Comparison of Alternative Exposure Metrics of Air Pollution for Use in Public Health Surveillance

Period of Performance: 06/01/2011-09/09/2011

Final Report for Subcontract to SciMetrika

Principal Investigator: Jim Mulholland, Professor, Environmental Enginerring, Georgia Tech

Co-Investigators: Ted Russell, Georgia Power Distinguished Professor; Yongtau Hu, Research Scientist II

I. BACKGROUND

Estimating exposure at highly resolved geographic and time scales is important for tracking health effects associated with air pollution. Exposure estimates derived from statistical receptor-based models and from mechanistic emissions-based models have been used to fill temporal and spatial gaps in ambient air monitoring data, providing input to the public health community involved in risk assessment. Collaboration of air quality scientists with expertise ranging from emission source to exposure with public health scientists is needed to develop new metrics of exposure and to interpret health risk data.

Issues impeding a full understanding of health effects of ambient air pollution from the existing body of literature are widely appreciated. Atmospheric processes driving transformation of the primary pollutants emitted from stationary and mobile sources lead to a dynamic ambient environment comprised of a multitude of agents with unique physical and chemical characteristics. Composition of the mixture thus varies over both time and space, and attempts to characterize the mixture are generally accompanied by measurement error that varies across the species of interest. Further, how humans move through the microenvironments and how their behaviors and activities alter their personal exposure to pollutants of ambient origin add further complexity. Finally, the physiological responses occur through a complex web of feedback loops on multiple time scales with interactions occurring among specific components and other individual-level factors such as other exposures and genetic constitution. Epidemiologic models focusing on single pollutants are prone to concerns about whether the pollutant is operating as a surrogate for an etiologic agent or group of agents. In the presence of differing levels of measurement error and/or unmeasured confounders, multi-pollutant epidemiologic models do not obviate this concern.

We have a long record of collaboration with Emory Public Health School researchers on a number of air pollution health studies in Atlanta, collectively referred to as Studies of Particles and Health in Atlanta (SOPHIA). This collaboration continues with the recent establishment of an EPA Clean Air Center at Emory and Georgia Tech – the Southeastern Center for Air Pollution and Epidemiology (SCAPE). In this work, we utilize our expertise in atmospheric modeling methods and air quality measurements as well as our experience in working with health researchers to investigate alternative approaches for developing exposure metrics of air pollution for use in public health surveillance.

II. OBJECTIVES AND METHODS

In this three-month study, we compared observation-based and emission-based approaches for spatiotemporal ambient air quality analysis that can then be used to develop geo-imputation methods to convert grid-level predictions to various geographic scales. We focus on three criteria pollutants, NO_2/NO_x (primary pollutant), O_3 (secondary pollutant), and $PM_{2.5}$ (mixed origin), and use the 20-county Atlanta metropolitan area as our study domain. The primary goals of the project were to evaluate the strengths and limitations of these distinctly different approaches for the application of public health tracking and to recommend a hybrid approach for future evaluation. In addition, by working closely with CDC staff, a secondary goal of this project was to share expertise with CDC Environmental Public Health Tracking (EPHT) researchers.

Modeling methods for providing spatially resolved air pollutant estimates using the tools listed below were investigated in this project.

1. Observation-based geo-statistical techniques (e.g. D^2-interpolation, semivariogram analysis).

2. CMAQ (community multiscale air quality modeling system): CMAQ modeling system has been designed to approach air quality as a whole by including state-of-the-science capabilities for modeling multiple air quality issues, including tropospheric ozone, fine particles, toxics, acid deposition, and visibility degradation. In this application, CMAQ provides spatial resolution to the 4 km scale.

3. AERMOD: A steady-state plume model that incorporates air dispersion based on planetary boundary layer turbulence structure and scaling concepts, including treatment of both surface and elevated sources, and both simple and complex terrain. Here, AERMOD provides information on near-source gradients.

III. RESULTS AND DISCUSSION

Georgia Tech researchers worked closely with CDC researcher Ambarish Vaidyanathan on all aspects of this work. A computer was purchased by CDC for CMAQ and AERMOD modeling and integrated with Georgia Tech computational facilities. In this project, correlation analyses of results of monitor-based spatial interpolation and emissions-based CMAQ modeling were used to better understand the strengths and limitations of these methodologies for public health surveillance. Spatial autocorrelation obtained from using these two approaches was assessed as the impacts of measurement error on time-series health studies have been found to depend on spatial autocorrelation (Goldman *et al.*, 2010 and 2011). Results are presented and discussed below. Instructional presentations used by Georgia Tech and CDC researchers are included as appendices for the receptor-based interpolation procedure (Appendix A), the CMAQ modeling system (Appendix B), and the AERMOD modeling software (Appendix C).

Receptor-based Model Evaluation

In previous work, a robust methodology was developed to compute population-weighted daily measures of ambient air pollution for use in time-series studies of acute health effects (Ivy *et al.*, 2008). As a part of this previous work, data from ambient monitors were spatially resolved over the 20-county metropolitan Atlanta area over the time period 1999-2004 to provide daily air pollutant fields of regulatory ambient concentrations (i.e. fields without the local gradients to sources such as roadways) for 11 pollutants: nitrogen dioxide (NO_2), nitrogen oxides (NO_x), carbon monoxide (CO), ozone (O_3), sulfur dioxide (SO_2), particulate mass of particles less than 10 µm in aero-dynamic diameter (PM_{10}), particulate mass of particles less than 2.5 µm in aerodynamic diameter ($PM_{2.5}$), and $PM_{2.5}$ components elemental carbon (EC), organic carbon (OC), nitrate (NO_3^-), and sulfate (SO_4^{2-}). A map showing locations of monitors is provided in Appendix A. Here, we briefly summarize the methodology for spatially resolving the ambient monitor data and then provide results of new analyses that address the accuracy and precision of predications as well as the spatial autocorrelation of results.

Ambient monitor data were log-transformed and normalized as follows.

$$\beta_{ik} = \frac{\ln(C_{ik}) - \mu_i}{\sigma_i} \tag{1}$$

Here, β_{ik} is the normalized value of the pollutant at monitor i for day k, μ_i is the mean of $\ln(C_{ik})$ values for a year at monitor i, and σ_i is the standard deviation of $\ln(C_{ik})$ values for year at monitor i. Thus, the distribution of β_i has an annual mean of zero and an annual standard deviation of one. The normalized values were then inverse distance-square weighted to the 660 census tracts as follows.

$$V_{jk} = \frac{\sum_i \beta_{ik} / D_{ij}^2}{\sum_i 1 / D_{ij}^2} \tag{2}$$

Here, V_{jk} is the interpolated normalized value for each day k at each census tract j, and D_{ij} is the distance from monitor i to census tract j. Normalized values, as opposed to the actual concentrations, were used to produce a smoother interpolated surface and increase the robustness of the metric when monitor data are missing. That is, without normalization, interpolation would result in average concentrations "floating" to regions where no monitors are located. In the case of a limited monitoring network of pollutants with concentrations that are much higher near the urban center than in surrounding rural areas (e.g., vehicular emission pollutants), direct interpolation would lead to unrealistic spatial distributions. The interpolation method used here is based entirely on the ambient monitor data and does not require the use of artificial boundary conditions. Moreover, without normalization the impact of missing data on these interpolations might be such that the results are only useful if data are available from all monitors. Such a reduction in completeness of the dataset might substantially decrease the power of a time-series health study. The normalized value at each census tract was then converted back to a concentration using descriptive models of the means and standard deviations as a function of distance from the urban center.

The normalized value at each census tract was then converted back to a concentration using descriptive models of the means and standard deviations as a function of distance from the urban center.

$$C_{jk} = e^{V_{jk}\sigma_j + \mu_j} \tag{3}$$

Here, μ_j is the modeled mean of $\ln(C_{jk})$ values for the year at census tract j and σ_j is the modeled standard deviation of $\ln(C_{jk})$ values for the year at census tract j. Logistic and linear functions were used to model the annual means and standard deviations, respectively, providing a smooth spatial surface in which local source impacts and biases due to differences in measurement methods are minimized. This procedure allows for daily anisotropic pollutant fields, but the annual average pollutant fields (means and standard deviations) are assumed to be isotropic (i.e., dependent on radial distance only). This assumption has been assessed in previous work (Wade *et al.*, 2006).

The monitor data and estimates calculated at monitor locations by the method described above are highly correlated, as expected, with R^2 values of 0.94 or greater for all pollutants. Some bias is introduced due to the smoothing of mean and standard deviation profiles over space. To evaluate model performance in predicting daily pollutant levels at particular locations in space, the correlation of monitor observations and model predictions calculated without using data from that monitor are shown as a function of distance to the urban center in Figure 1. In general, as distance from the urban center increases, the number of monitors decreases and the variability between monitors increases, resulting in decreasing predictive capability. For pollutants that are predominantly secondary in nature (i.e., formed in the atmosphere), such as O_3 and $PM_{2.5}$ total and SO_4^{2-} and NO_3^- component masses, high correlations (r > 0.8) are obtained even for sites within 65 km of the urban center. On the other hand, pollutants strongly associated with mobile sources, such as NO_2/NO_x, CO, and EC, are not well predicted at rural sites, with R values between 0.3 and 0.4 for the Yorkville site located approximately 64 km from the urban center. The ability to predict the SO_2 concentrations is particularly poor. Major sources of SO_2 in the Atlanta area are coal combustion point sources, in particular a coal-fired power plant located 11.5 km northwest of the urban center. When a plume from this plant impacts the Atlanta area, its width is narrow resulting in a spatially heterogeneous pollutant field that is not well characterized by the ambient monitors. The correlation of observations and predictions for OC, which has significant primary and secondary components, is intermediate.

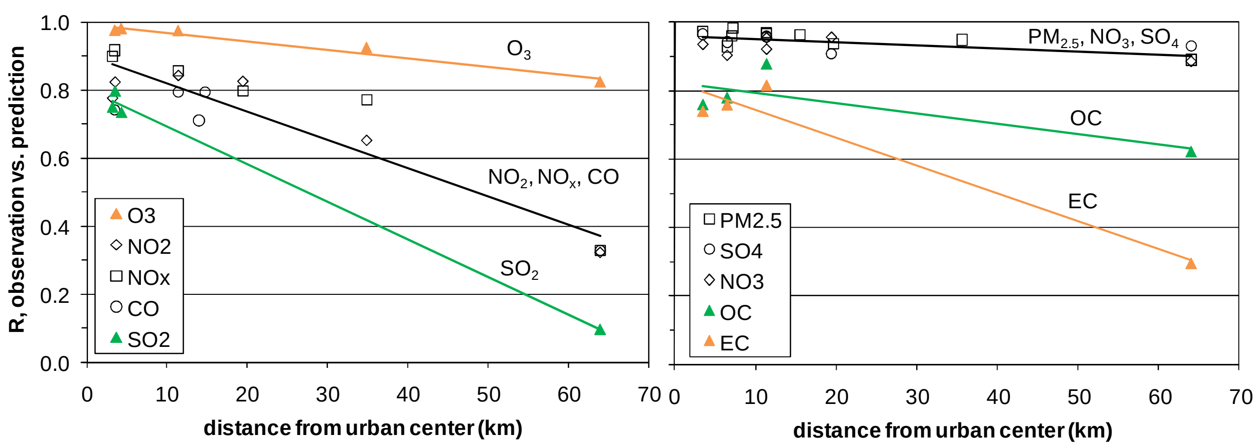


Figure 1. Correlation of monitor observations and model predictions without using data from that monitor as a function of distance from the urban center for pollutant gases (left) and $PM_{2.5}$ total and major component masses (right). Curves indicate spatial trends for single pollutants or groups of pollutants. For collocated monitoring sites, both sets of observations were removed for model prediction at those sites.

