GEOSTATISTICAL DATA FUSION ESTIMATION METHODS OF AMBIENT PM2.5 AND POLYCYCLIC AROMATIC HYDROCARBONS

Jeanette M. Reyes

A dissertation submitted to the faculty at the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Environmental Sciences and Engineering in the Gillings School of Global Public Health.

Chapel Hill
2016

Approved by:

Marc L. Serre

William Vizuete

Joachim Pleil

Michael Flynn

Amy Herring

ABSTRACT

Jeanette M. Reyes: Geostatistical Data Fusion Estimation Methods of Ambient PM2.5 and Polycyclic Aromatic Hydrocarbons
(Under the direction of Marc L. Serre)

Fine Particulate Matter (PM2.5) is a complex air pollutant associated with a host of adverse health effects. In epidemiologic studies there is a need to accurately predict exposures to reduce misclassification. Recently there has been a surge in data fusion methods which combine observed data with gridded modeled data like the regulatory Community Multiscale Air Quality (CMAQ) model. Substantial resources are allocated to the evaluation of CMAQ. However, this model has inherent error and uncertainty. Currently, CMAQ can only be operationally evaluated at locations where observed data exist, leaving potentially large spatial and temporal gaps in a given modeling domain. This study develops a framework for evaluating gridded air quality modeled data that can then be corrected for systematic error and combined with observed data in a geostatistical framework. First, this dissertation develops the novel Regionalized Air quality Model Performance (RAMP) method that performs a non-homogenous, non-linear, non-homoscedastic model evaluation at each CMAQ grid for a well-documented 2001 regulatory episode across the continental United States. The RAMP method comparatively outperforms other model evaluation methods with a 22.1% reduction in Mean Square Error (MSE). Secondly, the RAMP corrected CMAQ modeled data are combined with observed data in the modern Bayesian Maximum Entropy (BME) geostatistical framework which combines the accuracy of observed data with the spatial and temporal coverage of gridded modeled data. RAMP BME resulted in a 6-7 times increase in spatial refinement compared to using kriging alone. Lastly, the data rich PM2.5 environment is contrasted with the data poor environment of Polycyclic Aromatic Hydrocarbons (PAHs). The Mass Fraction (MF) BME method is developed through a relatively small number of paired PM2.5 and PAH values and is applied to PM2.5 observed locations where PAH have not been observed to create the first detailed spatial maps of PAH across North Carolina in 2005. The MF BME method reduces MSE by over 39% compared with using kriging alone. Accurate assessment of ambient air pollutants is essential in

public health to explore and elucidate true underlying relationships between pollutants and health endpoints.

ACKNOWLEDGMENTS

First and foremost, I would like to thank my advisor, Dr. Marc Serre, for his constant guidance and counsel. His mentorship throughout the years has taught me life-long lessons about research, technical writing, public health and professional collaborations. I would like to thank Dr. William Vizuete for his guidance in the WHIMS project along with his technical background and expertise with atmospheric chemistry. I would also like to thank my committee members Dr. Michael Flynn, Dr. Amy Herring and Dr. Joachim Pleil for their thoughtful and poignant comments that guided this dissertation.

Along with those at UNC, I would like to thank mentors at the EPA, especially Dr. Ana Rappold, Dr. Lisa Baxter and Dr. Lucas Neas for providing additional insight and research opportunities that both directly and indirectly bolstered this work.

I would like to thank past and present members of the BME lab for their insights and discussions including Dr. Yasuyuki Akita and Prahlad Jat along with my co-author for this work, Dr. Yadong Xu.

Lastly, I would like to thank all my friends and family for their support and their constant convincing that this work would be completed.

I would like to acknowledge the National Institute on Aging (NIA) under award number R01AG033078, the National Institute of Occupational Safety and Health (NIOSH) under grant 2T42/OH-008673 and the National Institute of Environmental Health Sciences (NIEHS) under grant T32ES007018 for their financial support.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1: INTRODUCTION

Epidemiologic studies investigating long term health effects of ambient air pollution exposures require accurate assessments to properly investigate underlying associations and health measures. Substantial efforts from sophisticated models have been used in past works to reduce misclassification that can otherwise obfuscate associations. The work presented here improves upon existing methods of exposure assessment. This work performs a model performance evaluation of gridded modeled Particulate Matter ≤ 2.5 micrometers (PM2.5) data across the continental United States (US) in 2001 which is then corrected of systematic error. The corrected modeled data is combined with observed PM2.5 data in the Bayesian Maximum Entropy (BME) geostatistical framework. Lastly, this work ends with the estimation of Polycyclic Aromatic Hydrocarbons (PAHs) across North Carolina in 2005. This is the first known work to create a flexible data fusion method combining observed and gridded modeled data for PM2.5 using the BME framework. This is also the first known work to create a full prediction map of PAH across North Carolina for 2005. We hypothesize that combining environmental air pollution data sets defined over different supports (i.e. over a point location versus over a grid) in a geostatistical framework will improve estimation accuracy compared to using a single data source for a given environmental parameter.

Model performance evaluation is needed to understand resulting model error and in an epidemiologic context when the resulting gridded estimations are being used to estimate exposure. There is a wealth of studies that explore performance evaluation of gridded Chemical Transport Models (CTMs), specifically the Community Multiscale Air Quality (CMAQ) model. CMAQ is the US Environmental Protection Agency's (EPAs) regulatory air quality model (Appel et al., 2013a, 2008; Carlton et al., 2010; Foley et al., 2015a, 2015b, 2010). It is used to assess attainment of regulated air pollution to ensure they are under the regulatory standard. Substantial resources go towards ensuring that CMAQ modeled values match as closely as possible with corresponding paired observed values, through operational, diagnostic and dynamic assessments (Dennis et al., 1996). The metrics used to evaluate operational

performance are numerous and multifaceted. These metrics becomes increasingly difficult to properly calculate when the region over which they are being calculated becomes increasingly smaller in size (Simon et al., 2012). Currently, an operational performance cannot be assessed in-between monitors and, in the limiting case, cannot be assessed for individual CMAQ grids. This information is needed to assess performance in-between monitors and to perform an error correction on individual CMAQ grids.

