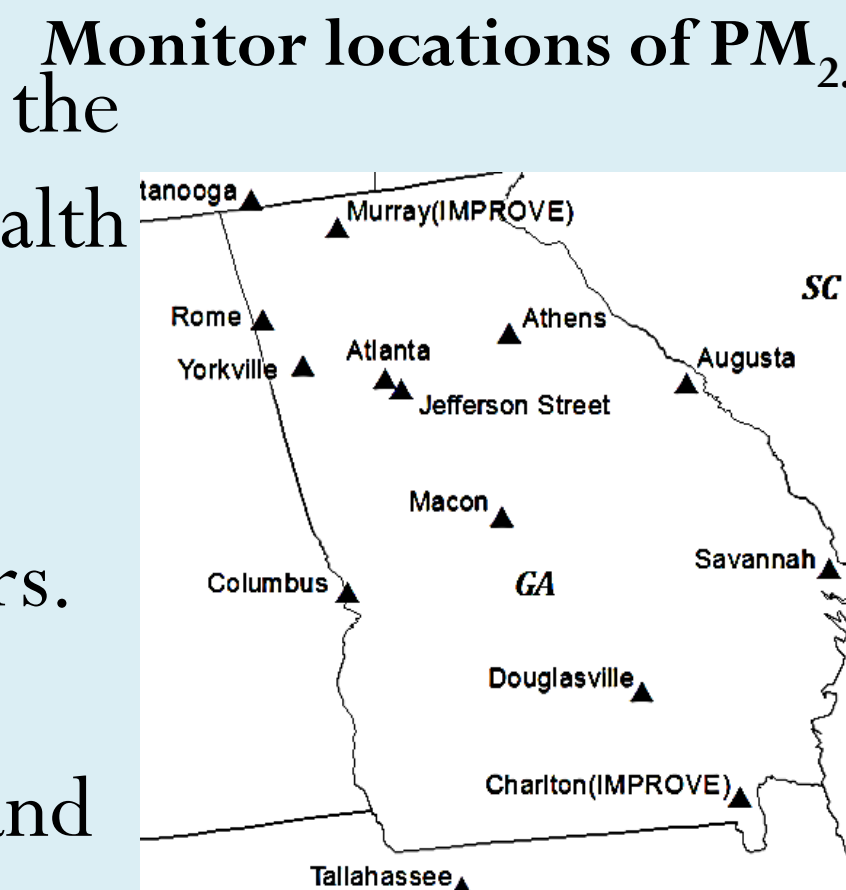# Spatiotemporal Error Assessment for Ambient Air Pollution Estimates obtained using an Observation-CMAQ Data Fusion Technique

XINXIN ZHAI, MARIEL FRIBERG, Heather Holmes, Yongtao Hu, James Mulholland, Armistead Russell

