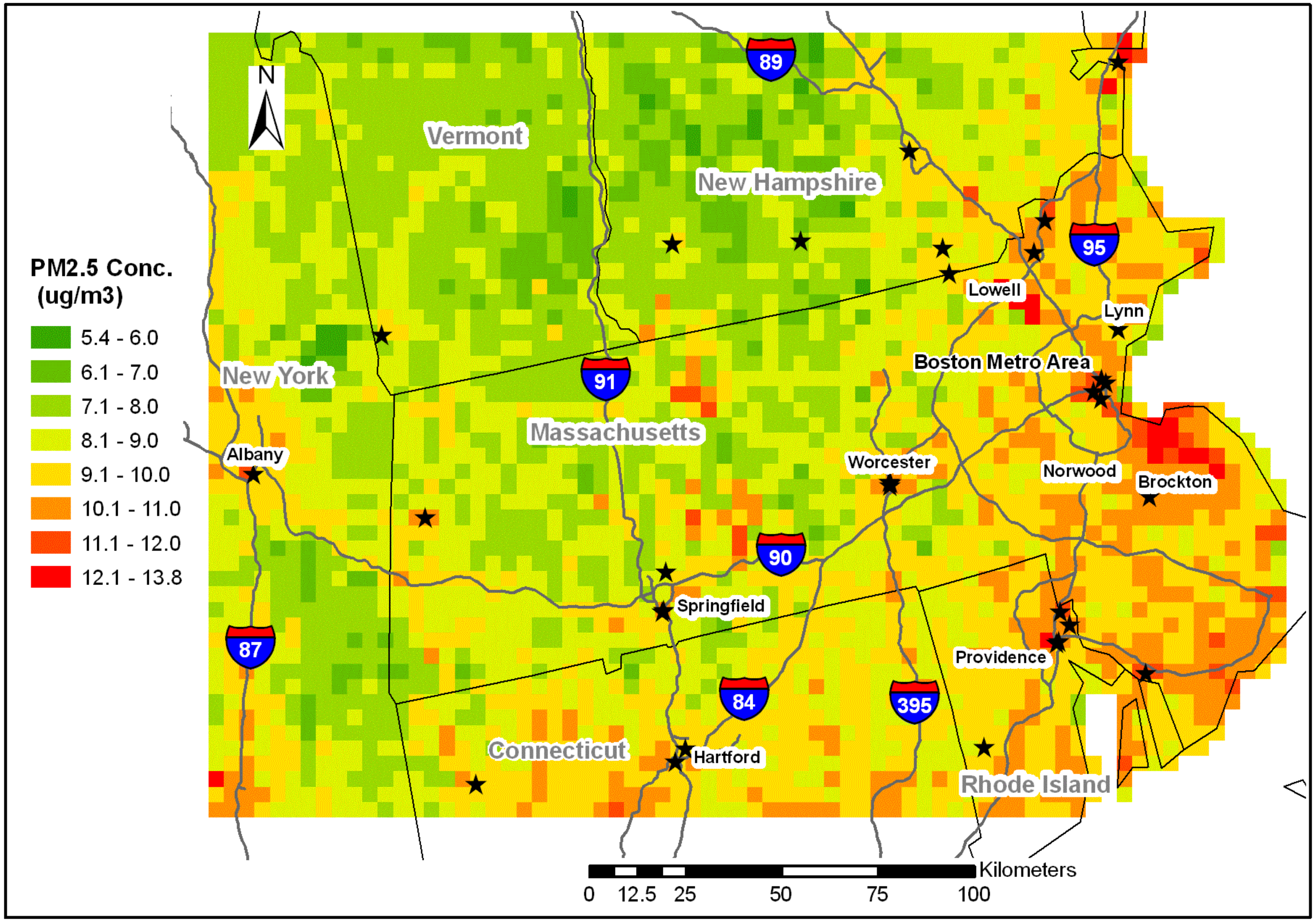
$$\frac{\sum \frac{abs(fitted - observed)}{observed}}{N}$$

- **Cross Validation by site to prevent overfitting**
- **CV R^2 ranges from 0.50 to 0.91, overall 0.79**