Most epidemiologic studies utilize data fusion methods. In recent years, these methods have increased in popularity. Data fusion methods typically combine two different data sources of a given air pollutant, where the data sources have different levels of support into a geostatistical framework. Different supports typically include observed monitoring data defined at a point location with modeling data defined over a grid. Data fusion methods allow for the accuracy typically associated with observed data with the spatial refinement and coverage associated with gridded modeling data. Popular approaches include Bayesian Melding and the Downscaler method (Berrocal et al., 2010a; Fuentes and Raftery, 2005). As sophisticated as these models can be, they still assume the relationship between modeled and observed data to be linear and homoscedastic. This can be limiting when there is a known difference between uncertainties of errors for difference ranges of PM2.5 concentrations. There are a variety of geostatistical methods that can implement a data fusion framework, including the BME framework.

BME is a mathematically rigorous geostatistical space/time framework originally developed by Christakos (Christakos and Serre, 2000; Christakos et al., 2001). BME is an extension of linear kriging in which information about a Space/Time Random Field (S/TRF) is divided into two knowledge bases: 1) a site-specific knowledge base characterizing the Space/Time Random Field (S/TRF) representing a process at a specific space/time, 2) a general knowledge base that comes in the form of a prior Probability Distribution Function (PDF) describing the random field. These knowledge bases are combined and the BME posterior PDF can be used to predict environmental parameters at unmonitored locations. Unlike kriging, BME is able to incorporate information that is non-Gaussian. In an environmental setting this can be essential when the distribution of the parameter is known to be highly skewed, when a sizable portion of a PDF is below zero, or when parameters are below a given detection limit (Messier et al., 2015). BME has been successfully implemented in water (Akita et al., 2007), air (Reyes and Serre, 2014) and disease parameters (Allshouse et al., 2011). BME data fusion has been

performed with ozone in previous work (de Nazelle et al., 2010; Xu et al., 2016). BME can also be used to inform mapping scenarios in data poor environments.

PM2.5 is a complex mixture of many different constituents. A component of PM2.5 is PAH. PAHs are created from incomplete fuel combustion with some species being carcinogenic (Bocskay et al., 2005; Menzie et al., 1992; Wolff et al., 2005). They can come from a variety of sources including wildfires. However, ambient PAHs are currently not regulated by the EPA and therefore, a nationwide monitoring network does not currently exist for them. PAH measurements can also be costly (Pleil et al., 2004). There is a need to investigate PAH concentrations over a large region in a cost effective way. Previous work has investigated the relationship between PAH and PM2.5 concentration around the World Trade Center after September, 11th (Allshouse et al., 2009). In particular, PM2.5 samples were analyzed for several different species of PAH. The relationship between the fractions of PAH to PM2.5 was investigated and applied to other areas in the sampling region. However, this was only applied to a relatively small region over a short period of time. Maps of PAH concentrations across North Carolina are currently lacking in the literature.

The theme tying all Chapter 2, 3 and 4 together are the benefits of combining multiple data sources together with the goal of these merged/combined data sources being more beneficial that any one data source individually. This dissertation explores two different types of data environments: a data rich environment and a data poor environment. Chapter 2 and 3 explore a data rich environment. A data rich environment lends itself to allowing flexible relationships between the two different data sources. The paired observed and modeled PM2.5 data are plentiful enough to develop a relationship that is non-homogenous, non-linear and non-homoscedastic. This relationship is flexible enough to be non-parametric. Guided with the goal of improved exposure assessment, Chapter 2 and Chapter 3 look at model performance evaluation and error correction of daily PM2.5 mass across the continental United States for 2001. This application explores the themes of performance evaluation and error correction in a data rich environment. Chapter 2 develops the Regionalized Air quality Model Performance (RAMP) method (Xu et al., 2016). Chapter 3 combines the RAMP-corrected CMAQ modeled data with daily observed PM2.5 mass into the BME geostatistical framework. This is compared to more typically used geostatistical methods like kriging and more popular data fusion methods like the Downscaler method.

By contrast, Chapter 4 explores data merging in a data poor environment. Paired observed PAH and PM2.5 are sparse across North Carolina. Therefore, the relationship between them needs to be parsimonious and as specific to the air shed of interest as possible. Chapter 2, 3 and 4 look at developing a relationship between different air pollution data sets that is regionalized and specific to the location of interest. Two relationships are explored between the paired PAH and PM2.5 data: a standard simple linear regression model and a model that explores the mass fraction between the two. These two methods are then compared with other typically used geostatistical methods like kriging and cokriging.

CHAPTER 2: REGIONALIZED PM2.5 COMMUNITY MULTISCALE AIR QUALITY MODEL PERFORMANCE EVALUATION ACROSS A CONTINUOUS SPATIOTEMPORAL DOMAIN[1]

2.1 Introduction

Particulate Matter ≤ 2.5 micrometers in diameter (PM2.5) is one of the six "criteria air pollutants" regulated in the United States (Boldo et al., 2006; Pope et al., 2009) due to its association with adverse health effects, including cardiovascular and respiratory disease and mortality (Beelen et al., 2007; Krewski et al., 2009; Pope et al., 2004). The Community Multiscale Air Quality (CMAQ) model is used for regulatory purposes to estimate PM2.5 and assess attainment. Substantial efforts are made to assess and understand the model performance of CMAQ ensuring that modeled values match with observed data (Appel et al., 2013b, 2008; Carlton et al., 2010; Foley et al., 2015a, 2015b, 2010). Past work evaluating model performance typically gives modeling performance statistics over an aggregated level (e.g. monitoring locations, regions of the country, monitoring networks, etc.) (Simon et al., 2012). For the modeling domain of the continental United States, metrics are typically calculated for the Eastern versus Western US, urban stations versus rural stations, summer versus winter monitoring, etc. (Appel et al., 2013a). Displaying model performance metrics at each monitoring site location across the US reveals that CMAQ performance changes in a non-homogenous manner (Appel et al., 2012). However, model performance at a specific unmonitored space/time location is typically not explored or known. Therefore current methods fail to assess geographical or temporal changes of model performance across the spatiotemporal continuum, particularly in-between monitors.

The goal of this work is to address this significant knowledge gap by introducing a method that assesses model performance at any space/time region of interest across the spatiotemporal continuum.

[1] This chapter was submitted as an article to the journal Atmospheric Environment. Reyes, Jeanette M., Xu,Yadong, Vizuete, William, Serre, L. Marc. Regionalized PM2.5 Community Multiscale Air Quality model performance evaluation across a continuous spatiotemporal domain.

Advantages for assessing model performance at any region across a continuum include being able to 1) exactly delineate geographical patterns of modeling errors and 2) correct systematic errors across the modeling domain for individual CMAQ grid concentrations.

