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

V. R. Kotamarthi
Argonne National Laboratory, Argonne, IL, USA

1 INTRODUCTION

CMAQ is considered primarily as a computer-based tool for scientific problem exploration and decision-making, and is widely used for regulatory, policy and research purposes. Therefore, it is important to understand how well CMAQ model output represents reality and how this relates to the model resolution. High resolution CMAQ model output might provide a better representation of pollutant concentrations in atmosphere, but in order to run CMAQ at high resolution one has to run CMAQ at low resolution first to establish initial and boundary conditions for the high 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 compare accuracy across resolutions. It is also intuitive to ask 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 analysis of variance. These statistical analyses are very simple to implement but the results are insightful, providing a summary description of the statistical characteristics of the difference between observations and CMAQ model output at different spatial resolutions.

*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.

In this paper, we present two case studies of hourly ozone concentration 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 may not have a smaller fractional bias or root normalized mean squared error values 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 shows that CMAQ is good at modeling diurnal effect, but poor at capturing spatial variations and space-time interactions.

2 STATISTICAL METHODS

2.1 *Fractional Bias and Root Normalized Mean Squared Error*

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 hour k on day j. Similarly, M_{ijk} is the model output from some version of the CMAQ model at location i hour

k on day j. FB and RNMSE are defined as

$$FB_i = \frac{\overline{M}_{i..} - \overline{X}_{i..}}{(\overline{M}_{i..} + \overline{X}_{i..})/2},$$

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

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 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 Variation*

By comparing observations with model output directly via calculating performance measures, we learn the on average performance of CMAQ model output. 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 time components to better understand the source of variability. For CMAQ model output there are no missing values, and, in the observations, less than 2% of the 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. We decompose Z_{ijk} as,

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

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. 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 versions of CMAQ can capture various aspects of the space-time variation in ozone and not as a basis for formal statistical inference. In model (1), α_i represents the site effect at location i; β_j is the j-th day effect and γ_k is the hourly effect (diurnal pattern). $(\alpha\beta)_{ij}$, $(\alpha\gamma)_{ij}$ and $(\beta\gamma)_{jk}$ are interaction terms: $(\alpha\beta)_{ij}$ is for the effect at site i on day j, $(\alpha\gamma)_{ik}$ is for the effect at site i hour k, and $(\beta\gamma)_{jk}$ is for the effect on day j at hour k. The variation due to each component can be calculated directly from the data. In this study, we group overall mean and γ_k as the diurnal effect, since this is the dominant source of variation in the data.

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 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 run two sets of nested 36 km, 12 km and 4 km resolution CMAQ (version 4.3) with different inputs for the planetary boundary layer (PBL) variable, which effects meteorological fields. The PBL values for the first nested CMAQ run are substantially lower than those for the second nested CMAQ run. Therefore, the first run is referred as the low PBL run and the second run as the high PBL run. Both sets of CMAQ model output are available 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 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. We define the area covered by one 36 km grid cell, which covers the city of

Chicago, as the Urban Region, and everywhere else in the spatial domain as the Rural Region. Observational data are available at 24 sites within the same spatial domain over this time period. The monitoring sites are numbered 1 to 24 from west to east and 8 monitors are located in the Urban Region.

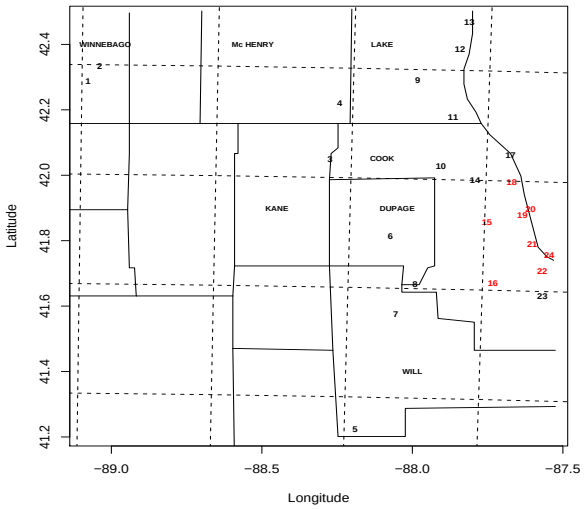

Figure 1: The spatial domain for the Chicago area study, where the number indexes the monitoring site. The red is for the site in Urban Region and the black is for the site in Rural Region. The 36 km gird cells are drawn by the grey dash lines.

