DATA FUSION OF MONITOR DATA AND CHEMICAL TRANSPORT MODEL SIMULATIONS

A Dissertation
Presented to
The Academic Faculty

by:

Francesca Metcalf

In Partial Fulfillment
of the Requirements for the Degree
Environmental Engineering in the
Georgia Institute of Technology

Georgia Institute of Technology
May 2017

DATA FUSION OF MONITOR DATA AND CHEMICAL TRANSPORT MODEL SIMULATIONS

Approved by:

Dr. James A. Mulholland, Advisor
School of Civil and Environmental Engineering
Georgia Institute of Technology

Dr. Armistead G. Russell
School of Civil and Environmental Engineering
Georgia Institute of Technology

Dr. Stefanie E. Sarnat
School of Environmental Health
Emory University

Date Approved: April 27, 2017

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

SUMMARY

Associations between air quality and acute health effects vary across pollutants and across spatial and temporal metrics of concentration. Studies that investigate these associations require data that are both spatially and temporally complete across many pollutants. The objective of this study was to create accurate and complete pollutant concentration fields by combining the benefits of observed data and a chemical transport model, CMAQ, while reducing the effects of their incomplete spatial and temporal coverage and limited accuracy, respectively. Using a previously developed approach, these spatially and temporally resolved pollution fields were created over the domain of Georgia for 12 pollutants (8-hr maximum O_3, 1-hr maximum NO_2, NO_x, CO and SO_2, and 24-hr average PM_{10}, $PM_{2.5}$ and five $PM_{2.5}$ subspecies) and four years (2009 - 2012). It was found that the results from this data fusion agree very well with observations as well as results from previous studies. Through a cross-validation analysis, it was found that the fusion is also able to estimate concentrations far from monitor locations with reasonable accuracy. SO_2 is predicted most poorly due to difficulties in capturing plumes from coal combustion. For the other 11 pollutants considered, R^2 values ranged from 35.8% to 83.8% from the cross-validation analysis. Because of their ability to capture spatial and temporal variations, concentration fields produced here are well suited for use in epidemiological studies.

Two one-step methods were also investigated. When implemented for NO_2 and $PM_{2.5}$ in 2010, these alternatives were not able to predict concentrations as well as the original method, but are computationally much more efficient. It was found that developing and using models of annual mean concentration fields can account for some of the mismatch between point measurements and 12-km gridded CMAQ simulations and thus improves predictions. For larger scale applications, such as over the entire U.S., it is recommended that a one-step method incorporating annual mean models be implemented to provide results for use in health studies.

CHAPTER 1: INTRODUCTION

1.1 Background and Motivation

Ambient air pollutants provide a major risk to human health, and the adverse effects that they bring about have been well documented. Among these effects are impaired respiratory function, cardiovascular disease, aggravation of respiratory diseases such as asthma, and even death. Because of these serious effects, the associations between air quality and morbidity and mortality have been a topic of close investigation (Ozkaynak, et al., 2009). The associations between health and ambient air quality can vary greatly between different pollutants, as well as different pollutant mixtures. Concentrations of pollutants, and therefore their effects on health, vary over both time and space. Therefore, health studies that investigate these associations require concentration data that are both spatially and temporally accurate across many pollutants. Highly resolved spatial and temporal information on these pollutants is necessary to gain insight into the concentrations of pollutants different people, such as those living in different regions or in neighborhoods of different socioeconomic conditions, are being exposed to and for how long (O'Lenick et al., 2017). It is also important to have this information for a wide range of pollutants because pollutants may interact with each other to result in effects that would not be seen with only individual pollutants.

Epidemiological studies often only use observational data collected from monitoring networks to assess personal exposure levels. These regulatory monitoring networks are known to give air quality measurements with low amounts of error. In a study done by Goldman, et al., correlations between collocated monitors at several sites and multiple pollutants showed a high degree of precision, with correlations ranging from 85% and 99.8% (Goldman, et al., 2010). However, when using only field data to estimate ambient concentrations across a wide area, it is often impossible to gain a complete spatial, temporal and chemical picture of all pollutants.

Monitors often have infrequent or inconsistent data collection. Many do not take readings every day, or do not take measurements consistently throughout the year, which causes temporal data to be incomplete. Although the measurements have low error when giving concentrations at the monitor location, they do not give any information on concentrations at locations removed from the monitor, and correlations between monitor measurements and actual concentrations decrease significantly with increasing distance from the monitor. The number and distribution of monitor locations is also very limited, causing spatial information to be sparse. The majority of monitors tend to be located in cities, and many pollutants have few monitors located away from urban areas, so there is very little information on the concentrations of these species in rural areas.

On the other hand, chemical transport model simulations are able to give complete concentration estimations that are highly resolved spatially and temporally. Models can give estimates for a vast array of pollutants, over large areas at small time-steps and fine resolutions. The mechanisms in these models are developed from lab and field measurement information on chemical species, reactions, rate constants and photochemistry that are converted to differential equations. These are coded into computer models with numerical solvers and are then used to estimate the fate of air pollutants. Because they are based on emission and meteorological data inputs and not on observational data, these models can give concentration estimates for places and times where there are no monitors taking measurements. Therefore, there are no gaps in the simulation outputs. However, these models have their own set of limitations as well. With increasing chemical species, reactions, time steps and spatial resolutions, computational demands increase significantly. Computing times and storage constraints considerably restrict how thorough and exhaustive the mechanisms of a model can be. Incomplete knowledge of atmospheric chemistry, especially of organic reactions, also limits the inclusion of exhaustive

chemical mechanisms. These gaps of information must be filled with assumptions and estimates that often limit the accuracy of these models (Stockwell, 2012).

In this study, the Community Multi-scale Air Quality (CMAQ) model is used. This is an air quality chemical transport model developed by the U.S. Environmental Protection Agency (EPA). It makes use of emission data from both anthropogenic and natural sources, meteorological data and a chemical-transport model to generate ground level concentration estimates for a wide range of pollutants (EPA, 2016). The model can provide concentration fields over time and space at various resolutions. However, results of the model are completely independent of observations and the model is known to have biases. Inaccuracies in emission or meteorological data or the specifications in the chemical-transport model can all strongly affect CMAQ's accuracy. CMAQ also does not effectively capture small-scale day-to-day variations in pollutant concentrations, and is known to over- or under-predict many species. As part of the Public Health Air Surveillance Evaluation (PHASE) project, the EPA along with Centers for Disease Control (CDC) used CMAQ to generate hourly concentration fields for the continental US at a resolution of 12 km. These data are now publically available.

In a study performed by the U.S. EPA, the performance of CMAQ version 4.5 was thoroughly examined through comparisons to observational data over the eastern United States. The study recorded the overestimation of ozone (O_3) when observed O_3 was low, and an underestimation while observations were high. These findings were consistent with those seen in previous versions of CMAQ (Appel et al., 2007). $PM_{2.5}$ was found to be overestimated in winter and fall, with a normalized mean bias of over -30%, while in the summer the normalized mean bias is only -4.6%. Particulate nitrate and ammonium are also largely over-predicted in the fall in the eastern United States, likely due to overestimations of seasonal ammonia emissions. In late spring and summer, carbonaceous aerosols were found to be under-represented in the eastern

United States, likely because of incomplete secondary aerosol formation pathways (Appel et al., 2008).

1.2 Objective

The objective of this study is to combine the benefits of observed data and chemical transport model simulations while reducing the effects of their limitations in order to create highly resolved and accurate pollutant concentration fields. Although estimations from chemical transport models have biases, the use of observations can correct for these inaccuracies, and the limitations of spatially and temporally incomplete monitor data can be overcome by blending these measurements with estimations from a spatiotemporally complete model. Through the interpolation of observed monitor data, while using the chemical transport model, CMAQ, as a guideline to provide more coherent spatial and temporal information, spatiotemporally resolved ambient air quality fields were created.

Friberg, et al. developed this fusion process as well as the process to evaluate the method's performance, and implemented the process over the domain of Georgia as well as four cities for the years 2002-2008 (Friberg et al., 2016). Here, the method was implemented over the same Georgia state domain for the years of 2009-2012 for 12 pollutants. Of these 12 pollutants, five are gases and seven are particulate matter (PM) species. The five gases of interest are carbon monoxide (CO), ozone (O_3), sulfur dioxide (SO_2), nitrogen dioxide (NO_2), and total nitrous oxides (NO_x). The seven particulate matter species are total $PM_{2.5}$, total PM_{10}, and $PM_{2.5}$ subspecies sulfate (SO_4^{2-}), nitrate (NO_3^-), ammonium (NH_4^+), elemental carbon (EC), and organic carbon (OC). In order to assess the performance and to quantify uncertainties from the data fusion process, the results were evaluated using data withholding. The fusion method that Friberg developed is a three-step process that performs very well in estimating daily

concentration fields, but is very computationally intensive and could be greatly simplified. Therefore, two alternative one-step methods were assessed to attempt to simplify the procedure while retaining the accuracy of the original method. The performances of the two alternative methods were evaluated using a comparison with the original method's results as well as cross validation through data withholding.

CHAPTER 2: 2009-2012 DATA FUSION, GA

2.1 Introduction

In this study, observed pollutant concentration data from ground-based monitoring networks were fused with chemical transport model simulations. CMAQ simulations resolved at 12km were used for the chemical transport model. Data fusion was conducted in order to create daily pollutant concentration fields over Georgia for four years (2009-2012) and 12 pollutants that are both accurate as well as spatially and temporally complete.

Fusion of observations and model simulations is desired because neither data set is able to provide complete and coherent concentration information on their own. Observational data acquired from ambient monitoring networks give highly accurate pollutant concentration measurements. However, monitors give very little spatial information, and are often temporally incomplete. Conversely, chemical transport simulations offer concentration estimates that are both spatially and temporally complete, and are based on emission and meteorological inputs rather than observations. However, model limitations including computational constraints, incomplete atmospheric chemistry knowledge, and inexact emission and meteorological inputs cause biases in the simulations and reduce accuracy.

In order to fuse the two data sets, two different pollutant fields are developed. The first field is based on the interpolation of observations while using CMAQ to capture spatial trends. The second field is based on CMAQ simulations that have been adjusted to the observations using the relationship between the two data sets. These two fields are then fused using a weighting factor based on how well the observation-based field predicts temporal variation compared to the CMAQ-based field at any point. This method allows for the creation of fields that capture the benefits of both data sets, while minimizing the effects of both of their

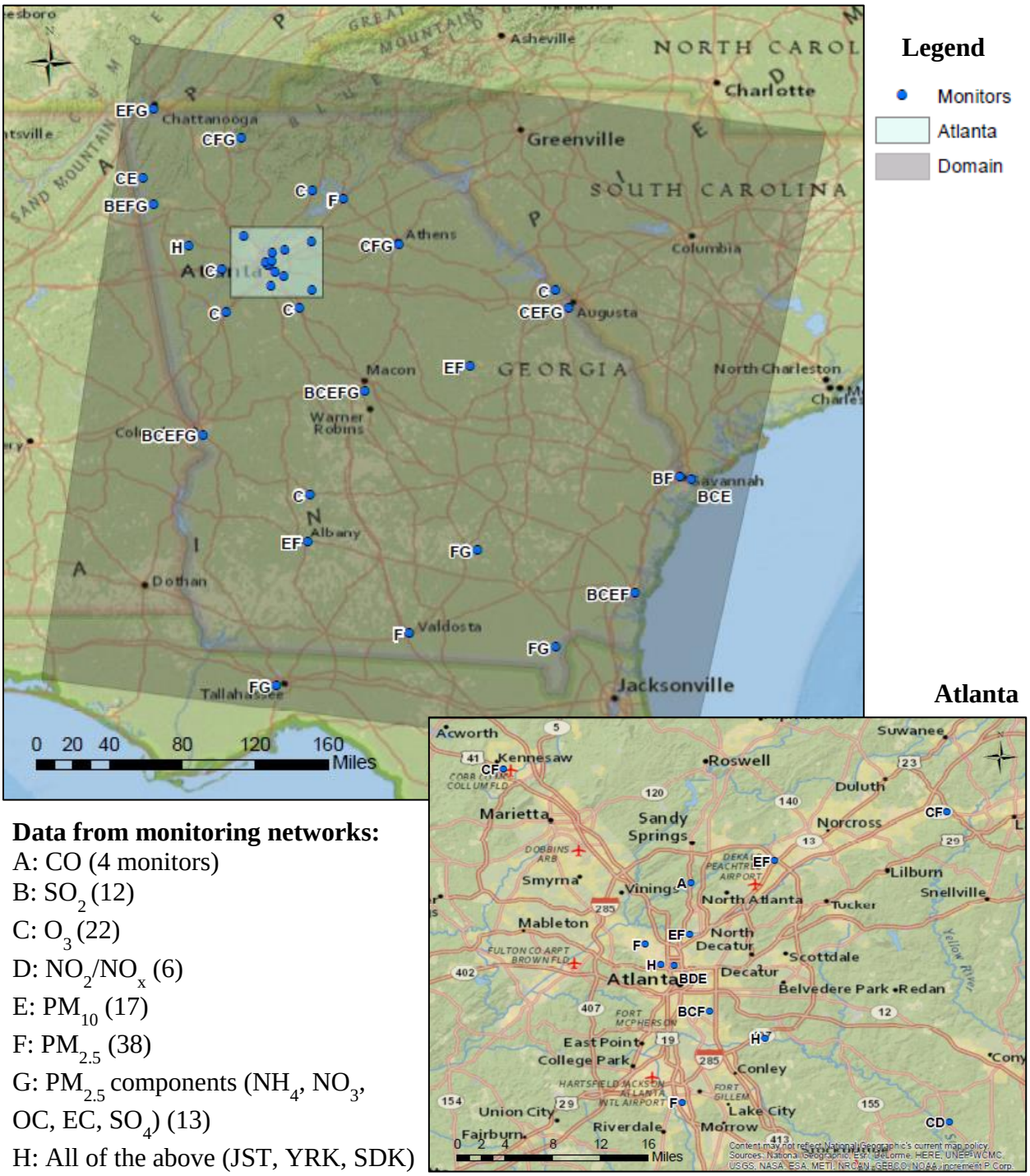
limitations. These concentration fields can be used in future health studies that investigate the associations between air quality and acute health effects.

2.2 Ambient Monitoring Networks Specifications

In this study, a total of 12 pollutants were observed: five gases and seven particulate matter species. The five gases of interest are carbon monoxide (CO), ozone (O_3), sulfur dioxide (SO_2), nitrogen dioxide (NO_2), total nitrous oxides (NO_x). The 7 particulate matter (PM) species are total $PM_{2.5}$, total PM_{10}, and $PM_{2.5}$ subspecies sulfate (SO_4^{2-}), nitrate (NO_3^-), ammonium (NH_4^+), elemental carbon (EC), and organic carbon (OC). The data for these 12 pollutants were obtained from monitors across Georgia, as well as from two monitors close to the Georgia border. These monitors are located in Chattanooga, Tennessee and Tallahassee, Florida. Data were obtained from the U.S. EPA's Air Quality System (AQS), the Southeastern Aerosol Research Characterization (SEARCH) network, the Interagency Monitoring of Protected Visual Environments (IMPROVE) and the Assessment of Spatial Aerosol Composition in Atlanta (ASACA) network.

Figure 1 shows the location of the monitors used. There were 64 total monitors used, however, each monitor only measured concentrations from a selection of the 12 pollutants. Only three of the monitors, Jefferson Street (JST), Yorkville (YRK) and South DeKalb (SDK), provided data for all 12 pollutants, with the exception of 2009, when only the JST and YRK monitors provided data for all species. These two are also the only monitors to take daily measurements for all species. Because JST is located in an urban location close to the center of Atlanta, while YRK is located in a more rural area, about 45 miles west of Atlanta, these two monitors are used to demonstrate the performance of the fusion method in different population densities as well as monitor densities. As seen in the map of Figure 1, most monitors are located

in urban locations, close or in the major cities of Georgia, especially Atlanta. Often, there are multiple monitors of the same pollutant located close to one another in these cities, while the few monitors in rural areas do not have other monitors in close proximity. Therefore, if monitors in rural areas are missing parts of their data sets, there are rarely other sources of observational data close by to give an approximation of pollutant concentrations in the area.



Data from monitoring networks:

A: CO (4 monitors)
B: SO_2 (12)
C: O_3 (22)
D: NO_2/NO_x (6)
E: PM_{10} (17)
F: $PM_{2.5}$ (38)
G: $PM_{2.5}$ components (NH_4, NO_3, OC, EC, SO_4) (13)
H: All of the above (JST, YRK, SDK)

Figure 1 – Georgia domain and location of ambient air quality monitors that supplied data during the 2009-2012 study time frame

The monitoring networks provided the best spatial coverage for $PM_{2.5}$, with up to 38 monitors providing data, and O_3 with up to 22 monitors across Georgia. There is very sparse

spatial coverage for NO_2, NO_x and CO, with as few as five monitors for NO_2 and NO_x and three monitors for CO. The measurement frequency and temporal coverage also varies greatly from pollutant to pollutant and among monitors. Most monitors record concentration data every day for the five gases, while many monitors only take measurements for particulate matter species concentrations every three or six days. Additionally, only three of the 22 O_3 monitors provide daily data year round. The other 19 O_3 monitors only provide daily data eight months out of the year: March to October, when ambient temperatures and ozone levels tend to be greatest.

Data are recorded on an hourly basis by the monitoring networks, while daily metrics were necessary for this study. In order obtain daily concentrations from hourly data, metrics based on the National Ambient Air Quality Standards (NAAQS) were calculated according to the units identified as primary exposure variables in health analyses (Friberg et al., 2016). These metrics were one-hour maxima for CO, NO_x, NO_2 and SO_2, eight-hour maxima for O_3, and 24-hour average for total PM mass and all PM species. On days with 16 hours or more of recorded hourly data, any missing hourly measurements were estimated by linear interpolation. On days with fewer than 16 hourly data points, the day was treated as missing. Because it is common to have days with fewer than 16 recorded hourly concentrations, the data sets for all pollutants were temporally incomplete, even when sampling frequency is taken into account. The completeness of the data sets as well as the total number of monitors and readings for each pollutant and year is summarized in Table 1. Completeness indicates the percentage of available versus expected daily readings based on sampling frequency for all monitors and days in the specified years of the study.

Table 1 – Summary of monitoring networks 2009-2012

Species	Year	Monitors	Readings	Sampling Frequency	Completeness
O_3 8-hr max (ppb)	2009-2010; 2012	22	16897	Daily: 3; Seasonally Daily (March to October): 19	97.9%
	2011	21	5420	Daily: 3; Seasonally Daily (March to October): 18	98.5%
NO_2 1-hr max (ppb)	2009	6	1876	Daily: 5; 1-in-3: 1	96.4%
	2010-2012	5	5351	Daily: 5	97.6%
NO_x 1-hr max (ppb)	2009	6	1885	Daily: 5; 1-in-3: 1	96.8%
	2010-2012	5	5345	Daily: 5	97.5%
CO 1-hr max (ppm)	2009	3	1059	Daily: 3	96.7%
	2010-2012	4	4272	Daily: 4	97.4%
SO_2 1-hr max (ppb)	2009	10	10120	Daily: 9; 1-in-3: 1	96.7%
	2012	9	3201	Daily: 9	97.2%
$PM_{2.5}$ 24-hr avg ($\mu g/m^3$)	2009-2010	38	13429	Daily: 13; 1-in-3: 21; 1-in-6: 4	89.0%
	2011	29	5201	Daily: 9; 1-in-3: 18; 1-in-6: 2	92.9%
	2012	24	3731	Daily: 5; 1-in-3: 17; 1-in-6: 2	92.7%
PM_{10} 24-hr avg ($\mu g/m^3$)	2009	14	1051	Daily: 1; 1-in-3: 2; 1-in-6: 11	82.3%
	2010	14	1173	Daily: 1; 1-in-3: 2; 1-in-6: 11	91.8%
	2011	15	1739	Daily: 4; 1-in-6: 11	92.2%
	2012	15	1984	Daily: 4; 1-in-6: 11	92.9%
SO_4 24-hr avg ($\mu g/m^3$)	2009-2012	13	5774	Daily: 2; 1-in-3: 3; 1-in-6: 8	91.2%
NO_3 24-hr avg ($\mu g/m^3$)	2009-2012	13	5727	Daily: 2; 1-in-3: 3; 1-in-6: 8	90.5%
NH_4 24-hr avg ($\mu g/m^3$)	2009-2010	11	2431	Daily: 2; 1-in-3: 1; 1-in-6: 8	90.8%
	2011-2012	13	2922	Daily: 2; 1-in-3: 3; 1-in-6: 8	92.2%
EC 24-hr avg ($\mu g/m^3$)	2009-2012	13	5811	Daily: 2; 1-in-3: 3; 1-in-6: 8	91.8%
OC 24-hr avg ($\mu g/m^3$)	2009-2012	13	5447	Daily: 2; 1-in-3: 3; 1-in-6: 8	86.0%

Table 2 describes the data obtained from the monitoring networks including the interquartile range (IQR). This information is given for all four years of the study and all 12 pollutants. In general, the characteristics of the observed concentrations are very similar throughout the four observed years. However, for most pollutants, average concentrations tend to increase slightly between 2009 and 2010, and then decrease slightly from 2010 to 2012.

