The University of Chicago
Center for Integrating Statistical and Environmental Science
www.stat.uchicago.edu/~cises

Chicago, Illinois USA

TECHNICAL REPORT NO. 40

STATISTICAL COMPARISON OF OBSERVED AND MULTI-RESOLUTION CMAQ MODELED HOURLY OZONE CONCENTRATIONS

Li Chen, Michael L. Stein, Alexis Zubrow

December 2006

This research is funded by
U.S. EPA - Science To Achieve Results (STAR) Program
Grant # R-82940201-0

Statistical Comparison of Observed and Multi-Resolution CMAQ Modeled Hourly Ozone Concentrations

Li Chen*
CISES, The University of Chicago, Chicago, IL, USA

Michael L. Stein
The Department of Statistics, The University of Chicago, Chicago, IL, USA

Alexis Zubrow
CISES, The University of Chicago, Chicago, IL, USA

Abstract. This paper compares observed hourly ozone concentrations to the Community Multiscale Air Quality (CAMQ) modeling system modeled hourly ozone concentrations at different spatial resolutions. Some performance measures, e.g., fractional bias (FB) and root normalized mean squared error (RNMSE), are calculated. Comparisons of CMAQ model runs to observations in the Chicago and Atlanta areas show that high resolution CMAQ model output does not necessarily provide smaller FB and RNMSE than lower resolution runs. However, when high resolution output is aggregated to reduce small scale spatial fluctuations, one generally obtains better agreement than either the unaggregated high resolution model output or the low resolution model output in terms of RNMSE. Decomposing the total variation into components depending on hour, day and location and their interactions allows better understanding of the statistical behavior of CMAQ model output at different resolutions. The temporal variation in observation is captured reasonably well by CMAQ model output, but spatial variation and space-time interactions are not.

Key Words: Aggregation, Analysis of variance, CMAQ, Fractional bias, Root normalized mean squared error

---

*Corresponding author: Li Chen, the Center for Integrating Statistical and Environmental Science (CISES), the University of Chicago, 5734 S. Ellis Ave. Rm 459, Chicago, IL 60637; email: lichen@uchicago.edu; telephone number: 773-834-9929; and fax number: 773-834-1045.

1 Introduction

The Community Multi-scale Air Quality (CMAQ) modeling system is a multi-resolution multi-scale air quality model that simulates all atmospheric and land processes that affect the transport, transformation and deposition of air pollutants and their precursors on both regional and urban scales (Ching and Byun 1999). Outputs from meteorology and emission modeling systems are taken as the input for CMAQ Chemical Transport Model (CCTM). Given initial and boundary conditions, CCTM simulates the major atmospheric chemistry, transport and deposition processes. As a result, the CMAQ model output estimates the pollutant concentrations and depositions given specified emission levels and meteorological conditions. For a more detailed discussion of CMAQ, see Byun and Schere (2006).

CMAQ is a tool for scientific problem exploration and decision-making and is widely used for regulatory, policy and research purposes. Given that running CMAQ at higher resolutions requires substantial computational resources, it is important to understand how well CMAQ model output represents reality and how this relates to the model resolution. Indeed, in order to run CMAQ at high resolution one nests the runs running CMAQ at low resolution first to establish initial and boundary conditions for a higher resolution run. Since one routinely then has CMAQ model output at different spatial resolutions, for example, 36 km, 12 km and 4 km, it would be valuable not only to evaluate the accuracy of CMAQ model output at each resolution, but also to compare accuracy across resolutions. It is also informative to assess how the CMAQ model output differs at different resolutions and what is gained from higher resolution CMAQ model output. To answer these questions, we carried out a series of statistical analyses, which include the calculation of some model performance measures and analyses of variance. These statistical analyses are simple to implement but the results provide useful insights into the statistical characteristics of the differences between observations and CMAQ model output at different spatial resolutions that cannot be obtained by the usual point-wise comparisons carried out in the model evaluation literature.

In this paper, we present two case studies of hourly ozone concentrations in parts per billion (ppb), one in the Chicago area (Illinois) and the other in the Atlanta area (Georgia). The results show that the high resolution model output does not consistently have smaller fractional biases or root normalized mean squared errors than the low resolution model output, but that aggregated high resolution model output does yield a better prediction on average in terms of these performance measures. The analysis of variance demonstrates that CMAQ effectively represents the diurnal effect, but poorly captures the spatial variability and space-time interactions, although these are a much smaller part of the total variation in the observations than the temporal effects.

This paper is structured as follows. The statistical methods are described in Section 2; two case studies are presented in Section 3; and Section 4 provides some discussion of other approaches to comparing model output to observations.

2 Statistical Methods

2.1 Fractional bias and root normalized mean squared error

A simple and direct quantitative way to compare model output with observations is to calculate bias and mean squared error. Bias measures systematic differences between model output and observations, and mean squared error is the average of the square of differences between CMAQ modeled values and observations. Bias and mean squared error are two basic performance measures, but when they are compared across locations, they might be misleading, since the observation level might be quite different at each location. Here we use fractional bias (FB) and root normalized mean squared error (RNMSE) instead, which are commonly used in the environmental modeling literature to characterize the accuracy of model output (Canepa and Irwin 2005).

Write X_{ijk} for the observed ozone at location i and hour k on day j. Similarly, M_{ijk} is the CMAQ model output from a specific resolution at hour k on day j interpolated at location i. FB and RNMSE are defined as

$$\begin{aligned} FB_i &= \frac{\overline{M}_{i..} - \overline{X}_{i..}}{(\overline{M}_{i..} + \overline{X}_{i..})/2}, \\ RNMSE_i &= \sqrt{\frac{\frac{1}{N_i}\sum_{j,k}(M_{ijk} - X_{ijk})^2}{\overline{M}_{i..}\overline{X}_{i..}}}, \end{aligned}$$

where $\overline{M}_{i..} = \frac{1}{N_i}\sum_{j,k} M_{ijk}$, $\overline{X}_{i..} = \frac{1}{N_i}\sum_{j,k} X_{ijk}$, and N_i is the total number of non-missing observations at location i. The missing observations and the corresponding model output are not included in this step. The FB measures how large the difference between observations and model output is relative to the average magnitude of the observed and modeled values. Thus, if the biases at two different locations are the same but the mean values are very different, the FB at the location with higher mean value is smaller than the other. FB ranges between -2 and +2. For a perfect model $FB = 0$, while if $FB > 0$ (< 0) the model output on average overestimates (underestimates) the observed concentration values. The RNMSE is normalized in a similar fashion. The smaller RNMSE is, the better model output agrees with observations. These normalizations make the values of FB and RNMSE more comparable across monitoring sites.

2.2 Analysis of variance

By comparing observations with model output directly via calculating performance measures, we learn about the average performance of CMAQ. But in fact, CMAQ model output provides us more information. From a statistical point of view, we would like to know the sources of the disagreement between CMAQ model output and observations. Therefore, we propose decomposing the total variation into space and time components to better understand the source of variability. There are no missing values in the CMAQ model output, and less than 2% of the observed data are missing. At this stage, we replaced the missing values using the procedure described in Appendix A. When the fraction of missing data is higher, the approach to handling missing values will be more critical.