Systematic errors are consistent deviations of modeled data from observed data. Systematic errors, once assessed, can be used to correct the modeled value. The remaining error, i.e. the random noise of the modeled value around the observed data, is the random error. While current CMAQ model performance evaluation methods are multifaceted (Dennis et al., 2010) and use a wide array of metrics to quantify performance (Kang et al., 2007; Thunis et al., 2012; USEPA, 2005; Venkatram, 2008), this work specifically focuses on set of metrics that investigate systematic and random errors. Hence, to achieve our goal, we introduce modeling error statistics that parse total errors into systematic and random errors. Few studies have apportioned error in this manner (Solazzo and Galmarini, 2016).

The method we introduce in this work to assess model performance across the spatiotemporal continuum is the Regionalized Air quality Model Performance (RAMP) method, which we use to study daily PM2.5 across the continental US. Our framework is a regionalized space/time extension of the Constant Air quality Model Performance (CAMP) method (de Nazelle et al., 2010) and parallels the work of Xu et al. (Xu et al., 2016). The CAMP method was originally used to account for the non-linear and non-homoscedastic relationship between modeled and observed ozone data in North Carolina for a particular ozone episode. The CAMP method assumes that model performance is homogenous across the state and does not change as a function of the space/time CMAQ grid locations. This assumption of homogeneity of model performance begins to break down as the modeling domain increases in size, particularly when this increase is substantial. The novel RAMP method introduced here for PM2.5 extends the CAMP method by accounting for the non-homogeneity of model performance in a regionalized fashion across the entirety of a modeling domain and fully characterizes the non-linear and non-homoscedastic relationship at any space/time region for any modeled value of interest.

This work demonstrates the use of the RAMP method by implementing a regionalized performance evaluation of daily PM2.5 mass predicted by CMAQ across the entirety of the continental United States. As an evaluation of the RAMP method, we have chosen a regulatory episode developed for the years 2001 and 2002. The model performance for this episode has been well documented and

thus provides an ideal case study of the RAMP method. The results of the RAMP analysis include maps showing the geographical variations of systematic and random errors at a fine spatial resolution displayed at the resolution of an individual CMAQ grid cell. These results provide new insights regarding model performance that complement existing performance evaluation methods. The RAMP results are helpful in making decision on resource allocation for further improvement in the air quality model. Furthermore, calculating systematic errors for individual CMAQ grids facilitate systematic error correction leading to maps of PM2.5 concentrations with improved mapping accuracy.

2.2 Materials and Methods

2.2.1 Observed and Modeled Data

Daily observed PM2.5 for each space/time location during 2000-2002 were constructed based on raw monitoring data from monitoring stations measuring either hourly or daily PM2.5 obtained from the EPA's Air Quality Systems (AQS) data base (US EPA, n.d.). Daily PM2.5 data were also constructed from CMAQ modeled data for years 2001 and 2002 using CMAQv4.5 across the contiguous United States on a 36 km grid. For more detailed information regarding the aggregation and pairing process of observed and modeled data see Appendix A.

2.2.2 Variable Definition

Random variables X are in upper case and known values are in lower case. Let $\hat{X}(\boldsymbol{p})$ be the random variable representing the observed concentration at a single space/time location $\boldsymbol{p} = (\boldsymbol{s}, t)$, $\hat{x}(\boldsymbol{p})$ be its known value (i.e. realization) at space/time location $\boldsymbol{p}$ and $\tilde{x}(\boldsymbol{p})$ be the CMAQ modeled value at space/time location $\boldsymbol{p}$. Because $\tilde{x}(\boldsymbol{p})$ covers the entirety of the domain, it is known everywhere. We define error as

$$E(\boldsymbol{p}) = \tilde{x}(\boldsymbol{p}) - \hat{X}(\boldsymbol{p}) \qquad \text{(Equ. 2-1)}$$

Error is defined as $e(\boldsymbol{p}) = \tilde{x}(\boldsymbol{p}) - \hat{x}(\boldsymbol{p})$ at locations where the observed data are known. The definition of error in this work is a deviation from what is typically used in the model performance literature. The differences in the nomenclature are explicitly stated in Appendix A (Table A.1 and Table A.2).

2.2.3 Systematic and Random Error Statistics

In this work metrics are geared towards dividing error in a dichotomous manner. Namely, metrics are divided into systematic and random errors. Systematic errors are consistent errors between observed

and modeled CMAQ data and can be removed through calculating the mean systematic error. Random errors are the residual errors remaining once the systematic error is removed. Random errors can be conceptualized as the random noise between CMAQ and observed data. Total error is the sum of systematic and random error. In the naming convention of a statistic the first letter(s) is used to identify the statistical operator as follows: M=mean, V=variance, S=Standard deviation, RMS=square Root of the Mean of Squared values. The last letter(s) is used to identify the value of interest as follows: E=Error (Equ. 2-1), SE=Squared Error=E^2, S=Standardized error=E/σ_E, NE=Normalized Error=E/$\hat{x}$ and R=square Root of error variance=$\sqrt{\sigma_E}$. Statistics that are calculated over an entire domain $\mathcal{D}$ are $ME(\mathcal{D}) = \frac{1}{n(\mathcal{D})}\sum e_i$ and $VE(\mathcal{D}) = \frac{1}{n(\mathcal{D})-1}\sum(e_i - ME(\mathcal{D}))^2$. $ME^2(\mathcal{D})$ quantifies the systematic error, $VE(\mathcal{D})$ quantifies the random error and $MSE(\mathcal{D}) = ME^2(\mathcal{D}) + VE(\mathcal{D})$ quantifies the total error. The equations of systematic, random and total error can be represented pictorially through use of a target analogy, probability distribution function of error and plotting observed data as a function of modeled data (Fig. 2-1a-i). Other statistics used in model performance evaluation include the square Root of the Mean of Squared Standardized errors (RMSS) and Mean of the square Root of variance (MR).