School of Civil and Environmental Engineering, Georgia Institute of Technology, Atlanta, GA
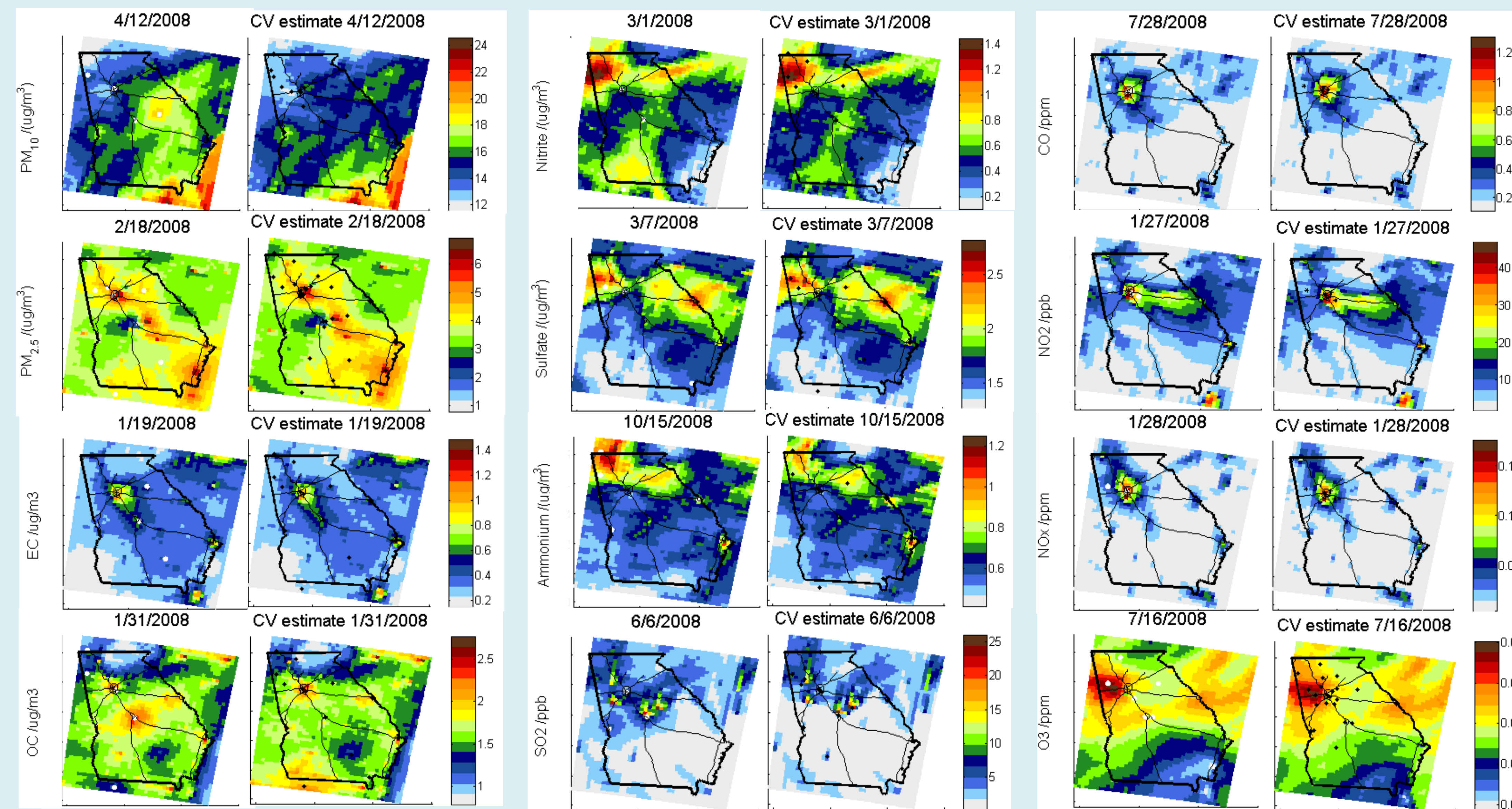
Southeastern Center for Air Pollution and Epidemiology (SCAPE)
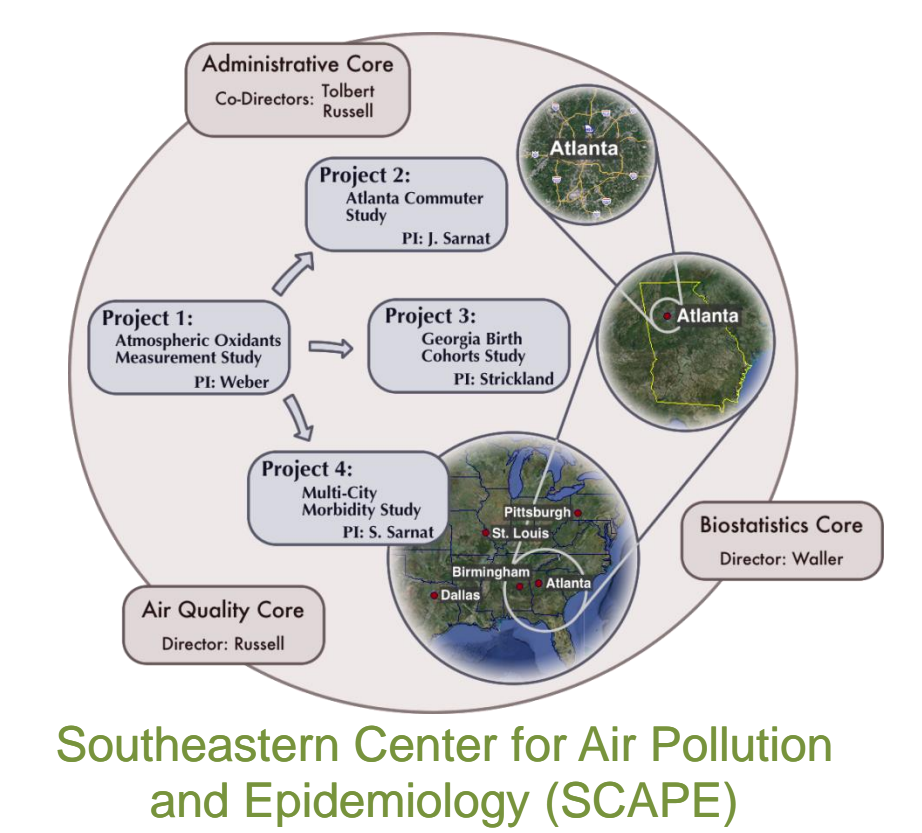
Motivation and Objective