Let Z be a general notation for the quantity of interest, e.g., in our case the differences between CMAQ modeled values and observations $M_{ijk} - X_{ijk}$. We decompose Z_{ijk}, the quantity of interest at location i and hour k on day j, into site (α), daily (β), diurnal (γ) effects and their interactions ($\alpha\beta$, $\alpha\gamma$, and $\beta\gamma$), specifically as,

$$Z_{ijk} = \mu + \alpha_i + \beta_j + \gamma_k + (\alpha\beta)_{ij} + (\alpha\gamma)_{ik} + (\beta\gamma)_{jk} + r_{ijk}, \tag{1}$$

for which every term sums to 0 when summed over any index, so that, for example, $\sum_i (\alpha\beta)_{ij} = 0$ for all j and $\sum_j (\alpha\beta)_{ij} = 0$ for all i. The term r_{ijk} represents the residual at location i and hour k on day j. This model has the form of the linear model in a standard analysis of variance (ANOVA) for a three factor model with usual (sum to 0) constraints, but these terms here are only viewed as numbers, not unknown parameters. We use this decomposition as a tool for summarizing how well different resolutions of CMAQ can capture various aspects of the space-time variation in ozone and not as a basis for formal statistical inference. The variation due to each component can be calculated directly from the data. See Appendix B for details.

We do this analysis of variance for the differences between CMAQ model output and observations, and compare to the corresponding decomposition for the observations. If CMAQ is able to capture some source of variation, we would expect to see smaller number in the decomposition of differences than in the observations.

3 Two Case Studies

3.1 Study one: Chicago area

3.1.1 Data

The first case study covers the Chicago metro area. We ran two sets of nested 36 km, 12 km and 4 km resolution CMAQ (version 4.3) with different parameterization of the planetary boundary layer (PBL) variable in the meteorological model (MM5, which is Mesoscale Model Version 5 developed in cooperation with Penn State University and the University Corporation for Atmospheric Research). Both sets of CMAQ model output were computed for the time period from June 24th to August 1st in 1996. This time period covers the whole month of July, in which the highest ozone concentrations would be expected. The lower the resolution, the larger the spatial domain. For example, the 36 km CMAQ model output covers the eastern United States, while the 4 km CMAQ model output only covers the greater Chicago area. The common spatial domain covered by all three resolutions is northeastern Illinois (Figure 1). The geographical features of this region are very diverse, including rural and urban areas as well as part of Lake Michigan, which has a significant impact on the meteorological fields. These geographical characteristics make it very challenging for CMAQ to produce accurate air pollutant predictions throughout the region.

The PBL values for the first nested CMAQ run are substantially lower than those for the second run. Therefore, the first run is referred as the low PBL run and the second run as the high PBL run. For the low PBL run at 36 km resolution, the daytime (from 6 am to 6 pm CST) average PBL values are around 1000 m over the land, while the values are around 1400 m for the high PBL run (Figure 2). The lower the PBL value, the less convective mixing occurs. These low values of PBL increase nitric oxide and nitrogen dioxide (known together as NO_X) to higher values, which tends to deplete ozone. As a result, the hourly ozone concentrations simulated by the low PBL run are much lower than in the high PBL run. As shown in Figure 3, the daytime average values of hourly ozone concentration at 36 km resolution given by the low PBL run are smaller than the high PBL run. The one exception is the grid cell which is defined as water by the 36 km resolution but, in fact, contains half water (southwest part of Lake Michigan) and half land (Chicago). We will call the area covered by this 36 km grid cell as the Urban Region, since it covers the city of Chicago, and everywhere else in the spatial domain as the Rural Region, even though much of this region is suburban.

Observational data are available at 25 sites within the same spatial domain over this time period. Observations at the top of the Sears Tower in Chicago are excluded because its measurements may be quite different than the other ground level ozone monitors. The remaining sites are numbered 1 to 24 from west to east. There are 8 monitors located in the Urban Region and we name them Urban sites, while the other 16 sites are defined as Rural sites. Figure 3 plots the daytime average values of observed hourly ozone concentrations at each monitoring site. For the Urban sites the average values vary quite dramatically. The 36 km CMAQ model output obviously cannot capture this spatial variability.

The observed data are averages over small spatial regions and can be treated as point measurements, but CMAQ model output at best represents averages over each grid cell, so they have much larger spatial support than the observations. One might hope that the high resolution CMAQ model output will have better agreement with observations, at least in part because the differences in spatial support are less severe. Figure 3 shows that high resolution CMAQ produces small scale spatial variations, but, unfortunately, they do not match well those in the observations as shown in the black box in Figure 3. To remove these local fluctuations we aggregate the high resolution CMAQ model output to obtain a new version of low resolution CMAQ model output. For example, each 36 km gird cell is matched up with 9×9 square of 4 km grid cells, so that they have same size and location. We take the spatial average of the 4 km CMAQ model output at these 9×9 grid cells at each time as a new version of 36 km CMAQ model output and this new version model output is at the same location as the original 36 km CMAQ model output. We refer to this new version of model output as aggregated CMAQ model output.

In order to compare CMAQ model output with observations, we interpolate the CMAQ model output to the locations of the monitoring sites. Shao et al. (2005) compared the behavior of naive interpolation using the nearest available grid cell and bilinear interpolation. They found that bilinear interpolation performs better than naive interpolation. They also found that the more computationally intensive thin plate spline has little or no improvement over bilinear interpolation. Therefore we use bilinear interpolation to interpolate the CMAQ model output. At each monitoring site there are observations, X; interpolated original CMAQ model output, M^{36}, M^{12} and M^{4} at 36 km, 12 km and 4 km resolution runs; and aggregated and then interpolated CMAQ model output, A^{4} and A^{12}, at 36 km resolution based on 4 km and 12 km model runs. Hereafter, when we say model output, it refers to interpolated CMAQ model output.

3.1.2 FB and RNMSE comparison

The values of FB at each location are shown in Figures 4 and 5 for the Rural and Urban Regions respectively. The FB values have a larger range for the Urban Region (-0.75, 0.46) than the Rural Region (-0.25, 0.45). Aggregating the high resolution model output generally moves the FB values in the positive direction by about 0.1 to 0.2 for Rural and about 0.2 to 0.6 for Urban, depending on the site and run. For the Rural sites the values of FB for A^{4} are generally smaller than the ones for A^{12}, while for the Urban sites the FB values for A^{4} are similar to the ones for A^{12}.

The RNMSE values at each site are also plotted in Figures 4 and 5. First of all, looking at the RNMSEs from M^{36}, M^{12} and M^{4} at each location, we notice that RNMSE from M^{4} is not the best at any of the Urban sites nor at 11 out of 16 Rural sites. Secondly, for the low PBL run, the values of RNMSE from A^{4} and A^{12} at each location are smaller than the RNMSEs from M^{4} and M^{12} respectively. For the high PBL run RNMSEs from A^{4} and A^{12} are better than M^{4} and M^{12} respectively at all Urban sites and 14 out of 16 Rural sites. Comparing RNMSEs from A^{4} to M^{36}, A^{4} is better than M^{36} for both low and high PBL runs at all sites except the Urban site 19. Similarly, A^{12} has better RNMSE than M^{36} at all Urban sites other than site 19 and at 14 out of 16 Rural sites.