An explanation of the limited predictive capability of the monitor-based interpolation methodology is the high degree of spatial autocorrelation in the air pollutant fields. In Figure 2, correlograms using data from 2004 only are shown for four pollutants: NO_2, EC, O_3, and $PM_{2.5}$. Model results represent Pearson correlation coefficients of each of the 660 census tract estimates with the central census tract; monitor results represent Pearson correlation coefficients with the central monitor (Jefferson St), with the value at distance zero obtained for collocated instruments. The model predications exhibit more spatial autocorrelation on average than the monitor observations. At monitor locations the model correlations approach that of the monitors, but away from monitors the model correlations are much higher.

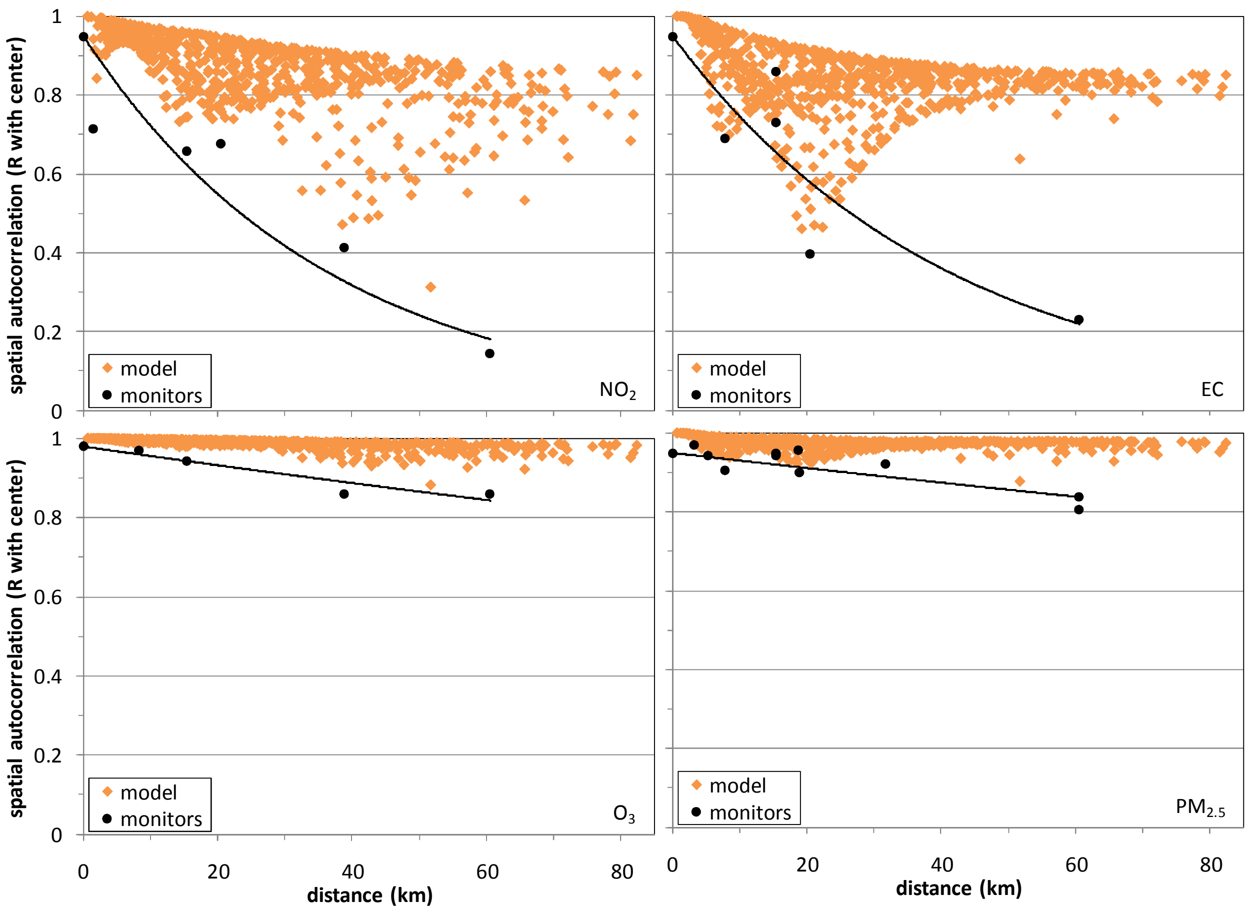


Figure 2. Spatial autocorrelation of monitor observations and interpolation model predictions for four pollutants, 2004 data.

These results demonstrate that the number and location of monitors limits the predictive capability of a monitor-based interpolation model, particularly for primary pollutants. Emission-based model results that incorporate dispersion effects are evaluated next to assess their predictive capabilities.

Emissions-based Model Evaluation

At a 4 km resolution, 48-hour forecasts of air quality using the CMAQ modeling system were obtained for the Atlanta metropolitan area for 2010. Daily metrics were computed from the hourly forecasts and compared with monitor data. In Table 1, Pearson correlation coefficients and percent bias between model estimates and observations at an urban location and a rural location are listed for five pollutant gases and $PM_{2.5}$ mass. The biases are much greater than for the monitor-based interpolation model, as expected since the CMAQ predictions do not use the monitor data. The correlation coefficients are similar to those found when data are withheld using the interpolation method (Figure 1).

Table 1. Comparison of observations and CMAQ model predictions at two locations, 2010 data.

	urban (JST)		rural (Yorkville)	
	correlation	bias	correlation	bias
NO_2	0.84	23%	0.72	140%
NO	0.80	-52%	0.32	138%
SO_2	0.59	329%	0.34	295%
CO	0.84	55%	0.55	-8%
O_3	0.73	26%	0.68	7%
$PM_{2.5}$	0.60	7%	0.61	7%

The spatial autocorrelation in the CMAQ predictions is compared with that in the monitor data in Figure 3. CMAQ derived spatial autocorrelation is not as great as that obtained when using the interpolation model (Figure 2), but it is still greater than that suggested by the monitor data for primary pollutants. Monitor data can include local source impacts, and dispersion in the CMAQ model is likely underestimated due to limited meteorological and surface feature inputs.

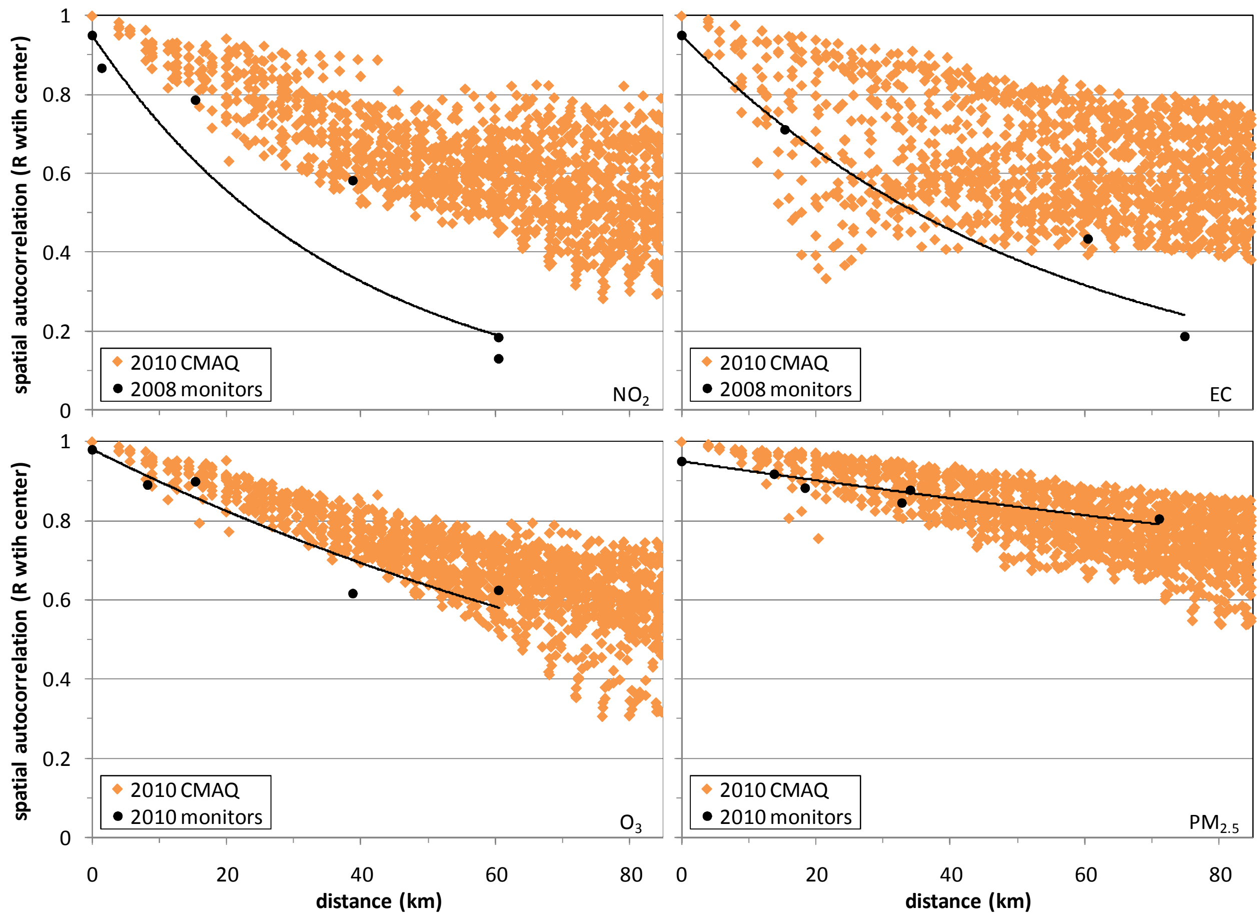


Figure 3. Spatial autocorrelation of monitor observations and CMAQ model predictions for four pollutants, 2010 data.

As further evidence of this limitation in CMAQ predictions, inter-pollutant correlations were computed and compared those based on monitor data. As shown in Table 2, CMAQ inter-pollutant correlations appear to be too high. In PM source apportionment studies that incorporate CMAQ modeling, we have shown that this limitation results in source impacts that vary little relative to each other, which limits the use of such an approach for time-series epidemiologic studies (Marmur *et al.*, 2006).

Table 2. Inter-pollutant Pearson correlation coefficients at Jefferson St for monitor data (a) and for CMAQ predictions (b).

	NO_2	NO	SO_2	CO	O_3	$PM_{2.5}$	SO_4	NO_3	NH_4	EC
(a) 2007 Jefferson St observations										
NO	0.621									
SO_2	0.310	0.249								
CO	0.698	0.909	0.217							
O_3	-0.318	-0.345	-0.230	-0.232						
$PM_{2.5}$	0.375	0.175	0.009	0.431	0.414					
SO_4	0.046	-0.089	-0.060	0.086	0.514	0.799				
NO_3	0.308	0.150	0.291	0.196	-0.442	0.078	-0.198			
NH_4	0.067	-0.093	-0.040	0.096	0.453	0.802	0.971	-0.037		
EC	0.707	0.685	0.163	0.835	-0.052	0.664	0.290	0.135	0.287	
OC	0.409	0.297	-0.001	0.529	0.201	0.837	0.372	0.150	0.374	0.715
(b) 2010 CMAQ predictions at Jefferson St										
NO	0.760									
SO_2	0.511	0.602								
CO	0.924	0.851	0.479							
O_3	-0.756	-0.666	-0.427	-0.733						
$PM_{2.5}$	0.835	0.773	0.569	0.837	-0.505					
SO_4	0.361	0.253	0.408	0.303	0.151	0.656				
NO_3	0.632	0.715	0.513	0.746	-0.713	0.691	0.073			
NH_4	0.722	0.679	0.572	0.755	-0.404	0.935	0.697	0.722		
EC	0.876	0.891	0.469	0.874	-0.637	0.872	0.384	0.632	0.742	
OC	0.710	0.595	0.330	0.658	-0.307	0.882	0.631	0.371	0.714	0.774

These results suggest that a hybrid approach that incorporates CMAQ modeling capabilities and monitor data might provide improved air pollutant fields. However, over-prediction of spatial autocorrelation will likely continue to be a limitation. A near-source model, such as AERMOD or CALINE, might provide more realistic spatial autocorrelation. Georgia Tech and CDC researchers have begun using these models and are exploring ways of developing hybrid approaches that include near-source modeling.