Monitor locations of PM$_{2.5}$

- Spatially resolved daily exposure estimates are needed to investigate the relationship between geo-coded health data and ambient air quality.
- Observations are available from a sparse network of ambient monitors. Emission-based CMAQ model simulations are available at 12 km and 4 km resolutions.
- To minimize bias over space while maximizing prediction of variance over time (i.e., Pearson R2), we developed a data fusion methodology to provide daily concentration fields for 12 pollutants (CO, NO$_2$, NO$_x$, SO$_2$, O$_3$, PM$_{10}$, PM$_{2.5}$ and PM$_{2.5}$ constituents SO$_4$, NO$_3$, NH$_4$, EC, and OC) in the state of Georgia from 2002 through 2010.
- To evaluate the method, we applied both cross validation (CV) technique by withholding 10% of the observations and withhold-one-site technique.

OBS-CMAQ Data Fusion Method [1]

$$C^* = F C_1 + (1-F) C_2$$

$$\langle C \rangle = \alpha \langle CMAQ \rangle^\beta$$

$$F = \frac{R_1(1-R_2)}{R_1(1-R_2)+R_2(1-R_1)}$$

$$R^* = F R_1 + (1-F) R_2$$

(fields above for 1-hr max NO$_2$, ppb, 1/6/08, 4-km)

annual mean regression: $\langle C \rangle = 4.2 \langle CMAQ \rangle^{0.5}$

$R_1 = 0.92\, e^{-D/450}$

R_2: CMAQ-OBS

C_1: annual avg CMAQ field scaled to daily OBS
C_2: daily CMAQ field scaled to annual avg OBS
C^*: weighted avg concentration field
F: daily weighting factor field
R: Pearson correlation coefficient
D: distance from nearest OBS (km)