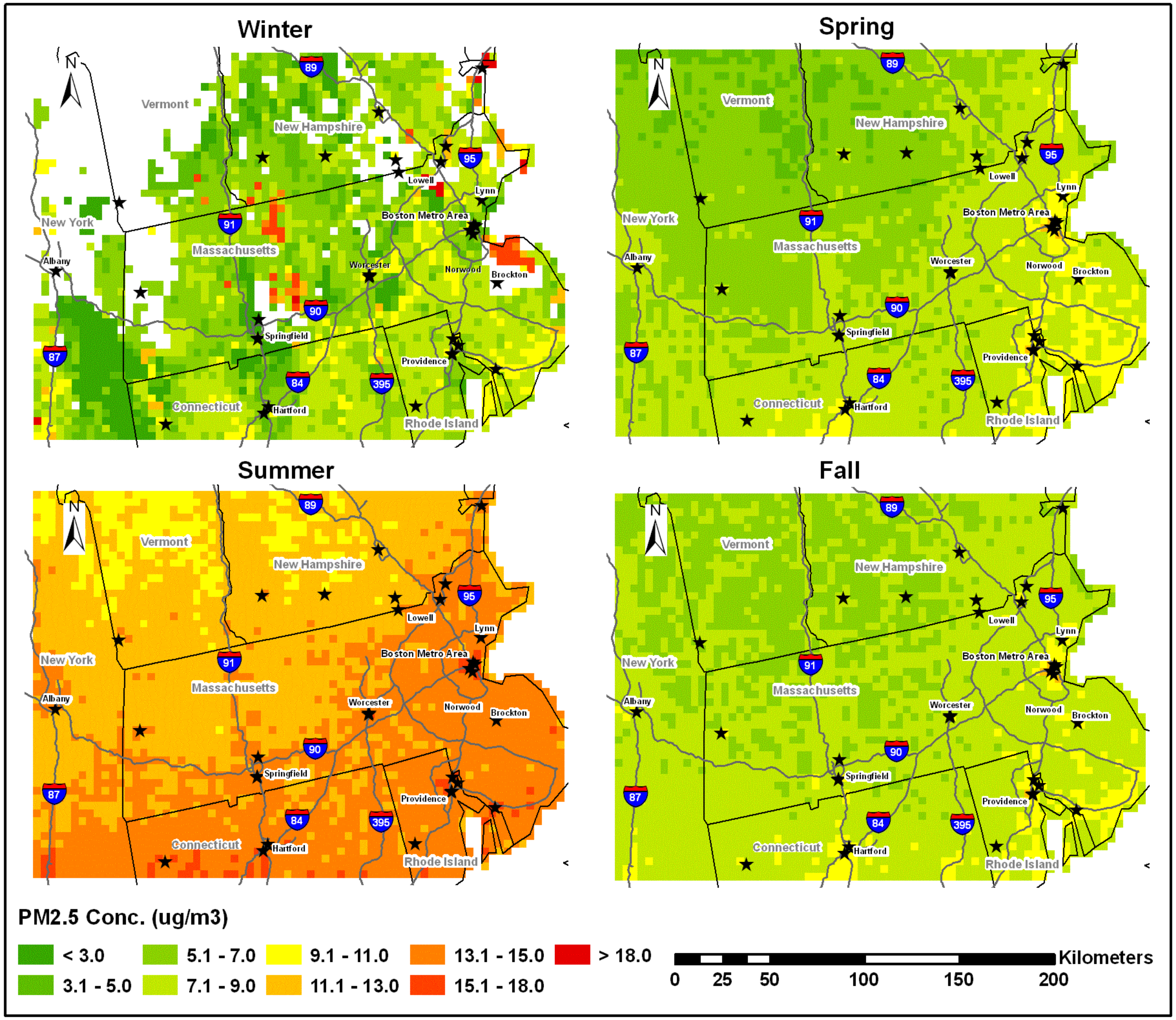
Domain Prediction: # Obs



Overall Mean Predicted $PM_{2.5}$ Distribution



Seasonal Mean Predicted $PM_{2.5}$ Distribution



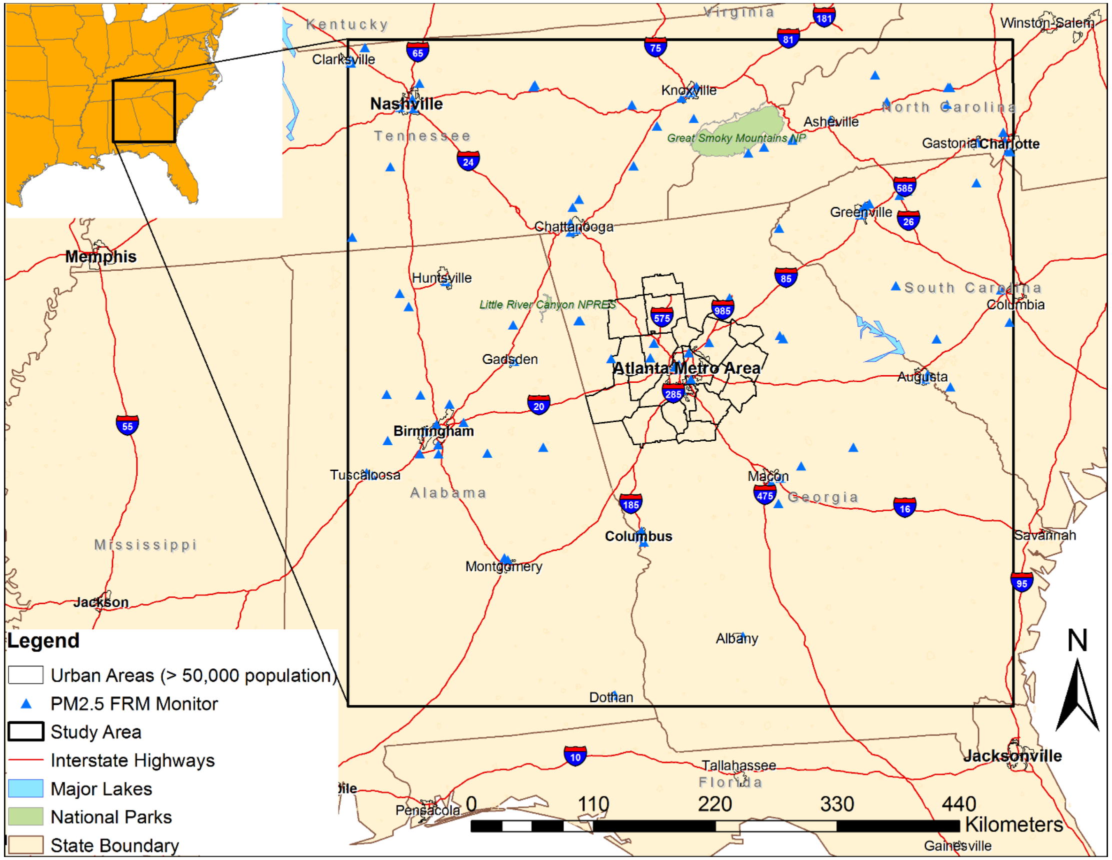
Model Applications Example –Trend Analysis

ROLLINS SCHOOL OF PUBLIC HEALTH
EMORY

Study objective: evaluate the long-term trend of PM2.5 levels in the Southeast (Hu et al. *ACP* 2014)

Study area: 600 x 600 km^2 centered at Metro Atlanta, covering most of GA, AL, NC, and part of SC

Modeling grid: 1 km



Model Structure

- Stage 1: LME (daily)

$$PM_{2.5,\, s,t} = (b_0 + b_{0,t}) + (b_1 + b_{1,t}) AOD_{s,t} + \sum (b_i + b_{i,t}) MetFields_{i,s,t} + b_2\, Forest_s + b_3\, Elev_s + b_4\, MajorRoad_s + b_5\, PointEmit_s + \varepsilon_{s,t}$$

- Stage 2: GWR (monthly)

$$PM_{2.5}\_resi_{st} = \beta_{0,s} + \beta_{1,s} AOD_{st} + \varepsilon_{st}$$

- Can be relatively easily expanded nationwide

Coverage

ROLLINS SCHOOL OF PUBLIC HEALTH
EMORY

	Prediction Days	Mean Daily Spatial Coverage
2001	288	49%
2002	269	48%
2003	296	52%
2004	293	50%
2005	308	54%
2006	316	59%
2007	337	55%
2008	327	54%
2009	314	47%
2010	332	57%