Table 2 – Data description of observed data from monitoring networks

Species	Year	Average	Minimum	Maximum	IQR	Standard Deviation
O_3 8-hr max (ppb)	2009	40.82	0.52	103.60	18.33	13.37
	2010	46.51	2.44	98.13	18.51	13.15
	2011	47.26	4.88	98.47	19.61	13.92
	2012	44.01	3.25	122.88	17	13.02
NO_2 1-hr max (ppb)	2009	15.69	0.80	68.27	21.01	14.33
	2010	16.73	0.93	77	23	15.69
	2011	16.16	0.63	82.53	22.61	15.16
	2012	14.87	0.92	70.1	19.99	13.61
NO_x 1-hr max (ppb)	2009	38.66	0.84	462	37.59	63.86
	2010	42.30	1	582	39.38	68.15
	2011	38.25	0.81	545.66	35.10	63.93
	2012	35.60	0.93	417.5	36.43	55.86
CO 1-hr max (ppm)	2009	0.48	0.10	3.3	0.44	0.37
	2010	0.51	0.10	2.51	0.47	0.36
	2011	0.50	0.10	2.12	0.47	0.32
	2012	0.51	0.11	1.92	0.48	0.33
SO_2 1-hr max (ppb)	2009	5.84	0.11	157	5	9.69
	2010	6.64	0.07	112	6	10.48
	2011	6.00	0.05	161.8	5.4	10.21
	2012	4.32	0.08	160.6	2.41	9.85
$PM_{2.5}$ 24-hr avg ($\mu g/m^3$)	2009	10.76	0.1	55.5	6.59	5.30
	2010	11.89	0.1	96.8	7.12	5.60
	2011	11.61	0.1	64.5	7.7	5.86
	2012	9.80	0.08	44.7	5.15	4.13
PM_{10} 24-hr avg ($\mu g/m^3$)	2009	19.33	1	94	11.46	9.42
	2010	18.70	1	62	11.73	8.84
	2011	18.83	1	65	11.90	8.50
	2012	16.42	1	68	9.40	6.91

Table 2 Continued

Species	Year	Average	Minimum	Maximum	IQR	Standard Deviation
SO_4 24-hr avg ($\mu g/m^3$)	2009	2.35	0.18	7.66	1.45	1.22
	2010	2.57	0.20	8.28	1.61	1.16
	2011	2.45	0.02	8.38	1.84	1.40
	2012	1.89	0.06	5.70	1.13	0.88
NO_3 24-hr avg ($\mu g/m^3$)	2009	0.55	0.01	7.16	0.46	0.68
	2010	0.61	0.03	6.14	0.49	0.67
	2011	0.46	0	6.26	0.30	0.51
	2012	0.41	0	4.13	0.29	0.38
NH_4 24-hr avg ($\mu g/m^3$)	2009	0.88	0.05	3.78	0.61	0.50
	2010	0.93	0.05	3.24	0.57	0.45
	2011	0.88	0	7.76	0.71	0.60
	2012	0.63	0	2.47	0.50	0.37
EC 24-hr avg ($\mu g/m^3$)	2009	0.57	0.01	4.81	0.45	0.49
	2010	0.64	0.04	4.52	0.48	0.57
	2011	0.63	0	6.33	0.46	0.54
	2012	0.58	0	3.24	0.44	0.45
OC 24-hr avg ($\mu g/m^3$)	2009	2.34	0.31	18.48	1.53	1.48
	2010	2.66	0.31	17.6	1.71	1.65
	2011	2.80	0.02	62.25	1.94	2.57
	2012	2.53	0.34	34.45	1.59	1.53

2.3 CMAQ Specifications

Chemical transport model simulations for the 12 considered pollutants were obtained from the PHASE project. In this project, the EPA and CDC generated publicly available CMAQ simulations for the continental U.S. CMAQ version 5.0.2 at a resolution of 12km was used to give concentration fields for the 2009-2012 study-period, while CMAQ version 4.5 was used in Friberg's 2002-2008 data fusion. There were large updates to CMAQ in the new model including improvements to gas-phase chemistry, aerosol chemistry and speciation, and transport processes. These, along with many other improvements, have allowed CMAQ to become a significantly more robust model (Adelman, 2012).

CMAQ simulations were created for the entire continental U.S., however only the domain of Georgia was considered here. The coordinates of the considered domain are (35.5409,

-85.4686) in the northwest corner, (34.8267, -79.9788) in the northeast corner, (30.4894, -86.2089) in the southwest corner and (29.8262, -81.0590) in the southeast corner. Like the monitoring networks, CMAQ model simulations were run to give hourly estimates of all pollutants. Daily concentration estimates were again calculated from hourly data based on the primary standard units of the NAAQS. However, because model simulations are able to offer complete data sets, no hourly data points were missing. Unlike for the monitoring network data sets, it was then unnecessary to interpolate any missing data points, and daily metrics could be calculated for every day within the considered timeframe for all pollutants.

Table 3 describes the data used from the CMAQ simulations in 2009-2012. Temporal minimum and maximum values were determined by finding the CMAQ field average for each day and taking the minimum and maximum of those values. The average values given are the mean of all CMAQ values over all days of the year. Also included in the table are the yearly average values seen at the Jefferson Street (JST) monitor and the Yorkville (YRK) monitor, as well as the corresponding average values seen from CMAQ in the cells in which the monitors are located.

Table 3 – CMAQ simulation description for 2009-2012

Species	Year	Average	Temp. Minimum	Temp. Maximum	CMAQ JST Avg.	Monitor JST Avg.	CMAQ YRK Avg.	Monitor YRK Avg.
O_3 8-hr max (ppb)	2009	43.16	17.41	61.41	35.65	37.57	42.03	41.18
	2010	46.54	21.23	72.88	40.81	43.04	48.51	47.08
	2011	45.66	21.51	67.06	39.87	44.47	45.23	44.57
	2012	43.68	19.69	69.09	37.72	43.56	45.01	44.34
NO_2 1-hr max (ppb)	2009	7.36	2.57	16.68	41.13	31.33	10.53	5.37
	2010	8.41	2.51	17.10	54.24	33.91	11.40	5.18
	2011	7.21	2.48	15.33	42.39	30.98	10.21	4.86
	2012	7.25	2.63	16.82	42.64	29.24	10.40	4.00
NO_x 1-hr max (ppb)	2009	8.25	2.72	22.79	74.68	82.41	11.89	5.83
	2010	10.64	2.62	29.65	179.91	82.91	13.11	5.74
	2011	8.00	2.51	19.67	70.94	70.84	11.26	5.29
	2012	8.12	2.70	23.04	79.04	66.67	11.19	4.45

Table 3 Continued

Species	Year	Average	Temp. Minimum	Temp. Maximum	CMAQ JST Avg.	Monitor JST Avg.	CMAQ YRK Avg.	Monitor YRK Avg.
CO 1-hr max (ppm)	2009	0.21	0.11	0.78	0.78	0.54	0.25	0.20
	2010	0.23	0.13	0.48	1.35	0.54	0.25	0.20
	2011	0.20	0.10	0.43	0.72	0.49	0.22	0.20
	2012	0.20	0.11	0.39	0.78	0.48	0.22	0.19
SO_2 1-hr max (ppb)	2009	1.91	0.35	6.06	3.35	8.50	2.84	4.22
	2010	2.00	0.35	5.80	3.27	10.69	2.51	3.39
	2011	1.60	0.36	5.09	3.16	8.92	2.06	2.90
	2012	1.13	0.28	3.85	2.02	2.57	1.50	1.76
$PM_{2.5}$ 24-hr avg ($\mu g/m^3$)	2009	8.07	1.39	66.09	13.37	10.22	8.75	9.22
	2010	8.95	1.82	41.60	19.54	12.15	10.27	10.25
	2011	8.19	1.46	29.22	13.70	11.16	8.42	10.46
	2012	7.48	1.89	32.87	13.89	9.43	8.32	8.90
PM_{10} 24-hr avg ($\mu g/m^3$)	2009	10.34	2.10	73.20	17.03	16.07	10.71	12.79
	2010	11.10	2.75	47.90	26.44	20.12	12.31	15.32
	2011	10.87	2.38	33.63	18.54	17.81	10.78	15.36
	2012	10.19	2.94	39.44	18.74	15.48	10.49	10.50
SO_4 24-hr avg ($\mu g/m^3$)	2009	1.95	0.43	4.31	2.32	2.50	2.12	2.45
	2010	2.23	0.66	5.76	2.61	2.68	2.40	2.54
	2011	1.75	0.46	4.28	2.15	2.61	1.88	2.61
	2012	1.48	0.46	3.03	1.84	1.95	1.68	1.99
NO_3 24-hr avg ($\mu g/m^3$)	2009	0.53	0.01	5.75	1.33	0.66	0.87	0.56
	2010	0.70	0.01	6.02	1.40	0.74	1.11	0.61
	2011	0.46	0.01	3.34	1.08	0.57	0.74	0.53
	2012	0.49	0.02	4.92	1.18	0.47	0.85	0.37
NH_4 24-hr avg ($\mu g/m^3$)	2009	0.71	0.15	2.16	1.05	0.99	0.90	0.98
	2010	0.87	0.23	2.66	1.12	1.02	1.09	1.03
	2011	0.61	0.12	1.61	0.89	0.94	0.76	0.98
	2012	0.53	0.11	1.99	0.81	0.69	0.73	0.72
EC 24-hr avg ($\mu g/m^3$)	2009	0.55	0.09	5.89	1.96	0.79	0.63	0.34
	2010	0.61	0.09	3.24	4.31	0.95	0.71	0.36
	2011	0.52	0.10	2.32	1.77	0.81	0.57	0.40
	2012	0.51	0.10	2.41	1.99	0.74	0.62	0.37
OC 24-hr avg ($\mu g/m^3$)	2009	1.83	0.19	27.39	3.11	2.79	1.86	2.04
	2010	1.94	0.19	13.18	4.80	3.10	2.14	2.39
	2011	1.87	0.18	10.07	3.28	2.93	1.71	2.56
	2012	1.70	0.26	9.92	3.47	2.76	1.72	2.41

In general, CMAQ behavior stays constant through the four observed years. It often follows the observed trends seen in Table 2, but as seen in the comparisons between the JST and

YRK monitor and CMAQ averages the trends do not follow observations consistently. In many instances observation averages increase or decrease, while CMAQ simulations do the opposite. Figure 2 shows the spatial trends of CMAQ through the four-year average spatial fields of the simulations. These fields have been normalized to their maximum average concentration.

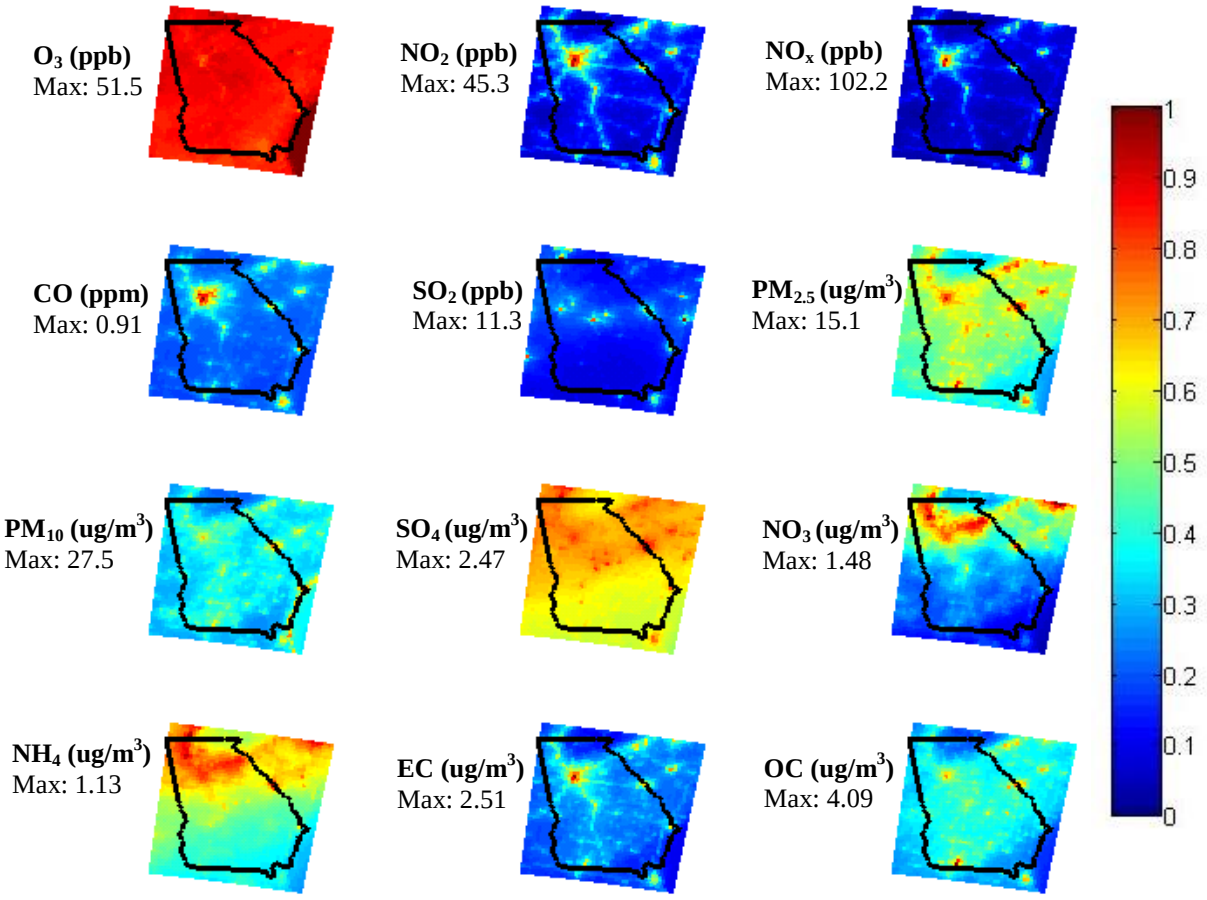


Figure 2 – Average CMAQ fields for all pollutants 2009-2012

Primary pollutants NO_2, NO_x, CO, SO_2 and EC tend to have steep concentration gradients. Their concentrations tend to be significantly higher directly around their primary sources than at locations removed from the sources. Other than SO_2, on-road emissions are the largest source of emissions for all of these primary pollutants. The effects of this can be seen in

their spatial fields, where the highest concentrations of these pollutants are observed over major urban centers and roadways. The largest source of SO_2 is coal-combustion power plants. The impact of these point sources can be seen in the SO_2 average spatial field.

The other seven pollutants considered here are secondary pollutants or have mixed origins. These pollutants tend to be more spatially homogeneous than primary pollutants. Although some pollutants show peaks of concentrations over urban centers, the concentrations drop more gradually than primary pollutants when moving away from the urban centers.

2.4 Data Fusion Method

The data fusion process developed by Friberg is a three-step process, which includes the creation of two different concentration fields that are then fused. The first concentration field (FC_1) is developed by first normalizing daily observations using the yearly average values of the respective monitor. These values are then interpolated using the kriging interpolation method, and the interpolated field is denormalized using the annual CMAQ field that has been adjusted to observations. This field allows the observations to dictate daily temporal variation, which monitors are able to capture very well. It also allows CMAQ to provide spatial trends, which it does much more realistically than observations. However, day-to-day concentration variability is predicted poorly at locations far from any monitors, and interpolation of spatially scarce monitor data causes spatial variability to be captured poorly.

The second field (FC_2) is developed using daily CMAQ simulations that have been adjusted to observations. In this field, temporal variation is independent of monitor location, and spatial variation is based only on CMAQ, and so is only dependent on emission and meteorological data. However, biases in CMAQ simulations restrict the accuracy of this field.

Finally, these two fields are fused using a weighting factor (W) to create a final fused concentration field (C^*). This weighting factor is based on the spatial correlation of observations and the correlations between observations and CMAQ. These two correlations dictate the performance of FC_1 and FC_2 respectively, so give information as to which field is more representative of what is actually occurring at any point. A summary of the entire method is presented in Figure 3.

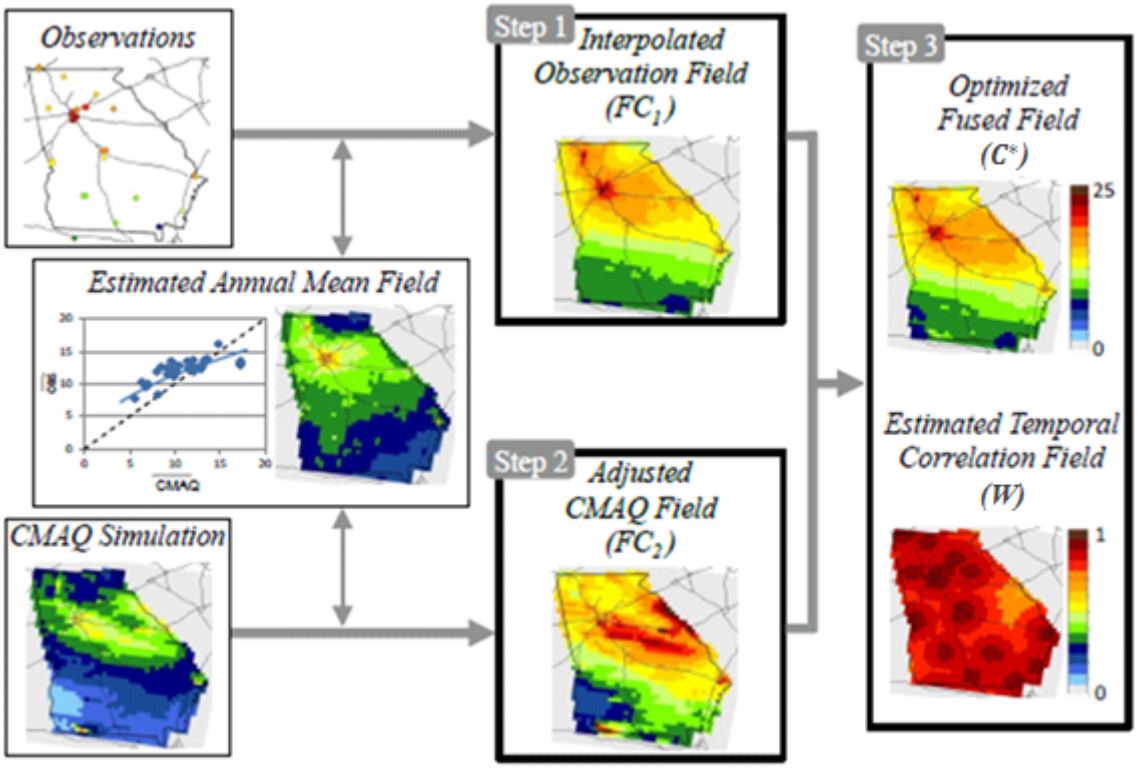


Figure 3 – Summary of data fusion method created by Friberg using 24-hr $PM_{2.5}$ (ug/m^3) fields for July 23, 2008 at a resolution of 12 km

2.4.1 Interpolated Observation Field (FC_1)

The interpolated observation field (FC_1) is created by first normalizing daily observation values (OBS) by the yearly averaged observed values at each monitor. These values are then

spatially interpolated using the krig interpolation method and are denormalized using the adjusted yearly average CMAQ field.

Kriging is a geostatistical interpolation method, which is based on the relationship between the observation points. It assumes that the distance between monitors is related to the correlation of the measurements, and this relationship is used to create variation in the surface between the monitors. Ordinary kriging assumes that there is an unknown, constant mean function that describes the spatial variation of pollutant concentrations. This produces a smooth, continuous surface of estimated values. However, this method does not take into account any chemical or physical processes, and the variation of concentrations is not actually completely smooth, so this leads to the creation of incorrect spatial features. By normalizing the daily observation values with yearly averages, a smoother data set is provided for interpolation. To further correct for the inaccurate spatial information resulting from interpolation, CMAQ is used to denormalize the created spatial field. Equation 1 describes the procedure of creating this interpolated observation field.

$$FC_1 = \left(\frac{OBS}{\overline{OBS}}\right)_{krig} \times \alpha\overline{CMAQ}^{\beta} \qquad (1)$$

Here, α and β are parameters developed from the relationships between observations and CMAQ data. Yearly observed averages at each monitor are compared to the yearly average CMAQ value at the cell in which the monitor is located. Only CMAQ values on days on which there are monitor readings are considered when calculating yearly CMAQ averages. Using regression, parameters that describe the relationship between the averages of the two data sets can be developed. These parameters attempt to correct for the annual biases CMAQ presents. One β value is developed for all four years for each pollutant, while α values are determined for each year. For the five gases, as well as EC, a β value of one is used, while non-linear fits

improved the prediction of other PM species. All α and β used for all four years can be found in the Appendix.