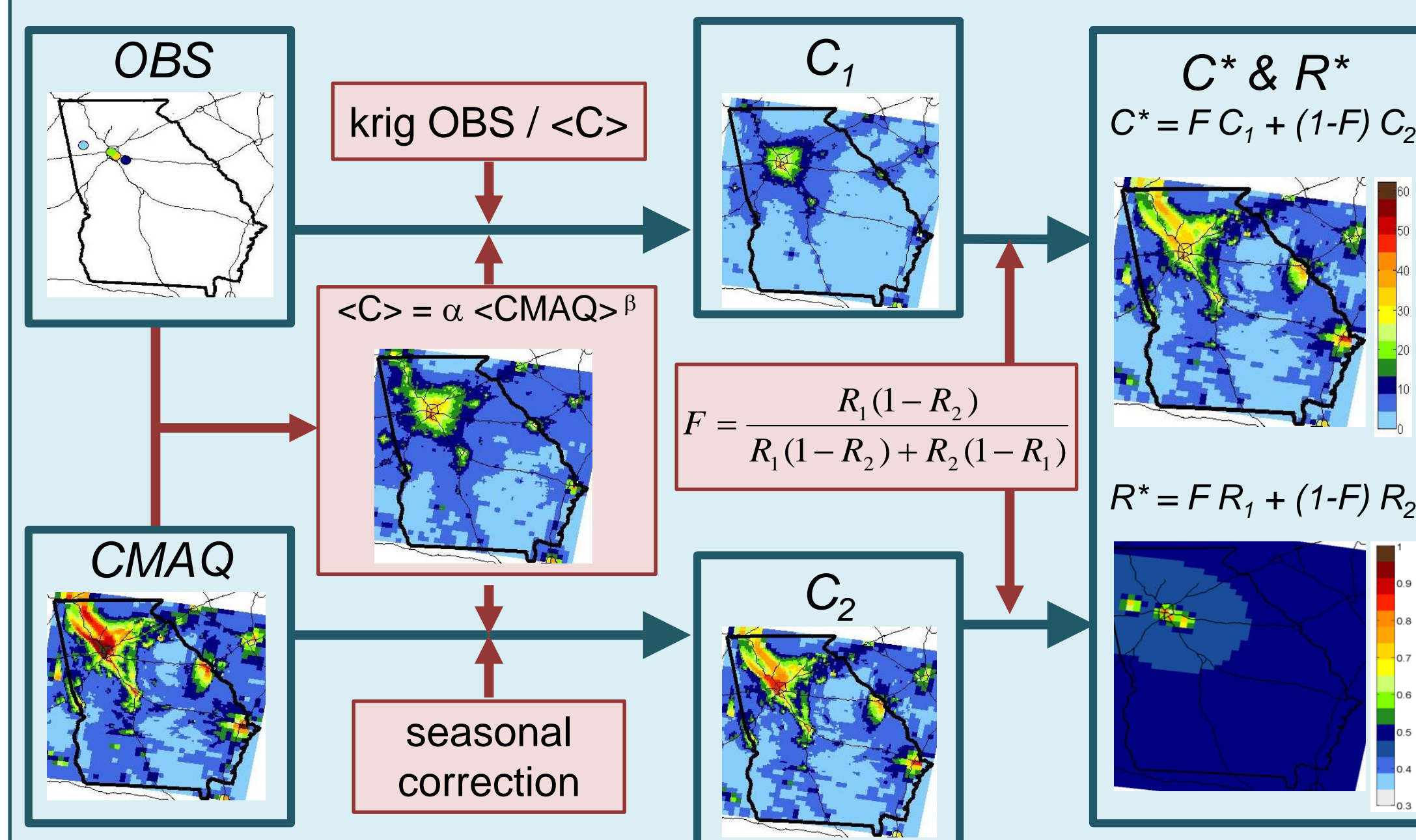## Error Assessment by Cross Validation (CV)

- For evaluation of the sensitivity to missing days over time and space, we randomly withheld 10% of the observations over each year and all sites and got estimates at chosen days to compare with observations.
- Both spatially and temporally, the missing 10% data has a slight impact on the fused data.