IV. CONCLUSION

The work performed under this Subcontract demonstrates the need for a hybrid modeling approach that incorporates actual observations to provide air pollutant fields that can be used for public health surveillance. Not only are the spatial resolved estimates needed, but uncertainties in these estimates must be provided as these uncertainties vary markedly between pollutants and over space. The three-month effort described here provides a starting point for CDC to develop the improved capabilities in air quality modeling that are needed for the public health tracking system being developed.

Related Work Cited in this Report

Goldman GT, Mulholland JA, Russell AG, Srivastava A, Strickland MJ, Klein M, Waller LA, Tolbert PE, Edgerton ES (2010). "Ambient Air Pollutant Measurement Error: Characterization and Impacts in a Time-Series Epidemiologic Study in Atlanta," *Environ. Sci. Technol.*, **44**:7692-7698.

Goldman GT, Mulholland JA, Russell AG, Strickland MJ, Klein M, Waller LA, Tolbert PE (2011). "Impact of Exposure Measurement Error in Air Pollution Epidemiology Time-Series Studies: Effect of Error Type," *Environ. Health*, **10**:61.

Ivy D, Mulholland JA, Russell AG (2008). "Development of Ambient Air Quality Population-Weighted Metrics for Use in Time-Series Health Studies," *J. Air & Waste Manage. Assoc.*, **58**:711-720.

Marmur A, Park S-K, Mulholland JA, Tolbert PE, Russell AG (2006). "Source Apportionment of $PM_{2.5}$ in the Southeastern United States Using Receptor and Emissions-Based Models: Conceptual Differences and Implications for Time-Series Health Studies," *Atm. Environ.*, **40**:2533-2551.

Wade KS, Mulholland JA, Marmur A, Russell AG, Harsell B, Edgerton E, Klein M, Waller L, Peel JL, Tolbert PE (2006). "Instrument Error and Spatial Variability of Ambient Air Pollution in Atlanta, Georgia," *J. Air & Waste Manage. Assoc.*, **56**:876-888.

Appendix A

Receptor-based Interpolation Instruction Presentation

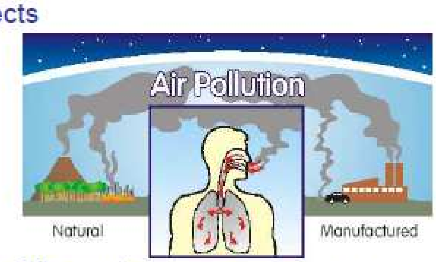
Population-Weighted Measures of Spatial Interpolated Data for use in Air Pollution Epidemiological Studies

Jim Mulholland

Georgia Institute of Technology
Civil and Environmental Engineering

Motivation

- Air pollution has acute health effects
  - Asthma
  - Respiratory Disease
  - Cardiovascular Disease



Source: http://www.atsdr.cdc.gov/

- Emory's retrospective study in the 20-county Atlanta metropolitan region
  - Over one million emergency department visits each year for respiratory and cardiovascular illnesses

SOPHIA
Study of Particles and Health in Atlanta

Objective

- Develop a spatial composite variable for each air pollutant that is:
  - independent of bias due to measurement method
  - representative of study area
  - robust when data are missing

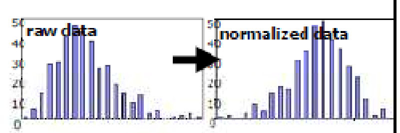
- Provide daily estimate of uncertainty to represent error in metric

- Provide daily estimate of spatial variance to represent variability in exposure

Methods

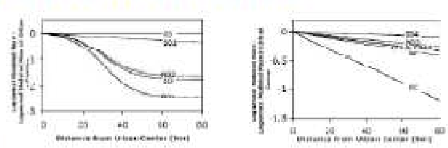
Overview

1. **Normalize station data: $C_{i,k} \Rightarrow \beta_{i,k}$**



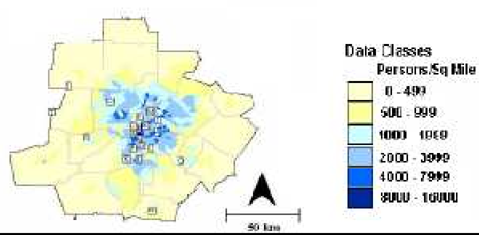
2. **Distance-square weight to census tract: $\beta_{i,k} \Rightarrow V_{j,k}$**

3. **De-normalize using functions of distance from central station: $V_{j,k} \Rightarrow C_{j,k}$**

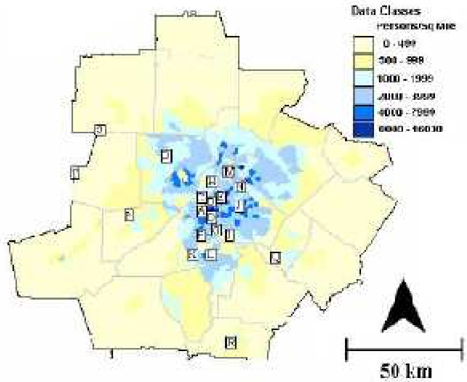


4. **Population-weight: $C_{j,k} \Rightarrow C_k$**

i = monitoring site index
j = census tract index
k = day index

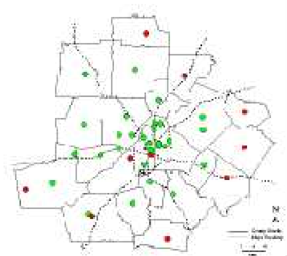


Methods



A	JS-Jefferson Street (1,2,3,4,5,6,7)	K	EP-East Point Health Center (6)
B	GT-Georgia Tech (1,2,5)	L	FP-Forest Park (6)
C	FS8-Fire Station #8 (5,6)	M	DHC-Doraville Health Center (5,6)
D	FC-Fulton County Health Department. (5)	N	Tr-Tucker (2,6,7)
E	ERS-East Rivers School (5,6)	O	Ke-Kennesaw (6)
F	FM Fort McPherson (6,7)	P	Do Douglasville (2)
G	CA-Confederate Avenue (1,4)	Q	Co-Conyers Monastery (2,4)
H	RR-Roswell Road (5)	R	Gr-Griffin (5)
I	SD-South Dekalb (2,3,4,5,7)	S	St-Stilesboro (1)
J	DT-Dekalb Tech (3)	T	Yo-Yorkville (1,2,3,4,5,6,7)

1=SO_2; 2=NO_2/NO_x; 3=CO; 4=O_3; 5=PM_{10}; 6=$PM_{2.5}$; 7-$PM_{2.5}$ composition



660 Census Tracts

4,112,198 Total Population

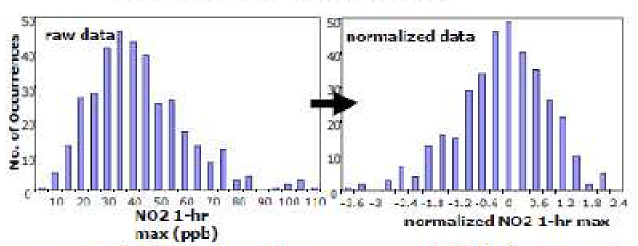
Methods

Normalize Station Data

$$\beta_{i,k} = \frac{\ln[C_{i,k}] - \mu_i}{\sigma_i}$$

μ_i = mean of $\ln[C_{i,k}]$ values for year at monitor i
σ_i = standard deviation of $\ln[C_{i,k}]$ for year at monitor i

- At each monitor, the distribution of β for the year is about the same (mean of 0 and standard deviation of 1)



- normal distribution parameters
  μ_{am} = 44.2 ppb
  σ_{am} = 18.4 ppb (25.8-62.6 ppb)
- lognormal distribution parameters
  μ_g = 40.8 ppb
  σ_g = 1.49 (27.4-60.9 ppb)

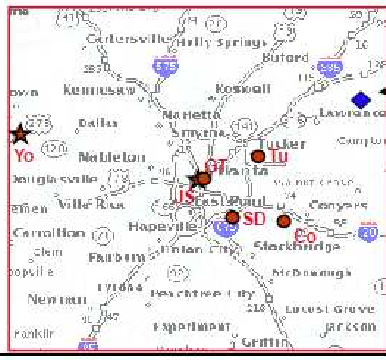
- normal distribution parameters
  μ_{am} = 0.0097
  σ_{am} = 0.984
- same distribution at all sites

Methods

Distance-square Weight to Census Tract

$$V_{j,k} = \frac{\sum_i \beta_{i,k} / D_{ij}^2}{\sum_i 1/D_{ij}^2}$$

j = index of census tract
i = index of monitoring station
k = index of day
D_{ij} = distance from monitor i to census tract j



example: census tract 1801.01	
population fraction	0.00246
distance from JS	63.6 km
distance from Yo	114.9 km
distance from GT	62.9 km
distance from Tu	43.4 km
distance from SD	59.7 km
distance from Co	56.5 km

Methods

De-normalize to Concentrations

$$C_{j,k} = e^{V_{j,k}\sigma_j + \mu_j}$$

μ_j = mean of ln[$C_{j,k}$] values for year at census tract j
σ_j = standard deviation of ln[$C_{j,k}$] for year at census tract j

• Model μ_j and σ_j as function of distance from urban center for each year

Logistic fit

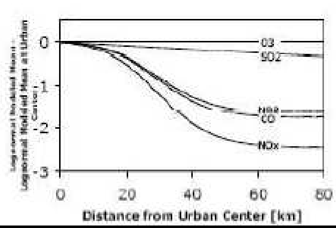
$$\mu_j = M - \frac{\Delta}{1+\exp\{-(x_j - x^*)/r\}}$$

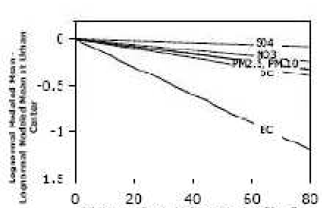
Linear fit

$$\sigma_j = mx + b$$

Constant fit

$$\sigma_j = b$$



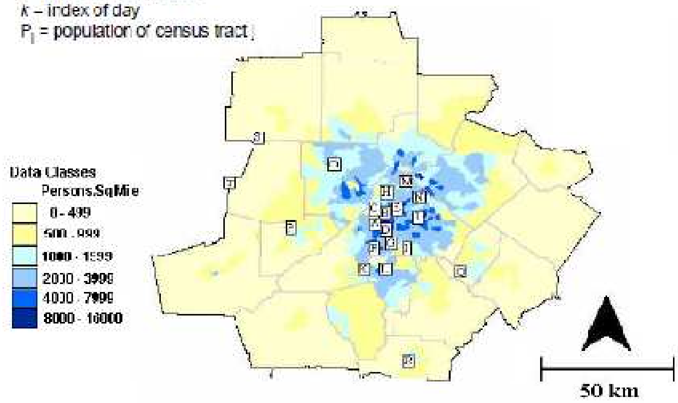


Methods

Population-Weight

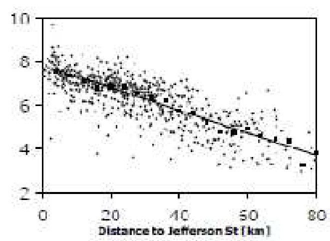
$$C_k = \frac{\sum_j C_{j,k} P_j}{\sum_j P_j}$$

j = index of census tract
k = index of day
P_j = population of census tract j



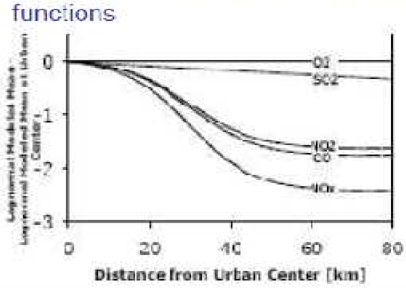
Results

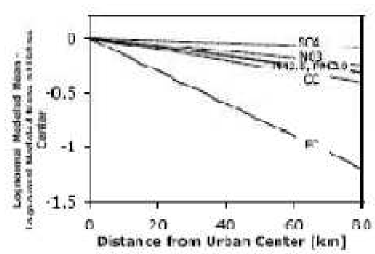
Models



- Population density exponentially decreases with respect to urban center
- Population density may be an indicator of spatially varying pollutant concentrations
- Traffic density might be a better indicator