Figure 4 shows the monitor yearly averages for the four years being considered against the CMAQ yearly averages in the cells that the monitors are located in on days for which there are monitor readings for all pollutants. These relationships are used to develop the α and β parameters, which are then used to adjust the annual CMAQ fields. These are used to denormalize the interpolated observation fields in the creation of the FC_1 spatial field. The plots shown here include the annual averages for 2009-2012, so if a monitor gave measurements for all four years it has four points on the plot. However, in the actual determination of the parameters, each year was done separately. It can be seen that CMAQ performance varies widely from pollutant to pollutant. CMAQ predictions align least with PM_{10}, and EC observed values, with total R^2 values of 0.02 and -0.45 respectively when the y-intercept is set to zero. CMAQ predictions and observations agree the most closely with NO_x, SO_4 and NO_3, with R^2 values of 0.78, 0.65 and 0.63 respectively. In general, CMAQ and observations tend to most closely agree with pollutants with the least spatial variation.

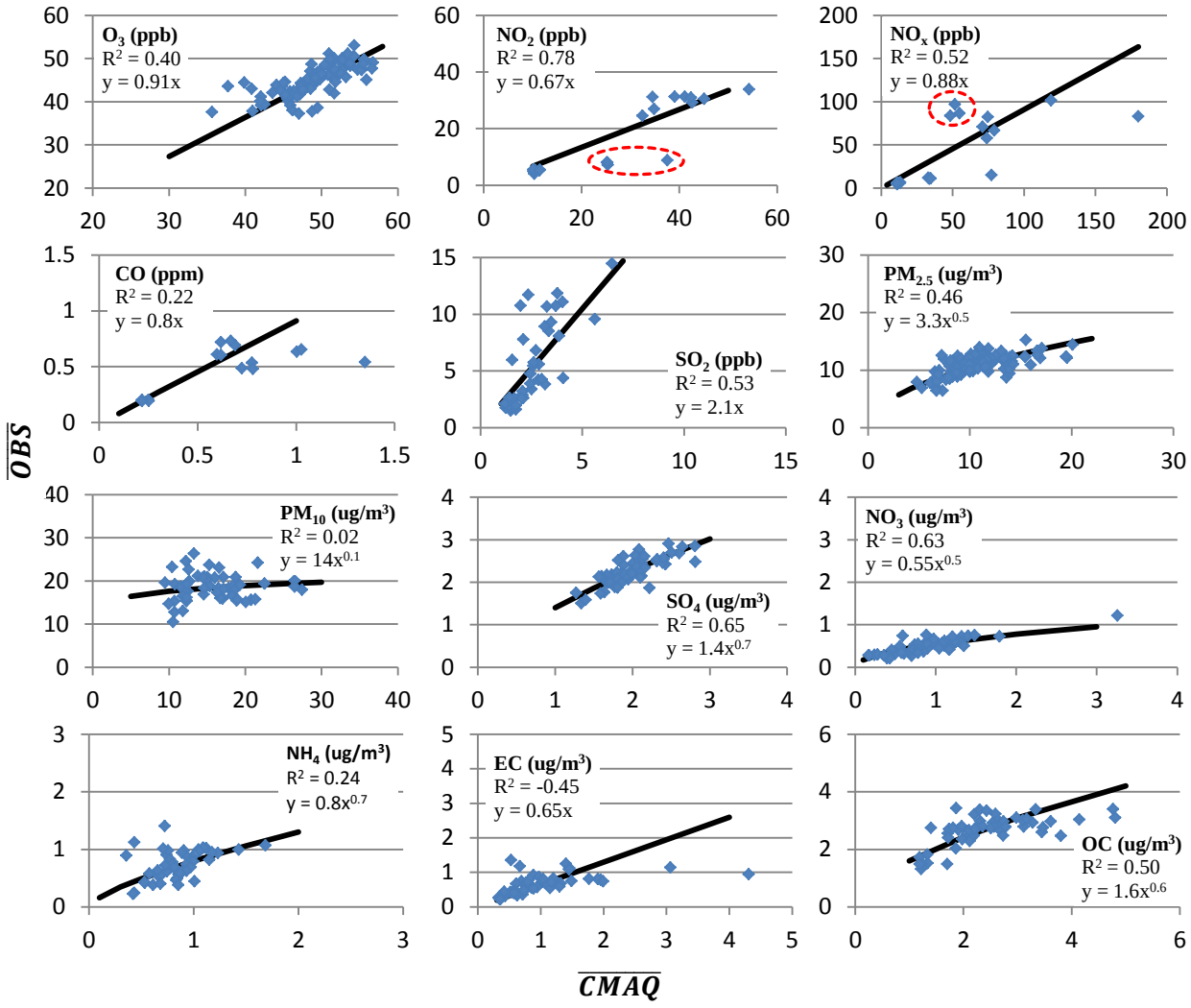


Figure 4 – CMAQ and observed annual averages used to develop adjustment parameters. β values and average α values for 2009-2012 are shown

To a certain extent, it is expected to see differences in CMAQ and monitor values. This is largely because the pollutant concentration estimate that CMAQ gives a cell is the average concentration for the entire 12 x 12 km cell, while monitors only give measurements at the one point within the cell that they are located. In many cases, some bias is desired in the comparisons of average monitor and CMAQ values, because if CMAQ and observations agreed perfectly, it would indicate that the CMAQ value might not be representative of the entire cell. The result of this difference can be seen in the outliers of the NO_2 and NO_x plots that are circled in red in Figure 4. The points on the NO_2 plot all correspond to averages from the Conyers monitor, while

the points on the NO_x plot all correspond to averages from the South DeKalb (SDK) monitor. The locations of both of these monitors are shown in Figure 5.

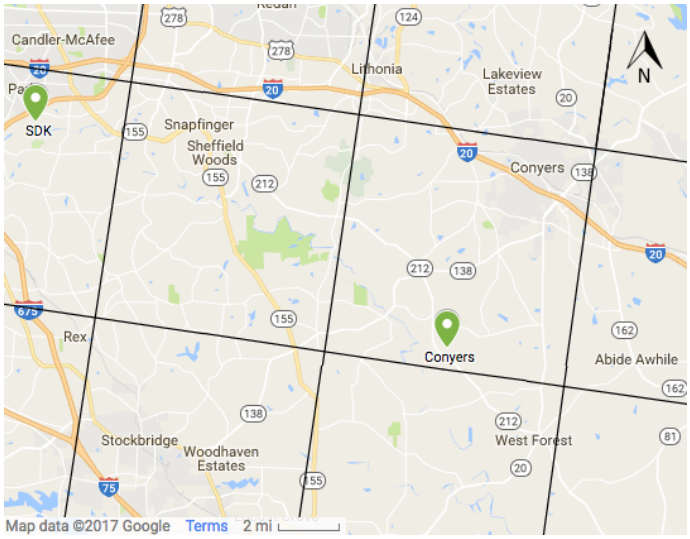


Figure 5 – Location of Conyers and South DeKalb (SDK) monitors in relation to their CMAQ 12-by-12km cell

On the map, it can be seen that the Conyers monitor is located at the south edge of its 12-by-12km CMAQ cell, denoted by the black lines, while Route I-20 runs through the north of the cell. In this case, it is not surprising that CMAQ values in the cell in which the Conyers monitor is located are higher than observations. The primary source of NO_2 in cities is from vehicular fossil fuel combustion, so major roadways, like I-20 are large sources of NO_2. However, because the Conyers monitor is located far from I-20, it will not pick up much of the NO_2 emitted from the interstate. The monitor is located in a more rural location within the cell and will have relatively low concentrations of NO_2, while the north likely has higher NO_2 concentrations. Therefore, the average cell concentration will be higher than the monitor concentration.

When looking at the outliers on the NO_x plot it can be seen that at the South DeKalb monitor the opposite trend occurs, where the average CMAQ values are much lower than observations. This monitor is located very close to Route I-20 as well as other major roadways,

and so is hit by high concentrations of NO and NO_2. However, concentrations of NO_x will not remain that high throughout the cell and the cell average should be lower than what is observed at the monitor.

The creation of the FC_1 field is dictated by daily observations, which are able to capture day-to-day variations in concentrations at locations close to monitors. However, far from monitors, this information is more difficult to capture. Spatial variability is also captured poorly through the interpolation of monitor data. Because monitors are limited in quantity and spatial distribution, and data sets from these monitors are incomplete, there are often very few data points being interpolated relative to the size of the domain. Kriging produces a smooth spatial field from daily observational data, which is based on the assumption that there is a continuous progression of concentrations between monitor locations. However, in actuality, concentrations will vary between monitors based on meteorology, emission source locations, chemistry, geography and other factors, which are not taken into account by interpolation. Therefore, there may actually be multiple high and low concentration regions between monitors that kriging will not capture, which leads to the creation of invalid spatial structures. However, by denormalizing the kriged field with the yearly-adjusted CMAQ field, more realistic spatial trends are created.

2.4.2 Adjusted CMAQ Field (FC_2)

The adjusted CMAQ field (FC_2) is created based on the daily CMAQ fields instead of daily observations. CMAQ annual fields that have been adjusted to the annual mean observations using the same α and β parameters previously developed are normalized by the unadjusted annual CMAQ field. This factor is used to scale the daily CMAQ fields to the observations. Equation 2 describes the procedure of creating the adjusted CMAQ field.

$$FC_2 = CMAQ \times \left(\frac{\alpha \overline{CMAQ}^{\beta}}{\overline{CMAQ}} \right) \quad (2)$$

The method developed by Friberg originally allowed for the inclusion of a seasonal correction factor in FC_2. However, here it was determined that this correction factor was unnecessary, and was not included. Because the creation of this spatial field is independent of monitor locations, FC_2 is able to capture the temporal variation of concentrations at points far from monitors better than FC_1. Because all spatial features are provided by CMAQ, and are then based on chemical and physical processes and well as emission and meteorological data instead of interpolation, spatial information is also more realistic. However, parameterized adjustments made to the annual CMAQ fields describe the yearly associations between the model and observations, and may not correctly account for the day-to-day differences in CMAQ and actual events. Therefore, this field is still susceptible to the biases in CMAQ simulations, which will limit the accuracy of this field.

2.4.3 Optimized Fused Field (C^*)

Once the two spatial fields, FC_1 and FC_2, have been created, they are combined to create a final, optimized fused field (C^*). Equation 3 shows the creation of the weighted average of these two fields.

$$C^* = W \times FC_1 + (1 - \mathrm{W}) \times FC_2 \quad (3)$$

The fusion of these two fields is dictated by a weighting factor (W), which is based on the correlations between monitor values (R_1) as well as correlations between monitor and CMAQ values (R_2). This weighting factor dictates the amount of weight the first field will be given, and varies both spatially and temporally depending on how close a point is to a monitor, and what

monitors have readings on a specific day. R_1 is a spatial field that depends on the distance a point is from a monitor. Because not all monitors give readings daily, this field changes over both time and space. As distances between monitors increase, the correlations between their measurements decrease. This relationship of decreasing correlation with increasing distance holds true for monitor measurements and points removed from any monitor as well. Equation 4 is the equation used in the development of R_1.

$$R_1 = E_i \times e^{-d/r} \quad (4)$$

Here, E_i is the y-intercept of the relationship between monitor correlations and distance (*d*). This results from instrument error and is typically close to one. The value *r* is the range of the relationship. This denotes the distance in kilometers that the correlations have decreased by a factor of *e*. The correlations used in the development in R_1 can be seen in Figure 6. This figure shows the Pearson correlation values between observations from monitors as a function of distance between monitors. The equations shown in the correlograms are the parameters used in the development of R_1. Correlations of monitor data from 2009 to 2013 were used in these correlograms. Because correlations change very little from year to year, the same parameters were used in all four years of the data fusion implemented here. The same parameters can be used if data fusion is implemented for the same domain in 2013.

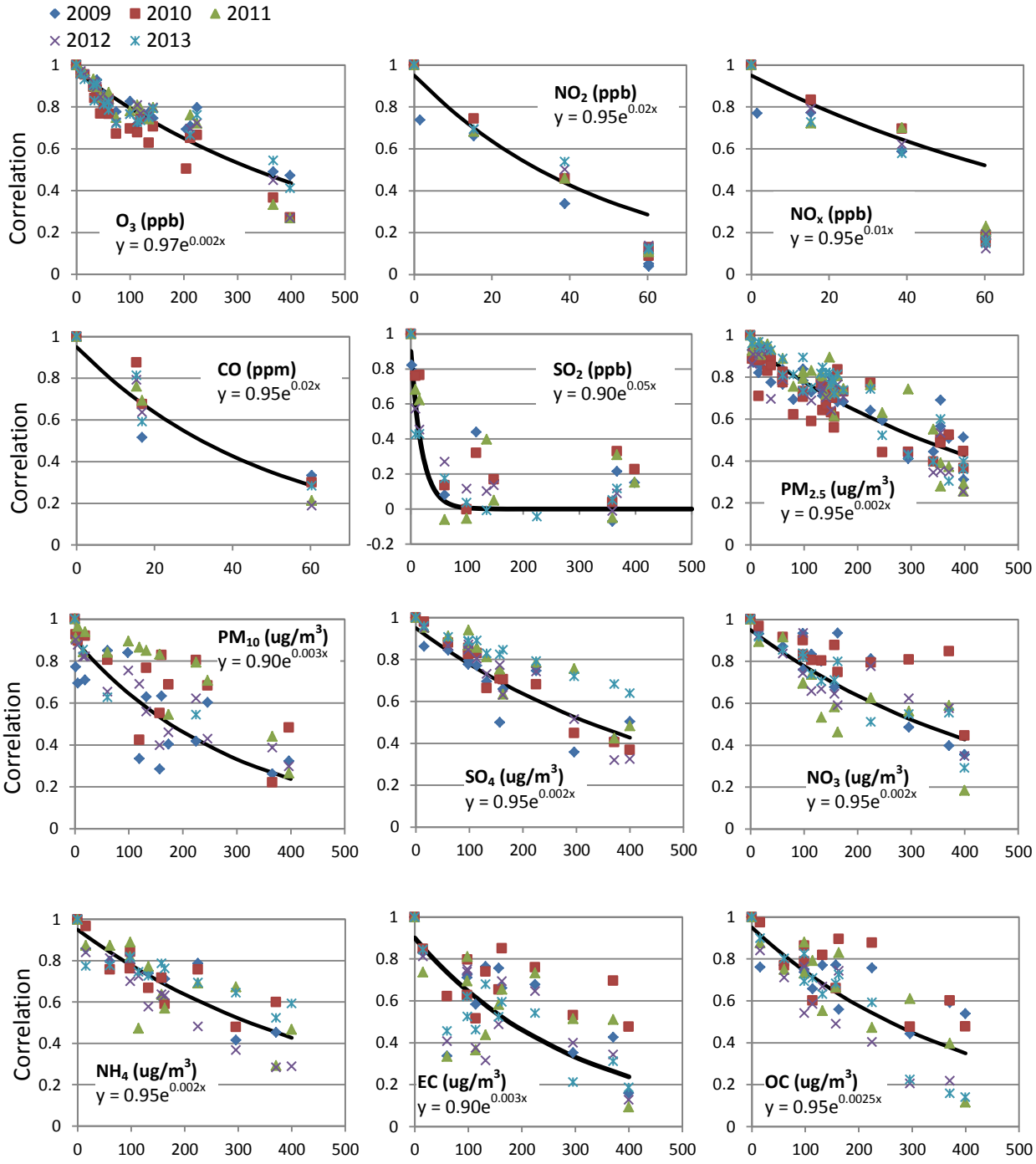


Figure 6 – Correlograms of observed monitor data 2009-2013. Curves and equations represent parameters used in fusion

R_2 is a constant for each year and pollutant. This value is the temporal correlation between the monitor values and CMAQ values at monitor locations on days that there are

monitor readings. Equation 5 shows the calculation of W based on R_1 and R_2. E_i and r parameter values used in R_1 and R_2 values for all years and pollutants can be found in the appendix.

$$W = \frac{R_1 \times (1 - R_2)}{R_1 \times (1 - R_2) + R_2 \times (1 - R_1)} \quad (5)$$

W is a value between zero and one, and the weight given to the FC_2 field will be one minus W. Because R_1 varies over both time and space, W also varies spatially and temporally. Because of the decreasing correlation between monitor measurements with increasing distance, R_1, and therefore W, are highest close to monitors and more weight is given to the FC_1 field at those points. This plays to the strength of the FC_1, which performs best at locations close to monitors. However, because of the limitations of interpolation, temporal and spatial information is captured poorly by FC_1 at points far from a monitor. At these locations, W will be lower, allowing FC_2, which performs better at these points than FC_1, to have more weight.

2.5 Results

The fusion process results in the formation of daily concentration fields based on both modeled and observed data. The average fields for all 12 pollutants over the four-year period can be seen in Figure 7. These fields have been normalized to their maximum average concentration.

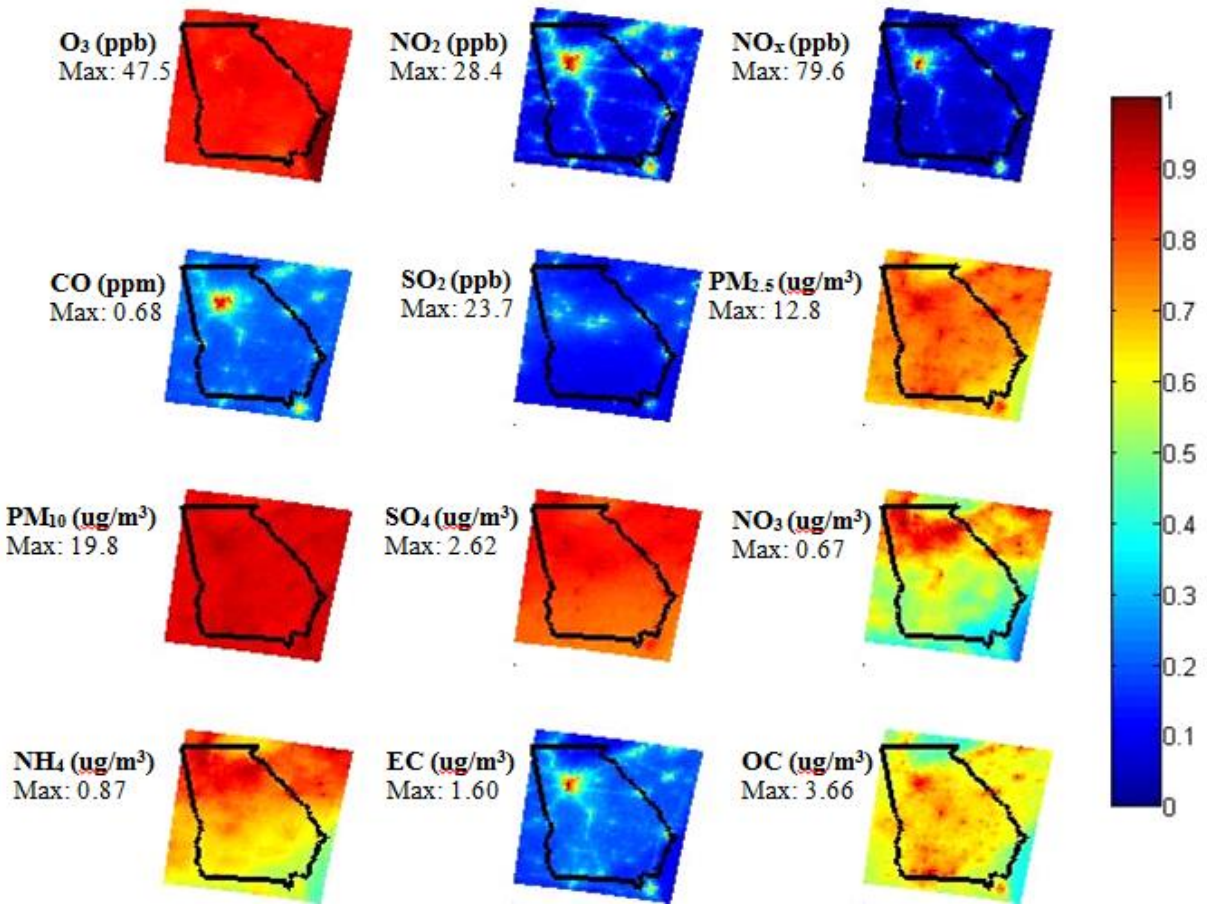


Figure 7 – Average optimized fused fields (C^*) for all pollutants 2009-2012 normalized to maximum concentrations

These fields look spatially similar to the CMAQ fields shown in Figure 2. Primary pollutants, NO_2, NO_x, CO, SO_2 and EC, still display steep urban-to-rural gradients, with the highest concentrations located over major urban centers and roadways, except in the case of SO_2. Highest concentrations here are located close the coal-fired power plants. Secondary or mixed origin pollutants, O_3, $PM_{2.5}$, PM_{10}, SO_4, NO_3, NH_4 and OC, are more spatially homogeneous than the primary pollutants. They also tend to be more spatially flat than seen in their average CMAQ fields. For these pollutants, there tends to be fewer and smaller areas of low concentrations and gradients moving away from urban centers tend to be more gradual. Maximum concentrations

for all pollutants also change significantly between the average CMAQ fields and C^* fields, reflecting the impact of the CMAQ adjustments used in the data fusion.