2.2.4 Constant Air quality Model Performance (CAMP)

The CAMP method (de Nazelle et al., 2010) performs a model performance analysis that accounts for the non-linearity and non-homoscedastic behavior of model performance with respect to the modeled value $\tilde{x}_k$. The CAMP method does this by modeling the mean $\lambda_1(\tilde{x}_k;\mathcal{D}) = M\left[\hat{X}|\tilde{x}_k;\mathcal{D}\right]$ and variance $\lambda_2(\tilde{x}_k;\mathcal{D}) = V\left[\hat{X}|\tilde{x}_k;\mathcal{D}\right]$ of the observed value $\hat{X}$ as function of a given model value $\tilde{x}_k$ across the domain $\mathcal{D}$ using the equations

$$\lambda_1(\tilde{x}_k;\mathcal{D}) \approx \frac{1}{n(\tilde{x}_k;\mathcal{D})}\sum \hat{x}_i \quad \text{(Equ. 2-2)}$$

$$\lambda_2(\tilde{x}_k;\mathcal{D}) \approx \frac{1}{n(\tilde{x}_k;\mathcal{D})-1}\sum\left(\hat{x}_i - \lambda_1(\tilde{x}_k;\mathcal{D})\right)^2 \quad \text{(Equ. 2-3)}$$

where $n(\tilde{x}_k;\mathcal{D})$ is the number of paired modeled $\tilde{x}_i$ and observed $\hat{x}_i$ values across the space time domain $\mathcal{D}$ such that $\tilde{x}_k - \Delta\tilde{x} \le \tilde{x}_i \le \tilde{x}_k + \Delta\tilde{x}$ where $\Delta\tilde{x}$ is a small tolerance corresponding to half of a decile of modeled values.

Previous work (de Nazelle et al., 2010) found that the relationship of $\lambda_1(\tilde{x}_k;\mathcal{D})$ and $\lambda_2(\tilde{x}_k;\mathcal{D})$ with respect to $\tilde{x}_k$ can be expressed through domain wide "S-curves" that are non-linear, indicating that model

performance is non-linear and non-homoscedastic. However the CAMP method does not investigate how $\lambda_1(\tilde{x}_k;\mathcal{D})$ and $\lambda_2(\tilde{x}_k;\mathcal{D})$ S-curves change across space and time.

2.2.5 Regionalized Air quality Model Performance (RAMP)

The Regionalized Air quality Model Performance (RAMP) method introduced here consists of extending the CAMP method (de Nazelle et al., 2010) by regionalizing the model performance to a space/time region $\mathcal{R}(\boldsymbol{p})$ associated with the space/time coordinate $\boldsymbol{p}$. In this work the region $\mathcal{R}(\boldsymbol{p})$ was selected such that it contains all paired modeled and observed data from the 3 closest stations within 180 days of $\boldsymbol{p}$, resulting in a regionalized S-curve (Fig. 2-1j). The 3 closest stations within 180 days were chosen for being as spatially specific as possible while still maintaining a stable pattern with the associated regionalized $\lambda_1(\tilde{x}_k;\mathcal{R}(\boldsymbol{p})) = M\left[\hat{X}|\tilde{x}_k;\mathcal{R}(\boldsymbol{p})\right]$ and $\lambda_2(\tilde{x}_k,\mathcal{R}(\boldsymbol{p})) = V\left[\hat{X}|\tilde{x}_k;\mathcal{R}(\boldsymbol{p})\right]$ parameters (see Appendix A for S-curve parameter optimization). The parameters are calculated as

$$\lambda_1(\tilde{x}_k;\mathcal{R}(\boldsymbol{p})) \approx \frac{1}{n(\tilde{x}_k;\mathcal{R}(\boldsymbol{p}))}\sum \hat{x}_i \quad \text{(Equ. 2-4)}$$

$$\lambda_2(\tilde{x}_k;\mathcal{R}(\boldsymbol{p})) \approx \frac{1}{n(\tilde{x}_k;\mathcal{R}(\boldsymbol{p}))-1}\sum\left(\hat{x}_i - \lambda_1(\tilde{x}_k;\mathcal{R}(\boldsymbol{p}))\right)^2 \quad \text{(Equ. 2-5)}$$

where $n(\tilde{x}_k;\mathcal{R}(\boldsymbol{p}))$ is the number of paired modeled and observed points within $\mathcal{R}(\boldsymbol{p})$ and around $\tilde{x}_k$.

An efficient numerical implementation of the calculation of $\lambda_1(\tilde{x}_k;\mathcal{R}(\boldsymbol{p}))$ and $\lambda_2(\tilde{x}_k;\mathcal{R}(\boldsymbol{p}))$ is performed as follows. All modeled/observed $(\tilde{x}_i,\hat{x}_i)$ pairs within $\mathcal{R}(\boldsymbol{p})$ are divided into deciles based off all the collected $\tilde{x}_i$ (Fig. 2-1j). The mean and variance of observed values in each decile $\tilde{x}_l$ are calculated to obtain $\lambda_{1,l}(\tilde{x}_l;\mathcal{R}(\boldsymbol{p}))$ and $\lambda_{2,l}(\tilde{x}_l;\mathcal{R}(\boldsymbol{p}))$, respectively. A linear interpolation between deciles is performed to obtain $\lambda_1(\tilde{x}_k;\mathcal{R}(\boldsymbol{p}))$ and $\lambda_2(\tilde{x}_k;\mathcal{R}(\boldsymbol{p}))$. If the S-curve contains less than 150 pairs, points from the nearest stations are pulled in until at least 150 pairs are obtained. When calculating the variance of the error correction of the modeled data (Equ. 2-5), it is assumed that nearby observed values used in the calculation are independent and identically distributed (i.e. $\hat{x}_i \sim iid$). Thus, $\lambda_1(\tilde{x}_k;\mathcal{R}(\boldsymbol{p}))$ and $\lambda_2(\tilde{x}_k;\mathcal{R}(\boldsymbol{p}))$ describe the mean and variance of observed concentration as a function of both $\tilde{x}_k$ and the space/time region $\mathcal{R}(\boldsymbol{p})$. For example, in Fig. 2-1j for the given $\mathcal{R}(\boldsymbol{p})$ and $\tilde{x}_k = 5.6\ \mu g/m^3$, $\lambda_1(\tilde{x}_k;\mathcal{R}(\boldsymbol{p})) = 7.9\ \mu g/m^3$ and $\sqrt{\lambda_2(\tilde{x}_k;\mathcal{R}(\boldsymbol{p}))} = 2.5\ \mu g/m^3$.