The overall FB and RNMSE values, which are the averages across sites, are calculated separately for the two regions and listed in Table 1. Among M^{36}, M^{12} and M^{4}, M^{4} has the smallest absolute FB only for the combination of the Rural Region with the high PBL CMAQ run. For the Rural

Region, the FB values of aggregated model output, from both nested CMAQ runs, are worse than the original model outputs. But for the Urban Region, aggregation helps to reduce the absolute value of FB, A^4 and A^{12} significantly improve the FB for this region. For the Urban Region, the high resolution model output produces some small scale spatial variation, but the fact that aggregation reduces the FB suggests that the modeled small-scale spatial fluctuations do not match the actual fluctuations.

Table 1. Fractional bias (FB) and root normalized mean squared error (RNMSE) for the Rural Region and the Urban Region in the Chicago area study.

	Region	PBL level	M^{36}	M^{12}	A^{12}	M^4	A^4
FB	Rural	low	0.031	0.007	0.086	-0.050	0.072
		high	0.202	0.197	0.248	0.135	0.226
	Urban	low	-0.265	-0.399	-0.089	-0.544	-0.080
		high	-0.265	-0.135	0.106	-0.244	0.099
RNMSE	Rural	low	0.531	0.560	0.509	0.580	0.502
		high	0.523	0.532	0.509	0.530	0.496
	Urban	low	0.886	0.930	0.690	1.091	0.681
		high	0.905	0.734	0.618	0.814	0.606

The overall RNMSE values in Table 1 show that for each run M^4 is not the best for any region among the original model output M^{36}, M^{12} and M^4. But the aggregated model output agrees better with the observations. In both regions A^4 is the best, although A^{12} is only slightly worse. Aggregation helps to reduce the RNMSE values and the reduction is substantial for the Urban Region.

The unaggregated high resolution model output does not have a consistently smaller FB or RNMSE value than the low resolution model output. If one stopped the analysis at this point, one might conclude that there is little point in carrying out high resolution runs, at least if the goal is to match observed ozone levels on an hourly basis. However, the aggregated model output based on the high resolution model output does have noticeably better prediction on average in terms of FB or RNMSE than either low resolution runs or unaggregated high resolution runs. One interpretation of the advantage of aggregation is that the high resolution runs may capture some of the spatial variation of the system, but it is making small scale errors in the locations of this variation. Therefore, by aggregating, we can take advantage of this variation and average over local regions to correct for the location errors.

3.1.3 Analysis of variance

Since the land use and geography across the spatial domain varies widely, we perform the analysis of variance for the Rural Region and the Urban Region separately. There are 16 sites in the Rural Region and 8 sites in the Urban Region, therefore we divide the variation of each effect by the number of sites to make them comparable. If model output captures what happens in the observations, the variation of the effects for differences between CMAQ model output and observations would be small. For each effect, we would hope to see smaller variation for the differences between model output and observations than the variation for observations.

The total variation is listed in Table 2. CMAQ does capture from 60% to 80% of the total observed variation depending on the region and run. In addition CMAQ does a better job capturing the space-time variation in the Rural Region than in the Urban Region. Among M^{36}, M^{12} and M^4, M^4 is the best only for the Rural Region at high PBL run. But A^4 is the best overall for both regions and both PBL scenarios.

The variation for each effect is also given in Table 2. We grouped overall mean and diurnal effect together as the hourly effect, since the diurnal effect is the dominant source of variation. Except for the hourly variation, the observed variations in the Urban Region are generally larger than the ones in the Rural Region, especially for the site effect and the interactions between site and hour and between day and hour. For the hourly effect, all versions of CMAQ model output capture the diurnal effect reasonably well for both PBL levels and both subregions. It is the main source for the reduction in the total variation. For the Rural Region, the high resolution model output does better than the low resolution model output, and aggregation does not help to capture more hourly variation. In contrast, for the Urban Region, aggregation does help to capture more diurnal variation. For the day to day variation, all versions of CMAQ model output capture some of this large scale temporal variation. Aggregation from high resolution to low resolution provides small improvements to the daily effect. Both day to day variation and the diurnal pattern are well modeled by CMAQ. Aggregation helps to capture both hourly and daily variations, particularly for the

Urban Region. For the interaction between daily and hourly effect, none of the versions of CMAQ model output capture this short scale temporal variation, except for the Rural Region at low PBL run.

Table 2. The analysis of variance for the Chicago area study ($\times 10^3$ per site).
RL and RH are for the Rural Region with low PBL and high PBL respectively.
UL and UH are for the Urban Region with low PBL and high PBL respectively.

	M^{36}	M^{12}	A^{12}	M^{4}	A^{4}	X	M^{36}	M^{12}	A^{12}	M^{4}	A^{4}	X
	Hour ($\mu+\gamma$)						Day (β)					
RL	45	41	45	36	38	864	35	40	34	45	35	91
RH	52	49	71	31	60		38	36	34	34	32	
UL	111	103	38	148	33	737	60	55	49	63	51	109
UH	88	23	15	45	14		70	57	51	56	45	
	Site (α)						Day×Hour ($\beta\gamma$)					
RL	8	11	10	7	8	15	48	51	47	50	47	57
RH	13	12	9	9	9		58	60	56	58	53	
UL	30	17	25	16	25	24	88	91	84	90	82	85
UH	39	18	25	16	25		104	100	98	97	93	
	Site×Day ($\alpha\beta$)						Site×Hour ($\alpha\gamma$)					
RL	21	24	20	25	19	13	5	3	3	4	3	3
RH	21	20	16	22	16		6	5	4	5	4	
UL	17	24	16	26	16	13	13	10	11	10	11	11
UH	18	23	15	26	14		14	11	11	10	11	
	Total						Residual					
RL	207	224	201	227	193	1079	45	54	43	61	43	35
RH	238	245	236	228	219		49	63	46	70	46	
UL	353	339	256	401	252	1010	33	40	33	47	33	31
UH	368	277	250	305	238		36	45	34	55	35	

For the site effect in the Rural Region, all versions of CMAQ model output from both runs capture some of the site variation and, not surprisingly, the high resolution model output does better than the low resolution model output. In the Urban Region, both 12 km and 4 km resolution model output captures a small fraction of the site effect, but aggregating makes the agreement worse. For the site and day interaction, none of the model runs are able to predict this effect well. Aggregation helps, but the agreement is still poor. For site and hour interaction, neither of the CMAQ runs capture this term. All model output predicts site related effects poorly. For the residuals based on differences between observations and original model output, it increases as the spatial resolution of CMAQ goes from 36 km to 4 km. But aggregation helps to reduce this quantity.

For some effects, the differences between model output and observations have larger variation than the observations, but never more than double. If model output and observations have the same amount of variation, but are statistically independent, then the expected variation in the difference between model output and observations will be two times the observed variation, which seems to occur in a few places in Table 2. For example, for the site and day interaction at the Urban Region with high PBL run, the differences between M^4 and observations have variation of 26×10^3 per site, which is twice the observed variation of 13×10^3 per site.