Spatial Distribution for Correlation Coefficient of Estimate vs OBS

- We estimated the correlation coefficient R2 between fused data and OBS. The correlations are better near to the monitors with sufficient data than the monitors with temporal gaps in measurements.
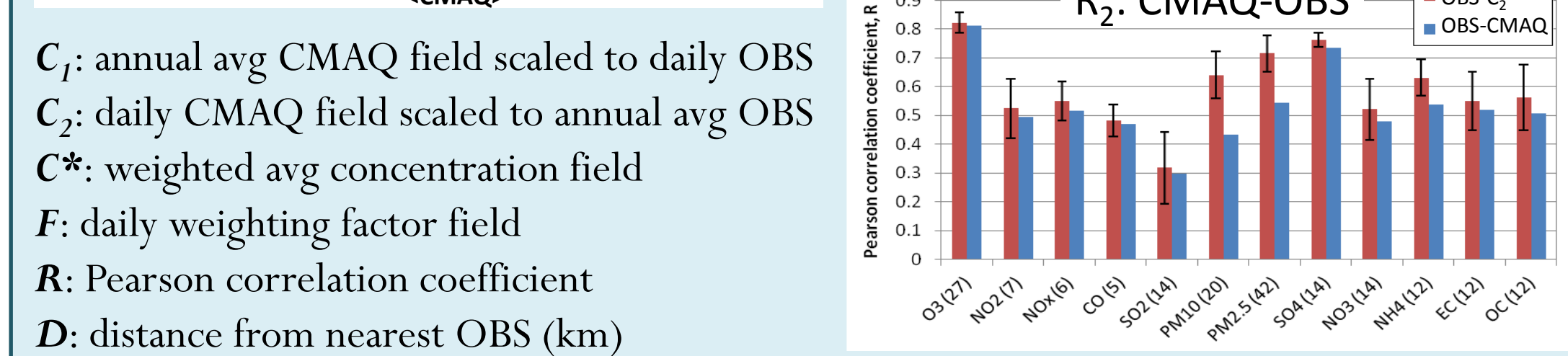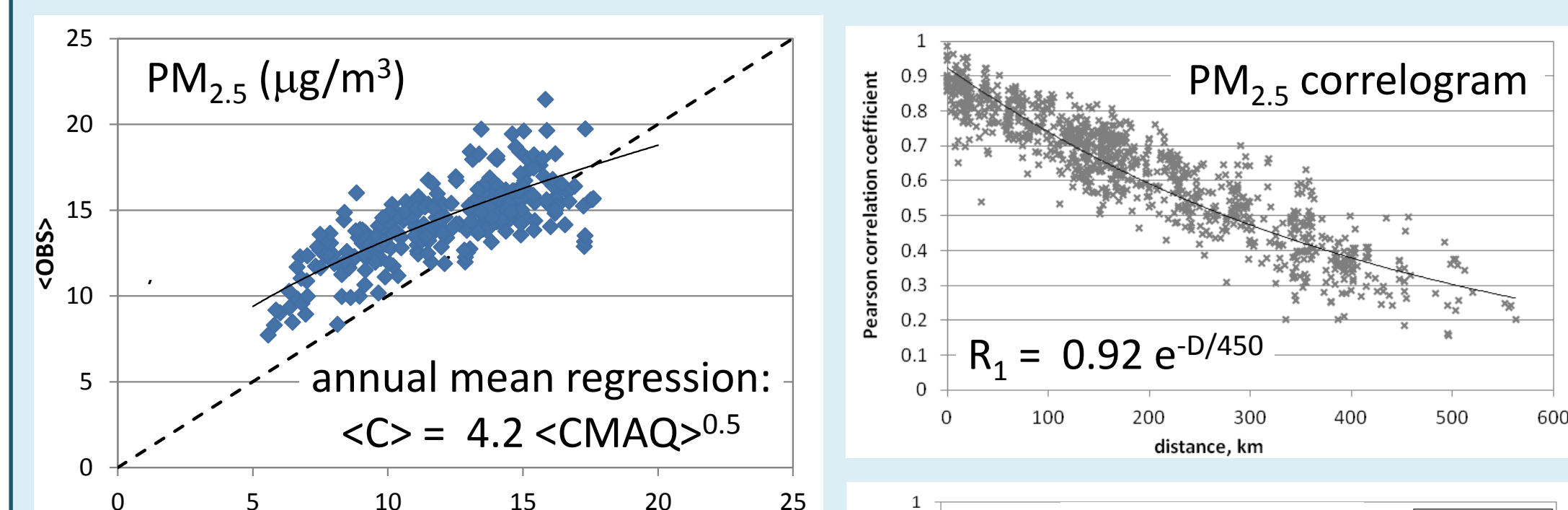
Estimated R^2 Annual Average