The fusion process performance can be further characterized in two ways. Performance is first statistically characterized through comparisons of the relationships between observations and CMAQ values and the observations and the fused field values. Three different metrics are used in this examination: R^2 values (RSQ), percent root mean squared error (RMSE) and percent mean bias. The fusion method is later evaluated using a cross-validation analysis. Figure 8 shows the R^2 values, RMSE and bias for all pollutants and years. These values were calculated from all monitor values along with CMAQ or C^* values at monitor locations on days with observed measurements. RMSE and bias have both been normalized to the mean observed concentration value for all monitors over the four considered years.

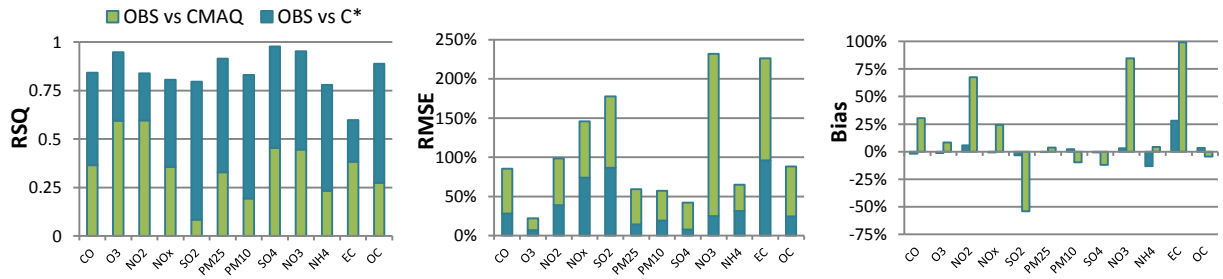


Figure 8 – Comparisons between observations and simulation values using R^2, RMSE and bias

At monitor locations, the R^2 values were significantly increased for all pollutants, ranging from a 41% increase from NO_2 to an 843% increase from SO_2. The normalized RMSE decreased significantly for all pollutants ranging between a 49% decrease for NO_x and 89% for NO_3. The bias is also reduced to close to zero at monitor locations, with C^* biases ranging between 28% and -13%, compared to CMAQ biases between 99% and -54%. RMSE values remain the highest and R^2 values remain the lowest for primary pollutants, while secondary pollutants tend to have

lower RMSE values and higher R^2 values. O_3 and SO_4, both secondary pollutants, agreed best with observations, while SO_2 and EC, both primary pollutants, showed the poorest agreement. In general, the fusion process performed very well when considering these statistics. However, these improvements are expected since the final optimized fields are created using observation values. Resulting C^* values should be more similar to observed values than CMAQ simulated estimations are because the model is independent of observations.

However, the fusion process does result in RSQ values of less than one and non-zero RMSE and bias values, showing that exact observation values are not reproduced during fusion in the cells in which the monitors are located. This is partly because cells may contain multiple monitors, and differences in the monitor readings would be reflected in the final cell value. Additionally, although the interpolated observation field (FC_1) is often weighted very heavily at monitor locations, there are cases in which W is not one, and the adjusted CMAQ field (FC_2) is weighted significantly. This is true especially for pollutants that have low R_1 values, such as EC and PM_{10}, because correlations between monitors are very low. This will cause CMAQ values to more heavily influence C^* values even in cells with monitors. Even when FC_1 is heavily weighted, the annual means of both observations and CMAQ are used in the creation of this field. This will cause the final C^* values to differ from the daily measured values as well. These deviations from observations are desirable because the concentration recorded at the location of a monitor may not be representative of the entire cell in which the monitor is located. By using average CMAQ values, the difference between monitor values and average cell values is more truthfully captured.

2.6 Discussion

The fusion of observational data and CMAQ, a chemical transport model, results in a final optimized field that captures the different strengths of the two data sets while minimizing their limitations. This is accomplished by the creation of two separate spatial fields, FC_1, which is based on the interpolation of observations and FC_2, which is based on CMAQ simulations that have been adjusted to observations. These fields are then combined using a weighting factor dependent on monitor-to-monitor correlations and monitor-to-CMAQ correlations. Because FC_1 is able to capture temporal trends at locations close to monitors, this field is favored in fusion at these points. FC_2 more accurately represents spatial trends at locations removed from monitors, so this field is favored at these locations. The created fields are spatially and temporally complete. This method allows for estimations to be created that are based on observations, physical and chemical processes, and emission and meteorological data. The estimations also are able to reflect concentration variations over time and space.

The estimates produced from this method agree very well with observations. They are significantly more correlated with observation values than CMAQ simulated values are with the same observations. This is to be expected since the estimations from fusion are dependent on observation values. Values do not match exactly with observation values due in part to the fact that some cells have multiple monitors. Additionally, because annual CMAQ and observation values are used in the model, fusion will not yield original observation values. These differences are appropriate because estimations are given as an average value for an entire 12 x 12km cell, while observations are only representative of one point within that cell. Later this method will be evaluated using a cross-validation method in order to assess how well the method can predict concentrations at points where there are no monitor values.

Trends seen from the implementation of this data fusion process here are consistent with results seen in Friberg's implementation of the process in 2002-2008. Spatially, all fused fields look very similar from both implementations, with primary pollutants displaying very steep urban-to-rural concentration gradients and secondary pollutants looking more spatially homogeneous. Friberg's characterization also found that RMSE values decreased and R^2 values increased significantly for all pollutants in the comparison between observations and CMAQ and C^*. These improvements are similar to those seen in the characterization seen here. R^2 values are highest for O_3 and SO_4 in the 2002-2008 data fusion as well as the 2009-2012 data fusion. It was previously found that the most significant improvement in R^2 values came from PM_{10}, $PM_{2.5}$ and NH_4, largely due to seasonal corrections. Although no seasonal corrections were used here, large improvements in R^2 values were still observed for these pollutants. RMSE values were highest for EC, SO_2 and NO_x in both implementations.

The major weaknesses of this method come from the limitations of the two data sets used. There is a scarcity of ambient air quality monitors within the Georgia domain, which causes observational data to be incomplete spatially. Because many of these monitors do not record concentrations daily, the data are also incomplete temporally. Conversely, CMAQ is spatially and temporally complete, but because of the inherent biases within the model, its accuracy is restricted.

Another drawback of this method comes from the complexity in its implementation. The development of the two correlations in the calculation of the weighting factor is computationally intensive. Additionally, FC_2 initially had the ability to include a seasonal correction, but no seasonal corrections were used in this case, which makes some steps redundant. Because of these factors, it is beneficial to simplify this procedure. Later, a one-step method that does not require

the development of a weighting factor or any correlations is evaluated against the three-step process used here.

2.7 Conclusion

Observational data from monitoring networks and CMAQ, a chemical transport model, were fused to create concentration fields that are resolved over time and space. The data fusion process was implemented for 12 pollutants during the years of 2009 to 2012 over Georgia. These spatially and temporally resolved fields agree well with observations, with secondary pollutants agreeing better than primary pollutants. The results also exhibit similar trends as to what was observed in past studies, when the same data fusion method was implemented for 2002-2008.

CHAPTER 3: EVALUATION OF DATA FUSION BY DATA WITHHOLDING

3.1 Introduction

In order to thoroughly evaluate the performance of the data fusion process and to validate the model, a cross-validation analysis was performed through observational data withholding. Previously, the final optimized spatial field (C^*) created from the fusion process was assessed by comparing the observational data with the final values of the simulations at the cells in which the monitors were located. This allows us to evaluate model performance at observation sites because monitors give accurate concentration measurements at the points that they are located. Very close to monitors, such as throughout the cell that the monitor is located, the monitor readings are closely related to the actual concentrations. However, although monitors give relevant information as to the concentrations of the cells that they are located in, the measurements should often not be the exact value of the cell's average concentration. This is because of the differences in concentrations across the cell. The monitor will only pick up concentrations from one point, while the cell value should be an average of all points within the cell. The comparison between observations and C^* values therefore is useful to assess how much the model uses the monitor values at those points.

However, this comparison would provide a biased evaluation of the performance of the entire method because the creation of the fused field depends heavily on the inputted observational data. Therefore, it is expected that the resulting fields agree much more with this data, compared to how well the CMAQ data, which does not depend on observational data, agrees with observations. Additionally, quantifying how well the model represents observational data at points where there is a monitor is only a partial assessment of the model. It is desirable to

evaluate how well the model estimates concentrations of pollutants at locations that there are no monitors.

To determine how well the model predicts values both near and far from observations, data withholding can be used. Here, a random 10% of all monitor data were removed for all 12 pollutants and all four years (2009-2012) considered in the initial data fusion. The fusion process was then conducted again, assuming that everything else remained constant. The final fused values that resulted at the points where data were withheld are then compared to the observational data at those points and times. Because the simulated results do not depend on the monitor data recorded at those location, this is a more unbiased and comprehensive picture of model performance across the domain.

The Jefferson Street (JST) and Yorkville (YRK) monitor data withholding performances are shown separately from performance from all monitors as well. These two monitors are the only monitors that measure all 12 considered pollutants on a daily bases for all four years and are situated in very different geographic locations. The JST monitor is located in an urban area, near the center of Atlanta, and close to multiple other monitors. Conversely, the YRK monitor is located west of Atlanta, in a relatively rural location with no other monitors within close proximity. These two monitors are singled out to demonstrate the impact of monitors in close proximity to the points being estimated.

3.2 Method

To perform data withholding, a random 10% of all monitor data were removed from the original observational data set for all pollutants and all four years. It was assumed that all parameters developed previously, such as the α and β values for the adjustment of CMAQ to observations and correlation values between monitors, were not impacted significantly.

Additionally, since these specifications influence the performance of the model, if these parameters were adjusted, the results would not truthfully represent the performance of the initial data fusion. Therefore, all values determined previously were kept constant.

In Yorkville, there are two monitors positioned within 100 feet of each other that both measure O_3, $PM_{2.5}$, NO_2 and NO_x. By having two monitors in such close proximity, data withholding at these sites becomes an ineffective evaluation. Because of their proximity, the two monitors will typically record very similar concentrations. When data are removed from one of the monitors, the second monitor is still available to provide observational data at the same location. The result of the data fusion at those points would reflect the values given by the second monitor, which will be strongly correlated with the monitor value that was withheld. This will cause an overestimation of the performance of the data fusion at these points. Therefore, in these cases with two monitors in close proximity, when one of the Yorkville monitor's data points is withheld, the second Yorkville monitor's values are also removed. This prevents the misrepresentation of estimations at this location.

Because additional data points were removed for four of the species after the initial random 10% were removed, more than 10% of the observational data were actually withheld for these species. However, the total amount of data withheld remained between 13.4% and 10% for the four species. When accounting for data removed from both Yorkville monitors, the range of data points withheld ranged from 2388 for O_3 to 534 for CO across all four years. Table 4 gives a more in depth description of the data withheld for the evaluation including the total number of points withheld for each pollutant and year, and the total percent of data withheld. Because the number of observations recorded each year, and the number of times one of the Yorkville monitors is removed, the total number of observations removed changes each year. These statistics show how representative the withheld points are of the total observed data set. When

compared to Table 2 from Chapter 2, it can be seen that for the most part, the characteristics of this data set are similar to the total data set.

Table 4 – Characteristics of withheld points

Species	Year	Monitors	# Withheld (% withheld)	Avg.	Min.	Max.	IQR
O_3 8-hr max (ppb)	2009	22	598 (10.7%)	40.70	5.75	81	18.75
	2010	22	608 (10.7%)	46.39	2.75	90.75	18.5
	2011	21	575 (10.6%)	46.88	5.5	85.38	19.09
	2012	22	607 (10.7%)	44.19	3.5	105.13	17.13
NO_2 1-hr max (ppb)	2009	6	258 (13.8%)	13.31	1	65.37	18
	2010	5	235 (13.0%)	14.96	0.93	65	16.99
	2011	5	240 (13.3%)	13.70	0.63	57	15.6
	2012	5	233 (13.3%)	12.59	1.3	52.6	19.18
NO_x 1-hr max (ppb)	2009	6	189 (10.0%)	32.84	0.9	347	28.89
	2010	5	181 (10.0%)	47.68	1	338	42.5
	2011	5	180 (10.0%)	45.53	1	514.97	45.6
	2012	5	175 (10.0%)	39.33	1.5	280.38	50.29
CO 1-hr max (ppm)	2009	3	106 (10.0%)	0.54	0.14	2.3	0.45
	2010	4	142 (10.0%)	0.55	0.13	2.27	0.56
	2011	4	144 (10.0%)	0.54	0.15	1.67	0.54
	2012	4	142 (10.0%)	0.54	0.11	1.6	0.51
SO_2 1-hr max (ppb)	2009	10	324 (10.0%)	6.54	0.28	100	6
	2010	10	330 (10.0%)	5.92	0.07	83	6
	2011	10	358 (10.0%)	7.20	0.12	115.6	6.9
	2012	9	320 (10.0%)	4.28	0.1	87.1	2.45
$PM_{2.5}$ 24-hr avg ($\mu g/m^3$)	2009	38	687 (10.2%)	11.04	0.4	38.1	6.44
	2010	38	690 (10.3%)	12.07	0.5	96.8	6.94
	2011	32	520 (10.0%)	11.53	1	41.7	7.82
	2012	25	373 (10.0%)	9.95	1.49	26	4.58
PM_{10} 24-hr avg ($\mu g/m^3$)	2009	15	105 (10.0%)	19.82	4.45	94	10.5
	2010	14	117 (10.0%)	19.04	1	62	11.48
	2011	15	174 (10.0%)	18.02	1	54	11.66
	2012	15	198 (10.0%)	16.45	2.82	39.5	9.27
SO_4 24-hr avg ($\mu g/m^3$)	2009	13	148 (10.0%)	2.37	0.57	6.67	1.53
	2010	13	136 (10.0%)	2.49	0.53	5.55	1.49
	2011	13	148 (10.0%)	2.44	0.65	7.57	1.67
	2012	13	145 (10.0%)	2.01	0.41	5.06	1.27
NO_3 24-hr avg ($\mu g/m^3$)	2009	13	146 (10.0%)	0.45	0.02	3.2	0.4
	2010	13	136 (10.0%)	0.59	0.03	3.21	0.47
	2011	13	147 (10.0%)	0.45	0.04	3.47	0.35
	2012	13	144 (10.0%)	0.37	0.05	2.06	0.3

Table 4 Continued

Species	Year	Monitors	# Withheld (% withheld)	Avg.	Min.	Max.	IQR
NH_4 24-hr avg ($\mu g/m^3$)	2009	11	124 (10.0%)	0.87	0.12	3.48	0.65
	2010	11	119 (10.0%)	0.89	0.08	2.14	0.55
	2011	13	146 (10.0%)	0.87	0.01	2.74	0.63
	2012	13	146 (10.0%)	0.66	0	1.98	0.5
EC 24-hr avg ($\mu g/m^3$)	2009	13	146 (10.0%)	0.58	0.05	2.38	0.49
	2010	13	135 (10.0%)	0.70	0.08	4.52	0.53
	2011	13	150 (10.0%)	0.63	0.08	3.25	0.4
	2012	13	150 (10.0%)	0.56	0.06	1.99	0.48
OC 24-hr avg ($\mu g/m^3$)	2009	13	127 (10.0%)	2.46	0.35	7.9	1.58
	2010	13	127 (10.0%)	2.87	0.53	17.6	1.62
	2011	13	145 (10.0%)	2.70	0.14	8.99	1.87
	2012	13	146 (10.0%)	2.42	0.65	6.28	1.58

In the assessment of the fused results produced from withholding, the values produced at the Jefferson Street and Yorkville monitors are shown separately. This displays the effects of monitoring clusters on the performance of the data fusion process. Table 5 describes the data that were withheld from these two sites individually for all years and all pollutants. This shows how well the points taken from these two locations represent the complete data set from those monitors. Although some pollutants have two monitors at the Yorkville site, only one of the monitors takes daily readings for all pollutants. If the withheld points happen to be largely outliers, their removal may skew the results and lead to an incorrect demonstration of model performance. This is the only Yorkville monitor considered here. The average values of the complete data set at these two monitors can be seen in Table 3 in Chapter 2. Although these statistics and the statistics seen in Table 5 do not match the complete data set's characteristics perfectly, for the most part the set of data withheld for all pollutants is representative of the entire data set. Therefore, this evaluation will yield a valid representation of the model performance.

Table 5 – Description of points withheld from the Jefferson Street and Yorkville Monitor

		Jefferson Street Monitor				Yorkville Monitor			
Species	**Year**	**# With-held**	**Avg.**	**Min.**	**Max.**	**# With-held**	**Avg.**	**Min.**	**Max.**
O_3 8-hr max (ppb)	2009	39	37.82	11.46	74.69	51	41.17	8.84	71.75
	2010	29	40.16	13.64	72.24	59	46.32	16.16	75.47
	2011	45	43.47	14.23	82.34	49	42.68	11.85	74.26
	2012	47	41.66	13.07	67.09	57	45.27	22.23	72.01
NO_2 1-hr max (ppb)	2009	38	30.96	10.65	65.37	73	6.05	1.49	36.99
	2010	39	33.29	12.10	58.90	60	5.56	0.93	21.10
	2011	37	30.10	7.93	55.67	63	4.96	0.63	26.85
	2012	31	29.27	10.47	46.86	61	3.79	1.51	17.32
NO_x 1-hr max (ppb)	2009	33	62.98	11.19	276.75	37	6.52	0.90	28.53
	2010	38	75.38	13.22	320.86	36	5.59	1.04	31.92
	2011	40	75.95	9.22	514.97	36	5.89	1.55	27.13
	2012	39	88.69	10.28	280.38	33	3.21	1.50	5.98
CO 1-hr max (ppm)	2009	43	0.47	0.21	1.10	26	0.22	0.14	0.60
	2010	31	0.68	0.19	2.27	38	0.20	0.13	0.31
	2011	30	0.50	0.17	1.67	36	0.21	0.15	0.31
	2012	33	0.49	0.17	1.33	30	0.21	0.11	0.60
SO_2 1-hr max (ppb)	2009	29	8.64	0.98	29.44	33	5.79	0.28	59.07
	2010	34	9.38	0.20	32.34	40	3.39	0.07	13.70
	2011	29	8.46	0.33	23.83	37	3.30	0.12	18.99
	2012	38	2.85	0.14	14.61	33	2.17	0.12	9.92
$PM_{2.5}$ 24-hr avg ($\mu g/m^3$)	2009	40	11.11	4.94	22.15	40	8.46	3.06	21.78
	2010	36	12.33	2.13	23.53	40	11.07	4.03	22.33
	2011	38	11.73	2.88	24.26	42	9.35	3.02	17.66
	2012	28	8.47	1.49	19.78	34	9.38	2.49	22.08
PM_{10} 24-hr avg ($\mu g/m^3$)	2009	6	11.95	7.34	19.93	7	13.85	4.45	22.50
	2010	9	20.95	9.97	35.49	13	14.70	6.41	27.96
	2011	35	17.69	7.16	41.59	14	15.02	1.93	41.65
	2012	32	14.08	2.82	25.55	39	12.07	3.33	25.99
SO_4 24-hr avg ($\mu g/m^3$)	2009	36	2.52	0.57	6.49	41	2.42	0.62	6.56
	2010	39	2.60	0.92	5.45	34	2.66	0.53	5.55
	2011	36	2.55	0.87	7.57	30	2.75	0.70	7.23
	2012	33	1.88	0.52	3.98	37	1.97	0.41	3.30
NO_3 24-hr avg ($\mu g/m^3$)	2009	37	0.63	0.05	2.35	36	0.47	0.02	2.25
	2010	31	0.70	0.11	3.21	34	0.50	0.10	1.79
	2011	29	0.57	0.12	2.36	32	0.47	0.11	1.39
	2012	29	0.38	0.07	1.84	33	0.36	0.11	1.21

Table 5 Continued

		Jefferson Street Monitor				Yorkville Monitor			
Species	**Year**	**# With-held**	**Avg.**	**Min.**	**Max.**	**# With-held**	**Avg.**	**Min.**	**Max.**
NH_4 24-hr avg ($\mu g/m^3$)	2009	33	0.99	0.26	1.95	34	1.00	0.33	3.48
	2010	34	1.04	0.46	2.14	26	0.98	0.49	1.76
	2011	38	1.00	0.23	2.01	26	1.05	0.28	2.33
	2012	37	0.72	0.28	1.62	28	0.75	0.19	1.98
EC 24-hr avg ($\mu g/m^3$)	2009	35	0.73	0.15	1.72	35	0.35	0.05	0.74
	2010	50	0.92	0.29	3.04	26	0.44	0.09	1.16
	2011	30	0.74	0.23	1.57	36	0.39	0.13	1.47
	2012	37	0.70	0.27	1.99	36	0.37	0.15	0.68
OC 24-hr avg ($\mu g/m^3$)	2009	32	3.27	1.57	6.92	32	2.24	0.80	7.90
	2010	26	2.66	1.62	6.53	38	2.28	0.86	4.21
	2011	38	2.78	0.82	6.10	33	2.70	0.57	5.98
	2012	38	2.51	1.13	4.75	30	2.24	1.01	5.07

3.3 Results

Data fusion implemented with data withholding results in daily concentration fields. These fields are based on CMAQ simulated data and observational data that have had 10% of all original values removed. Three different metrics are used to evaluate the performance of this simulation in order to determine how well the model is able to predict pollutant concentrations. The metrics used for the evaluation are the R^2 value, percent root mean squared error (RMSE) and percent mean bias. These values were calculated for all pollutants over all four years using the withheld observational data points and the resulting C^* values at the time and locations that the observations were withheld. These values can be seen for all pollutants in Figure 9. RMSE and bias has been normalized to the average withheld monitor value for each pollutant.