3.2 Study two: Atlanta area

3.2.1 Data

The second case study considers a region around Atlanta. A set of 32 km, 8 km, and 2 km resolution CMAQ model output were run by the US Environmental Protection Agency (EPA) for the period from August 1st to August 24th in 1999. The CMAQ model output for the first three days are treated as a burn-in period and are not included in the analysis. The spatial domain covered by all three resolutions is the Atlanta metro area (Figure 6). Observations from the Clean Air Status and Trends Network (CASTNET) and the Aerometric Information Retrieval System (AIRS) are available at the 12 sites labeled 1 to 12 from west to east. Similar to the first study, we aggregate 8 km and 2 km resolution model output to 32 km resolution. Then bilinear interpolation is applied to all available model output. Therefore, we have observations X, interpolated model output, M^{32}, M^8 and M^2, and aggregated then interpolated model output, A^8 and A^2, at each location.

3.2.2 FB and RNMSE comparison

Figure 7 plots the values of FB and RNMSE at each location for the five sets of CMAQ output. The FB values have a range (-0.1, 0.54). Among these 12 sites, 11 of them have positive FBs, which means all versions of CMAQ model output on average overestimate the observed ozone values. The FB value from M^{32} is the worst at all sites. Comparing to M^{32} and M^8, M^2 gives better FB values at 8 out of 12 sites. Aggregation does not help much to reduce FB. For RNMSE at each location, both A^2 and A^8 are better than M^{32}, and A^8 is better than M^8. The RNMSEs from A^2 are better than M^2 at all sites except for site 7. Aggregation does help to improve RNMSE at the most of sites. The overall FBs and RNMSEs are listed in Table 3. The value of FB from high resolution is better than that from low resolution, though the FB from M^2 is only slightly better than M^8. Aggregation does not improve FB. For RNMSEs, M^2 has the smallest value among M^{32}, M^8 and M^2. Moreover, RNMSEs from A^8 and A^2 are better than M^8 and M^2 correspondingly, and A^2 is the best among them all. Overall, the 2 km model output agrees with the observations best among 32 km, 8 km and 2 km model outputs in terms of both FB and RNMSE. Aggregation helps to reduce RNMSE but not FB.

Table 3. Fractional bias (FB) and root normalized mean squared error (RNMSE) for the Atlanta area study.

	M^{32}	M^8	A^8	M^2	A^2
FB	0.377	0.298	0.302	0.291	0.302
RNMSE	0.512	0.517	0.483	0.490	0.459

3.2.3 Analysis of variance

The variance decomposition is performed on the observations and the differences between model output and observations; results are listed in the Table 4. For the total variation, similar to the Chicago runs, all versions of CMAQ model output capture about 75% of the observed total variation. M^2 is the best among the unaggregated model outputs, and A^2 is the best among all. All the versions of CMAQ model output capture the diurnal effect reasonably well, which is the dominant effect. Aggregation does not help to capture more hourly variation. For the daily effect, M^2 is better than M^{32} and M^8, and A^2 has similar capacity as M^2 to model day to day variation. For the day-hour interaction, no model output captures much of the variation. For site related effects, all of the model outputs perform poorly, except for the site-hour interaction which is partially captured by A^2 and A^8. Even though aggregation helps to reduce the variation of site related effects, most of them are still bigger than the variations in the observations.

The sum of squared residuals increases as the resolution increases. Aggregation helps to reduce the sum of squared residuals compared to their base, but A^2 still gives slightly larger values than M^{32}.

In summary, for the Atlanta area study all 12 monitoring sites are located close to major highways. They act similarly to the Rural sites in the Chicago area study.

Table 4. The analysis of variance for the Atlanta area study ($\times 10^3$).

Effect	M^{32}	M^8	A^8	M^2	A^2	X
Hour ($\mu + \gamma$)	3421	2159	2352	2013	2305	19719
Day (β)	357	525	494	269	250	407
Site (α)	271	463	275	389	249	198
Day× hour ($\beta\gamma$)	494	489	456	451	416	593
Site×Hour ($\alpha\gamma$)	349	335	204	269	175	305
Site×Day ($\alpha\beta$)	228	339	239	311	216	188
Residual	682	1135	730	1144	702	606
Total	5802	5445	4750	4846	4313	22016

4 Discussion

There are many ways to compare model output to observations. Overall performance measures, such as FB and RNMSE, provide the evaluation of the on average performance. But they provide no insight into what aspects of the spatial-temporal variation in the observations the model matches the pattern in the observations. Fuentes and Raftery (2005) present a Bayesian framework to jointly model the observations and model output in terms of the underlying truth. This method not only estimates the bias parameters in the model output and the parameters in the covariance structure of model output and observations, but also simulates the truth given model output and observations. This approach is very appealing, but it requires strong statistical assumptions and only applies to spatial processes at one time point. Jun and Stein (2004) propose to compare the space-time correlation structures of observations and numerical model output in evaluating numerical models. This empirical method is attractive and easy to implement in principle. But when observations are only available from a few monitoring sites, the results might be hard to interpret. The analysis of variance that we present in this paper is both easy to implement, and provides information of spatial-temporal aspects of the variation that can not be obtained from overall summary statistics.

Interpreting the observation and evaluating the effectiveness of CMAQ require a close examination of the monitor sites. A particular challenge to any air quality model is to capture the differences in ozone levels at site 19 and 20 in the Chicago area study. These sites are about 3 km apart, but site 19 is on a busy street close to the Sears Tower and site 20 is located in the Jardine water plant, which is next to Navy Pier in Chicago and is largely surrounded by water. The traffic pattern at these two sites are quite different. As a result, the observed hourly ozone concentrations at these two sites differ strikingly. As shown in Figure 8, the observed ozone values at monitor 20 are much higher than the ones at monitor 19, especially during the day time. In the 4 km resolution CMAQ modeling system, these two sites are in adjacent grid cells in the east-west direction. The interpolated 4 km model outputs from the high PBL run at both sites are similar. For monitor 20, the 4 km model output underestimates the hourly ozone concentration persistently. This might be caused by problems with the emissions input. Nevertheless, this small scale local feature is going to be difficult for CMAQ to resolve, even when run at high resolution.

The aggregation we have used here is a simple smoothing technique, which takes the spatial average as the corresponding value. We then did bilinear interpolation to sites. Other smoothing methods would be employed, but we have explored some other possibilities and found that the results do not change much.

The analysis of variance for both studies shows that CMAQ does an excellent job at modeling both the diurnal pattern and day to day variation, which are the dominant components in the total variation. Aggregation helps CMAQ to capture more diurnal variation in the Urban Region in the Chicago area study. Moreover, it helps in capturing the daily effect for both studies. The interaction between daily and hourly effects is a relatively important component, but none of the models describe this source of variation well. For site-related effects, all the model runs perform poorly. Even CMAQ runs at high spatial resolution are not able to capture well small scale spatial features in the observations. In the Chicago area study, the site effects at the Urban Region from both runs are similar. For these two runs, they have the same emission input, but different parameterizations of PBL in MM5. It would be valuable to have more runs under different PBL scenarios given the same emissions. If the site effect from each run has a similar pattern as shown in the study, it might indicate that the inability of CMAQ to capture the spatial pattern is a sign of a problem with emissions rather than meteorology or chemistry.