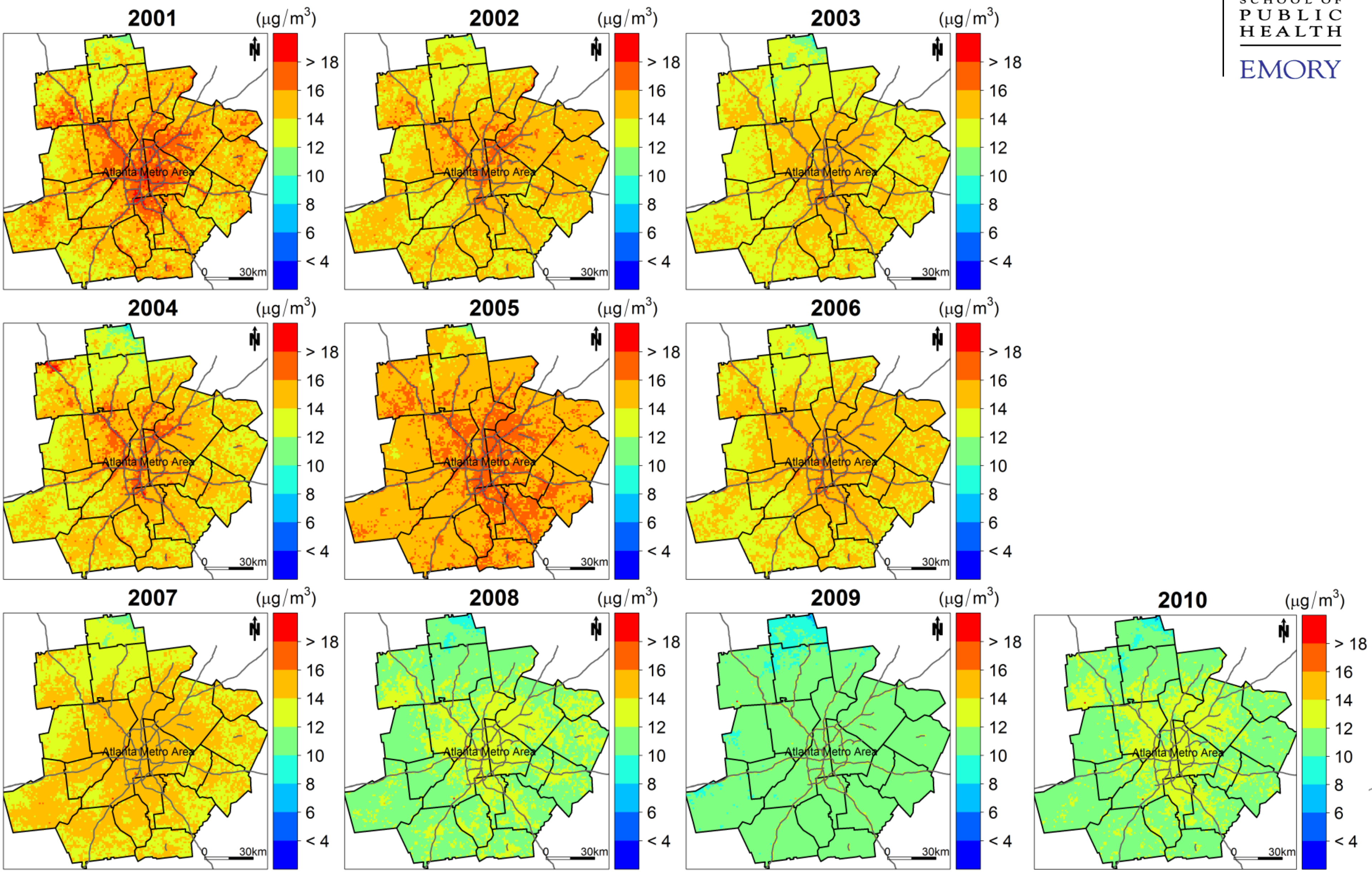
Model Performance

ROLLINS SCHOOL OF PUBLIC HEALTH
EMORY

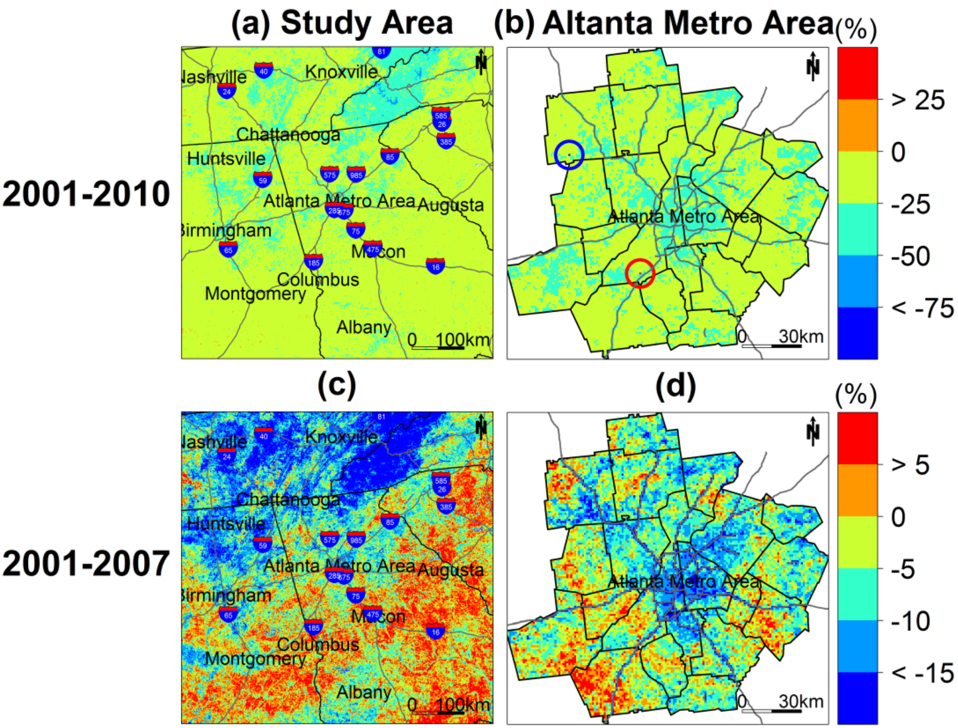
	Model Fitting		Cross Validation	
Year	R^2	MPE ($\mu g/m^3$)	R^2	MPE ($\mu g/m^3$)
2001	0.78	2.50	0.67	3.01
2002	0.84	2.10	0.75	2.62
2003	0.85	1.95	0.76	2.42
2004	0.85	1.97	0.77	2.40
2005	0.84	2.23	0.78	2.64
2006	0.85	2.02	0.78	2.43