Error Assessment by CV

- We compared the spatial-temporal correlation with the withheld OBS and CMAQ estimates, estimated Data Fusion (DF-WH) results from CV, and the DF results.
- R2 values indicate that the data fusion method increases the temporal correlation with measurements while the CV results indicate that the estimates are more accurate than CMAQ at locations where there are no OBS.
- RMSE in Table 2 shows that the average bias decreases 30% to 80% for all species compared to CMAQ.
- CV analysis tends to overestimate the R^2 when the monitors are clustered. The mean of the estimate R^2 of all grid cells indicates that the overall correlation is actually lower than the withheld dataset.
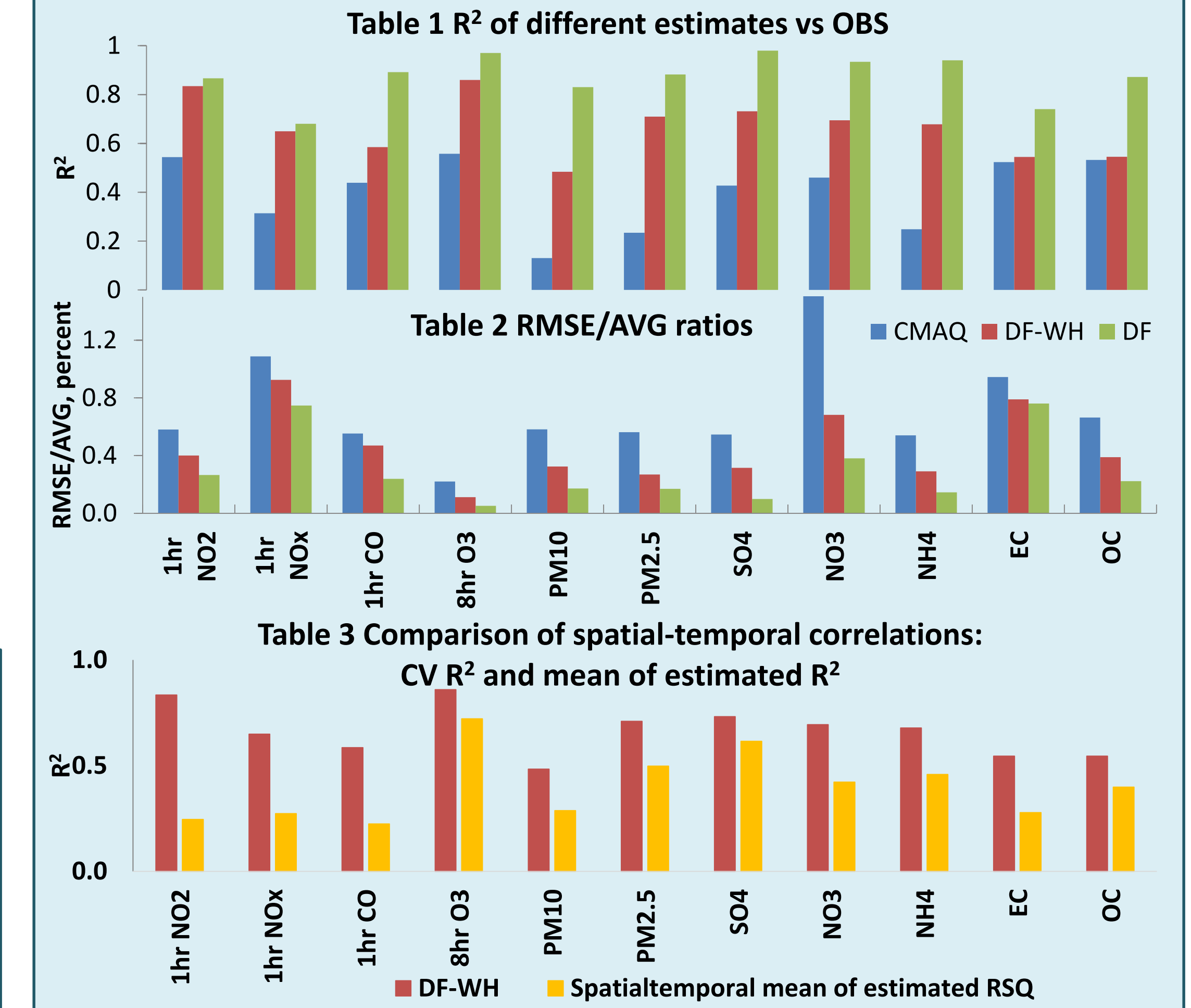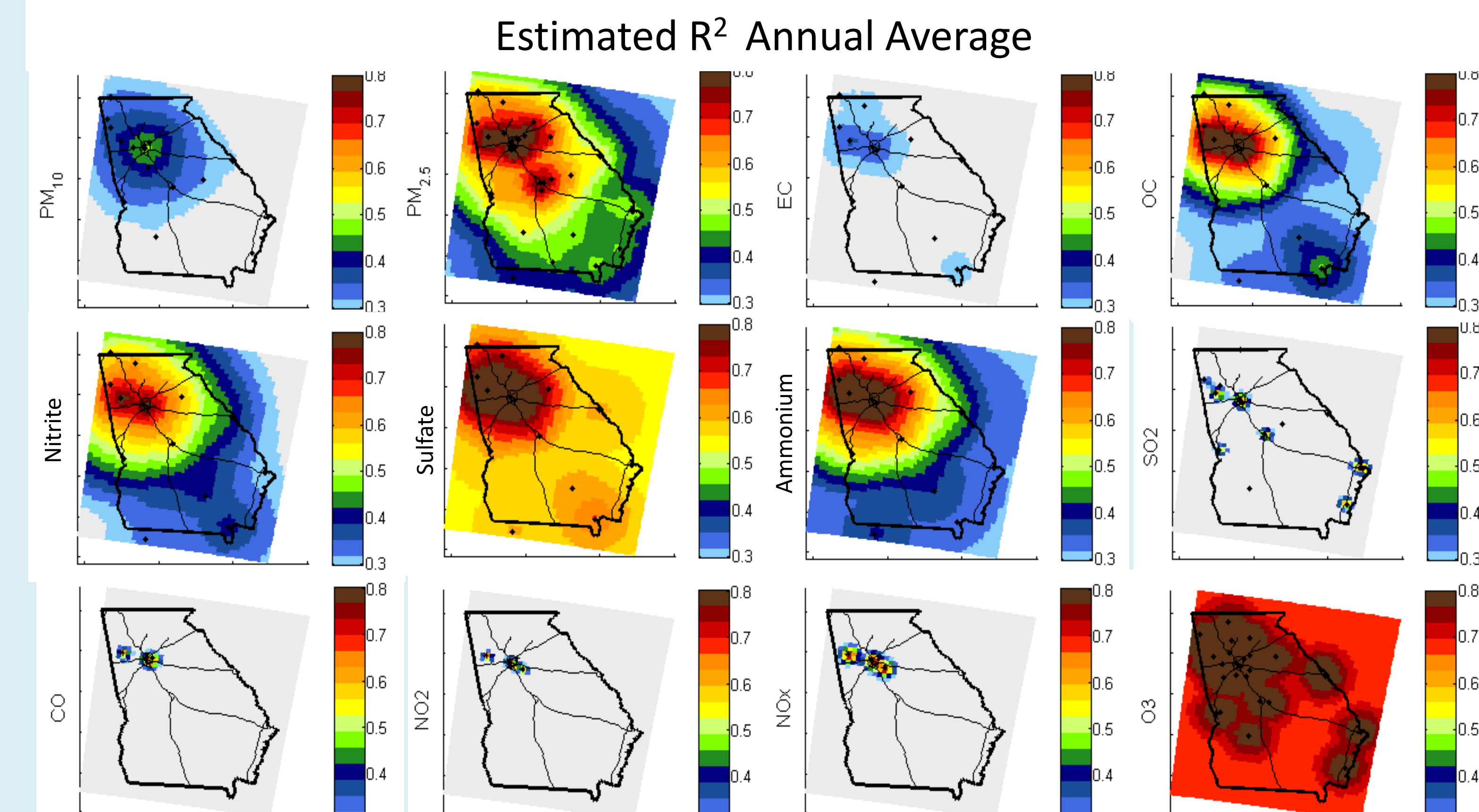
Table 1 R^2 of different estimates vs OBS

Table 2 RMSE/AVG ratios

Table 3 Comparison of spatial-temporal correlations: CV R^2 and mean of estimated R^2

References

1. Friberg, et al. "Spatiotemporal air quality metrics developed for Georgia for use in health studies." 107th Air & Waste Management Association's Annual Conference & Exhibition. Long Beach, California, June 2014.

Acknowledgements