Both studies suggest that different versions of CMAQ model output have different capacities in capturing different aspects of space-time variations. Thus, given that low resolution output is a prerequisite for obtaining high resolution model output, it makes sense to use the output at all resolutions when comparing model output to observations or when attempting to combine model output and observations as in Fuentes and Raftery (2005). For example, the model for the analysis of variance (1) can be used to represent the underlying true spatial-temporal process for hourly ozone concentrations. Each effect in the model can be estimated from different versions of CMAQ model output. For instance, in the Chicago area study, 4 km CMAQ model output can be used to estimate the diurnal effect for the Rural Region, while the aggregated 4 km CMAQ model output is used for the Urban Region. A combination of CMAQ model outputs at different spatial resolutions can also be employed to estimate the effects in model (1).

Acknowledgement

Although the research described herein has been funded wholly or in part by the United States Environmental Protection Agency through STAR Cooperative Agreement #R-82940201-0 to the University of Chicago, it has not been subjected to the Agency's required peer and policy review and therefore does not necessarily reflect the views of the Agency, and no official endorsement should be inferred.

Appendix A

Suppose the observed hourly ozone concentration X_{ijk} at location i and hour k on day j is missing. This missing value is replaced with $\hat{X}_{ijk}$, which is defined as

$$\hat{X}_{ijk} = \bar{X}_{...} + (\bar{X}_{i..} - \bar{X}_{...}) + (\bar{X}_{.j.} - \bar{X}_{...}) + (\bar{X}_{..k} - \bar{X}_{...}).$$

The missing value is replaced with the sum of overall mean, ith site effect, jth day effect and kth hour effect.

Appendix B

For model (1), let n_s, n_d and n_h be the number of monitors, days and hours, then the variation of each effect can be calculated as following:

$$\begin{aligned} SS_{site} &= n_d n_h \sum_{i=1}^{n_s} (\bar{z}_{i..} - \bar{z}_{...})^2, \\ SS_{day} &= n_s n_h \sum_{j=1}^{n_d} (\bar{z}_{.j.} - \bar{z}_{...})^2, \\ SS_{hour} &= n_s n_d \sum_{k=1}^{n_h} (\bar{z}_{..k})^2, \\ SS_{site \times day} &= n_h \sum_{i=1}^{n_s} \sum_{j=1}^{n_d} (\bar{z}_{ij.} - \bar{z}_{i..} - \bar{z}_{.j.} + \bar{z}_{...})^2, \\ SS_{site \times hour} &= n_d \sum_{i=1}^{n_s} \sum_{k=1}^{n_h} (\bar{z}_{i.k} - \bar{z}_{i..} - \bar{z}_{..k} + \bar{z}_{...})^2, \\ SS_{day \times hour} &= n_s \sum_{j=1}^{n_d} \sum_{k=1}^{n_h} (\bar{z}_{.jk} - \bar{z}_{.j.} - \bar{z}_{..k} + \bar{z}_{...})^2, \end{aligned}$$

where SS is for "Sum of Squares", $\bar{z}_{...} = \frac{1}{n_s n_d n_h} \sum_{i=1}^{n_s} \sum_{j=1}^{n_d} \sum_{k=1}^{n_h} z_{ijk}$, $\bar{z}_{i..} = \frac{1}{n_d n_h} \sum_{j=1}^{n_d} \sum_{k=1}^{n_h} z_{ijk}$, $\bar{z}_{ij.} = \frac{1}{n_h} \sum_{k=1}^{n_h} z_{ijk}$ and $\bar{z}_{.j.}$, $\bar{z}_{..k}$, $\bar{z}_{i.k}$ and $\bar{z}_{.jk}$ are defined in the same fashion.

References

Byun, D. and K. Schere (2006). Review of the governing equations, computational algorithms, and other components of the Model-3 Community Multiscale Air quality (CMAQ) modeling system. *Applied Mechanics Reviews 59*, 51–77.

Canepa, E. and J. Irwin (2005). Evaluation of air pollution models. In P. Zannetti (Ed.), *Air Quality Modeling - Theories, Methodologies, Computational Techniques, and Available Databases and Software. Vol II - Advanced Topic.* The EnviroComp Institute and the Air & Waste Management Association.

Ching, J. and D. Byun (1999). Introduction to the Models-3 Framework and the Community Multiscale Air Quality Model (CMAQ). *Available at http://www.epa.gov/asmdnerl/CMAQ/CMAQscienceDoc.html.*

Fuentes, M. and A. E. Raftery (2005). Model evaluation and spatial interpolation by Bayesian combination of observations with outputs from numerical models. *Biomet-*

rics 61, 36–45.

Jun, M. and M. L. Stein (2004). Statistical comparison of observed and cmaq modeled daily sulfate levels. *Atmospheric Environment 38*, 4427–4436.

Shao, X., M. L. Stein, and J. Ching (2005). Statistical comparisons of methods for interpolating the output of a numerical air quality model. *Environmental Science Journal of Statistical Planning and Inference (Accepted)*.

Figure captions

Figure 1. The spatial domain for the Chicago area study. The number indexes the monitoring site. The 36 km grid cells, which covers this region, are drawn by grey dash lines.

Figure 2. The day time average of 36 km PBL values in meters in the Chicago area study. The numbers in black indicate that the corresponding grid cell is classified as land and the blue is for water. Here the grid cells correspond to the grid cells drawn in Figure 1.

Figure 3. The day time average of hourly ozone concentration in the Chicago area study. The domain shown here is the area in the darker grey box in Figure 2. The large box with bold number is for 36 km model output, while the small box is for 4 km model output. The numbers in black indicate that the corresponding grid cell is classified as land and the blue is for water. The number in red is the day time average of observations. The black box is an example to show that the spatial variation modeled by 4 km CMAQ model output does not match the variation in observations.

Figure 4. Fractional bias (FB) and root normalized mean squared error (RNMSE) values for the Rural sites in the Chicago area study. At each location index, plot on left is for the low PBL run and the one on right is for the high PBL run.

Figure 5. Fractional bias (FB) and root normalized mean squared error (RNMSE) values for the Urban sites in the Chicago area study. At each location index, plot on left is for the low PBL run and the one on right is for the high PBL run.

Figure 6. The spatial domain for the Atlanta area study. The number indexes the monitoring site.

Figure 7. Fractional bias (FB) and root normalized mean squared error (RNMSE) values for monitors in the Atlanta area study.

Figure 8. Time series plots for hourly ozone concentration at the monitors 19 and 20 in the Chicago area study. The black solid line is for the observations at monitor 20 and the grey solid line is for the observations at monitor 19. The black dot line and the grey dot line are for the interpolated 4 km high PBL run model output at monitor 20 and monitor 19 respectively.