2007	0.79	2.26	0.71	2.64
2008	0.74	1.93	0.67	2.21
2009	0.71	1.73	0.62	2.00
2010	0.73	1.90	0.66	2.15

Spatial Trend in Metro Atlanta

ROLLINS SCHOOL OF PUBLIC HEALTH
EMORY

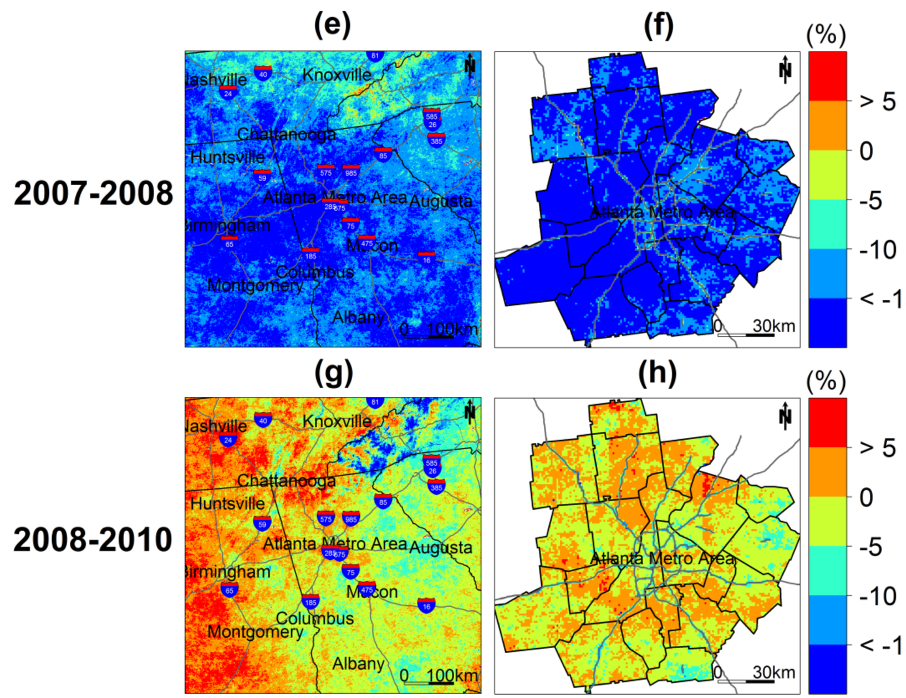


Non-linear Time Trend



Over the decade, relative decrease up to 50%

During first 7 years, up to 15% decrease in the north and Metro Atlanta, increase > 5% in the south