There is a correspondence between the parameters $\lambda_1(\tilde{x}_k;\mathcal{R}(\boldsymbol{p}))$ and $\lambda_2(\tilde{x}_k;\mathcal{R}(\boldsymbol{p}))$ and systematic and random errors. From Equ. 2-1 we have $\hat{X} = \tilde{x} - E$, which, once substituted into $\lambda_1(\tilde{x}_k;\mathcal{R}(\boldsymbol{p}))$ and $\lambda_2(\tilde{x}_k,\mathcal{R}(\boldsymbol{p}))$ yields

$$\lambda_1(\tilde{x}_k;\mathcal{R}(\boldsymbol{p})) = M[\tilde{x} - E\ |\tilde{x}_k;\mathcal{R}(\boldsymbol{p})] = \tilde{x}_k - M[E\ |\tilde{x}_k;\mathcal{R}(\boldsymbol{p})] = \tilde{x}_k - ME(\tilde{x}_k;\mathcal{R}(\boldsymbol{p})) \quad \text{(Equ. 2-6)}$$

$$\lambda_2(\tilde{x}_k;\mathcal{R}(\boldsymbol{p})) = V[\tilde{x} - E|\tilde{x}_k;\mathcal{R}(\boldsymbol{p})] = V[E(\boldsymbol{p})|\tilde{x}_k;\mathcal{R}(\boldsymbol{p})] = VE(\tilde{x}_k;\mathcal{R}(\boldsymbol{p})) \quad \text{(Equ. 2-7)}$$

where $ME(\tilde{x}_k;\mathcal{R}(\boldsymbol{p}))$ and $VE(\tilde{x}_k;\mathcal{R}(\boldsymbol{p}))$ are the mean and variance, respectively, of the error associated with an arbitrary value $\tilde{x}_k$ predicted within region $\mathcal{R}$ associated with $\boldsymbol{p}$.

We also define $\lambda_1^{RAMP}(\boldsymbol{p}) = \lambda_1(\tilde{x}(\boldsymbol{p});\mathcal{R}(\boldsymbol{p}))$ and $\lambda_2^{RAMP}(\boldsymbol{p}) = \lambda_2(\tilde{x}(\boldsymbol{p});\mathcal{R}(\boldsymbol{p}))$ as the mean and variance of observed concentration when $\tilde{x}_k = \tilde{x}(\boldsymbol{p})$, where $\tilde{x}(\boldsymbol{p})$ is the CMAQ modeled value at $\boldsymbol{p}$. By replacing $\tilde{x}_k$ with $\tilde{x}(\boldsymbol{p})$ in Equ. 2-6 and Equ. 2-7, we obtain

$$\lambda_1^{RAMP}(\boldsymbol{p}) = \tilde{x}(\boldsymbol{p}) - ME(\tilde{x}(\boldsymbol{p});\mathcal{R}(\boldsymbol{p})) \quad \text{(Equ. 2-8)}$$

$$\lambda_2^{RAMP}(\boldsymbol{p}) = VE(\tilde{x}(\boldsymbol{p});\mathcal{R}(\boldsymbol{p})) \quad \text{(Equ. 2-9)}$$

Equ.2-8 and Equ. 2-9 provide a physical interpretation of systematic and random errors. The systematic error $ME(\tilde{x}(\boldsymbol{p});\mathcal{R}(\boldsymbol{p}))$ is the error correction that can be applied to the modeled value $\tilde{x}(\boldsymbol{p})$ in region $\mathcal{R}(\boldsymbol{p})$ to produce a corrected modeled estimate $\lambda_1^{RAMP}(\boldsymbol{p})$, and the random error quantified by $VE(\tilde{x}(\boldsymbol{p});\mathcal{R}(\boldsymbol{p}))$ characterizes the residual uncertainty associated with the systematic error-corrected modeled estimate. In this work $ME(\tilde{x}(\boldsymbol{p});\mathcal{R}(\boldsymbol{p}))$ and $VE(\tilde{x}(\boldsymbol{p});\mathcal{R}(\boldsymbol{p}))$ can be approximated by $ME(\tilde{x}(\boldsymbol{p});\mathcal{R}(\boldsymbol{p})) \approx \frac{1}{n(\boldsymbol{p})}\sum e_i$ and $VE(\tilde{x}(\boldsymbol{p});\mathcal{R}(\boldsymbol{p})) \approx \frac{1}{n(\boldsymbol{p})-1}\sum(e_i - ME(\tilde{x}(\boldsymbol{p});\mathcal{R}(\boldsymbol{p})))^2$, respectively, where for a given $\boldsymbol{p}$, $n(\boldsymbol{p})$ is equal to the number of paired modeled and observed points in $\mathcal{R}(\boldsymbol{p})$.

The RAMP method provides the statistical distribution of observed air pollution as

$$\hat{X}(\boldsymbol{p})|\tilde{x}(\boldsymbol{p}) \sim N(\lambda_1^{RAMP}(\boldsymbol{p}), \lambda_2^{RAMP}(\boldsymbol{p})) \quad \text{(Equ. 2-10)}$$

where $\lambda_1^{RAMP}(\boldsymbol{p})$ and $\lambda_2^{RAMP}(\boldsymbol{p})$ are the mean and variance of observed values given the modeled value $\tilde{x}(\boldsymbol{p})$.