Figure 1

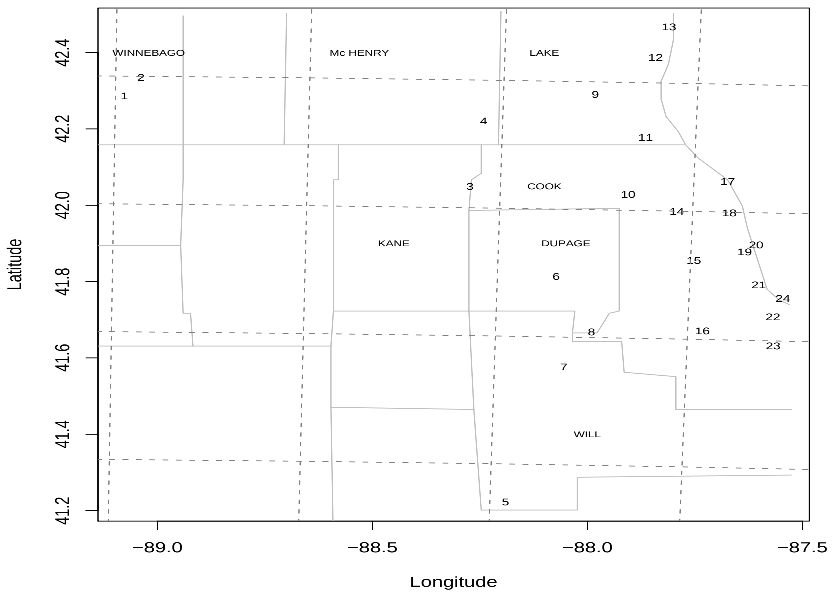


Figure 2

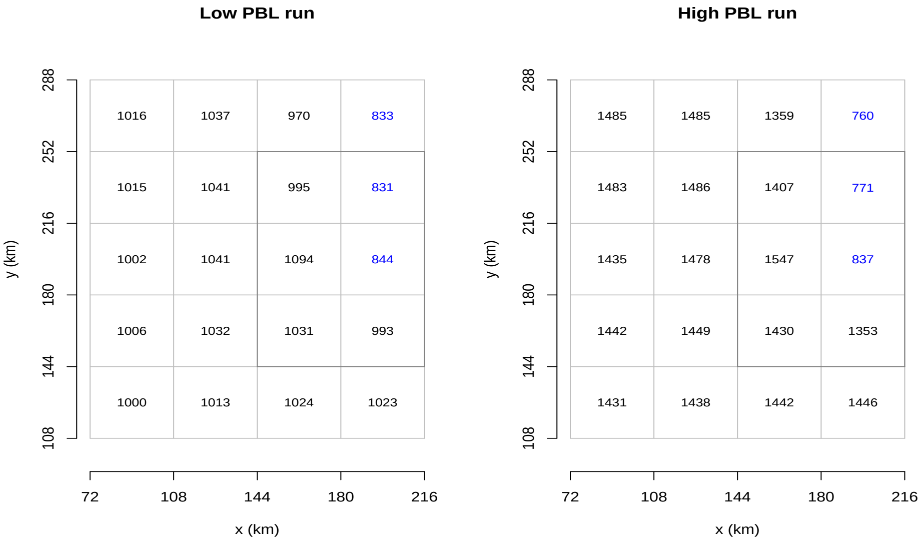


Figure 3

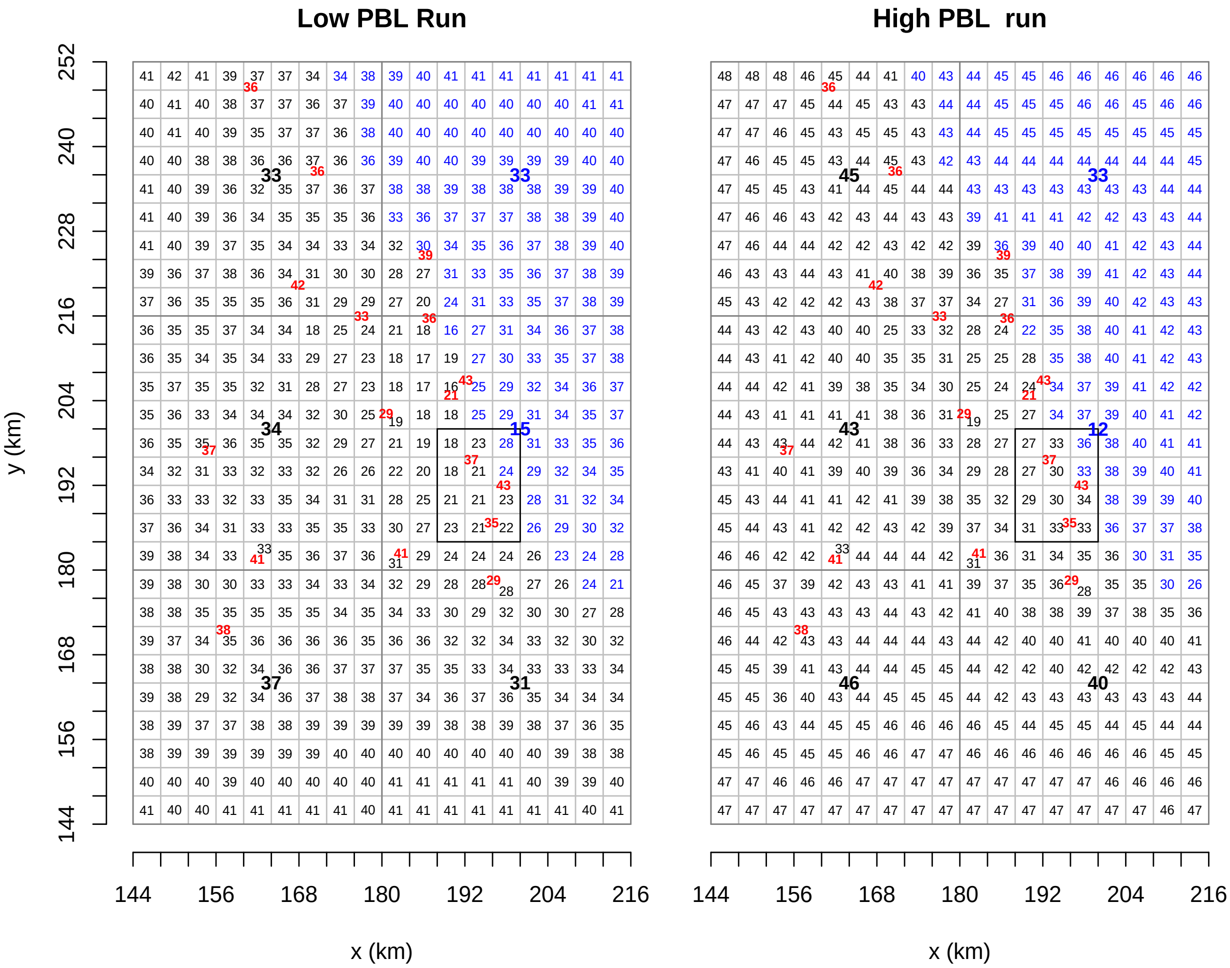
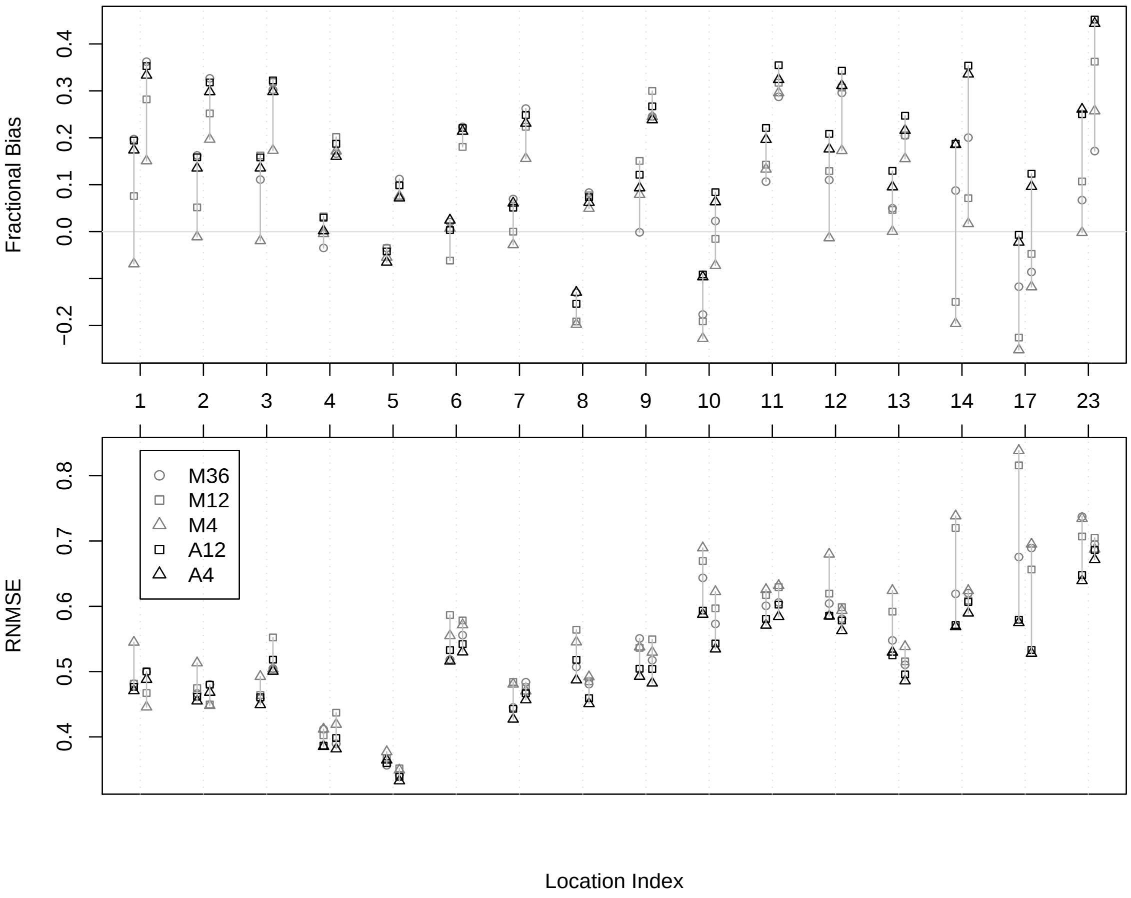