Between 2007 and 2008, universal decrease except in the mountain region

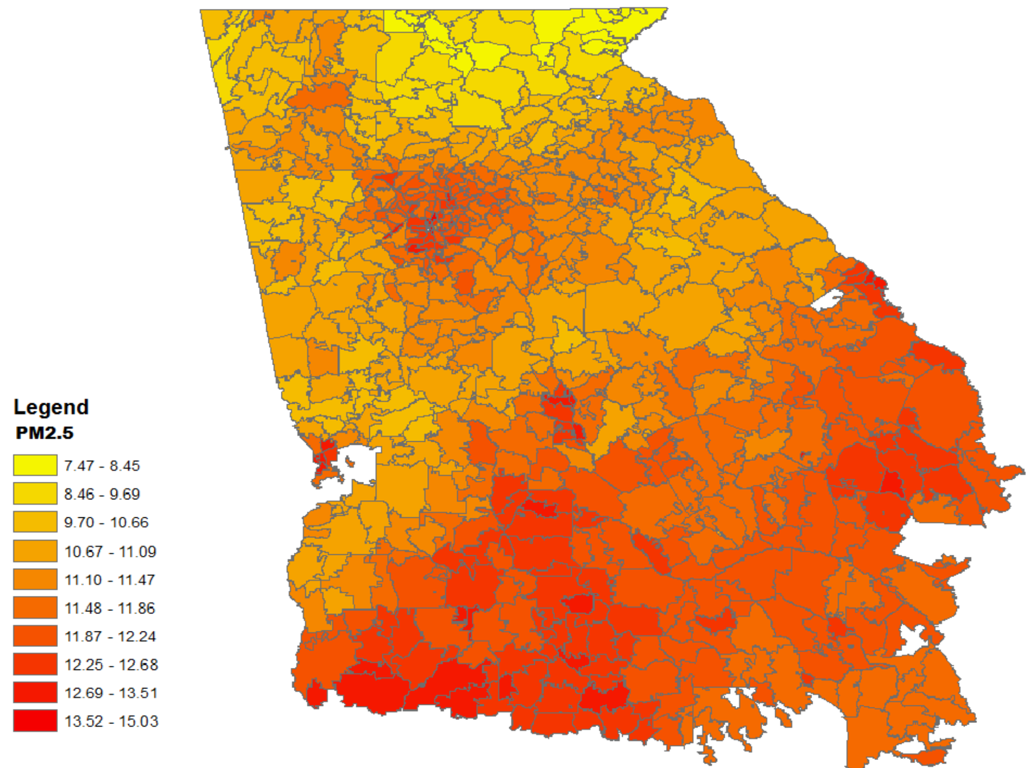
Between 2008 and 2010, small increase in most areas except in mountain regions