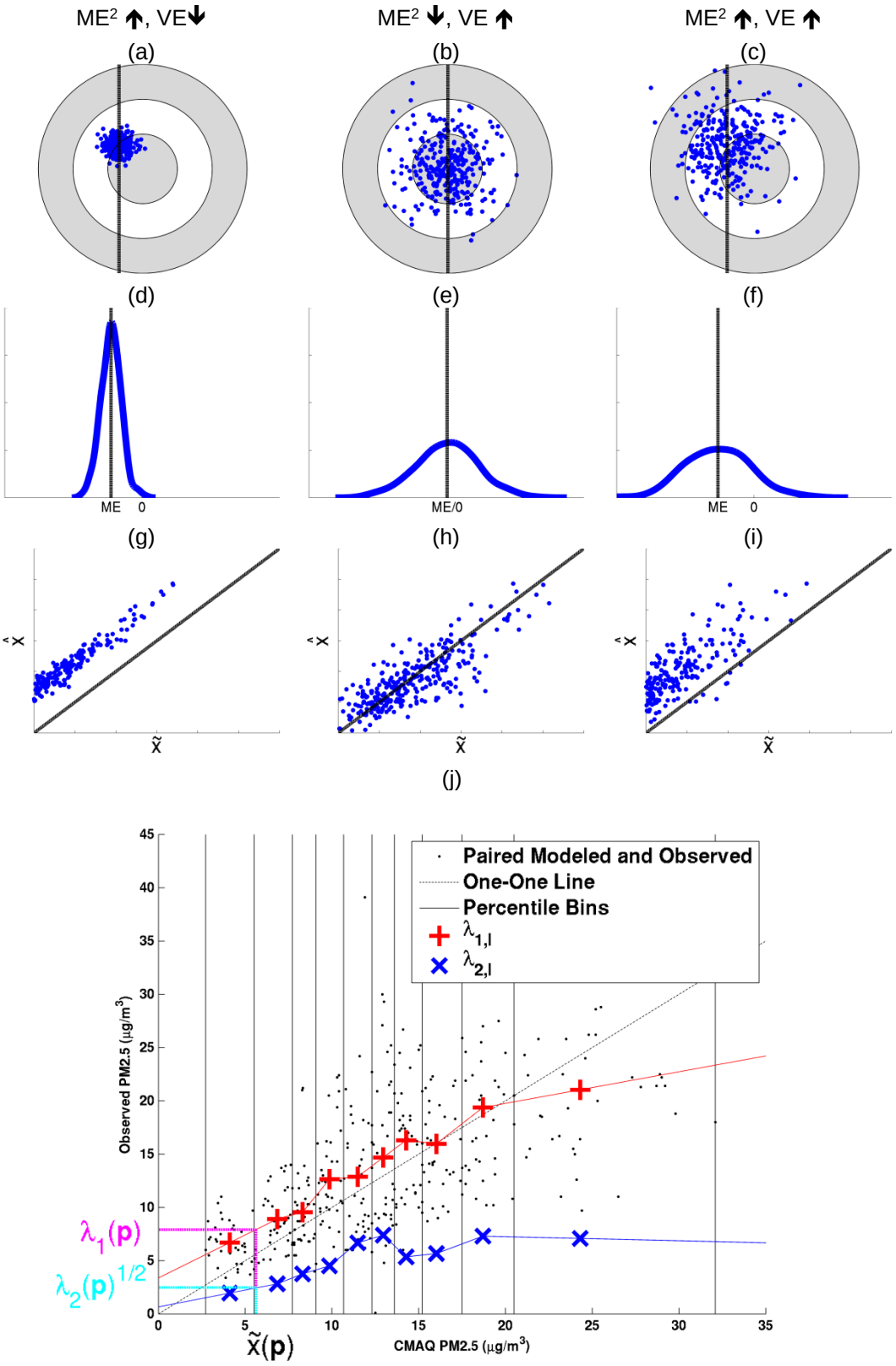


Figure 2.1. Visual representations of systematic and random error. Panels (a)-(i) show different scenarios of high and low systematic and random errors. The left column (plots (a), (d), (g)) displays three different visual representations of estimates with large systematic error (i.e. high ME^2) and low random error (i.e.

low VE). The middle column (plots (b), (e), (h)) displays representations of estimates with low systematic error and large random error. The right column (plots (c), (f), (i)) displays representations of estimates with large systematic error and large random error. The top row displays error using a target analogy, where estimates should ideally land on the target. The middle row displays the distribution of error via a PDF. The bottom row displays a group of paired modeled and observed concentrations around a given location. The modeled values are displayed on the independent axis as $\tilde{x}$ and the observed values are displayed on the dependent axis as $\hat{x}$. The solid line is the one-to-one line. Plot (j) shows the RAMP analysis of an arbitrary CMAQ grid location on 07/01/2001 for daily PM2.5. The black dots are all the paired modeled and observed daily PM2.5 concentrations within a space/time region $\mathcal{R}(\boldsymbol{p})$ consisting of the 3 closest stations to the CMAQ grid location of interest within 180 days of 07/01/2001, with modeled data on the independent axis and observed data on the dependent axis. The vertical black lines identify the 10 bins used to stratify all the paired data in which each bin contains one decile of all the paired points. The dotted black line is the one-to-one line between the modeled and observed data. The red + marker in each bin denotes $\lambda_{1,l}(\tilde{x}_l, \mathcal{R}(\boldsymbol{p}))$, the average of paired observed values within the l-th decile bin. The blue x marker in each bin denotes the square root of $\lambda_{2,l}(\tilde{x}_l; \mathcal{R}(\boldsymbol{p}))$, the standard deviation of paired observed values within that bin. As shown in the figure, the + and x markers are linearly interpolated to obtain the $\lambda_1(\boldsymbol{p})$ and $\sqrt{\lambda_2(\boldsymbol{p})}$ values, respectively, corresponding to the CMAQ modeled data $\tilde{x}(\boldsymbol{p})$ within $\mathcal{R}(\boldsymbol{p})$.

2.6 Validation and Stochastic Simulation

Validation is performed by comparing the accuracy of the model correction performed by three approaches: the Constant, CAMP, and RAMP correction methods. The Constant correction method is defined through

$$\lambda_1^{Constant}(\boldsymbol{p}) = \tilde{x}(\boldsymbol{p}) - ME(\mathcal{D}), \quad \text{(Equ. 2-11)}$$

with associated error variance

$$\lambda_2^{Constant}(\boldsymbol{p}) = VE(\mathcal{D}), \quad \text{(Equ. 2-12)}$$

i.e. the correction $ME(\mathcal{D})$ and its associated error variance $VE(\mathcal{D})$ are constant across the entirety of the domain with respect to both modeled value $\tilde{x}_k$ and location $\boldsymbol{p}$. The CAMP method assumes that the model performance of CMAQ is represented by domain wide S-curves $\lambda_1(\tilde{x}_k; \mathcal{D})$ and $\lambda_2(\tilde{x}_k; \mathcal{D})$ (Equ. 2-2,2-3) that are a function of the modeled value $\tilde{x}_k$, but not a function of space/time location $\boldsymbol{p}$. In the CAMP method the correction for $\tilde{x}(\boldsymbol{p})$ is performed by substituting $\tilde{x}_k$ with $\tilde{x}(\boldsymbol{p})$ in the domain-wide S-curves, i.e. using the correction

$$\lambda_1^{CAMP}(\boldsymbol{p}) = \lambda_1(\tilde{x}(\boldsymbol{p}); \mathcal{D}) = \tilde{x}(\boldsymbol{p}) - ME(\tilde{x}(\boldsymbol{p}); \mathcal{D}) \quad \text{(Equ. 2-13)}$$

with associated error variance

$$\lambda_2^{CAMP}(\boldsymbol{p}) = \lambda_2(\tilde{x}(\boldsymbol{p}); \mathcal{D}) = VE(\tilde{x}(\boldsymbol{p}); \mathcal{D}). \quad \text{(Equ. 2-14)}$$

The RAMP correction on the other hand is done using Equ. 2-8 and 2-9. The corrected $\lambda_1(\boldsymbol{p})$ values for the Constant (Equ. 2-11), CAMP (Equ. 2-13) and RAMP (Equ. 2-8) methods are compared by calculating

performance statistics between paired $\lambda_1(\boldsymbol{p})$ and $\hat{x}(\boldsymbol{p})$ values for 2001. The performance of $\lambda_2(\boldsymbol{p})$ is assessed through standardized errors as shown in Table A.2 (i.e. $\frac{\lambda_1(\boldsymbol{p})-\hat{x}(\boldsymbol{p})}{\sqrt{\lambda_2(\boldsymbol{p})}}$).