- Mean CO and NO_2/NO_x concentrations were best fit by logistic functions
- Other pollutant concentrations were best fit by exponential functions

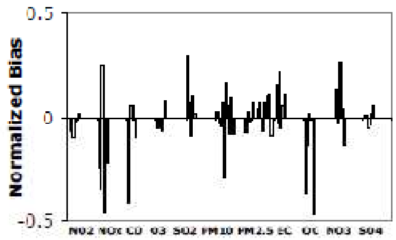


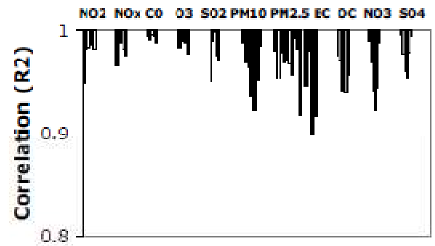


Results

Model Performance

$$NBias = \frac{1}{y}\frac{\sum_{i=1}^{N}(y_i - x_i)}{N}$$

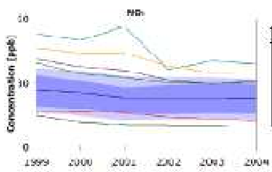




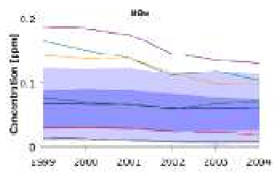
- Normalized bias is high for non-representative monitors since they are smoothed over by descriptive models
  - South Dekalb is located near a roadway
  - EC and OC are measured differently at SEARCH and AQS sites
- High correlation between monitor data and concentrations modeled at nearest census tract

Results

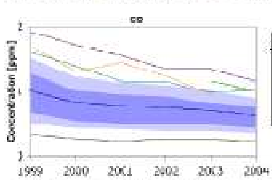
Overall



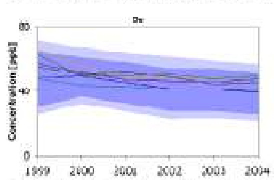
NO2	1-hr max
JS	0.866
GT	0.919
SD	0.86
Tu	0.511
Ga	0.678
Yo	0.091



NOx	1-hr max
JS	0.942
GT	0.95
SD	0.876
Tu	0.585
Ga	0.745
Yo	0.149

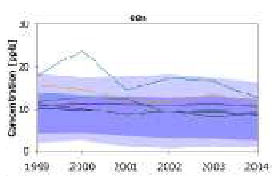


CO	1-hr max
JS	0.887
RR	0.81
SD	0.871
DT	0.875
Yo	0.27



O3	8-hr max
JS	0.984
DA	0.991
SD	0.981
Co	0.938
Yo	0.853

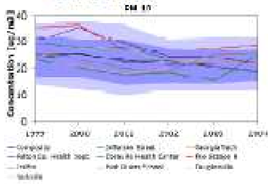
- Primary pollutants - population wted values are lower than central monitors and less correlated with distant monitors
- Secondary pollutants - population wted values are similar in magnitude and highly correlated with all monitors
- All pollutants are highly correlated (R>0.9) to data from the central monitors



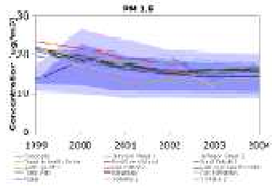
SO2	1-hr max
JS	0.395
GT	0.962
GA	0.363
Sr	0.301
Yo	0.14

Results

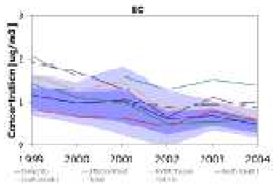
Overall



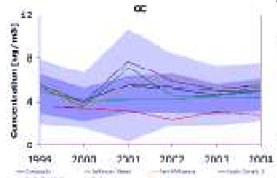
PM 10	24-hr avg
JS	0.951
GT F	0.555
FS8	0.733
FC	0.808
ERS	0.848
DHC	0.812
Dn	0.717
Gr	0.754
Yo	0.734



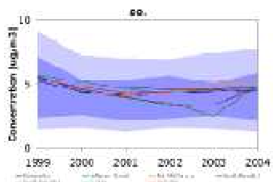
PM 2.5	24-hr avg
JS	0.951
FCH	0.750
ERS	0.919
FM	0.914
SD	0.901
EP	0.825
FP	0.89
DHC	0.922
Tu	0.739
Ke	0.903
Yo	0.861



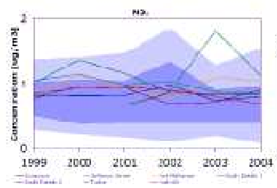
EC	24-hr avg
JS	0.948
FM	0.64
SD1	0.728
SD2	0.872
Tu	0.602
Yo	0.513



OC	24-hr avg
JS	0.908
FM	0.717
SD1	0.771
SD2	0.929
Tu	0.71
Yo	0.776



NO3	24-hr avg
JS	0.987
FM	0.754
SD1	0.855
SD2	0.962
Tu	0.894
Yo	0.787

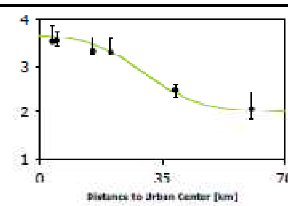


SO4	24-hr avg
JS	0.939
FM	0.825
SD1	0.739
SD2	0.957
Tu	0.746
Yo	0.89

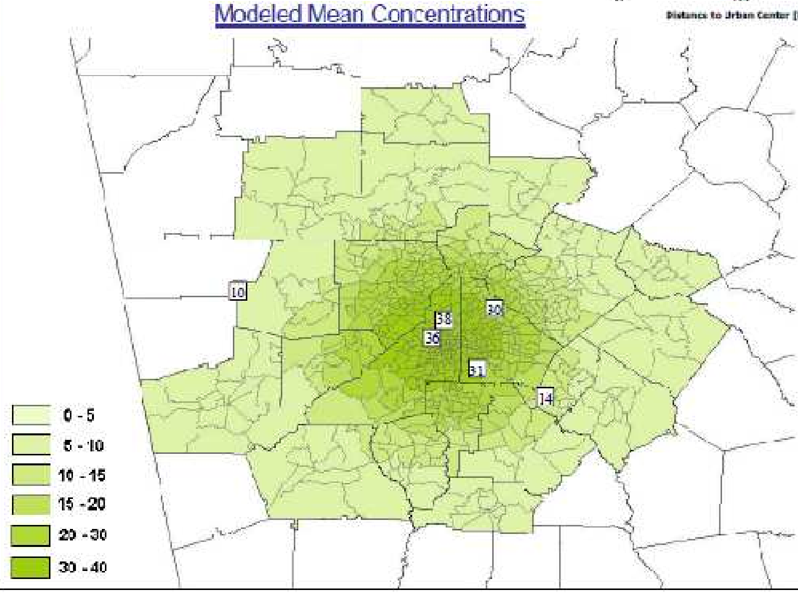
- Lower PM mass correlations for the GT PM_{10} monitor and FS8 $PM_{2.5}$ and PM_{10} monitors
- GT PM_{10} monitor is TEOM others are FRM
- FS8 is near railyard

Results

NO_2 [1-hr max]

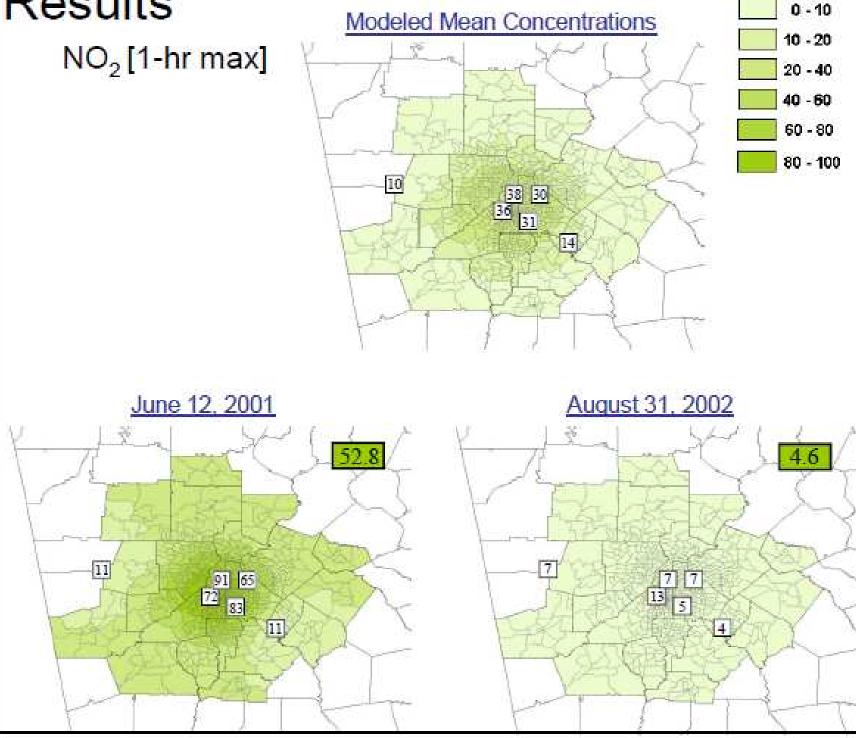


Modeled Mean Concentrations



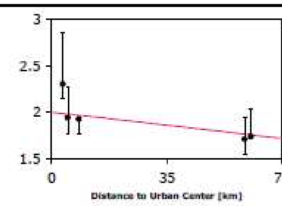
Results

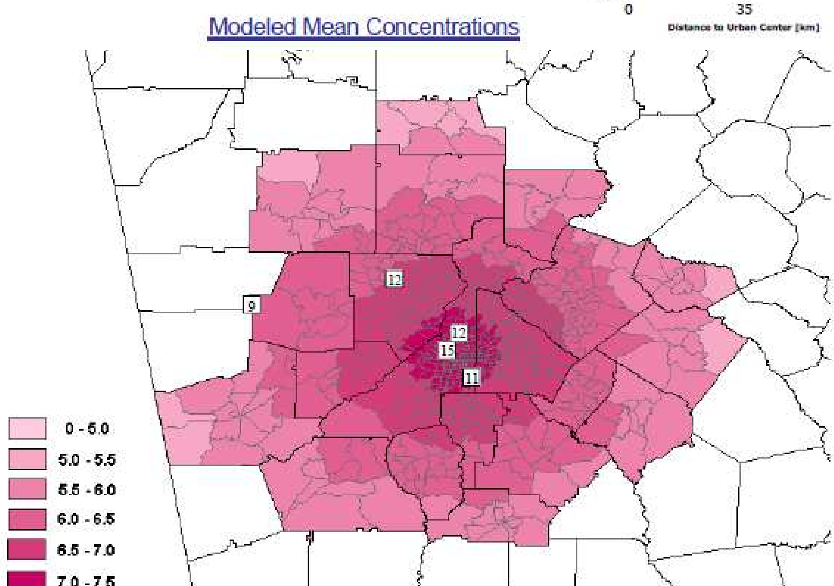
NO$_2$ [1-hr max]