ROLLINS SCHOOL OF PUBLIC HEALTH
EMORY

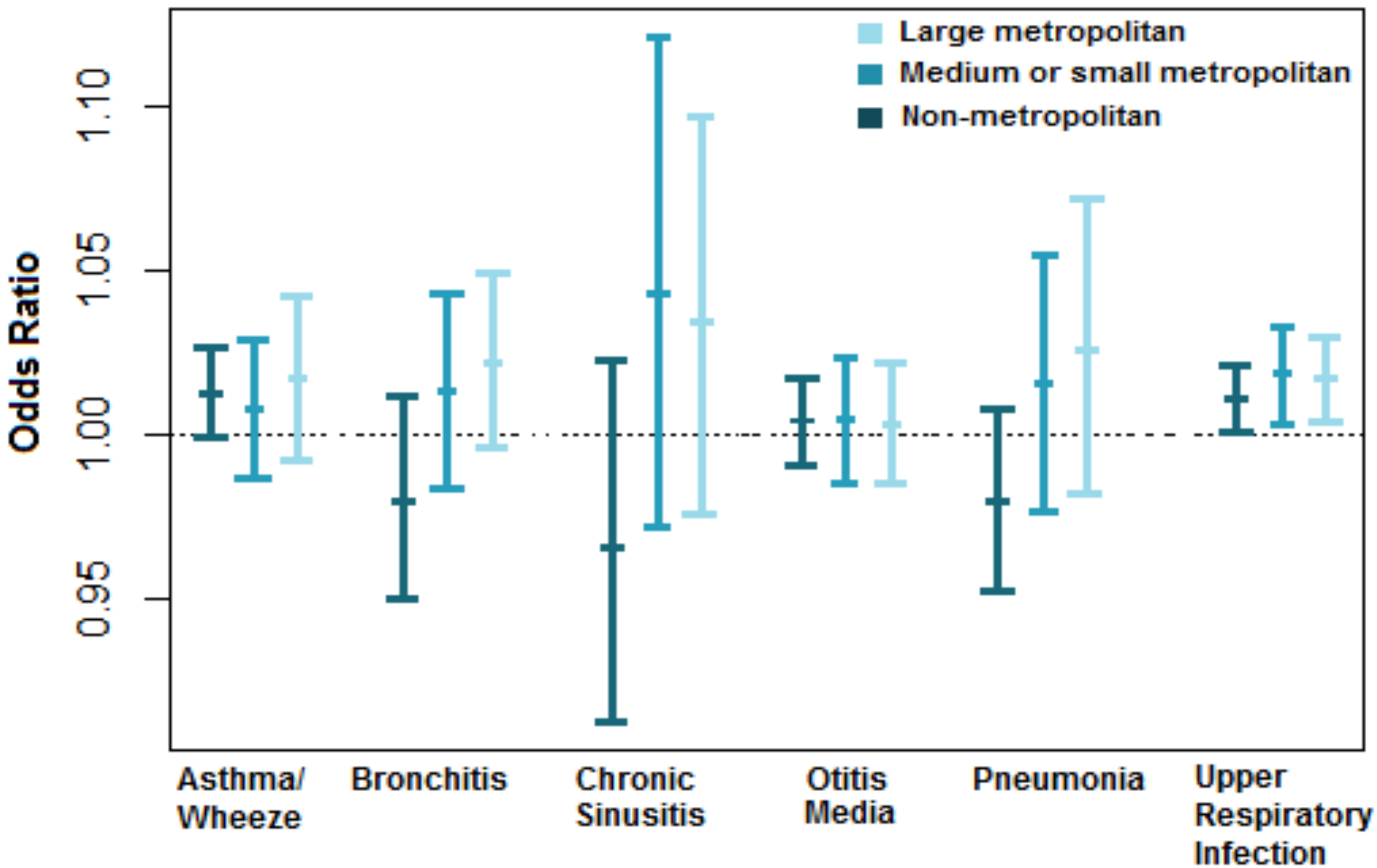
Model Applications Example – Air Pollution Health Effects

Study objective: Estimate associations between daily PM--$_{2.5}$ concentrations and ED visits for six pediatric conditions in Georgia **(Strickland et al. *EHP* 2015)**

Health Data: Individual-level data on pediatric ED visits in GA during Jan 2002 through Jun 2010, aggregated to ZIP codes



Satellite Data Extend Study Population to Rural Areas



Additional Readings

ROLLINS SCHOOL OF PUBLIC HEALTH
EMORY

ORIGINAL ARTICLE

Long- and Short-Term Exposure to $PM_{2.5}$ and Mortality
Using Novel Exposure Models

Itai Kloog,[a] Bill Ridgway,[b] Petros Koutrakis,[a] Brent A. Coull,[c] and Joel D. Schwartz[a]

Atmospheric Environment 74 (2013) 227–236

Contents lists available at SciVerse ScienceDirect

Atmospheric Environment

journal homepage: www.elsevier.com/locate/atmosenv

Estimating spatio-temporal resolved PM_{10} aerosol mass concentrations using MODIS satellite data and land use regression over Lombardy, Italy

Francesco Nordio[a,*], Itai Kloog[a], Brent A. Coull[b], Alexandra Chudnovsky[a], Paolo Grillo[c], Pier Alberto Bertazzi[c], Andrea A. Baccarelli[a], Joel Schwartz[a]