We also conduct a stochastic simulation to test how well each method reproduces the simulated truth. The maps of $\lambda_1(\boldsymbol{p})$ and $\lambda_2(\boldsymbol{p})$ obtained in this work are defined as being the true mean and variance of observed values. We also select $\tilde{x}(\boldsymbol{p})$ from this work as being the true modeled values. We randomly generate $\hat{x}^*(\boldsymbol{p}) \sim N(\lambda_1(\boldsymbol{p}), \lambda_2(\boldsymbol{p}))$ and then we re-calculate ${\lambda_1}^*(\boldsymbol{p})$ and ${\lambda_2}^*(\boldsymbol{p})$ using the Constant, CAMP and RAMP methods based only on paired $\tilde{x}(\boldsymbol{p})$ and $\hat{x}^*(\boldsymbol{p})$. Lastly, ${\lambda_1}^*(\boldsymbol{p})$ and ${\lambda_2}^*(\boldsymbol{p})$ are compared with $\lambda_1(\boldsymbol{p})$ and $\lambda_2(\boldsymbol{p})$ visually through maps and through statistical metrics to evaluate how well ${\lambda_1}^*(\boldsymbol{p})$ and ${\lambda_2}^*(\boldsymbol{p})$ are able to capture the spatial variability in the true mean, $\lambda_1(\boldsymbol{p})$, and variance, $\lambda_2(\boldsymbol{p})$, of observed values.

2.3 Results and Discussion

2.3.1 Model Performance Evaluation Results Demonstrating the RAMP Analysis

A demonstration of the RAMP method was performed using daily PM2.5 concentrations predicted by CMAQ at the 36 km grid level for 2001 across the continental United States. The version of CMAQ used to calculate the PM2.5 was v4.5, which is the most recent version of CMAQ available for 2001 across the continental US. Although newer versions of CMAQ exist for later years, it was critical to analyze model performance in 2001 due to an ongoing epidemiological study focused on novel neurodegenerative PM2.5 health end points and its association with loss of brain mass in older women (Chen et al., 2015). This study is based on a cohort from the Women's Health Initiative-Memory Study (WHI-MS) and exposure data was reconstructed from 1999 to 2006. From an epidemiologic perspective, a model performance evaluation that can distinguish systematic from random error is especially important for a model version with known deficiencies (Foley et al., 2010). This new information can inform subsequent error correction of systematic errors and data fusion with data from other sources.

Results of the RAMP analysis can be visualized for 07/01/2001 (Fig. 2.2). The RAMP results indicate that there are very clear geographical patterns in $ME^2(\boldsymbol{p}) = ME^2(\tilde{x}(\boldsymbol{p}); \mathcal{R}(\boldsymbol{p}))$ (Fig. 2.2a) and $VE(\boldsymbol{p}) = VE(\tilde{x}(\boldsymbol{p}); \mathcal{R}(\boldsymbol{p}))$ (Fig. 2.2b). This indicates that both systematic errors and random errors are non-homogenous as demonstrated by the > 10 fold variation in $ME^2(\boldsymbol{p})$ and $VE(\boldsymbol{p})$ across the continental

United States on that day. The maps shown in Fig. 2a-b allow for the identification of regions with high systematic and random errors. This is critical information needed to better understand the spatial uncertainty of model performance of the CMAQ predicted values $\tilde{x}(\boldsymbol{p})$ across a given day (Fig. 2.2c). The RAMP analysis also produces $\lambda_1^{RAMP}(\boldsymbol{p})$ (Fig. 2.2d), which can directly be used in place of CMAQ values. In addition to the results shown in Fig. 2.2, the RAMP analysis produces a rich set of more detailed model performance metrics (see Appendix A).

The domain wide model performance of CMAQ is assessed by the performance statistics $ME(\mathcal{D})$, $\sqrt{VE(\mathcal{D})}$, $MSE(\mathcal{D})$ and $r(\mathcal{D})$ calculated over a domain $\mathcal{D}$ corresponding to the continental United States in 2001. These statistics are shown in the first column of Table 2.1. Due to the influential nature of highly skewed standardized errors, all data were removed whose standardized errors were either less than the 0.1 percentile or greater than the 99.9 percentile, constituting 348 data points. As shown in Table 2.1, the mean error for CMAQ is $ME(\mathcal{D}) = -1.05(\mu g/m^3)$, indicating that CMAQv4.5 has systematic errors that on average underestimates PM2.5 by $1.05\ \mu g/m^3$ across the continental United States in 2001. Interestingly, $\sqrt{VE(\mathcal{D})} = 7.77(\mu g/m^3)$, indicating that random errors are much larger than systematic errors. These systematic and random errors result in a total error of $MSE(\mathcal{D}) = 61.5(\mu g/m^3)^2$ and with precision quantified by a correlation $r(\mathcal{D}) = 0.589$ between observed and modeled values.