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 then 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. High resolution CMAQ does produce small scale spatial variations, but, unfortunately, they do not match well those in the observations. Then 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 4 km grid cells. We then 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 monitoring sites. Shao, Stein, and Ching (2005) compared the behavior of naive interpolation using the nearest available grid cell and bilinear interpolation. They found that bilinear interpolation is generally quite a bit better than naive interpolation. They also found that the more computationally intensive thin plate spline performs no better than bilinear interpolation. Therefore we use bilinear interpolation to interpolate the CMAQ model output. So 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, and interpolated aggregated CMAQ model output A^4 and A^{12} at 36 km resolution based on 4 km and 12 km resolution.

3.1.2 FB and RNMSE comparison

To summarize the result, the overall FB and RNMSE values, which are the averages across sites within a region, are calculated for the combination of region and PBL level: the Rural Region with low PBL level (RL), the Rural Region with high PBL level (RH), the Urban Region with low PBL level (UL) and the Urban Region with high PBL level (UH).

For the FB values among M^{36}, M^{12} and M^4 in Table 1, M^4 has the smallest absolute FB only for RH. 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, and among all the model output, A^4 is the best. 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 fluctuation.

Table 1. Fractional bias for the Chicago area study.

	M^{36}	M^{12}	A^{12}	M^4	A^4
RL	0.031	0.007	0.086	-0.050	0.072
RH	0.202	0.197	0.248	0.135	0.226
UL	-0.265	-0.399	-0.089	-0.544	-0.080
UH	-0.265	-0.135	0.106	-0.244	0.099

The overall RNMSE values in Table 2 show that for each run M^4 is not the best for any combination of region and PBL level among M^{36}, M^{12} and M^4. But the aggregated model output has smaller RNMSE than both the unaggregated high resolution model output and the low resolution model output, and A^4 is the best among all the model output, although A^{12} is only slightly worse.

Table 2. Root normalized mean squared error for the Chicago area study.

	M^{36}	M^{12}	A^{12}	M^4	A^4
RL	0.531	0.560	0.509	0.580	0.502
RH	0.523	0.532	0.509	0.530	0.496
UL	0.886	0.930	0.690	1.091	0.681
UH	0.905	0.734	0.618	0.814	0.606

The high resolution model output may not consistently have a smaller FB or RNMSE value than the low resolution model output, so 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. 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.

3.1.3 ANOVA

The analysis of variance is performed for the Rural Region and the Urban Region separately. There are 16 sites in the Rural Region and 8 sites in the Urban Region, so 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 observations and model output than the variation for observations.

The analysis of variance (ANOVA) is given by Figure 2. 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. It also shows that CMAQ does a better job capturing the space time variation in the Rural Region than in the Urban Region.

For the hourly effect, for both PBL levels and both regions, all versions of CMAQ model output capture this diurnal effect reasonably well. 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. On the contrary, 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

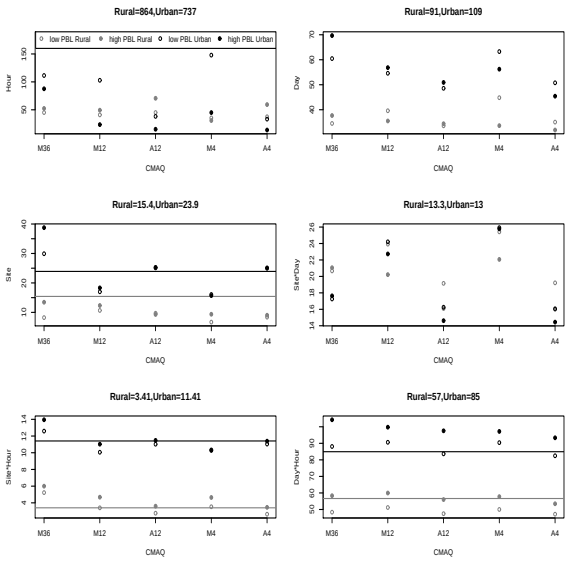

Figure 2: ANOVA for the Chicago area study ($\times 10^3$). In each plot, the observed variations are listed in the title. The variations for the differences are plotted using the following symbols: gray open circle for RL, gray dot for RH, black open circle for UL and black dot for UH.

to low resolution improves agreement a little bit for this 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 low PBL run.

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, the sum of squared residuals per site increases as the spatial resolution of CMAQ goes from 36 km to 4 km, as shown in Table 3. But aggregation helps to reduce

this quantity.

Table 3. Sum of squared residuals per site for the Chicago area study ($\times 10^3$).

	M^{36}	M^{12}	A^{12}	M^4	A^4	X
RL	44.6	54.5	42.8	60.6	42.6	34.9
RH	48.9	62.7	45.6	69.6	45.7	
UL	33.5	40.0	33.0	46.8	33.0	30.9
UH	35.7	44.6	34.5	54.5	34.8	

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, which were run by the US Environmental Protection Agency (EPA), is available for the time period from August 1st to August 24th in 1999. The spatial domain covered by all three resolutions is the Atlanta metro area (Figure 3). Observations from Clean Air Status and Trends Network (CASTNET) and Aerometric Information Retrieval System (AIRS) within the same spatial domain over this time period 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 interpolated aggregated model output, A^8 and A^2.