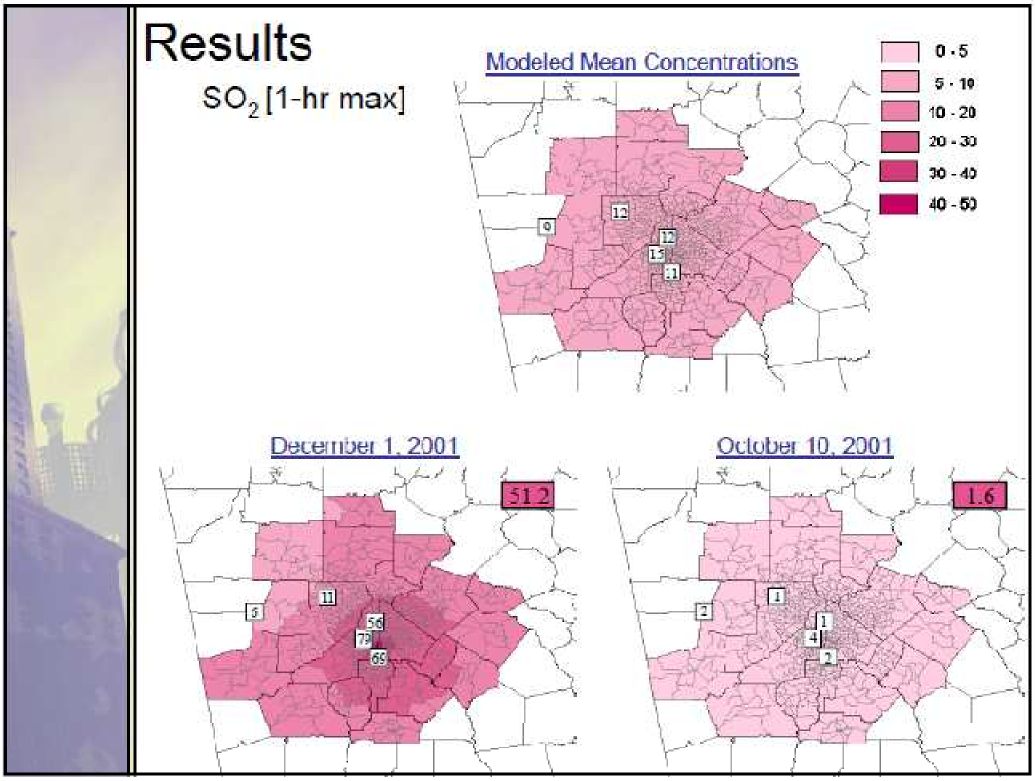


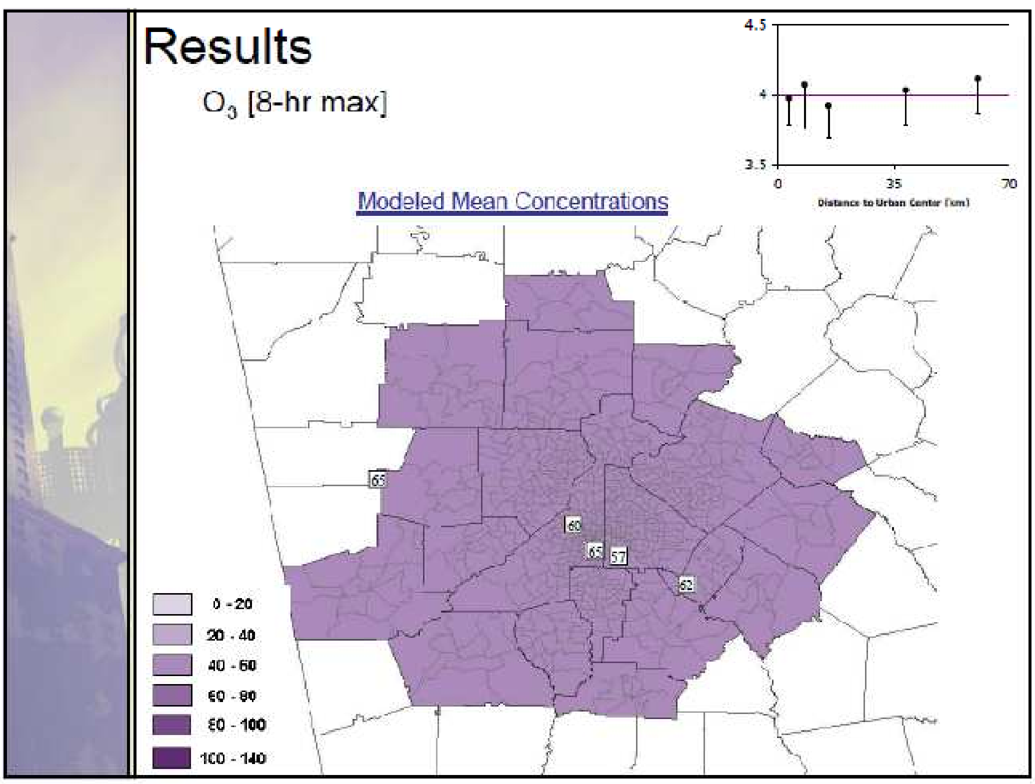
Results

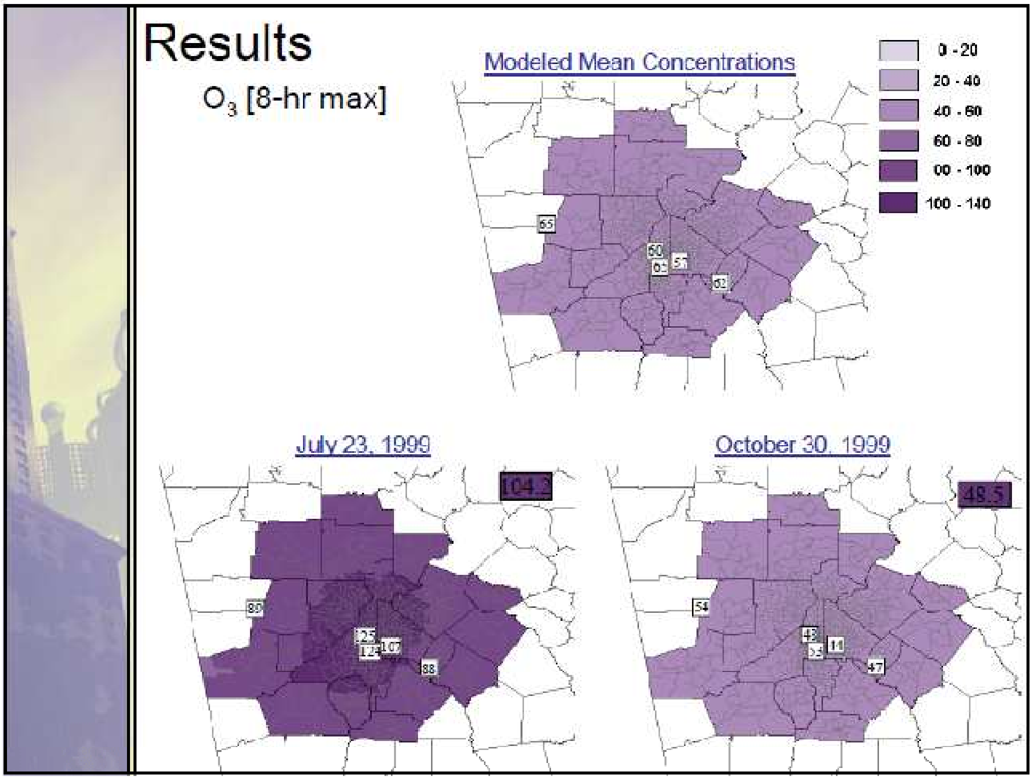
SO$_2$ [1-hr max]

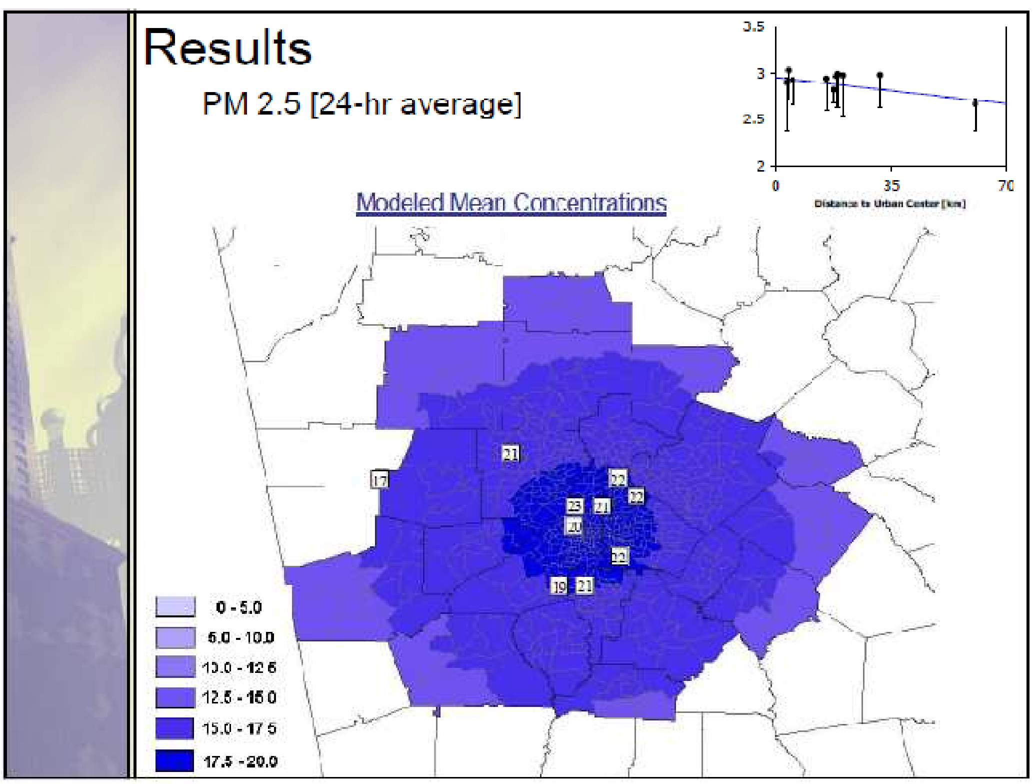





Results
SO2 [1-hr max]
Modeled Mean Concentrations
0 - 5
5 - 10
10 - 20
20 - 30
30 - 40
40 - 50
9
12
12
15
11
December 1, 2001
51.2
5
11
56
79
69
October 10, 2001
1.6
2
1
1
4
2


Results
O3 [8-hr max]
4.5
4
3.5
0
35
70
Distance to Urban Center (km)
Modeled Mean Concentrations
65
60
65
57
62
0 - 20
20 - 40
40 - 60
60 - 80
80 - 100
100 - 140