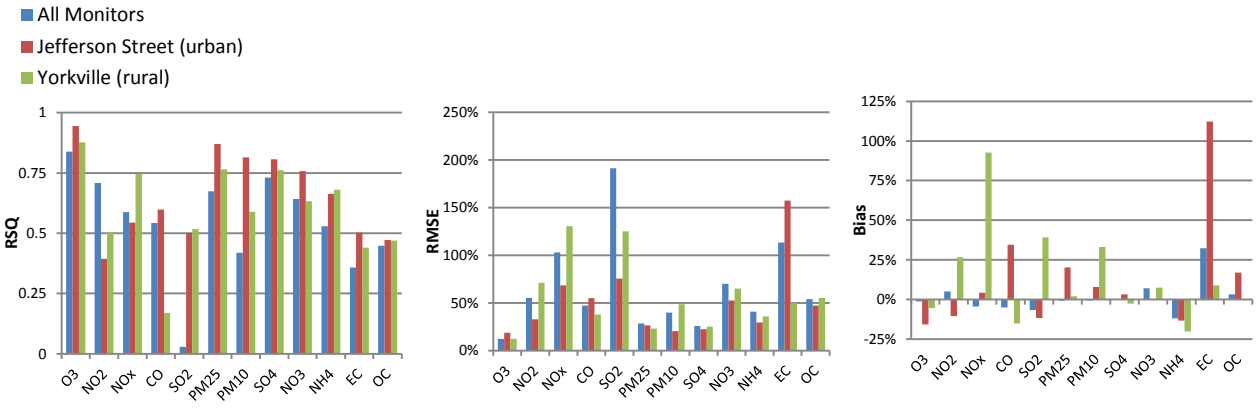


Figure 9 – R^2, RMSE and bias of all 12 pollutants at all withheld monitor locations, the withheld Jefferson Street monitor, and the withheld Yorkville monitor

Not surprisingly, the R^2, RMSE and bias for the fused field compared to withheld observations are worse than in the initial data fusion seen in Figure 8. However, all metrics remain reasonable for all 12 species. This shows that the model is able to adequately predict pollutant concentrations at locations where there are not monitors. Data withholding results for all monitors indicate an average bias between 32.2% and -11.9% for all pollutants. R^2 values range from 0.03 for SO_2 to 0.84 for O_3 and RMSE values range from 191.4% to 12.35%. This model most accurately predicts O_3, a secondary and relatively spatially homogeneous pollutant, with the highest R^2 of all species and the lowest RMSE.

The model tends to struggle the most in correctly predicting SO_2. SO_2 largely comes from coal-fired power plants, and the highest concentrations of this pollutant are found in the plumes coming from these plants. SO_2 predictions result in the highest RMSE and lowest R^2 values because the plumes it comes from are both difficult to measure and model accurately. Remarkably, although SO_2 performance is very poor at all monitors, it performs very well at both the YRK and JST monitors. The JST monitor is close to other SO_2 monitors, so the high performance seen at this point is likely due to the relevant data coming from these other monitors. On the other hand, the YRK monitor does not have any monitors located close by that

could offer information about concentrations at this location. However, the YRK monitor is located about 20 miles directly south of Plant Bowen, one of the largest coal-fired power plants in North America. CMAQ is likely able to model the plume from this plant accurately over short distances, and so when the Yorkville monitor is hit by the plume, CMAQ is able to model what actually occurs relatively well. This would allow for the reasonable performance of the final simulations at this point.

EC also performs poorly in this evaluation, yielding the highest bias and RMSE values at the JST monitor. On most days, EC only has two monitors giving measurements: the JST and YRK monitors, while all other monitors only give measurements every three or six days. Unlike other PM species, EC is a primary pollutant with steep urban-to-rural concentrations. EC is therefore harder to model in CMAQ and its performance is negatively impacted when there are fewer monitors more than other PM species. It more accurately predicted at the YRK monitor than the JST because in this rural area, EC concentrations are relatively low and constant, while at the JST monitor, concentrations can vary widely.

Also included in Figure 9 is an evaluation for only the JST monitor and YRK monitors, which take daily measurements for all 12 considered pollutants. Evaluating these two monitors individually allows for the demonstration of the effect of clusters of monitor on the prediction of concentrations. The JST monitor is located in Atlanta's city center and positioned close to other monitors, while Yorkville is in a more rural area, about 45 miles west of Atlanta, and does not have any other monitors nearby. In general, concentrations at the JST monitor are predicted more accurately than at Yorkville because the monitors close to the JST monitor give relevant information as to what is occurring at Jefferson Street even when there are no data from the JST monitor.

The benefit of monitor clusters holds true at most points across the domain and can be displayed through correlograms of the observations and simulated values. The relationship between distance to the closest monitor and how well the data fusion model is able to predict concentrations can be seen in Figure 10. These correlograms were created for all 12 pollutants and four years by relating the correlation between simulated values and withheld observed data points to the distance to the closest monitor after the observation point has been removed. Each monitor location may have distances to the closest recording monitor that vary day to day based on what monitors gave measurements that day and what other data points were removed. Correlations were computed for distances with 10 or more data points withheld over all four years.

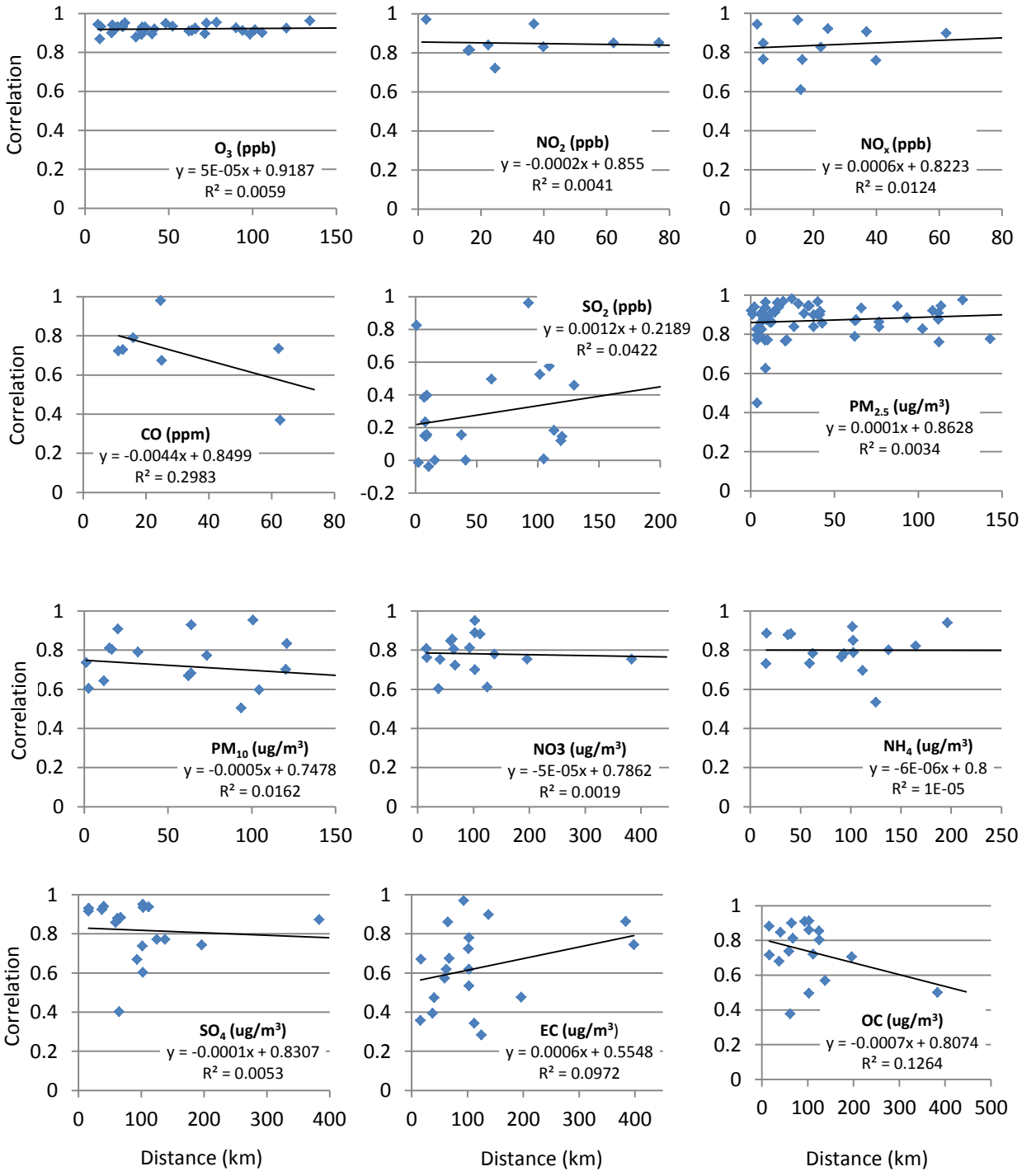


Figure 10 – Correlograms of withheld data for all years and all monitors with more than ten points of data withheld

In general, correlations between simulated values and observations decrease or stay very steady with increasing distances to monitors. This shows that the model is able to most

accurately predict concentrations when there is a monitor close by. However, correlations do remain high for most pollutants, even with great distances. This shows that, although the model does not perform as well, it is still able to predict concentrations far from monitors with reasonable accuracy.

Two pollutants, SO_2 and EC have correlations that increase significantly with distance. These two species seemed to perform the worst in Figure 9 as well. SO_2 correlations increase with distance because of the difficulties associated with capturing plumes from power plants, which carry high concentrations of SO_2. If a plume is hitting a monitor close by to a point, because plumes are thin, it is likely that the same plume is not hitting the point. When that point is being hit by the plume, it is likely that the monitor is not. This leads to very low or even negative correlations between observations and simulations at small distances.

EC also has higher correlations at higher distances. At distances below 200 km, correlations do not increase significantly from smaller distances. However, there are two points with distances slightly below 400 km with relatively high correlations. These points are located in rural areas, far from most of the EC monitors, which are mostly located in or near Atlanta. Because these areas are rural, and as a primary pollutant, EC has very steep urban-to-rural gradients, the concentrations of EC here are constantly low. CMAQ is likely able to capture these low concentrations more accurately than the higher concentrations seen in more urban areas.

3.4 Discussion

The cross-validation results indicate that the data fusion method used here is able to reasonably estimate pollutant concentrations, even when there are no monitors present at the considered location. However, through the comparison of results at the Jefferson Street and

Yorkville monitors, as well as in the correlograms, it can be seen that predictions become less correlated with observations with increasing distances from monitors. Monitor clusters, such as those found around the JST monitor, may exaggerate the performance of the data fusion method due to available additional observational data from nearby monitors. Data withholding at the YRK monitor resulted in lower R^2 values and higher RMSE values than at the JST monitor because of the lack of nearby observations.

Secondary pollutants are estimated better than primary because they tend to be spatially more homogeneous, with lower urban-to-rural gradients. This can be seen by the poor performance of EC, a primary pollutant, and the high performance of O_3, a secondary pollutant in the data withholding process. SO_2, another primary pollutant, performs the most poorly in the data fusion process. This is due to the difficulty of capturing ground-level concentrations of plumes from coal-fired power plants at both monitors and in CMAQ model simulations. The number and distribution of monitors in Georgia provides the most significant limitation of this study, as can be seen in the poor prediction of SO_2 and EC.

The results seen here are consistent with the trends seen in the cross-validation analysis performed by Friberg for the same pollutants and domain but for the years 2002 to 2008. Similarly to what was observed in this study, Friberg found that the data fusion process was better able to predict concentrations at the Jefferson Street monitor than at the Yorkville monitor due to the lack of nearby observations at Yorkville. It was also found that secondary pollutants, such as O_3 and SO_4 are better predicted than primary pollutants such as CO and NO_x because secondary pollutants tend to be more spatially homogeneous. In Friberg's study, O_3 was predicted the best across all monitors, with the lowest RMSE and highest R^2 of all pollutants, while SO_2 was predicted the worst, which is what was observed in this analysis as well. Data withholding resulted in an R^2 of 13.7% for SO_2 at all monitors, which is better than the 2.9%

seen here. However, the R^2 values for SO_2 at the Jefferson Street and Yorkville monitors were much higher in 2009-2012 than 2002-2008. EC also performed better in the 2002-2008 study, with a data withholding R^2 of 53.0%, compared to the 35.8% seen here. The average mean bias remained below 30% across all pollutants in Friberg's withholding, which holds true in this implementation for all pollutants except for EC, which had a mean bias of 32.2%. Previously, the pollutant with the largest mean bias was NO_3; however, this pollutant performs significantly better here, with a bias of only 7.1%.

3.5 Conclusion

A cross-validation analysis was used in order to evaluate the performance of the data fusion process implemented in Chapter 2. It was found that when observed data were withheld, the model is able to reasonably reproduce monitor values. This method performs best with secondary pollutants with low concentration gradients, and at locations near monitors. Results seen here demonstrate trends that are very similar to the results seen in previous studies.

CHAPTER 4: ALTERNATIVE DATA FUSION METHOD DEVELOPMENT

4.1 Introduction

The three-step data fusion method utilized in the data fusion discussed in Chapter 2 and 3 was developed by Friberg, et al. This process is based on the fusion of two spatial fields. One of the fields (FC_1) is based on the interpolation of observations and the yearly CMAQ spatial field. The second field (FC_2) is based on CMAQ, which has been adjusted to average observations. A seasonal correction can also be included in this step, however, in this study it was determined that no seasonal correction was needed for any of the pollutants. These two fields are then averaged together using a weighting factor. This weighting factor is based on the correlations between monitor readings as a function of distance between monitors and correlations between monitor readings and CMAQ values. These correlations can be difficult to determine, especially if implementing the data fusion process over a large domain with a high quantity of monitors. Therefore, determining a method that does not depend on these correlations would be beneficial.

In this study, multiple one-step methods are characterized and evaluated. These methods attempt to simplify data fusion implementation while maintaining the accuracy of the original method. All methods reduce the redundancies caused by excluding the seasonal correction term, and because only one field is developed, no weighting factor is needed, so the monitor-monitor and monitor-CMAQ correlations are not needed. The results of the methods are assessed against the results of the original method as well as observations, and the two processes are later evaluated using data withholding in Chapter 5.

To create the FC_1 field, daily observations are normalized using yearly average observation values. These values are then interpolated, and the field is then denormalized using the adjusted yearly CMAQ field for more realistic spatial trends. In the first one-step method

assessed here, observations are normalized with adjusted daily CMAQ fields. These values are interpolated and then denormalized by the adjusted daily CMAQ fields. The resulting field is the final fused field (C^*). In this method, observations are kept constant at the cell in which they are located. Performance of the data fusion approach depends on observations at locations that capture the spatial variation in what is being interpolated through kriging. However, what is being kriged may not vary smoothly since an observation may not be representative of a cell. Moreover, it is not desirable for C^* to match the observation when it is known that the observation is not representative of the cell. Therefore, this one-step method was revised to allow for values to be interpolated that vary more smoothly, and for final values to be more representative of the cell, instead of only the observations. In order to do this, a term, which is based on the yearly CMAQ adjusted values and yearly average observation values, is included in the calculation of the values to be interpolated.

Observational data from ground-based monitoring networks as well as simulations from a chemical mass transport model, CMAQ, resolved at 12km, were used to create daily pollutant concentration fields that are accurate as well as spatially and temporally complete. These data sets are the same as those used in the original data fusion process discussed in Chapters 2 and 3. The two alternative methods were implemented over the same Georgia domain as the original method. However, because they were implemented only to examine their effectiveness, data fusion was only conducted for 2010, and not all 12 of the original pollutants were used. Only one gas, NO_2, and one particulate species, $PM_{2.5}$, were considered.

4.2 Method

In this study, two pollutants were observed: one gas, nitrogen dioxide, 1-hour max NO_2 and one particulate matter species, 24-hour average $PM_{2.5}$. These two pollutants were chosen in order to evaluate the data fusion processes on two pollutants with very different spatial coverage from monitors. $PM_{2.5}$ has very thorough spatial coverage from monitors. In 2010, 38 monitors across the domain reported $PM_{2.5}$ measurements. Thirteen of these monitors provided daily data, 21 monitors provided data every three days, and the rest recorded measurements every six days. On the other hand, there were only five monitors total in 2010 measuring NO_2. All five of these monitors reported hourly concentrations. NO_2 is also a primary pollutant, while $PM_{2.5}$ is both a primary and secondary pollutant. $PM_{2.5}$ tends to be much more spatially homogeneous than NO_2, while NO_2 has much more dramatic urban-to-rural concentration gradients.

The data for these pollutants were obtained from monitors across Georgia, as well as from one $PM_{2.5}$ monitor in Chattanooga, Tennessee and one in Tallahassee, Florida. Data were obtained from the U.S. EPA's Air Quality System (AQS), the Southeastern Aerosol Research Characterization (SEARCH) network, the Interagency Monitoring of Protected Visual Environments (IMPROVE) and the Assessment of Spatial Aerosol Composition in Atlanta (ASACA) network.

Chemical mass transport model simulations were obtained from the PHASE project. CMAQ version 5.0.2 was used to give daily concentration fields for 2010 at a resolution of 12km. Both the observational and CMAQ data used are the same data that were used in the original data fusion implementation discussed in Chapters 2 and 3.

To keep the assessment as fair as possible, many things were kept constant from the original method. This includes the domain of the study, parameters used to adjust CMAQ to

observations, and the interpolation method used. For information on kriging, the interpolation method used here, or the determination of CMAQ adjustment parameters, see section 2.4.1.

4.2.1 Alternative Method A

The original data fusion process is a three-step method that includes the fusion of two spatial fields. Here, a one-step method based on the simplification of the three-step method, which would allow the fusion to be much easier to perform, is characterized. In the original method two fields are developed, the first one based on the interpolation of observations (FC_1), and CMAQ adjusted to observations (FC_2). These two fields are then fused using a weighting factor, which is based on monitor-to-monitor and monitor-to-CMAQ correlations. In this one-step method, these correlations do not need to be determined. Instead, the observations are interpolated using adjusted daily CMAQ values. More authentic spatial trends are created by denormalizing the interpolated field with the same adjusted daily CMAQ values. Equation 5 describes the procedure of creating these final fused fields in one step.

$$C^* = \left(\frac{OBS}{\alpha CMAQ^{\beta}}\right)_{krig} \times \alpha CMAQ^{\beta} \qquad (6)$$

α and β parameters are found from the relationship between CMAQ and observations. This method greatly simplifies the data fusion process. However, because observations are normalized with the same values that they are denormalized by, the adjusted CMAQ values, the value of a cell with a monitor in it will have the concentration of the monitor. Because monitors only measure concentrations at one point, while a cell value should be the average of the concentrations of all points within the cell, monitor values are not always a truthful representation of the cell. Because it is assumed in this method that the monitor and cell values should agree completely, spatial variation will not be captured accurately through interpolation.

For interpolation through kriging to be effective, the points being kriged should vary smoothly, which they may not.

4.2.2 Alternative Method B

Performance of the data fusion approach depends on observations that accurately represent concentration trends in the surrounding area and so can capture the spatial variation in what is being interpolated. However, observations may not be representative of the 12km cell in which the monitor is located. This can lead to a spatial field with features not consistent with a 12km resolved average field. Therefore, it is not always favorable for the fused field to match the observations when it is known that the observation is not representative of the cell, which is what occurs in Method A. Therefore, a revision to this one-step method was made in order to attempt to create values to be interpolated that would vary smoothly. Equation 6 described this revised one-step method.

$$C^* = \left(\frac{OBS}{\alpha CMAQ^{\beta}} \times \frac{\alpha \overline{CMAQ}^{\beta}}{\overline{OBS}} \right)_{krig} \times \alpha CMAQ^{\beta} \qquad (7)$$

By including a ratio of yearly average adjusted CMAQ values at their respective monitor locations and yearly average observation values, values are created that will vary more smoothly over space. This inclusion will also more truthfully capture the differences between monitor and cell averages. These differences usually come from monitors not picking up concentrations of pollutants from sources far from or downwind of the monitor, but still within the cell. Therefore, even though the monitor reading is precise at its location, it is not the correct average value for the entire cell. These sources, and consequently the relative difference between monitor and cell

values, are fairly constant over time. This bias can then be appropriately described and corrected for through the yearly average ratios of observations and CMAQ.