Environmental Science & Technology

Article

pubs.acs.org/est

Estimating Ground-Level $PM_{2.5}$ in China Using Satellite Remote Sensing

Zongwei Ma,[†,‡] Xuefei Hu,[‡] Lei Huang,[†] Jun Bi,[*,†] and Yang Liu[*,‡]

Atmospheric Environment 102 (2015) 260–273

Contents lists available at ScienceDirect

Atmospheric Environment

journal homepage: www.elsevier.com/locate/atmosenv

How well do satellite AOD observations represent the spatial and temporal variability of $PM_{2.5}$ concentration for the United States?

Jing Li[a,b,*], Barbara E. Carlson[a], Andrew A. Lacis[a]

Environment International 51 (2013) 150–159

Contents lists available at SciVerse ScienceDirect

Environment International

journal homepage: www.elsevier.com/locate/envint

Acute health impacts of airborne particles estimated from satellite remote sensing ☆

Zhaoxi Wang[a,*,1], Yang Liu[b,1], Mu Hu[c,1], Xiaochuan Pan[c], Jing Shi[a], Feng Chen[d], Kebin He[e], Petros Koutrakis[a], David C. Christiani[a]

Remote Sensing of Environment 163 (2015) 180–185

Contents lists available at ScienceDirect

Remote Sensing of Environment

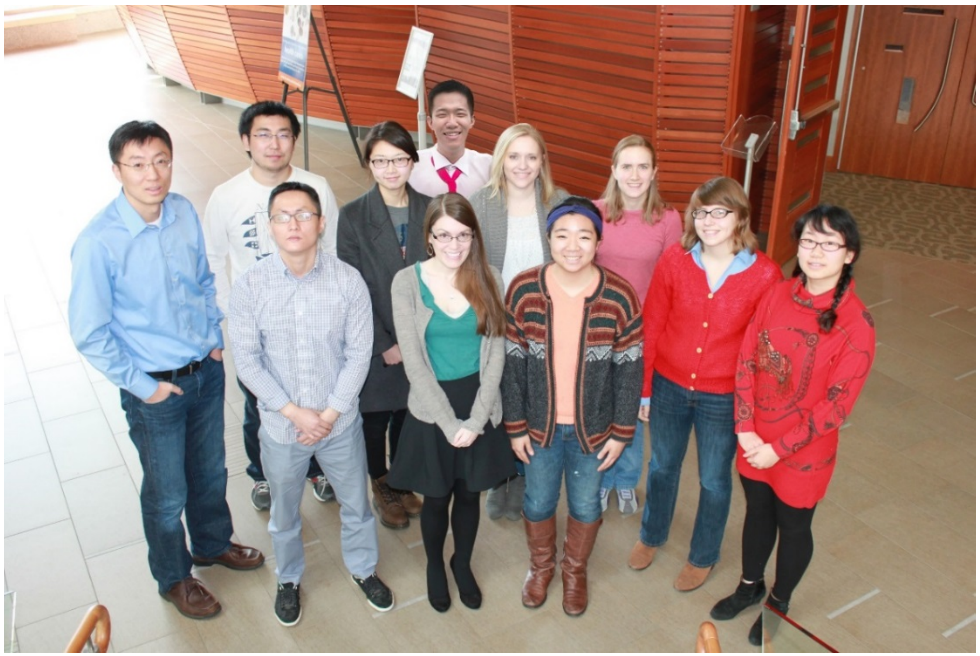
journal homepage: www.elsevier.com/locate/rse

ssessment of $PM_{2.5}$ concentrations over bright surfaces using MODIS satellite observations

Meytar Sorek-Hamer[a], Itai Kloog[b], Petros Koutrakis[c], Anthony W. Strawa[d], Robert Chatfield[d], Ayala Cohen[e], William L. Ridgway[f], David M. Broday[a,*]

The Emory Environmental Remote Sensing Group Welcomes Collaboration Opportunities

ROLLINS SCHOOL OF PUBLIC HEALTH
EMORY



Research interests:

1. Satellite remote sensing applications
2. Multi-scale $PM_{2.5}$ exposure modeling
3. Atmospheric CTM applications
4. Climate and health

Contact: yang.liu@emory.edu
http://web1.sph.emory.edu/remote-sensing/home.html