Low PBL Run
High PBL run
y (km)
252
240
228
216
204
192
180
168
156
144
144
156
168
180
192
204
216
x (km)
x (km)

Figure 4



Fractional Bias
0.4
0.3
0.2
0.1
0.0
−0.2
1
2
3
4
5
6
7
8
9
10
11
12
13
14
17
23
RNMSE
0.8
0.7
0.6
0.5
0.4
M36
M12
M4
A12
A4
Location Index

Figure 5

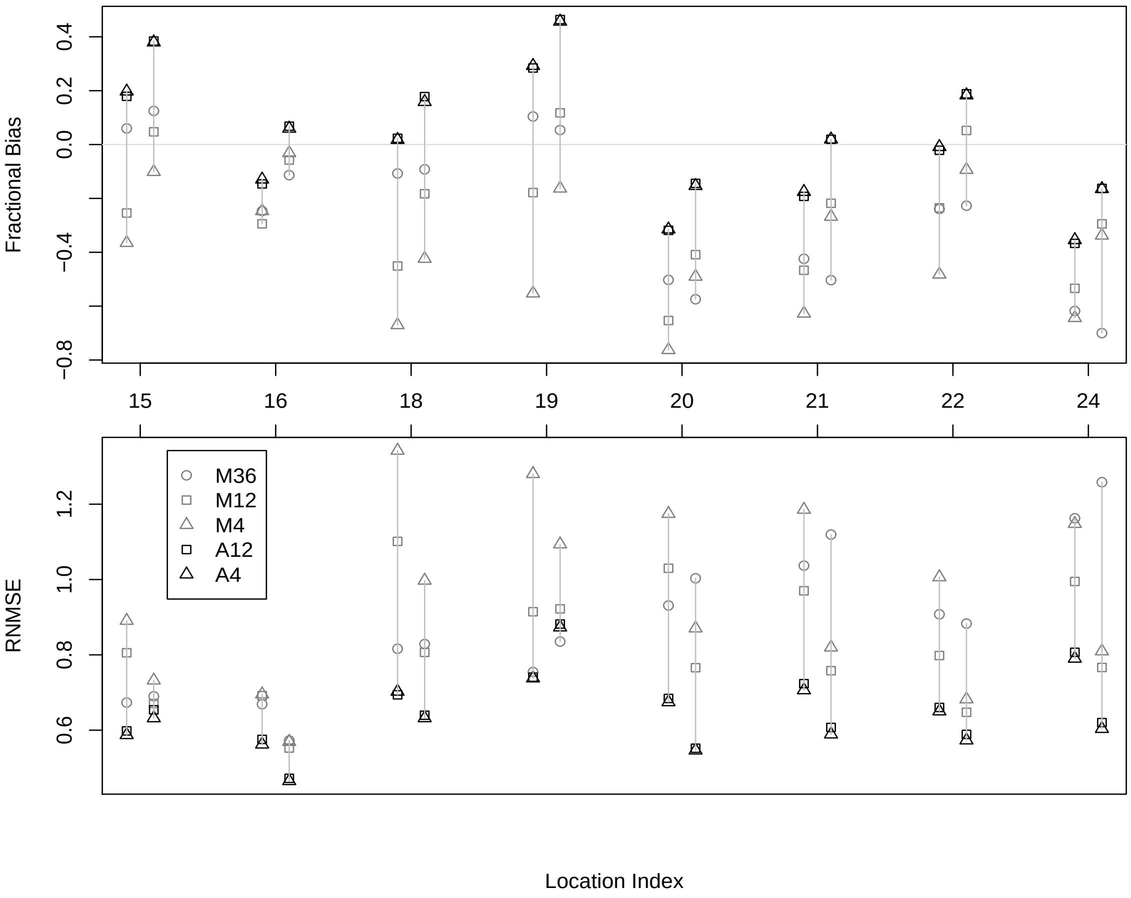

Fractional Bias
0.4
0.2
0.0
−0.4
−0.8
15
16
18
19
20
21
22
24
RNMSE
1.2
1.0
0.8
0.6
M36
M12
M4
A12
A4
Location Index