4.3 Results

Both one-step fusion processes generated daily concentration fields based on both modeled and observed data. In order to characterize and evaluate the performance of these two alternative methods, various comparisons can be made. For both methods, final fused values can be compared with observations, as well as with final fused values from data fusion using the original method, to characterize the performance of the data fusion model. A cross validation analysis is performed using data withholding to evaluate the performance of the data fusion model, which will be discussed in Chapter 5.

4.3.1 Method A Results

The one-step Method A was applied daily in 2010 for two pollutants: 1-hr max NO_2 and 24-hr $PM_{2.5}$. Figure 11 shows the yearly average CMAQ field and average observed values for these pollutants in 2010, and both the original and alternative fused fields. At first glance, it can be seen that both methods produce very similar results. Looking closely, however, the original final fused field for $PM_{2.5}$ yields slightly higher average maximum values and slightly lower minimum values, while the alternative produces slightly less spatial variation and tends to be more spatially flat. Although the values for $PM_{2.5}$ tend to be in the same range from both methods, the alternative method yields more high values than the original. The original NO_2 fused field yields lower average maximum and minimum values than the alternative field. The spatial trends from the two methods tend to be similar, but generally, the alternative yields overall higher values.

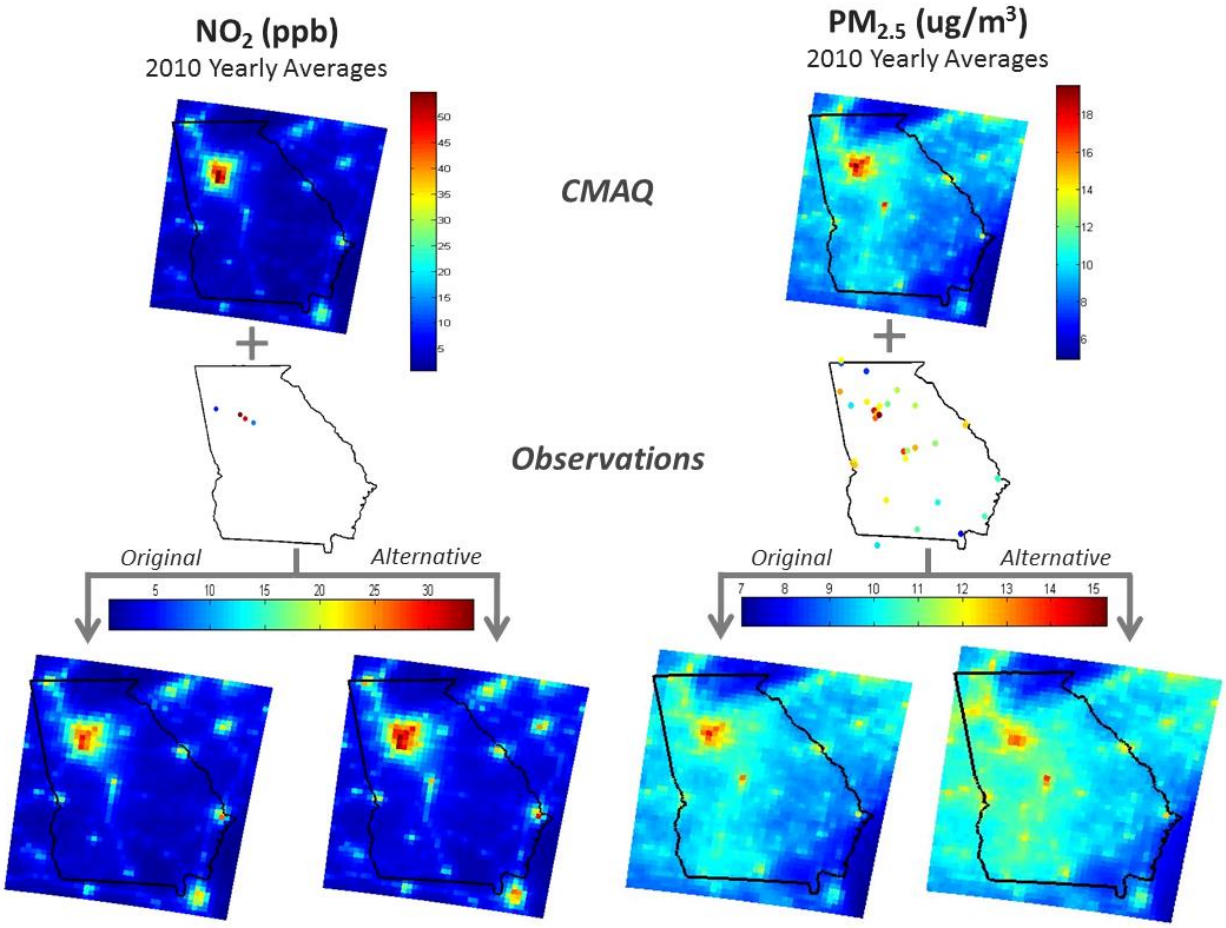


Figure 11 – Yearly average CMAQ and observations for NO_2 and $PM_{2.5}$ in 2010, and original and alternative Method A final fused fields

The performance of Method A can be statistically considered through comparisons of the relationships between observations and the corresponding fused field values for both the original method and the alternative method. Three different metrics are used in this examination: R^2 values, percent root mean squared error (RMSE) and percent mean bias. Table 6 shows these three metrics for NO_2 and $PM_{2.5}$ in 2010 for the alternative method and for 2009-2012 for the original method. These values were calculated from all monitor values along with all C^* values at monitor locations on days with observed measurements. RMSE and bias have been normalized to the mean observed concentration value over the considered years for each method.

Table 6 – R^2, RMSE and bias values for the original and alternative Method A

	Original		Method A	
	NO_2	**$PM_{2.5}$**	**NO_2**	**$PM_{2.5}$**
R^2	0.84	0.91	0.99	0.97
RMSE	39%	14%	2.4%	8.7%
Bias	5.7%	-0.46%	0.0%	0.0%

As seen in the table, the alternative method yields values that correspond to monitor values much better than the original method. It completely eliminates all bias at monitor locations, and has much lower RMSE values compared to the original values and much higher R^2 values- just about one for both pollutants. This is because this alternative method holds monitor values constant in the cells in which they are located, unless there are multiple monitors in one cell. NO_2 has fewer instances of multiple monitors in one cell, so its R^2 is higher and error is lower than $PM_{2.5}$. However, the fact that there are high R^2 values and low errors and biases does not necessarily signify the method's performance is better. As previously explained, there should be some bias because the monitor is not always representative of the cell.

Figure 12 show a comparison of all C^* values generated from the original method and the one-step alternative, Method A. The two methods produce results that are well correlated, but with plenty of scatter. Implementing Method A for $PM_{2.5}$ yields a tighter and more correlated fit with less variance than NO_2. The R^2 values between the two data sets are 0.81 and 0.86 for NO_2 and $PM_{2.5}$ respectively, while percent RMSE is 46% and 18% for NO_2 and $PM_{2.5}$ respectively.

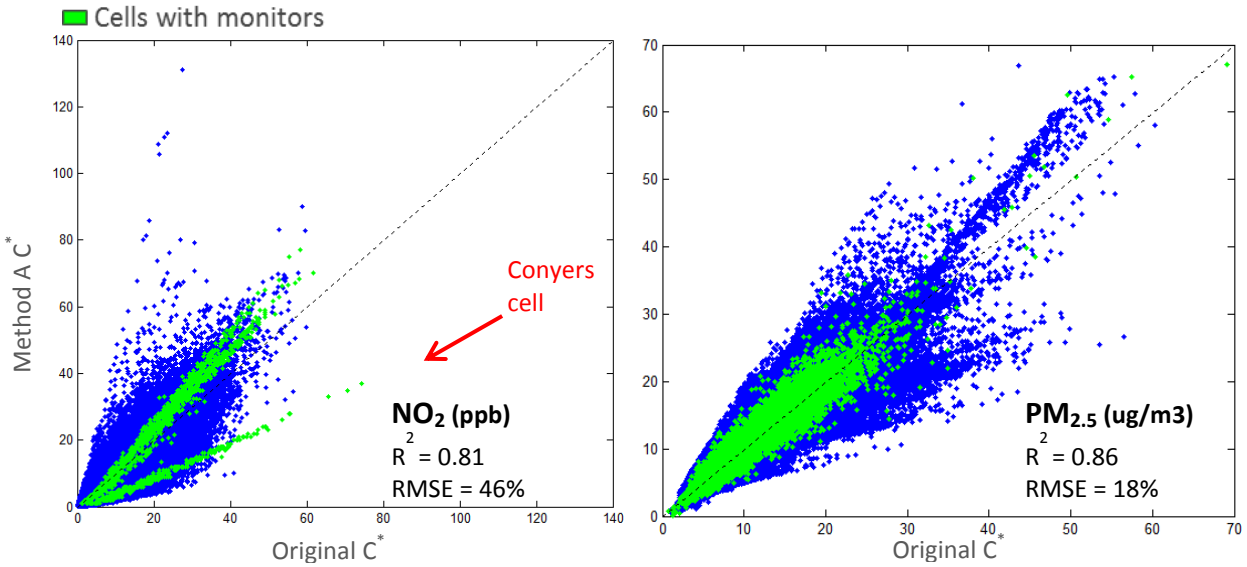


Figure 12 – Plot of all C^* values for NO_2 and $PM_{2.5}$ in 2010 for the original method and alternative Method A

Not only does the $PM_{2.5}$ plot have a tighter fit than NO_2, it also has less over all bias. This agrees with what was seen in the average fused fields in Figure 11, where the final fused fields looked very similar for $PM_{2.5}$, but NO_2 values were overall higher in the alternative method than the original. In both graphs in Figure 12, at cells that have monitors, which are shown in green, the C^* values are strongly linearly related. This suggests that the alternative method captures temporal variation similarly to the original method, but there are some spatial differences between the two. On the NO_2 plot, however, there is high bias between monitor values, while monitor values for $PM_{2.5}$ are centered on the identity line. The low green line for NO_2 corresponds to the cell in which the Conyers monitor is located. The alternative method gives lower NO_2 values than the original method in this case because the observations at that monitor are lower than what would likely be seen across the entire cell. The Conyers monitor is located at the south edge of the CMAQ grid, while a highway runs through the north of the grid. The monitor here does not pick up the high NO_2 concentrations coming from the highway, but the original method is able to reflect these differences.

The other four NO_2 monitors are biased high in Method A. These higher values at cells with monitors will lead to higher values across the domain after interpolation. This explains why the average spatial NO_2 field tends to have higher values from the alternative method than from the original method.

4.3.2 Method B Results

In order to create a smoother interpolation surface and to account for differences between monitor and cell values, the one-step Method A was revised by including a ratio between yearly CMAQ averages and yearly observation averages in the values to be interpolated. This Method B was applied daily in 2010 for two pollutants: 1-hr max NO_2 and 24-hr $PM_{2.5}$. Figure 13 shows the yearly average CMAQ and observations for NO_2 in 2010, along with the original and both alternative fused fields. Figure 14 shows the same fields for $PM_{2.5}$.

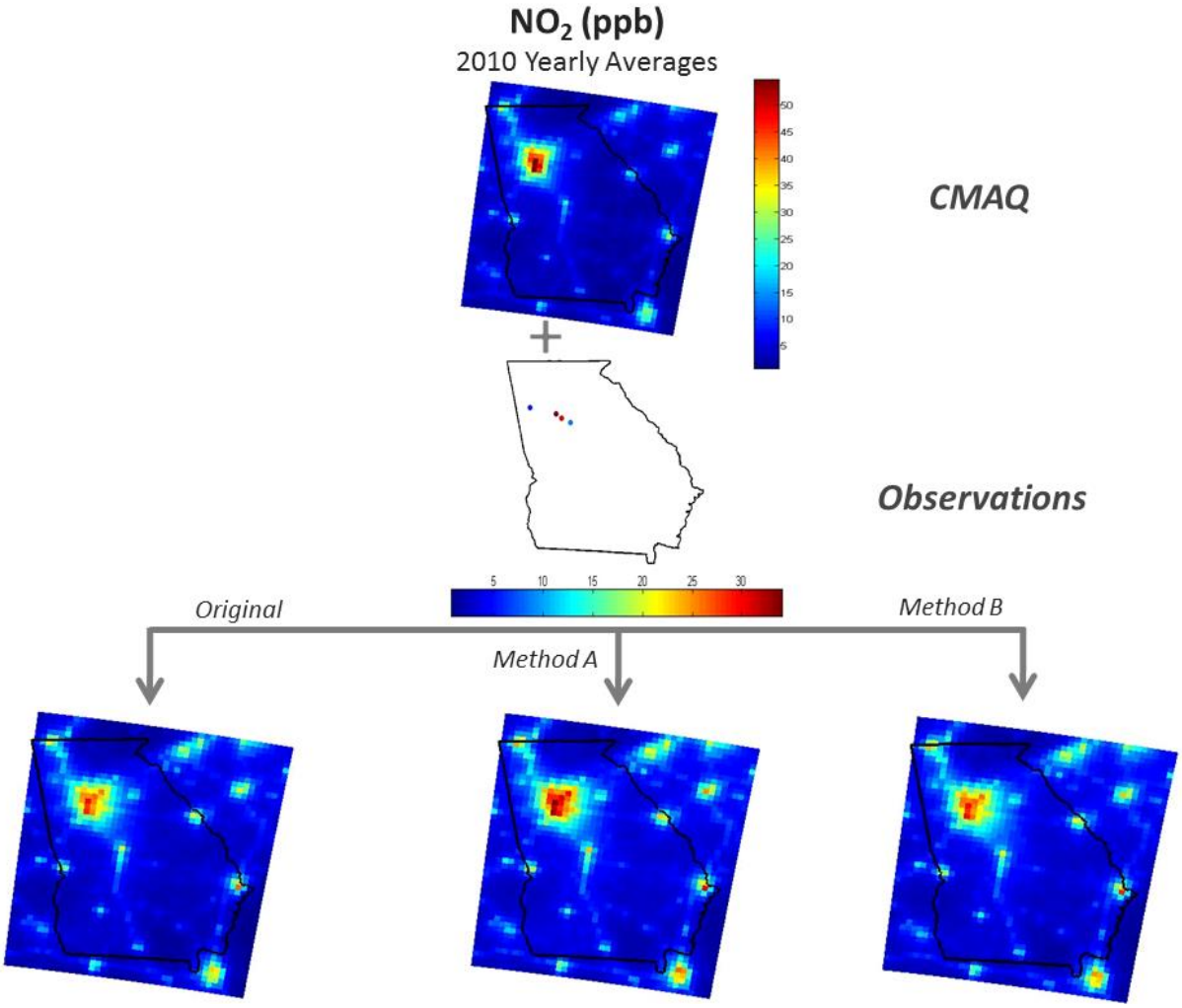


Figure 13 – Yearly average CMAQ and observations for NO_2 in 2010, and final fused fields yielding from the original data fusion method as well as Method A and Method B

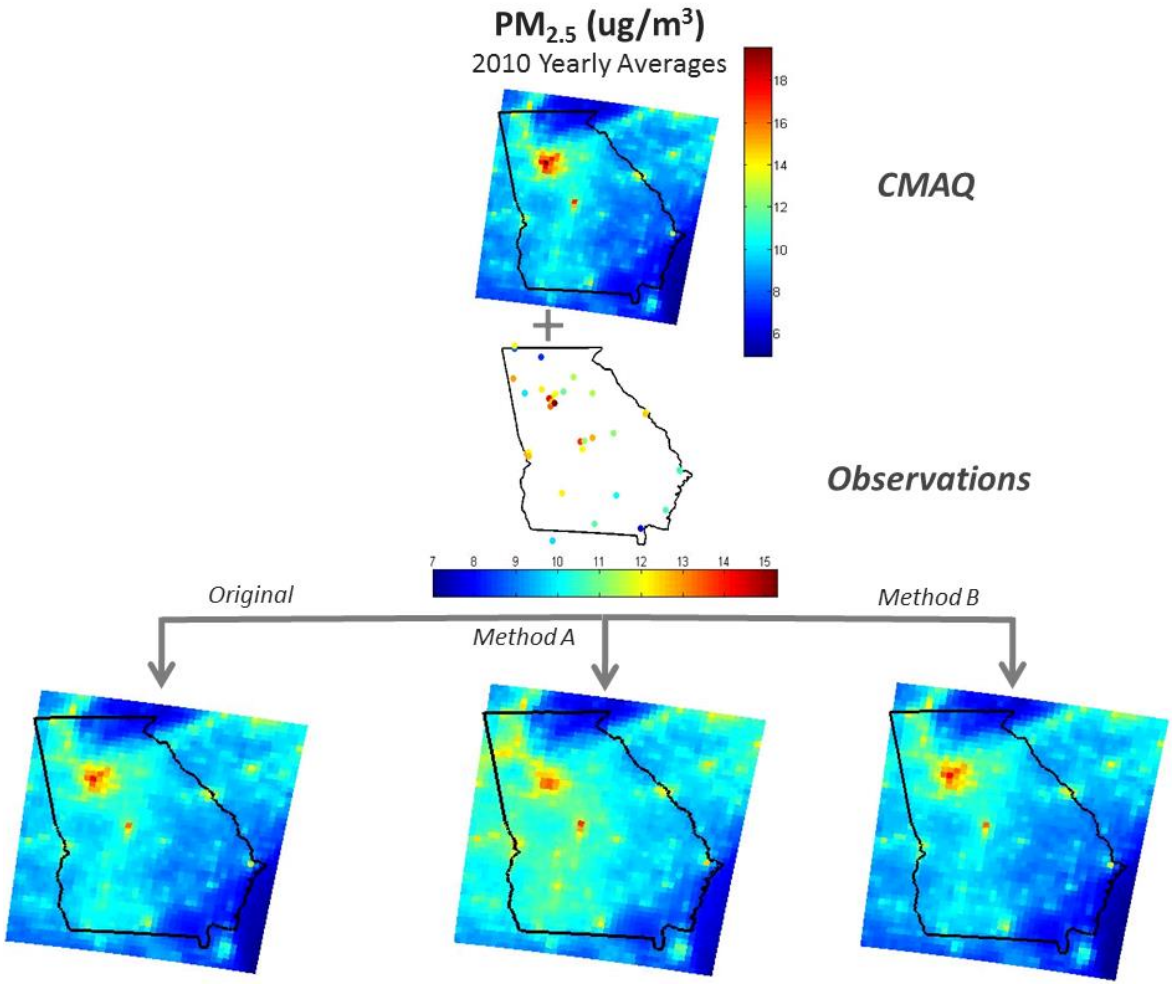


Figure 14 – Yearly average CMAQ and observations for $PM_{2.5}$ in 2010, and final fused fields yielding from the original data fusion method as well as Method A and Method B

For both NO_2 and $PM_{2.5}$, Method B yields average results that look more similar to the original method's results than Method A. Method A yielded an average field for NO_2 that had much higher values overall than in the original method. Spatial trends from Method B more closely match the original method in both distribution and magnitude. This holds true for the $PM_{2.5}$ spatial fields as well. Although Method A yielded results that generally were in the same range as the original method in this case, there were more high values from Method A, especially in the southwestern portion of the domain, which is predominantly rural. However, one of the only monitors in this region is located in Albany, Georgia, close to a major road. This location

likely has relatively high $PM_{2.5}$ concentrations compared to the surrounding area. This high value propagated throughout the region in interpolation in Method A, and created higher values than those seen in the original method. Because values are corrected for the differences between cell and monitor values in Method B, this southwest area had lower final concentrations that matched more closely to the original method's concentrations.

By comparing relationship between observed concentrations and the corresponding fused field values for both the original method and the alternative method, the performance of Method B can also be statistically characterized. Three different metrics are used in this examination: R^2 values, percent root mean squared error (RMSE) and percent mean bias. Table 7 shows these three metrics for NO_2 and $PM_{2.5}$ in 2010 for the alternative method and for 2009-2012 for the original method. These values were calculated from all monitor values along with all corresponding C^* values. RMSE and bias are normalized to the mean observed concentration value over the considered years for each method.

Table 7 – R^2, RMSE and bias values for the original and alternative Method B

	Original		**Method B**	
	NO_2	**$PM_{2.5}$**	**NO_2**	**$PM_{2.5}$**
R^2	0.84	0.91	0.78	0.91
RMSE	39%	14%	44%	14%
Bias	5.7%	-0.46%	3.0%	-0.54%

As discussed earlier, Method A yielded results that matched observations perfectly except in the case that there were multiple monitors in one cell. This eliminated all bias, produced R^2 values very close to one and very low RMSE values. This is not desired however, since cell

values will not realistically always be equal to monitor readings. Method B yields results with higher bias and RMSE values and lower R^2 values than Method A. These values are also very close to the values returned from the original method. The $PM_{2.5}$ values from both methods agree almost perfectly, while Method B results in slightly lower bias and R^2 values and higher RMSE values than the original method for NO_2. This suggests that this method does perform better than Method A with NO_2, but still not as well as the original method, since slightly higher bias in this case signifies that the model is able to capture the differences in monitor and cells.