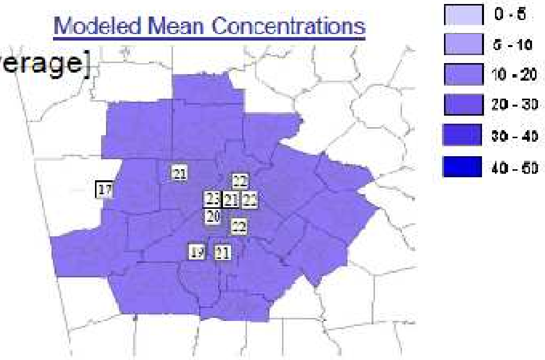

Results
O_3 [8-hr max]
Modeled Mean Concentrations
0 - 20
20 - 40
40 - 60
60 - 80
80 - 100
100 - 140
July 23, 1999
October 30, 1999
48.5


Results
PM 2.5 [24-hr average]
3.5
3
2.5
2
0
35
70
Distance to Urban Center [km]
Modeled Mean Concentrations
0 - 5.0
5.0 - 10.0
10.0 - 12.5
12.5 - 15.0
15.0 - 17.5
17.5 - 20.0

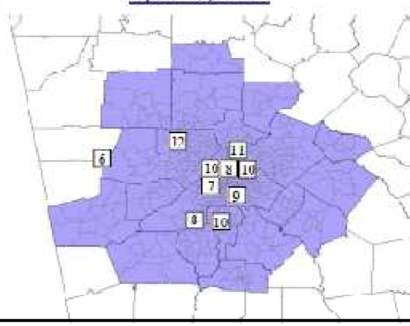
Results

PM 2.5 [24-hr average]

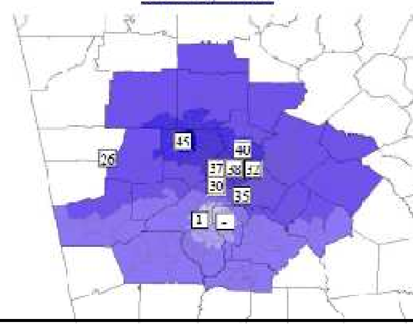
Modeled Mean Concentrations



April 12, 1999



June 5, 1999



Results

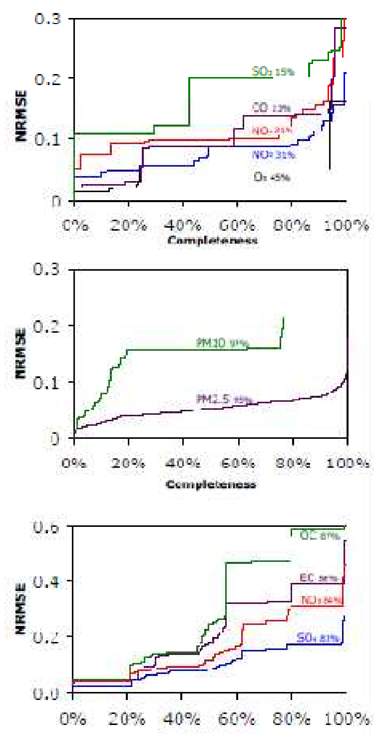
Error vs. Completeness

$$NRMSE = \frac{1}{\bar{y}}\sqrt{\frac{\sum_{i=1}^{N}(y_i - x_i)^2}{N}}$$

- O_3 can be computed with least error

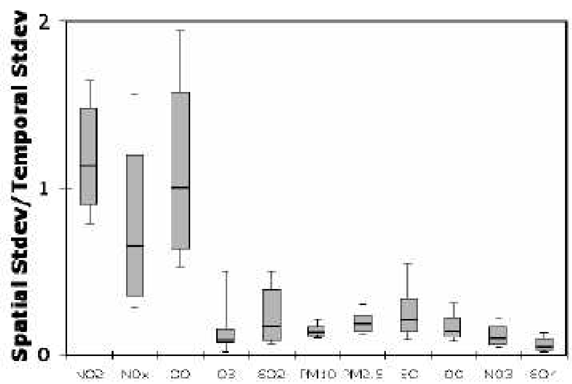
- SO_2 has the highest NRMSE of all gases - due to the spatial heterogeneity of SO_2 ambient levels

- For $PM_{2.5}$ components, the largest NRMSE is associated with EC and OC



Results

Spatial Variance



- Spatial variations of NO_2, NO_x and CO, whose primary source is vehicle emissions, are high relative to their temporal variations
- Spatially resolved measures for these pollutants might provide better indicators of exposure

Conclusions

1. Method provides a daily representative measure of ambient air pollution
2. Measure performs well for secondary and primary pollutants - SO_2 is the worst because it is difficult to spatially characterize
3. Descriptive models perform well --> high correlation and low bias between the actual monitor data and the modeled data
4. Measure allows for the improvement of completion of data sets at the cost of a low root mean square error

Appendix B

CMAQ Instruction Presentation

WRF-SMOKE-CMAQ

Yongtao Hu

yh29@mail.gatech.edu, ES&T 3230,404-385-4558
School of Civil and Environmental Engineering

Where you can find more information

- The Advanced Research WRF (ARW)
  - http://www.mmm.ucar.edu/wrf/users/
- SMOKE
  - http://www.smoke-model.org
- CMAQ
  - http://www.cmaq-model.org/

What you will find there: Expanded model names

Installation package with source code, supporting libraries and test case

Documentations: user manual, scientific description (equations)

Helpdesk, user email list and user forum: questions and answers

Is CMAQ modeling hard to learn: yes and no

- You have to be well prepared.
- The knowledge/skills you should have:
  - Linux/Unix: basic and advanced commands, c-shell programming, editor tool: vi or emacs
  - Compilers and etc.: pgi and ifort, makefile, compiler options and optimization, concurrent version system (CVS)
  - Language and libraries: fortran (mainly), c (a little), ioapi, netcdf, and mpi (for parallel).
  - Fortran programming skills with IOAPI (NETCDF libraries)
  - Visualization and other post-processing tools: PAVE, NCL etc.
  - Science back ground: atmospheric physics and chemistry, weather and pollution
  - Specific knowledge about grid models: single grid calculation
  - Map projection
  - Emissions inventory

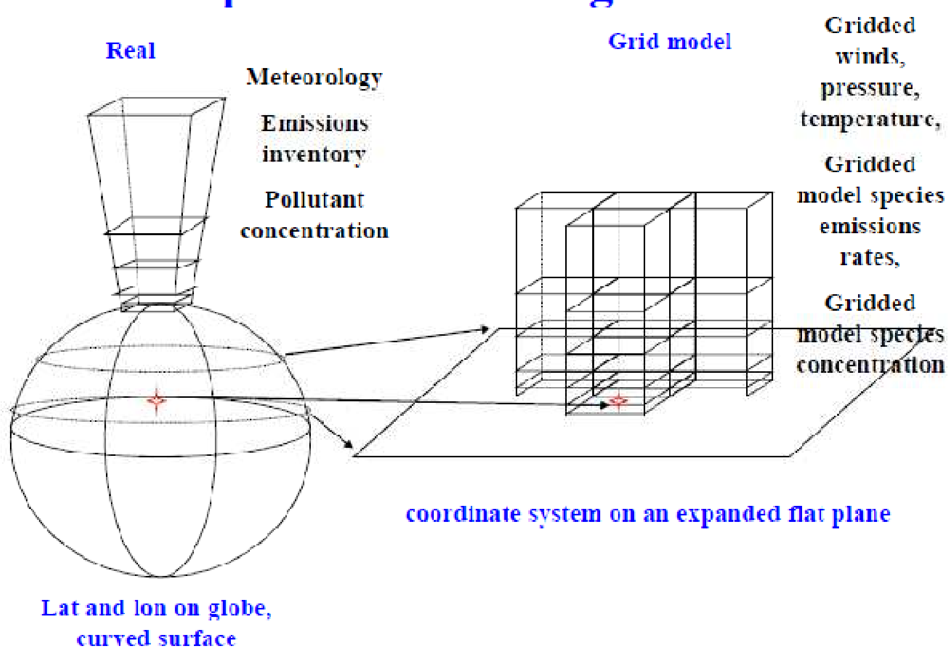
Why do you need CMAQ modeling?

- Think about it before you start to learn hard.
- Is it the only tool that can solve your issue? What is your research goal.
- What about other choices: global models, dispersion models or even a box model etc.
- How much time you can spend on it. Consider collaboration. Split met, emission and air quality modeling tasks.

Installation of the models

- Recommend Redhat Linux, that's what the community has the most experience with.
- Recommend ifort v10 and older. Not newer. PGI is more expensive.
- Install NetCDF 3 not 4. Install IOAPI 3.0. Install MPICH2.
- Follow README to install WPS and WRF, choose compile option for parallel with ifort
- SMOKE, install the executables, it should work on both 32 and 64bit platforms.
- CMAQ, the easiest.
- Tips: search through the website for compiler option on a specific OP system, asks the community mail list.

One more piece of knowledge before start



How to start a CMAQ modeling? (1)

- Overall scientific design of the project should have solved some issues:
  - If CMAQ should be used
  - Size of the modeling domain
  - what grid resolution should be used, 4-km or 1km?
  - Episode selection: two weeks or years, history or future.

How to start a CMAQ modeling? (2)

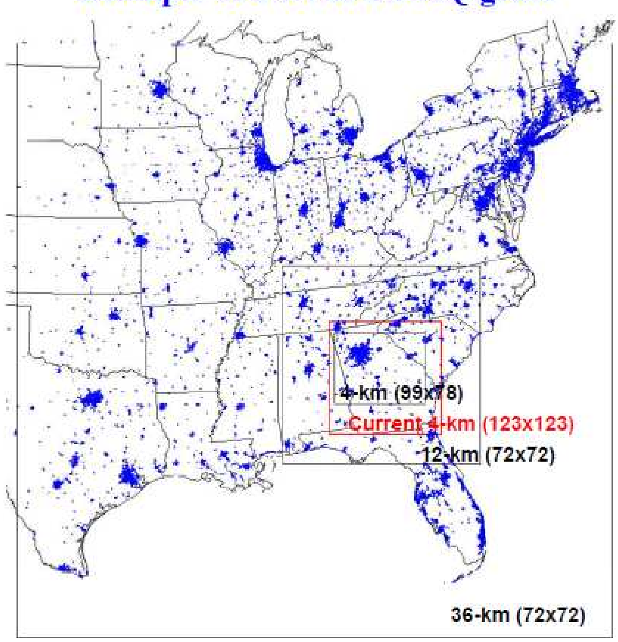
- Input datasets collection:
  - WRF
    - Large (global) scale input meteorological fields to WRF: history (global and regional reanalysis), forecast (global GFS, regional NAM etc), future (GISS model-E).
    - Landuse data: history data came with WRF, future
    - Stationary meteorological observation data for 4DDA for historical episode
  - SMOKE
    - Emissions inventory
    - Spatial surrogates for allocate emissions into grid cells
    - Temporal (monthly, weekly, and diurnal) variation of emisisons
    - Chemistry speciation: splits of VOC and PM2.5
  - CMAQ
    - Initial and boundary conditions for model species.