Figure 6

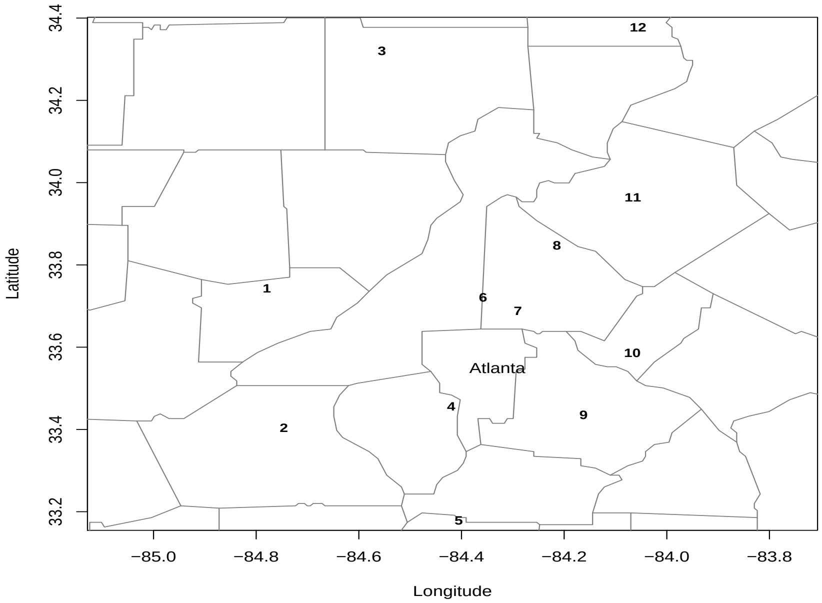


Figure 7

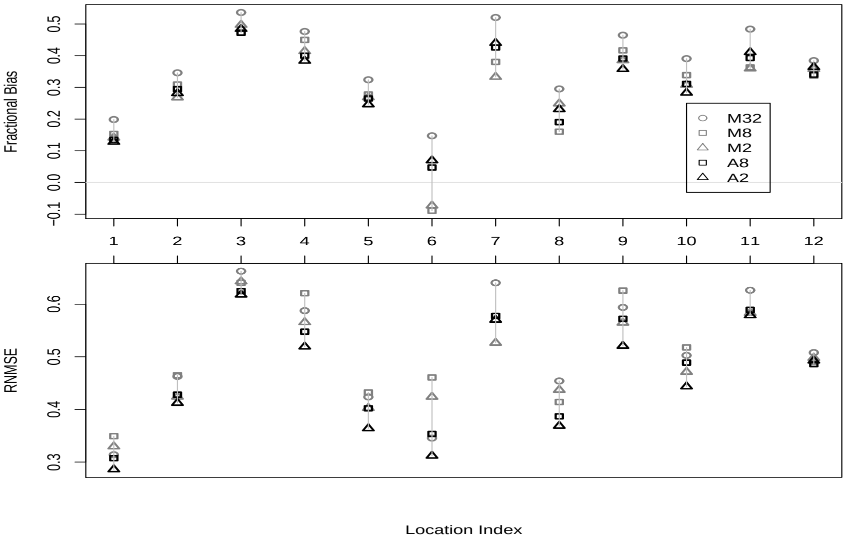
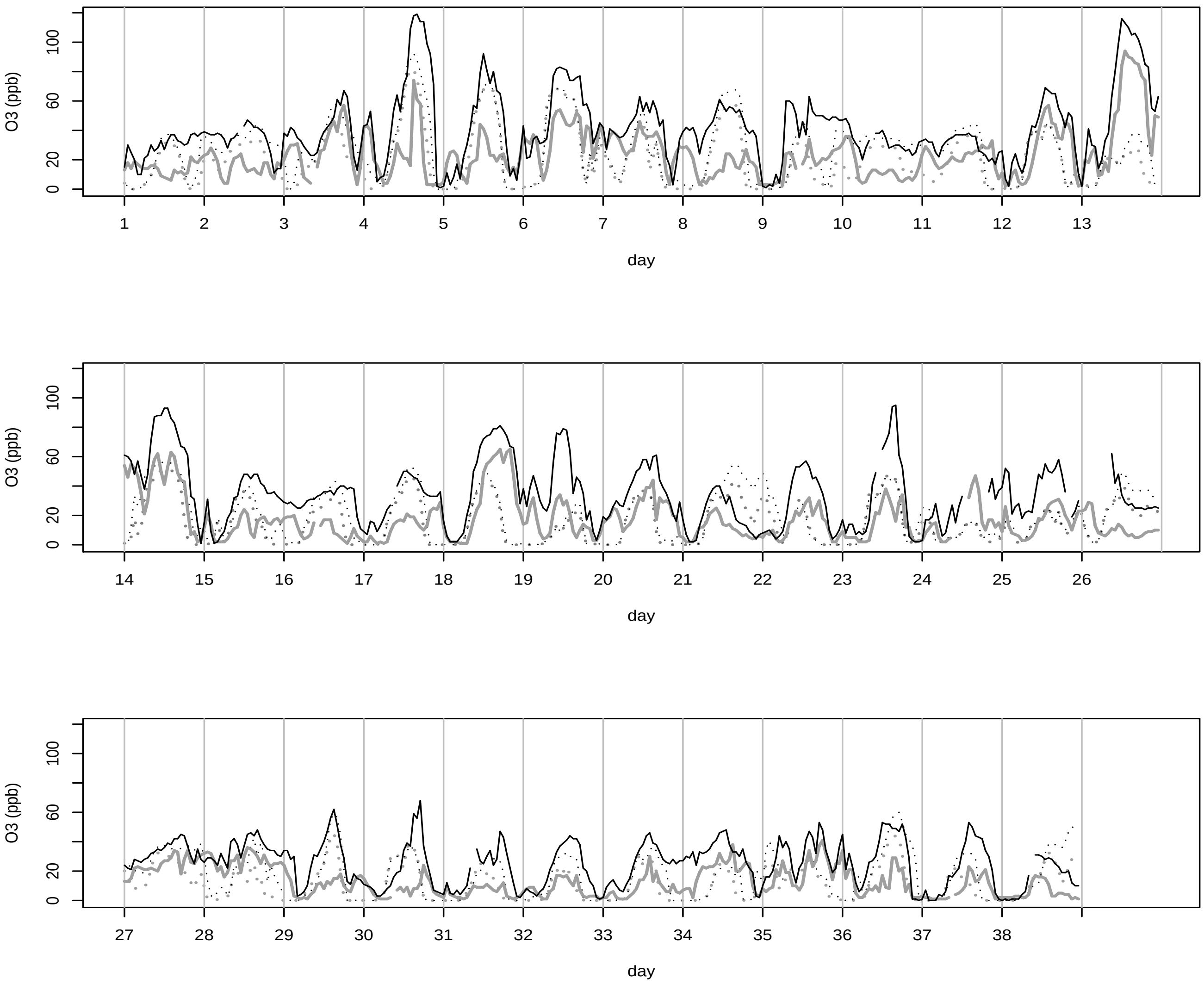


Figure 8


O3 (ppb)
0
20
60
100
1
2
3
4
5
6
7
8
9
10
11
12
13
day
O3 (ppb)
0
20
60
100
14
15
16
17
18
19
20
21
22
23
24
25
26
day
O3 (ppb)
0
20
60
100
27
28
29
30
31
32
33
34
35
36
37
38
day