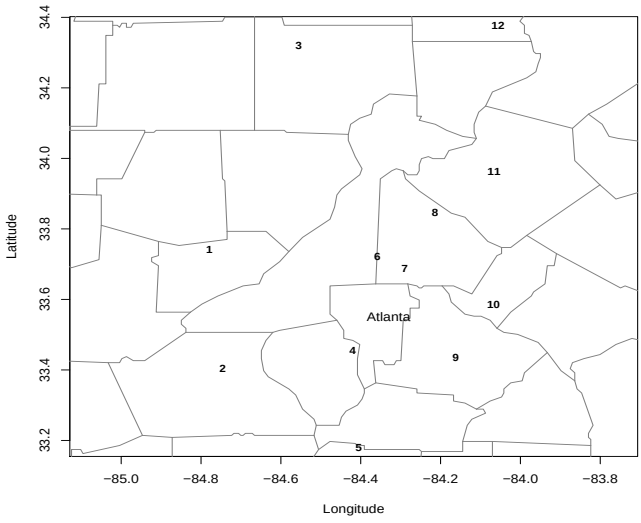

Figure 3: The spatial domain for the Atlanta area study, where the number indexes the monitoring site.

3.2.2 FB and RNMSE comparison

The overall FB and RNMSE are listed in Table 4. The value of FB from high resolution is better than the value from low resolution, though the FB from M^2 is only slightly better than M^8. Aggregation does not improve FB. For RNMSE, 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 has the best RNMSE. Overall, the 2 km model output agrees the observations best among 32 km, 8 km and 2 km model output in terms of both FB and RNMSE. Aggregation helps to reduce RNMSE, but not FB.

Table 4. Fractional bias and root normalized mean squared error for the Atlanta data 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 ANOVA

The ANOVA decomposition is performed on the observations and the differences between model output and observations; results are listed in Table 5. 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. For site related effects, all of the model outputs perform poorly. 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 decreases as the resolution changes from 2 km to 32 km. Aggregation helps to reduce the sum of squared residuals compared to their base, but A^2 still gives slightly larger values than M^{32}.

Table 5. ANOVA for the Atlanta area study ($\times 10^3$).

Effect	M^{32}	M^8	A^8	M^2	A^2	X
hour	3421	2159	2352	2013	2305	19719
day	357	525	494	269	250	407
site	271	463	275	389	249	198
hour×day	494	489	456	451	416	593
site×hour	349	335	204	269	175	305
site×day	228	339	239	311	216	188
Residuals	682	1135	730	1144	702	606

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 do not help us to understand whether the spatial-temporal variation given by the model matches with the pattern in the observations. 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 easy to implement, but also provides information of spatial-temporal aspects of the variation that can not be obtained from overall summary statistics.

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. The results of Atlanta area study have the similar pattern as shown in the Rural Region in Chicago area study.

High resolution CMAQ model output does not necessarily predict hourly ozone concentration better than low resolution CMAQ model output in terms of RNMSE. But the aggregated model output does help to improve the on average performance, especially for the Urban Region in the Chicago area study. Aggregation is a simple smoothing technique, which takes the spatial average as the corresponding value. In this paper we aggregated from high resolution to low resolution, for example, from 4 km to 36 km. We also did an experiment to obtain new versions of high resolution model output by aggregation. For example, the new model output at each 4 km grid cell is produced by aggregating from 4 km to 36 km centered at this cell. The results based on this new version model output are similar to the results presented in the previous Section. More sophisticated smoothing method, e.g., kernel based methods, can be employed to smooth the model output, but in this study we stick to the simple averaging method because it makes comparisons across model resolutions more intuitive.

The analysis of variance for both studies shows that CMAQ has great capability to model both the diurnal pattern and day to day variation, which are the dominant components in the total variation. Aggregation helps to capture more diurnal variation in the Urban Region in Chicago area study. Moreover, it helps in capturing the daily effect for both studies. The interaction between daily and hourly effect is a relatively important component, but none of the models describe this source of variation well. This might be caused by the poor meteorological/emission inputs. For site-related effects, all the model runs perform poorly. Even CMAQ runs at high spatial resolution are not able to capture well such small scale spatial features as shown in observations. 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.

APPENDIX A

Suppose the observed hourly ozone concentration X_{ijk} at location i, day j, and hour k 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.

References

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.

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. *Technical Report at the Center for Integrating Statistical and Environmental Science 23*.