Grid design: consistency between WRF and SMOKE/CMAQ

- Horizontal
  - Make sure the same map projection and coordinate system used for both WRF and SMOKE/CMAQ:
    - In WRF, mother domain's center is the center of the coordinate system.
    - Easier to define grid in SMOKE/CMAQ: parameters for map projection, origin of the coordinate system, origin of the grid defined, col and row number, grid cell spacing.
    - Start with define CMAQ grids and then decide how to set mother domain and the offset of the daughter domain in WRF.
  - Check the grid definition with plotting in ArcGIS or NCL.
  - Slightly larger domain for WRF, e.g. three cell more on each side
  - Note a difference between: node and col/row, dot point and cross point.
  - Avoid grid lateral boundary across any population dense area.
  - Nesting of the grids to get high resolution

X

Example of a set of CMAQ grids



Grid design: consistency between WRF and SMOKE/CMAQ

- ➢Vertical
  - ➢Choose a type: Z or P or sigma-p
    - ➢Sigma-p: p-p^{top}/(p^{botom}-p^{top})
  - ➢Denser layers in the PBL
  - ➢34 layers/35 levels in WRF is preferable.
  - ➢Can be less layers in CMAQ than WRF, but should select direct from the defined WRF layers instead of any interpolation

Example of vertical layers setting

WRF					24 layer CMAQ				
Level	Sigma	Height (m)	Pressure (mb)	Depth (m)	Level	Sigma	Height (m)	Pressure (mb)	Depth (m)
35	0	18663	50	2034	25	0	18663	50	3749
34	0.0332	16629	82	1715					
33	0.0682	14914	115	1515	24	0.0682	14914	115	2890
32	0.1056	13399	150	1375					
31	0.1465	12024	189	1255	23	0.1465	12024	189	2400
30	0.1907	10769	231	1145					
29	0.2378	9624	276	1045	22	0.2378	9624	275	2000
28	0.2871	8579	323	955					
27	0.3379	7624	371	870	21	0.3379	7624	371	1660
26	0.3895	6754	420	790					
25	0.4409	5964	469	715	20	0.4409	5964	469	1360
24	0.4915	5249	517	645					
23	0.5406	4604	564	580	19	0.5406	4604	564	1100
22	0.5876	4024	608	520					
21	0.6323	3504	651	465	18	0.6323	3504	651	880
20	0.6742	3039	690	415					
19	0.7133	2624	728	370	17	0.7133	2624	728	700
18	0.7494	2254	763	330					
17	0.7828	1924	794	293	16	0.7828	1924	794	552
16	0.8133	1631	823	259					
15	0.841	1372	849	228	15	0.841	1372	849	228
14	0.8659	1144	873	200	14	0.8659	1144	873	200
13	0.8882	944	894	174	13	0.8882	944	894	174
12	0.9079	770	913	150	12	0.9079	770	913	150
11	0.9252	620	929	128	11	0.9252	620	929	128
10	0.9401	492	943	108	10	0.9401	492	943	108
9	0.9528	384	953	90	9	0.9528	384	955	90
8	0.9635	294	963	74	8	0.9635	294	965	74
7	0.9723	220	974	60	7	0.9723	220	974	60
6	0.9796	160	981	48	6	0.9796	150	981	48
5	0.9854	112	986	38	5	0.9854	112	985	38
4	0.99	74	991	30	4	0.99	74	991	30
3	0.991	44	994	24	3	0.991	44	994	24
2	0.9971	20	998	20	2	0.9971	20	998	20
1	1	0	1000		1	1	0	1000	

Run WRF

- **WPS**: pre-process:
  - geogrid.exe: get landuse, terrain etc. static grid information
  - ungrib.exe: obtain large scale first-guess met fields for the period
  - metgrid.exe: horizontal interpolation onto defined grid cells
  - share "namelist.wps": set inputs, define grid, set duration
- **WRF**:
  - real.exe: vertical interpolation, create direct input files for WRF
  - wrf.exe: meteorology forecast/prediction
  - share "namelist.input": set parameters, physics options, set timestep and duration, define grids etc.

Example of namelist.wps and namelist.input (portion) for WRF

```
&share
 wrf_core = 'ARW',
 max_dom = 4,
 start_date = '2009-04-29_12:00:00','2009-04-29_12:00:00',
 end_date   = '2009-04-29_18:00:00','2009-04-29_18:00:00',
 interval_seconds = 10800
 io_form_geogrid = 2,
 opt_output_from_geogrid_path = '/cee/ted_raid8a/data/ythu/WRF_3.2.1/WPS/geo_wf',
 debug_level = 0
/

&geogrid
 parent_id         =   1,   1,  2,  3
 parent_grid_ratio =   1,   3,  3,  3
 i_parent_start    =   1,  24, 35,  8
 j_parent_start    =   1,   4, 24,  8
 e_we              =  55,  79, 79, 172
 e_sn              =  43,  79, 79, 148
 geog_data_res     = 'modis_30s+10m','modis_30s+10m','modis_30s+5m','modis_30s+2m',
 dx = 108000,
 dy = 108000,
 map_proj = 'lambert',
 ref_lat   =  40.00,
 ref_lon   = -97.00,
 truelat1  =  33.0,
 truelat2  =  45.0,
 stand_lon = -97.0,
 geog_data_path = '/cee/ted_raid8a/data/ythu/WRF_3.2.1/WPS/geog'
/

&ungrib
 out_format = 'WPS',
 prefix = 'FILE',
/

&metgrid
 fg_name = 'FILE'
 io_form_metgrid = 2,
/

&mod_levs
 press_pa = 201300 , 200100 , 100000 ,
             95000 ,  90000 ,
             85000 ,  80000 ,
             75000 ,  70000 ,
             65000 ,  60000 ,
             55000 ,  50000 ,
             45000 ,  40000 ,
             35000 ,  30000 ,
             25000 ,  20000 ,
             15000 ,  10000 ,
              5000 ,   1000
/
```

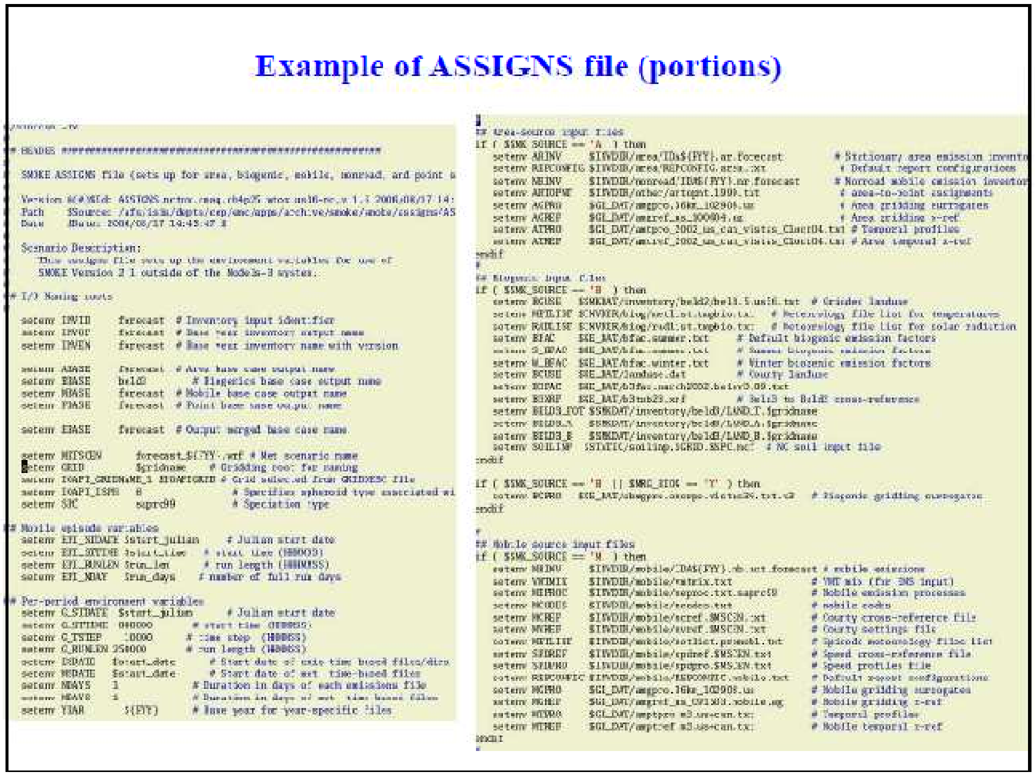
```
&physics
mp_physics                  = 2,      2,      2,
ra_lw_physics               = 1,      1,      1,
ra_sw_physics               = 1,      1,      1,
radt                        = 10,     10,     10,
sf_sfclay_physics           = 1,      1,      1,
sf_surface_physics          = 2,      2,      2,
bl_pbl_physics              = 1,      1,      1,
bldt                        = 0,      0,      0,
cu_physics                  = 1,      1,      0,
cudt                        = 5,      15,     60,
isfflx                      = 1,
ifsnow                      = 0,
icloud                      = 1,
surface_input_source        = 1,
num_soil_layers             = 4,
num_land_cat                = 20,
mp_zero_out                 = 0,
maxiens                     = 1,
maxens                      = 3,
maxens2                     = 3,
maxens3                     = 16,
ensdim                      = 144,
/

&fdda
grid_fdda                   = 1,      0,      0,
gfdda_inname                = "wrffdda_d<domain>",
gfdda_end_h                 = 240,    240,    240,
gfdda_interval_m            = 180,    180,    180,
fgdt                        =  0,     0,      0,
if_no_pbl_nudging_uv        =  0,     0,      0,
if_no_pbl_nudging_t         =  1,     1,      1,
if_no_pbl_nudging_q         =  1,     1,      1,
if_zfac_uv                  =  0,     0,      0,
 k_zfac_uv                  =  8,     8,      8,
if_zfac_t                   =  0,     0,      0,
 k_zfac_t                   =  8,     8,      8,
if_zfac_q                   =  0,     0,      0,
 k_zfac_q                   =  8,     8,      8,
guv                         = 0.0003, 0.0003, 0.0003,
gt                          = 0.0003, 0.0003, 0.0003,
gq                          = 0.0003, 0.0003, 0.0003,
if_ramping                  = 0,
dtramp_min                  = -60.0,
io_form_gfdda               = 2,
grid_sfdda                  = 1,
sgfdda_inname               = "wrfsfdda_d<domain>"
sgfdda_end_h                = 240
sgfdda_interval_m           = 360
io_form_sgfdda              = 2
guv_sfc                     = 0.0003
gt_sfc                      = 0.0003
```

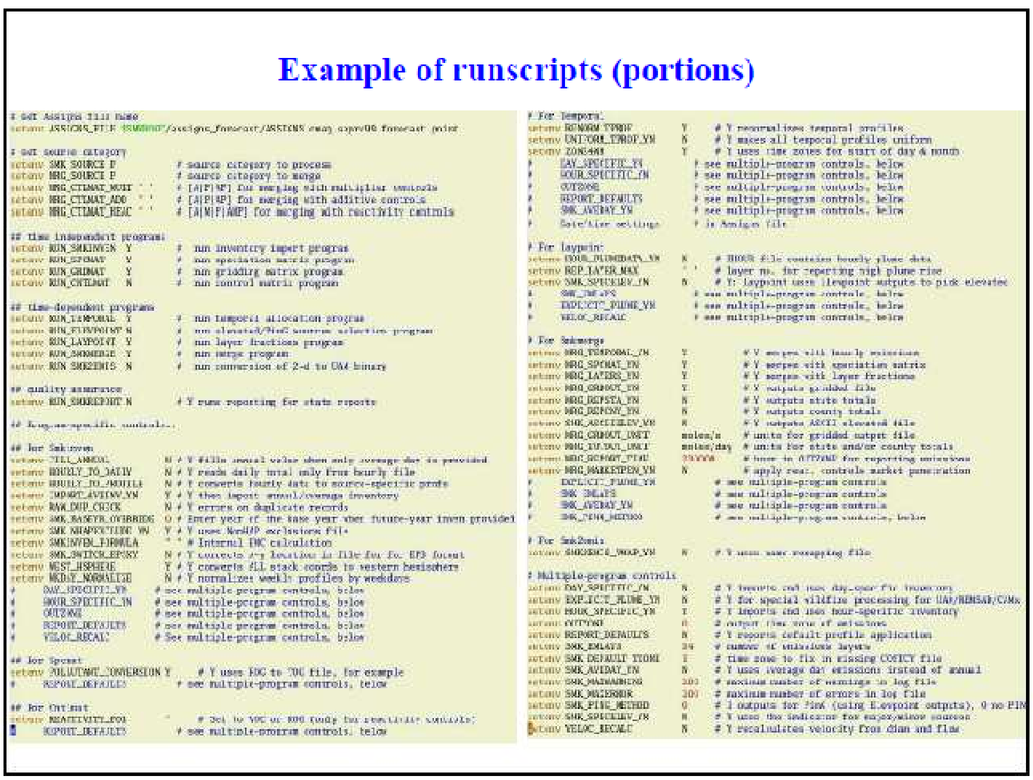
Model Configurations (1):