Figure 15 shows a comparison of all C^* values generated from the original method and the one-step alternative, Method B for NO_2 and $PM_{2.5}$ in 2010. Like Method A, Method B yields results for $PM_{2.5}$ that are more correlated with original C^* values and with less variance than NO_2. The R^2 and percent RMSE values for $PM_{2.5}$ stayed about constant, at 0.86 and 17% respectively, from those from the implementation of Method A, as seen in Figure 12. The R^2 value for NO_2 is also approximately constant; however, the RMSE is slightly higher, rising from 46% to 63%.

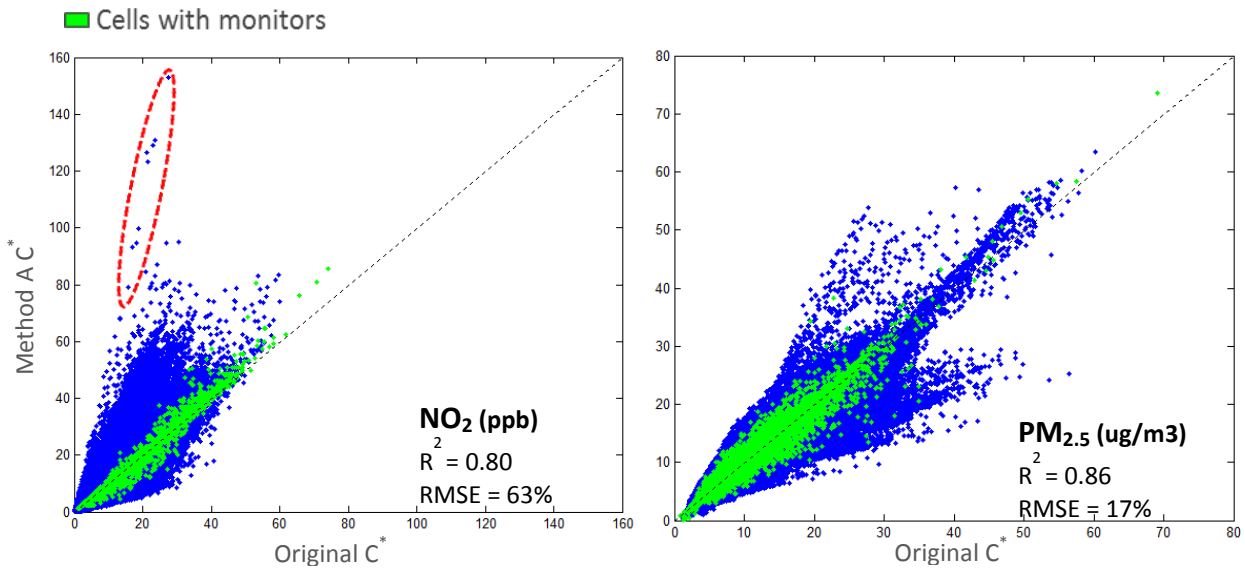


Figure 15 – Plot of all C^* values for NO_2 and $PM_{2.5}$ in 2010 for the original method and alternative Method B

Although there is a higher overall RMSE value from Method B for NO_2 than Method A, the cells with monitors tends to agree much more with the original method than previously. The bias seen at those cells was largely eliminated, and these points for Method B largely lay on the identity line. Monitor points for $PM_{2.5}$ see little change in Method B from Method A, however many points from cells without monitors move closer to the identity line in Method B.

The points seen on the NO_2 plots that are significantly higher in the alternative method than the original are from days and points where the closest monitor value is significantly higher than the estimated CMAQ value at the monitor location. The most extreme of these points are circled in red on the NO_2 plot in Figure 15, and similar points can be seen in Figure 12. This ratio creates a high value to be interpolated, and the entire surrounding area is therefore given high values during interpolation as well. When denormalizing the interpolated field with the adjusted CMAQ field, areas that have CMAQ simulated values that were not underestimated to the same extent that they were at the monitor location will then have very high concentration estimates. This will lead to some of the highly biased points seen in the plots. Because the ratio of average observation and adjusted CMAQ values is greater than one for NO_2 in 2010, the addition of this term will cause the interpolated values to be higher than those interpolated using Method A. Therefore, the highly biased points seen from Method B are higher than the same points from Method A. This error is corrected for in the original method because this method gives less weight to the interpolated field far from monitor locations.

$PM_{2.5}$ has many more monitors that are spatially distributed more evenly than NO_2 monitors, so individual $PM_{2.5}$ monitors do not dictate the interpolation behavior over large areas of the domain to the same extent that NO_2 monitors do. Because there are more interpolation points, the different performances of CMAQ across the domain are more easily captured, and it

is less likely that the interpolation will attempt to correct for an under- or overestimation of CMAQ in a region that does not need the correction.

4.4 Discussion

The fusion of observational data and CMAQ, a chemical transport model, results in a final field that seeks to capture the different strengths of the two data sets while minimizing their limitations. In the original method created by Mariel Friberg, this was accomplished through the creation of two spatial fields. One of these fields is based on the interpolation of observations and one is based on CMAQ simulations that have been adjusted to observations. These fields are then combined using a weighting factor that depends on monitor-to-monitor correlations and monitor-to-CMAQ correlations.

Two alternative methods that are much easier to implement and do not require these correlations were examined here. In the first method, Method A, observations normalized to adjusted daily CMAQ values are interpolated. Invalid spatial trends that are created through interpolation are corrected for by denormalizing the interpolated field with adjusted daily CMAQ values. The final fields created by this method have similar spatial trends as the fields created through the original method. On average, NO_2 values were higher in Method A, while $PM_{2.5}$ values tend to be very close to the original method. A major weakness of this method is the fact that this method holds monitor values constant in the final fields. Although this causes the results at cells with monitors to match very closely with observations, it causes inaccuracies in the final field. Monitor values are not always representative of the entire cell, so some differences between monitor and cell values are desired.

Method B attempts to address this difference through the inclusion of the ratio of average observation and adjusted CMAQ values in the term to be interpolated. This method matches the

original method's spatial trends in both shape and magnitude more closely than Method A. The bias seen in cells with monitors is also greatly reduced when compared to the same points generated from the original method.

Results that are likely erroneous can be seen from both of these alternative methods because they both depend on daily CMAQ values. The processes assume that CMAQ is performing similarly far from monitors as at monitors. However, realistically, CMAQ performance can vary greatly spatially, especially between rural and urban areas. Therefore, if CMAQ is over- or underestimating a point with a monitor, these two methods assume it has the same performance at locations far from a monitor. This leads to the creation of inaccurate spatial trends far from monitors. These methods do not perform as well at points far from monitor locations compared to the original method because the original method trusts adjusted CMAQ values at these locations more than observations.

The major limitation of the two alternative data fusion processes is the availability and spatial distribution of monitors. Within the Georgia domain, there is a scarcity of ambient air quality monitors for many pollutants. In this case, NO_2 had an especially sparse network of monitors, with only five across the entire domain. Additionally, because many monitors do not record concentrations daily, the data are also incomplete temporally. When there are abundant monitors, it is less common for incorrect spatial trends to be created, as seen by the better performance of $PM_{2.5}$ than NO_2 in both alternative methods. With more monitors, there are more interpolation points, and the different performances of CMAQ across the domain are more easily captured. It is less likely that the interpolation will attempt to correct for an under- or overestimation of CMAQ in a region that does not need the correction. The ability for these two methods to predict values both far and near to monitors is evaluated through data withholding in Chapter 5.

4.5 Conclusion

Two alternative data fusion methods that are able to greatly simplify the data fusion process were implemented for NO_2 and $PM_{2.5}$ in 2010. These one-step methods produce results that are similar to observations as well as the results produced by the original method. Because Method A holds values at monitors constant, while Method B is able to capture the differences between monitor and cell values, Method B produces more representative results than Method A.

CHAPTER 5: EVALUATION OF ALTERNATIVE DATA FUSION METHODS BY DATA WITHHOLDING

5.1 Introduction

In order to evaluate the performance of the two alternative data fusion processes, cross-validation analyses were performed through observational data withholding. Data withholding was implemented for two pollutants, $PM_{2.5}$ and NO_2, for 2010. Previously, the two final optimized spatial fields (C^*) created from the alternative fusion processes were characterized by comparing the final values of the two alternative simulations with the final simulations yielded from the original method, as well as by comparing the final simulation values at monitor sites with observation values.

In the first alternative method, Method A, observations are normalized to adjusted daily CMAQ values, which are interpolated and then denormalized by the adjusted daily CMAQ field. In this method, cells with monitor data are always given a final concentration of the observation. It was found that $PM_{2.5}$ produces results that more closely agree with values produced from the original method than NO_2, although both are slightly biased when compared to the original results. In the second method, Method B, a ratio of average observation and adjusted CMAQ values is included in the term to be interpolated, which allows the method to capture differences in monitor and cell values. Spatial trends of NO_2 appear to match the original more closely, although RMSE values between final values of this method and the original increase slightly from Method A. $PM_{2.5}$ performance stays about constant through the two alternative methods.

These comparisons only characterize how the alternative methods perform, but do not evaluate performance. The original method is not a perfect model, and its results are known to have some errors and biases. Therefore, it is difficult to distinguish the correct source of error when comparing the results of the alternative methods with the results of the original method. As

a result, this comparison only offers a relative depiction of the alternative methods. Comparing C^* values with observations also cannot provide an evaluation of the model because the creation of the fused field depends heavily on the inputted observational data. Therefore, it is expected that the resulting fields agree well with this data. This is especially true for values from Method A, since this method holds monitor values constant in the cells with monitors. Because monitors only give concentrations at the point that they measure, while cell concentrations should be an average of the concentrations at every point within the cell, realistically, cell values should not always match observed values. Therefore, very high correlations between C^* values and observations do not signify the perfect performance of a data fusion method. For a complete evaluation of the alternative methods, data that is independent of the results must be used. Data withholding is able to meet this requirement.

Additionally, evaluating how well the model represents observational data at cells with monitors is an incomplete assessment of the model. It is desirable to quantify the methods at locations with no monitors. To determine how well the two alternative methods are able to predict values both near and far from observations, data withholding can be used. Here, a random 10% of all monitor data were removed for $PM_{2.5}$ and NO_2 in 2010, the pollutants and timeframe initially considered in the two alternative methods' data fusion. This 10% withholding approach only addresses the performance of the method when data are missing from a point. Another approach to data withholding would be to remove all data from one monitor entirely. However, NO_2 does not have enough monitors to allow this to be a complete assessment.

After the 10% of data were removed, the two different alternative data fusion processes were then conducted again, assuming that everything else remained constant. The final values resulting at the points where data were withheld can then be compared to the observational data at those points and times. Because the simulated results do not depend on the monitor data

recorded at those locations and times, this is a more unbiased and comprehensive picture of the two models' performances at points at various distances to monitors and compared to the original method's performance.

The Jefferson Street (JST) and Yorkville (YRK) monitor data withholding performances are shown separately from performance from all monitors as well. These two monitors measured both considered pollutants on a daily bases during 2010 and are situated in very different geographic locations. The JST monitor is located in an urban area, near the center of Atlanta, and close to multiple other monitors. Conversely, the YRK monitor is located west of Atlanta, in a relatively rural location with no other monitors within close proximity. These two monitors are singled out to demonstrate the impact of monitors in close proximity to the points being estimated.

5.2 Method

To perform data withholding, a random 10% of all monitor data were removed from the observational data set for $PM_{2.5}$ and NO_2 in 2010. The withholding of data did not significantly affect the previously developed parameter values used to adjust CMAQ to observations. Therefore, all values determined previously were kept constant. This also allows for the most transparent evaluation.

$PM_{2.5}$ and NO_2 both have two monitors located in Yorkville positioned within 100 feet of each other. By having two monitors in such close proximity, data withholding at these sites becomes an ineffective evaluation. Because of their proximity, the two monitors will typically record very similar concentrations. When a data point is removed from one of the monitors, the second monitor is still available to provide observational data at the same location. The result of the data fusion at those points would reflect the values given by the second monitor, which will

be strongly correlated with the monitor value that was withheld. This will cause an overestimation of the performance of the data fusion at these points. Therefore, in these cases with two monitors in very close proximity, when one of the Yorkville monitor's data points is withheld, the second Yorkville monitor's value is also removed. This prevents the misrepresentation of estimations at this location. Because additional data points were removed after the initial random 10% were taken out, more than 10% of the observational data were actually withheld for these species. This causes a different number of points to be removed for each pollutant in each of the different data fusion methods. However, the total amount of data withheld remained between 13.5% and 10.3% for the two species and two alternative methods. The total number of observation points in 2010 for NO_2 is 1805 from five monitors while there are 6714 total observation points from 38 monitors for $PM_{2.5}$. More information on the data withheld can be seen in Table 8, while the same information can be found on the total observation set in Table 2, and the information on the data withheld using the original data fusion method can be found in Table 4.

Also shown in the table is information on the data withheld from the Jefferson Street (JST) and Yorkville (YRK) monitors. The YRK monitor information is given for only one of the two monitors located in Yorkville. The YRK monitor shown in the analysis here is a SEARCH monitor that measures all considered pollutants daily, while the other monitor only measures four of the 12 originally considered pollutants at various measuring frequencies.

Because the points removed for the data withholding are completely random, and because extra points need to be removed from the Yorkville monitor, the sets of data removed for each method are different. If the chosen set of data points happen to generally not be of the whole sample, data withholding can present results that capture these differences but may present them in a way to make it seem like the model was performing better or worse than it really was. The

differences in data points removed from each set may help explain the different results seen from the data fusion with data withholding.

Table 8 – Characteristics of data withheld for both alternative methods

	Method A		Method B	
	NO_2	$PM_{2.5}$	NO_2	$PM_{2.5}$
# Observations Withheld	243	685	237	694
Minimum	0.93	0.1	1	0.5
Maximum	68	65.2	75	50.3
Average	15.4	12.2	15	11.8
IQR	23.37	7.38	22.18	7.42
# Observations withheld JST	33	38	34	31
JST Minimum	10.19	3.75	7.81	3.75
JST Maximum	58.90	38.43	67.28	27.37
JST Average	34.75	14.20	36.95	12.06
# Observations withheld YRK	69	46	60	48
YRK Minimum	0.93	2.91	1.13	4.03
YRK Maximum	27.93	18.61	27.93	18.88
YRK Average	5.93	9.37	4.31	10.34

In general, the data sets removed for both pollutants are very similar for both methods of data fusion. However, in general at the JST monitor values tended to be higher for NO_2 in and lower for $PM_{2.5}$ in Method B than Method A. For both pollutants, but especially for NO_2, many more points were removed from the YRK monitor than from the JST monitor. In addition to

being randomly withheld, the YRK monitor data often times had to be removed deliberately because the other Yorkville monitor data were chosen at random to be removed. Compared to the overall data set seen in Table 2, the data removed for the evaluation of the two alternative methods are very similar to the data withheld in the evaluation of the original method.

5.3 Results

The two alternative data fusion methods were implemented using data withholding, resulting in the creation of daily concentration fields over the domain of Georgia. These fields are based on CMAQ simulated data and observational data that had 10% of all original values removed. In Figure 16, three different metrics are used to evaluate the performance of the two simulations in order to determine how well the models are able to predict pollutant concentrations. The metrics used for the evaluation are the R^2 value, percent root mean squared error (RMSE) and percent mean bias. These values were calculated for both considered pollutants using the withheld observational data points and the resulting C^* values at the time and locations that the observations were withheld. These values are given for both of the alternative methods as well as the original method for comparison. RMSE and bias has been normalized to the average withheld monitor value for each pollutant

Also included in Figure 16 is an evaluation for only the points withheld from the JST and YRK monitors, which take daily measurements for both pollutants. Evaluating these two monitors individually demonstrates the effect of monitor clustering on the prediction of concentrations. The JST monitor is located in Atlanta's city center and positioned close to other monitors, while Yorkville is in a more rural area, about 45 miles west of Atlanta, and does not have any other monitors nearby.

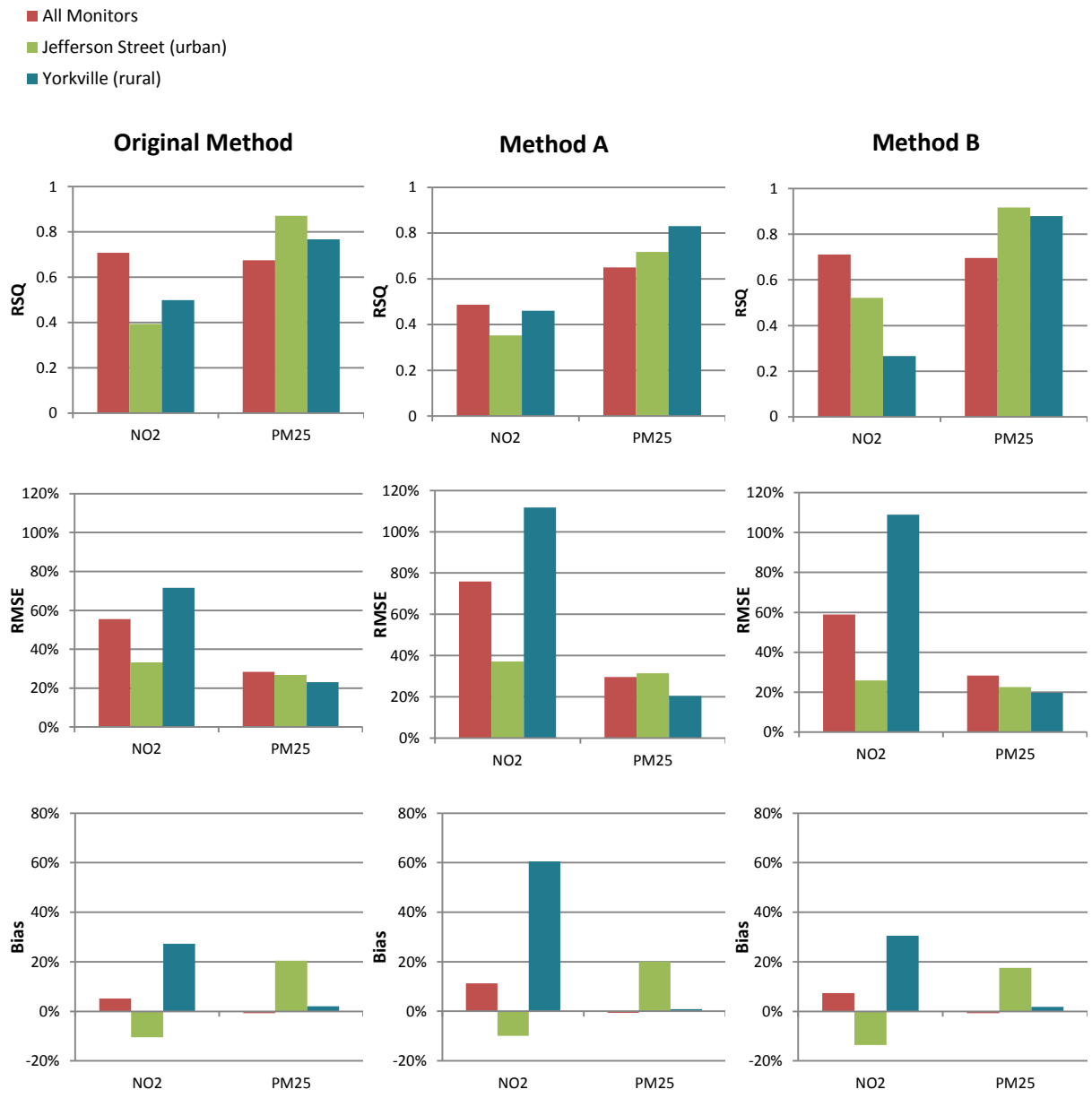


Figure 16 – R^2, RMSE and bias of $PM_{2.5}$ and NO_2 at all withheld monitor locations, the withheld Jefferson Street monitor, and the withheld Yorkville monitor using the two alternative methods, Method A and B, as well as the original method for comparison

Overall, similar trends are seen in the results of data withholding from the two alternative methods as from when it was performed originally. However, both of the alternative methods predict variation slightly less well than the original method. There are slightly lower R^2 values and higher error and bias than the original method.

When looking at values for all monitors, R^2 values are the highest for Method B and lowest for Method A for both pollutants. The R^2 in Method B match very closely to those seen in the original method. It rose less than 0.01 for NO_2 and from 0.67 to 0.70 for $PM_{2.5}$, while in Method A values dropped from 0.71 to 0.49 for NO_2 and from 0.67 to 0.65 for $PM_{2.5}$. The same trends hold true for RMSE and bias as well. In all cases when looking at all monitors, the results from Method B tend to be very close to the original method's values, while Method A yields results that are slightly worse than the original values. However, even though the R^2 values are slightly lower, and RMSE and bias are slightly higher, all values remain reasonable. This signifies that both of the alterative models are able to adequately predict pollutant concentrations at locations where there are not monitors. However, Method A cannot predict these concentrations as well as the other two methods.