- WRF (through namelists):
  - Landuse data: USGS or MODIS, in WPS geogrid.
  - Simulation duration: related WPS metgrid
  - Vertical definition conducted using real.exe
  - Timestep: 6xgrid spacing.
  - Physics options: PBL scheme, microphysics, cumulus cloud scheme, radiation scheme etc.
  - Land surface model: Noah LSM, Pleim-Xiu.
  - 4DDA: grid nudging, spectral nudging
  - Output timestep: hourly
  - Split output files day by day
  - MCIP: passing variables through

Run SMOKE

- A series of core programs depending on source type
  - SMKINVEN: read in raw inventory, criteria pollutants
  - SPCMAT: speciation to model species
  - GRDMAT: spatial allocation to grid cells
  - TEMPORAL: temporal split to each hour
  - SMKMERGE: generate gridded hourly emission rates for model species in ioapi (netcdf) format
- ASSIGNS file: set filenames for inputs, intermediate and outputs files
- Runscripts:
  - Smk_ar.csh, smk_bg.csh, smk_pt.csh, smk_nr.csh, smk_mb.csh: Choose programs need to run, set choices for each program, set other parameters.


Example of ASSIGNS file (portions)

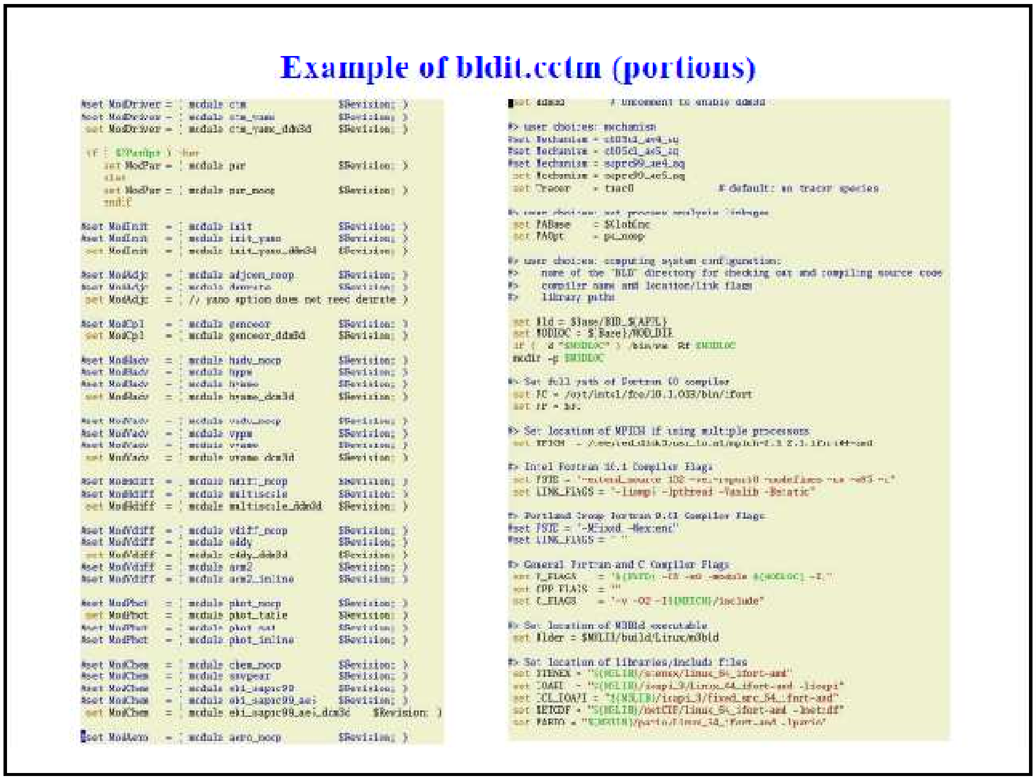

Example of runscripts (portions)

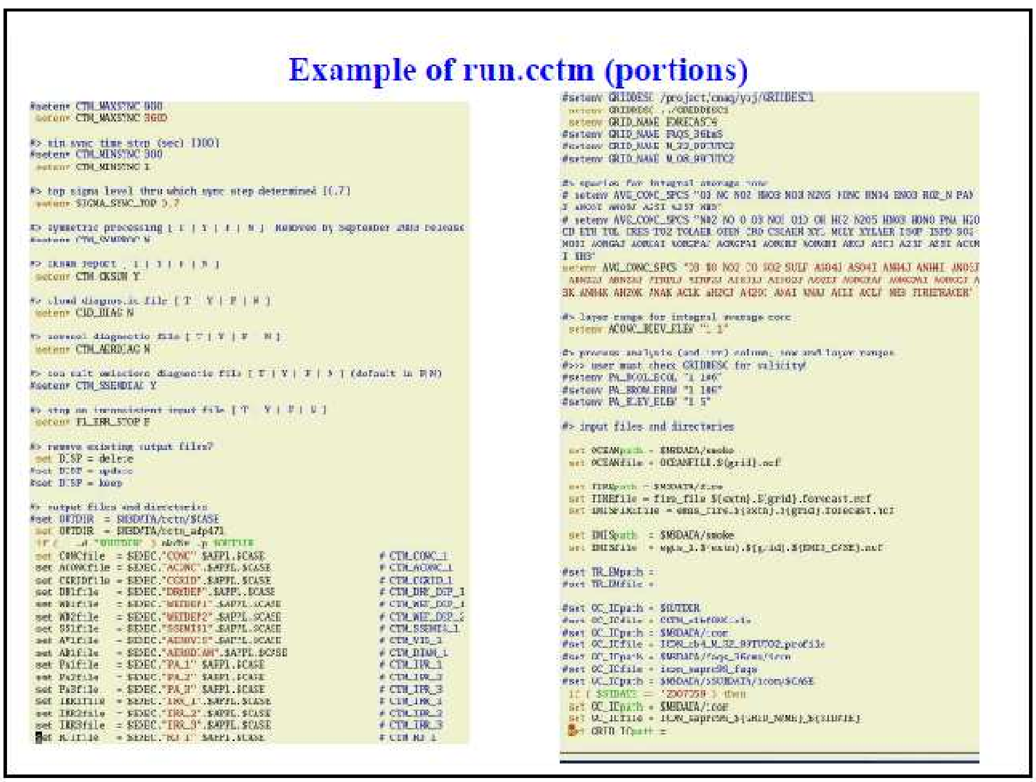
Model Configurations (2):

- ➢SMOKE (through assign file but hidden):
  - ➢Gas phase mechanism: SAPRC99 or SAPRC07, CB4 or CBO5, chemical speciation profiles and cross reference file.
  - ➢Grid definition: GRIDESC, pre-prepared spatial surrogate files for each defined grid

Run CMAQ

- ➢preprocess:
  - ➢WRF and SMOKE: meteorology and emissions
  - ➢MCIP: pass or further derive met variables needed by CMAQ, met fields in ioapi format
  - ➢ICON: generate initial air quality conditions in ioapi format
  - ➢BCON: generate lateral boundary conditions in ioapi format
  - ➢JPROC: generate look-up table of photolysis rates
- ➢CCTM (CMAQ):
  - ➢bldit.cctm script: configure module assembles
  - ➢run.cctm script: set filename and paths for inputs and outputs, environment variables for model choices, air quality modeling


Example of bldit.cctm (portions)
#> user choices: mechanism
#> user choices: computing system configuration:
#> Set full path of Fortran 90 compiler
#> Set location of MPICH if using multiple processors
#> Intel Fortran 10.1 Compiler Flags
#> General Fortran and C Compiler Flags
#> Set location of libraries/include files


Example of run.cctm (portions)
#> input files and directories

Model Configurations (3):

- CMAQ (through bldit script):
  - Gas phase mechanism: SAPRC99 or SAPRC07, CB4 or CBO5, different include file will be compiled.
  - Advection scheme: YAMO
  - Cloud module: ACM
  - Aerosol module: AERO4, AERO5
  - ICON, BCON, JPROC: mechanism consistency
  - Run CMAQ day by day, to control size of the outputs

Kick off your jobs:

- C-shell scripts for SMOKE and CMAQ
- Check your jobs: commands “top”, “ps –elf”
- Monitor the log files
- A useful commend: screen, a always alive terminal for keeping your jobs running without any interrupts.
- Another useful commend: crontab, for auto start of your jobs: forecast.

Post processing, evaluation and analysis:

- First check the log files if there is no error messages
- PAVE for spatial visualization, check all major met, emissions and air quality variables.
- IOAPI tools for data extraction, 8-hr or 24-hr averaging, sum up the models species to measured pollutants
- WRF evaluation: TDL data and METSTAT

http://www.camx.com/down/support.php

- CMAQ evaluation: STN, EPA AQS data, SEARCH, IMPROVE etc., with AMET (www.cmascenter.org) or self made codes.

Appendix C: AERMOD Instruction Notes

AERMOD is a steady-state plume model that incorporates air dispersion based on planetary boundary layer turbulence structure and scaling concepts, including treatment of both surface and elevated sources, and both simple and complex terrain. AERMOD was developed by a collaborative working group of scientists from American Meteorological Society (AMS) and Environmental Protection Agency (EPA). AERMOD estimates the contributions from point, area and volume sources, and is primarily used for regulatory compliance modeling. AERMOD model is PC-c compatible and requires a minimum of 2 MB of RAM, a math processor, and MS-DOS version 3.2 or higher. Good working knowledge on programming/editing batch scripts is required to run the command-line version of AERMOD.

AERMOD modeling system consists of two data input preprocessing systems, namely the terrain preprocessor (AERMAP) and the meteorology preprocessor (AERMET). AERMAP processes complex terrain using USGS Digital Elevation Data. AERMOD requires two types of meteorology inputs that are processed using AERMET. One file consists of surface scalar parameters, and the other file consists of vertical profiles of metrological data. Executing a simple AERMOD run involves setting up a runstream file, which involves the following steps.

1. Selecting modeling options: This is the first step and involves creating an output directory, selecting a pollutant and averaging period for that pollutant.
2. Specifying source inputs: This step involves identifying the location, type of source and other source specific parameters.
3. Specifying a receptor network: This step is necessary to identify a cartesian grid receptor network or a discrete receptor location.
4. Specifying the meteorological input: This step provides surface and vertical profile information necessary to run AERMOD.
5. Selecting output options: Includes keywords to produce output files, table options for generating plots.

Issues with procuring/running AERMOD:

The command line version is distributed free of cost by EPA, however, running AERMOD on command line involves the following drawbacks:

1. Processing input files, especially when there are multiple sources involved, is time consuming.
2. Very difficult to trace errors.
3. The visualization options that come with the command line version are very limited.

AERMOD that is built on a graphics user interface (GUI) is available through vendors such as Trinitiy Consultants and Lakes Environmental. The GUI versions currently in circulation are Breeze and Aermod View and a single user license costs somewhere between $1400-1700.

References: http://www.epa.gov/scram001/dispersion_prefrec.htm
http://www.epa.gov/ttn/scram/dispersion_prefrec.htm#aermod