In general, concentrations in the JST monitor grid cell are predicted better than in the YRK monitor cell because the monitors close to the JST monitor cell give relevant information as to what is occurring at Jefferson Street even when there are no data from the JST monitor. This can be seen especially well in the performance of NO_2. For all three methods, R^2, RMSE and bias values are significantly worse at YRK than at JST in every case but one. The performance of Method A for NO_2 at the YRK monitor is particularly poor. Because the data set removed from this monitor was very similar for Methods A and B, the differences seen here are primarily due to differences in the performance of the processes. This is likely because the closest monitor to Yorkville after the Yorkville monitor data point has been removed is in a larger city. When interpolating, concentration information is taken from this monitor to estimate a concentration at Yorkville, even though the information is likely not relevant. In the original method, this is not an issue, because the interpolated field is weighted less when the point is far

from a monitor, and adjusted CMAQ values are weighted more. In these cases, CMAQ gives a more accurate picture of concentrations than distant monitors do.

The relationship between distance to the closest monitor and how well the three models are able to predict concentrations can be seen in Figure 17. These correlograms were created for both pollutants and three methods by relating the correlation between simulated values and withheld data points to the distance to the closest monitor data point after the observation point had been removed. Each monitor location had various distances to the closest recording monitor based on what monitors gave measurements for that day and what other data points were removed. Correlations were computed for distances with 10 or more data points.

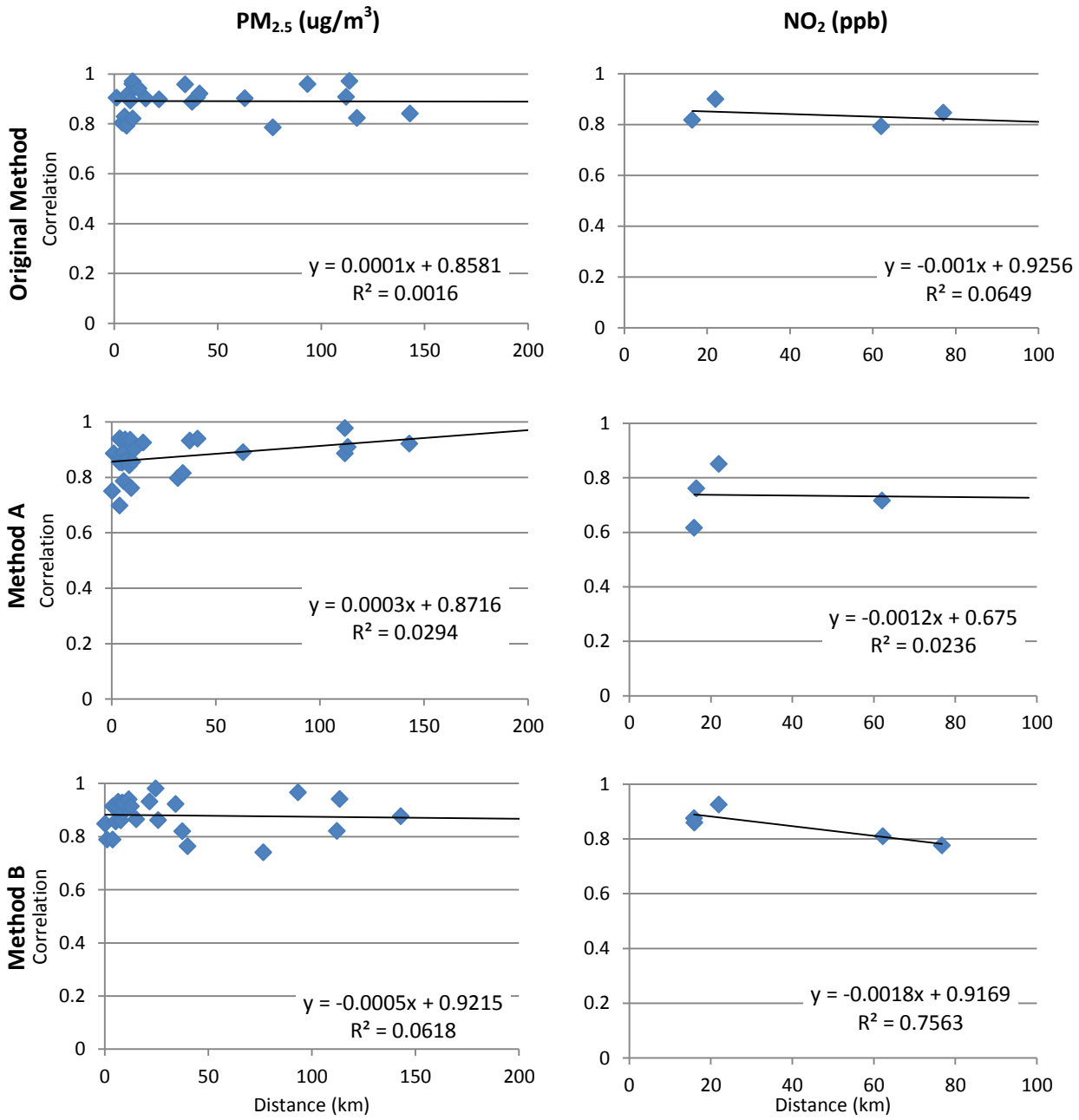


Figure 17 – Correlograms resulting from data withholding using the original method, Method A and Method B for NO_2 and $PM_{2.5}$

These graphs demonstrate that for all three methods, as distance from a monitor increases, it generally becomes more difficult for the models to accurately predict the concentration at a point, although correlations stay high overall. In general, NO_2 correlations are lower in Method A than the other two methods, and the y-intercepts for the original method and

Method B are significantly higher than the Method A. However, the slope of Method B is slightly steeper than the originals. This shows that the original method can predict concentrations far away from monitors the best, while Method A performs the worst in these predictions. Method A, while still not as accurate as the original method, performs much better than Method B.

The correlograms for $PM_{2.5}$ are similar for all three methods in that correlations tend to remain very high for all methods, even with great distances. In Method A, the correlations even increase slightly with distance. The increase is very small though and may be because if a monitor value is removed from a location that is far from any other monitor, it is likely in a rural area. CMAQ may be able to capture $PM_{2.5}$ concentrations in rural areas more accurately than in urban areas, where there are likely more monitors in close proximity.

$PM_{2.5}$ also tends to be spatially flat, having low concentration gradients across the domain, and only slight differences between concentrations in urban and rural areas. Therefore, even if there are no monitors close by, the model can accurately predict concentrations at any point because concentrations are similar over large distances. Therefore, the correlograms tend to produce a very flat relationship between increasing distance and correlations between simulated values and observed values. NO_2 is a primary pollutant and tends to have much higher concentrations in urban areas than rural areas. Over long distances, its concentration can change drastically. Therefore, a monitor located far away gives little relevant information as to concentrations that are actually occurring at a point, and correlations over distance drop more significantly.

5.4 Discussion

The cross-validation results indicate that the one-step alternative data fusion method, Method A, can adequately predict concentrations of pollutants over time and space, while making the data fusion method much easier to implement. However, its performance is consistently lower than that of the original method and Method B. This is largely because this method assigns cells with monitors the value given by the monitor. However, because monitor values are not always representative of cell values, this creates inaccuracies in interpolation, which then spreads to the entire fused field.

The other alternative method, Method B, is just as simple to implement as Method A. It does not require the determination of the correlations needed in the original method and completes the data fusion in one step instead of the original three steps. However, unlike Method A, it is able to capture differences in monitor and cell values to create a smooth surface for interpolation. It is also able to predict concentration values at points both near and far from monitors just as well as the original method. It is therefore recommended for this method to be used when implementing data fusion in the future.

A major shortcoming of the one-step methods is that both alternatives assume that CMAQ performance is constant throughout the domain on any given day. This means that if the ratio between a daily observation and the daily CMAQ value at that point is very high or very low, the model assumes that this ratio holds true at locations far from the monitor. However, CMAQ performance tends to vary not just over time but also through space. CMAQ performance may be similar in a small area; however, over large distances it is likely to change, especially when going from an urban to a more rural area. When interpolating, observations are first normalized to the CMAQ value of its cell, and this relationship is spread over large distances during the creation of a spatial field from this data. This causes biases in results in

locations far from monitor locations, especially for pollutants like NO_2, which vary considerably over space.

Therefore, the largest limitation in these two methods is the lack of monitors. With more monitors and a more even spatial distribution, performance would not have to be assumed over such large distances. Another limitation from the data withholding method is that the performance of the methods was evaluated at locations where there are data some of the time, but not all of the time. Another approach to the data withholding would be to evaluate performance where there are no data. This could be accomplished by withholding all observations from one monitor, instead of a portion of data from all monitors.

5.5 Conclusion

Two alternative methods that are able to greatly simplify Friberg's data fusion method were evaluated through data withholding. While both methods are able to reasonably predict concentrations at points with no monitor data, overall, Method B performed slightly better than Method A. In the evaluation, Method A produced results with higher bias and RMSE and lower R^2 values than Method B, while Method B produced results that were very similar to the original method's data withholding results.

CHAPTER 6: CONCLUSION

It is important for studies that investigate the associations between acute health effects and ambient air pollution to utilize accurate pollutant concentration data. Because the concentrations of pollutants vary greatly over both time and space, the data must also be highly resolved spatially and temporally. In this study, two different sources of data, observational data from monitoring networks and chemical mass transport model simulations were fused in order to create spatiotemporally resolved daily ambient air quality fields. This was completed over the state of Georgia for four years (2009-2012) and 12 pollutants: five gases and seven particulate matter species. The five gases were ozone (O_3), carbon monoxide (CO), nitrogen dioxide (NO_2), total nitrous oxides (NO_x), and sulfur dioxide (SO_2). The seven particulate matter (PM) species were total $PM_{2.5}$, total PM_{10}, and $PM_{2.5}$ subspecies sulfate (SO_4^{2-}), nitrate (NO_3^-), ammonium (NH_4^+), elemental carbon (EC), and organic carbon (OC).

Observational data and chemical transport model simulations were used in this fusion because of their complementary strengths and weaknesses. Observational data from monitoring networks give measurements with low amounts of error, but are not able to give much spatial information. At distances far from a monitor, the monitor measurement and the actual concentration at that point become poorly correlated. Monitors are also very limited in both number and distribution. Most of the monitors tend to be located in cities, and some pollutants have as few as four monitors across the entire domain considered here. This leads to a significant absence of spatial information on pollutants from observations. Lastly, the monitor data tend to be temporally incomplete. Some monitors only give measurements every other day, or every six days, or in the case of ozone, only 6 months out of the year.

Emission based models on the other hand are spatially and temporally complete. In this study, CMAQ simulations were used, which can give hourly concentration estimates for all

major pollutants at very fine resolutions. However, these models tend to have biases that restrict the accuracy of the estimates. Model parameters and specifications are imperfect, and estimates are strongly affected by the accuracy of the emission and meteorological inputs. The models also do not often capture day-to-day variability very well.

In this study, observed data and CMAQ data were fused to produce pollutant concentration fields that are more accurate and complete than either of the two data sets alone. The fields were created by combining the strengths of observed data and CMAQ, while reducing the effects of their weaknesses. In order to fuse the two data sets, two pollutant fields were first developed. The first field is based on the interpolation of observations using kriging as the interpolation method. An adjusted CMAQ field is then used to capture more realistic spatial trends. The second field is based on CMAQ simulations that have been adjusted to the observations. These two fields are fused using a weighting factor that is based on the correlation between monitor measurements and the correlation between CMAQ estimates and monitor measurements. The first field is able to capture temporal trends at locations close to monitors, while the second field is more capable at capturing spatial variations at locations removed from monitors. By combining these two fields, a final field is created that is able to capture variations over both time and space throughout the domain.

The estimates produced from this method agree very well with observations, which is an expected outcome since the estimates depend on observation values. Cross-validation was used to thoroughly evaluate the performance of the data fusion process. The data withholding results suggest that the data fusion method used is able to accurately estimate concentrations of pollutants at points where there are no observational data. However, the model performs much better at locations close to monitors, as can be observed through the comparison of estimates at the Jefferson Street and Yorkville monitor locations. The model also performs better with

secondary pollutants compared to primary pollutants because secondary pollutants, such as O_3, tend to be more spatially flat than primary pollutants, such as EC and SO_2. SO_2 is captured most poorly in the data fusion process due to the poor ability for both CMAQ and monitors to capture the plumes coming from coal-fired power plants. Those using the results of data fusion should be advised that errors are high for SO_2, even after the fusion process. However, overall this data fusion process is able to provide highly resolved pollutant concentration fields that are spatially and temporally complete and accurate.

The three-step process implemented in the initial data fusion was developed by Mariel Friberg. The correlations determined in the calculation of the weighting factor can be difficult to develop, especially if implementing the data fusion process over a large domain with a high quantity of monitors. In order to simplify the data fusion process, two one-step methods that do not depend on these correlations were evaluated. In the first one-step method, Method A, observations are normalized with the adjusted daily CMAQ values at the observation points. These values are then interpolated and denormalized using the adjusted daily CMAQ field. In this method, observations are kept constant at the cell in which they are located. Observation values may not be representative of a cell, and when these incorrect values are interpolated, it leads to biases across the domain. Therefore, a revision was made to the one-step method which takes into account the differences between monitor and cell concentrations, and values are created that allow for a smoother interpolation. In order to do this, the ratio of average observations and adjusted CMAQ values is included in the interpolated term.

The two alternative methods were implemented for 2010 for two pollutants, NO_2 and $PM_{2.5}$, over the same domain as previously used. The methods were characterized by comparing their results with the results of the original method as well as observations. It was found that in Method A, $PM_{2.5}$ results looked very similar to the original method's, however NO_2 was

generally biased high. This is because most NO_2 monitors record measurements that are higher than the concentrations seen in the monitor's cell, and this method is unable to capture those differences. Method B is able to address these differences, and gives more realistic final fields. Results from this method match the original method's spatial trends in both shape and magnitude more closely than Method A for both NO_2 and $PM_{2.5}$.

In the evaluation of the two methods through data withholding, it was found that both methods are able to adequately predict concentrations of pollutants located both near and far from monitor locations. However, the performance of Method A was consistently lower than that of the original method, while Method B performed very similarly to the original method. Similarly to the original data withholding, both of the models generally perform better at locations close to monitors. They also both perform better with $PM_{2.5}$, largely a secondary pollutant, than NO_2, a primary pollutant.

6.1 Limitations

The main limitation causing error in the two alternative methods is the assumption of constant CMAQ behavior over large areas. However, CMAQ behavior varies considerably over space, especially when moving from urban to rural areas. If a monitor value is significantly higher than the estimated CMAQ value at the monitor location, it will cause a high value to be interpolated, and high values will be given to the entire surrounding area as well. Close to the monitor this is appropriate because CMAQ is likely under-predicted in the surrounding area as well. However, this will not necessarily hold true at points farther from the monitor, and may cause these points to be over-predicted in the final fused field.

The primary factor affecting the severity of this limitation, as well as the performance of all data fusion methods, is the abundance and distribution of ambient air quality monitors. All

methods perform best close to monitors, because they provide the most accurate concentration data as well as temporal variation. More monitors provide more information on concentrations across the domain, as well as more interpolation points. With more data points, the different performances of CMAQ across the domain are more easily captured, and interpolation will result in a much more truthful spatial field. As the number of monitors increases, it becomes less likely that the interpolation will try to correct for an under- or overestimation of CMAQ in a region that does not need the correction. However, because monitors are both temporally and spatially incomplete, the overall performance of data fusion suffers. The performance of CMAQ is also a source of error throughout the data fusion processes. Although it is both spatially and temporally complete, its biases restrict how much confidence can be put in the model outputs.

6.2 Future Work

When implementing data fusion in the future, the one-step alternative, Method B, could be applied to other years as well as other pollutants. It is also recommended to investigate the performance of this method over domains larger than Georgia, such as over the entire United States. Additionally, it would be beneficial to further explore the reasons behind the high values yielded by the two alternative methods that are not present in the original method's results. Although without observations it is difficult to know for certain the correct values that should be generated at these points, it is likely that the alternative methods are largely overestimating these values. The one-step methods could be further revised in order to prevent these unwanted extreme values that are not produced in the original method.

6.3 Conclusion

Spatiotemporally resolved ambient air quality fields that were created here are consistent with observations as well as emissions and meteorology. The fields give daily pollutant concentration fields that are well suited to be used in future epidemiological studies investigating acute health effects due to air quality in the Georgia area. The alternative methods investigated here are both much simpler to implement than the original method, and both create results that are similar to results produced from the original method. However, overall, Method B is able to predict pollutant concentrations better than Method A, and about equally as well as the original method. Therefore, it is recommended that this method be used in future data fusion studies, especially those with very large domains or a high number of monitors. The results of this method would also be able to capture the spatial and temporal variations of pollutants needed in health studies.

APPENDIX

Table A1 – β and α CMAQ adjustment parameter values for all pollutants 2009-2012

	O_3 8-hr max	NO_2 1-hr max	NO_x 1-hr max	CO 1-hr max	SO_2 1-hr max	PM_{10} 24-hr avg	$PM_{2.5}$ 24-hr avg	SO_4 24-hr avg	NO_3 24-hr avg	NH_4 24-hr avg	EC 24-hr avg	OC 24-hr avg
β	1	1	1	1	1	0.1	0.5	0.7	0.5	0.7	1	0.6
α_{09}	0.90	0.67	1.00	0.90	2.1	16	3.2	1.4	0.6	0.9	0.6	1.5
α_{10}	0.90	0.54	0.52	0.50	2.1	14	3.2	1.4	0.6	0.8	0.6	1.5
α_{11}	0.95	0.67	1.0	0.90	2.1	14	3.6	1.5	0.5	0.9	0.7	1.7
α_{12}	0.90	0.67	1.0	0.90	2.1	13	3.2	1.3	0.5	0.6	0.7	1.6

Table A2 – Parameters used in R_1 correlations and R_2 values for all years and pollutants

	O_3	NO_2	NO_x	CO	SO_2	PM_{10}	$PM_{2.5}$	SO_4	NO_3	NH_4	EC	OC
E_i	0.97	0.95	0.95	0.95	0.90	0.90	0.95	0.95	0.95	0.95	0.90	0.95
r (km)	500	50	100	50	20	300	500	500	500	500	300	400
R_2, 2009-2010	0.78	0.60	0.61	0.61	0.35	0.42	0.61	0.63	0.64	0.57	0.75	0.75
R_2, 2011-2012	0.76	0.52	0.55	0.48	0.28	0.51	0.58	0.67	0.59	0.52	0.66	0.54

REFERENCES

Adelman, Z.CMAQ Version 5.0 Technical Documentation. *Community Modeling and Analysis System (CMAS)*. **2012**.

Appel, K. W., Bhave, P. V., Gilliland, A. B., Sarwar, G., and Roselle, S. J. Evaluation of the community multiscale air quality (CMAQ) model version 4.5: Sensitivities impacting model performance; Part II - particulate matter. *Atmos. Environ*. **2008**, *42*, 6057–6066.

Appel, K. W., Gilliland, A. B., Sarwar, G., and Gilliam, R. C. Evaluation of the Community Multiscale Air Quality (CMAQ) model version 4.5: Sensitivities impacting model performance part I - ozone. *Atmos. Environ*. **2007**, *41*, 9603–9615.

EPA. EPA's Community Multiscale Air Quality Modeling System (CMAQ); Tools for Controlling Air Pollution & Studying Climate Change. U.*S. Environmental Protection Agency, Office of Research and Development.* **2016.**

Friberg, M.D., Zhai, X., Holmes, H.A., Chang, H.H., Strickland, M.J., Sarnat, S.E., Tolbert, P.E., Russell, A.G., and Mulholland, J.A. Method for fusing observational data and chemical transport model simulations to estimate spatiotemporally resolved ambient Air pollution. *Environ Sci Technol*. **2016**, *50*, 3695–705.

Goldman, G. T., Mulholland, J. A., Russell, A. G., Srivastava, A., Strickland, M. J., Klein, M., Waller, L. A., Tolbert, P. E., and Edgerton, E. S. Ambient Air Pollutant Measurement Error: Characterization and Impacts in a Time-Series Epidemiologic Study in Atlanta. *Environ. Sci. Technol*. **2010**, 44, 7692–7698.

Ozkaynak, H., Glenn, B., Qualters, J. R., Strosnider, H., Mcgeehin, M. A., and Zenick, H. Summary and findings of the EPA and CDC symposium on air pollution exposure and health. *J. Exposure Sci. and Environ. Epidemiol*. **2009**, 19, 19–29.

O'Lenick, C. R., Winquist, A., Mulholland, J. A., Friberg, M. D., Chang, H.H., Kramer, M.R., Darrow, L.A., and Sarnat, S.E. Assessment of neighborhood-level socioeconomic status as a modifier of air pollution–asthma associations among children in Atlanta. *J. Epidemiol. Community Health*. **2017**, *71*, 129–36.

Stockwell, W.R., Lawson, C.V., Saunders, E., and Goliff, W.S. A Review of Tropospheric Atmospheric Chemistry and Gas-Phase Chemical Mechanisms for Air Quality Modeling. *Atmosphere* **2012**, *3*, 1-32