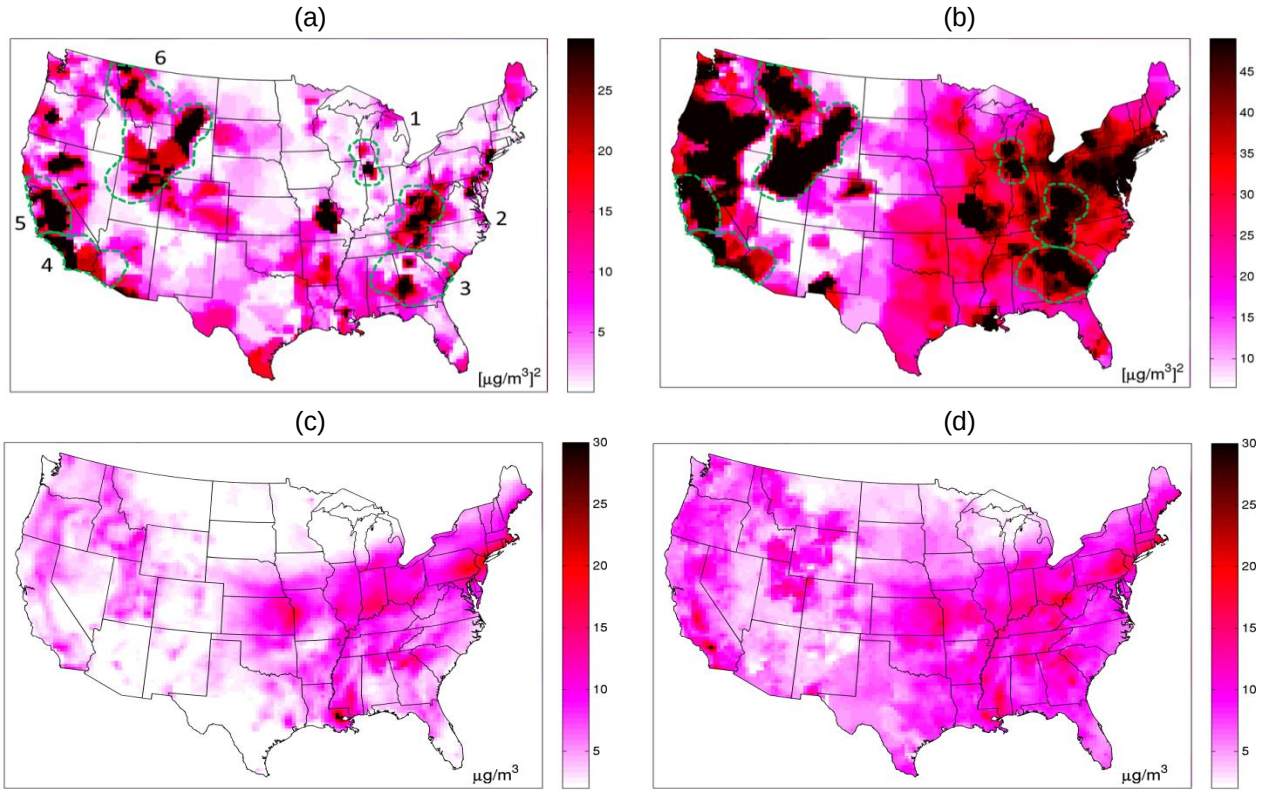


Figure 2.2. Maps of RAMP error and RAMP error correction of CMAQ. Daily PM2.5 across the continental United States on 07/01/2001 displaying (a) RAMP $ME^2(\boldsymbol{p})$, (b) RAMP $VE(\boldsymbol{p})$, (c) CMAQ concentration $\tilde{x}(\boldsymbol{p})$ and (d) $\lambda_1^{RAMP}(\boldsymbol{p})$. Plots (c) and (d) are in $\mu g/m^3$ and (a) and (b) are in $(\mu g/m^3)^2$. Plot (b) shows 6 regions of large random error delineated in the dashed green line with the same regions delineated and labeled in (a). Delineated regions include (1) the Great Lakes, (2) the Appalachian Mountains, (3) the South East, (4) Southern California, (5) Northern California and (6) the Rocky Mountains.

2.3.2 Validation Results

The validation statistics of three model performance evaluation methods (Constant, CAMP and RAMP) are shown in Table 2.1. These methods have increasing sophistication. The Constant method assumes that model performance is constant across $\mathcal{D}$, the CAMP method accounts for non-linear and non-homoscedastic model performance and the RAMP method accounts for non-linear, non-homoscedastic and non-homogeneous model performance shown in Table 2.1. The validation statistics are calculated using a corrected CMAQ value $\lambda_1(\boldsymbol{p})$ and associated error variance $\lambda_2(\boldsymbol{p})$ given by (Equ. 2-11, 2-12), (Equ. 2-13, 2-14), and (Equ. 2-8, 2-9) for the Constant, CAMP and RAMP methods, respectively.

Validation of $\lambda_1(\boldsymbol{p})$ is performed by comparing the $ME(\mathcal{D})$, $\sqrt{VE(\mathcal{D})}$, $MSE(\mathcal{D})$ and $r(\mathcal{D})$ performance statistics of the raw CMAQ estimate (the first column of Table 2.1) with $\lambda_1(\boldsymbol{p})$ for each of the

three performance evaluation methods (the last three columns of Table 2.1). The magnitude of $ME(\mathcal{D})$ drops from -1.05 $(\mu g/m^3)$ for CMAQ to 0.03 $(\mu g/m^3)$, 0.03 $(\mu g/m^3)$ and -0.02 $(\mu g/m^3)$ for the Constant, CAMP and RAMP methods, respectively. This was expected by design due to each method eliminating systematic errors across $\mathcal{D}$. The model performance evaluation methods differ in their abilities to reduce random errors, as demonstrated by the $\sqrt{VE(\mathcal{D})}$ statistic. The $\sqrt{VE(\mathcal{D})}$ statistic progressively reduces from 7.77 $(\mu g/m^3)$ for CMAQ to 7.18 $(\mu g/m^3)$, 6.58 $(\mu g/m^3)$ and 6.34 $(\mu g/m^3)$ for the Constant, CAMP and RAMP methods, respectively. This translates in a total error that is lower for RAMP ($MSE = 40.1$ $(\mu g/m^3)^2$) than for CAMP ($MSE = 43.3$ $(\mu g/m^3)^2$) and the Constant method ($MSE = 51.5(\mu g/m^3)^2$). This corresponds to a 22.1% reduction in MSE from the Constant to the RAMP method. This finding is further confirmed by the correlation between observed and $\lambda_1(\boldsymbol{p})$ values, which progressively increases from $r = 0.589$ for CMAQ to $r = 0.698$ for RAMP. These results demonstrate that $\lambda_1(\boldsymbol{p})$ calculated by the RAMP method is more accurate than the raw CMAQ output or the CMAQ corrected values obtained from the other model performance evaluation methods.

Validation of $\lambda_2(\boldsymbol{p})$ is performed by comparing the $VS(\mathcal{D})$, $RMSS(\mathcal{D})$ and $MR(\mathcal{D})$ performance statistics across the different model performance evaluation methods. The VS and RMSS are the variance and root mean squared error, respectively, of the Standardized error S, where $S = (\tilde{x}(\boldsymbol{p}) - ME(\mathcal{D}) - \hat{x}(\boldsymbol{p}))/\sqrt{VE(\mathcal{D})}$ for the Constant method and $S = (\lambda_1(\boldsymbol{p}) - \hat{x}(\boldsymbol{p}))/\sqrt{\lambda_2(\boldsymbol{p})}$ for the CAMP and RAMP methods. The standardized errors should ideally have a standard normal distribution, hence VS and RMSS should be close to 1. VS is 0.766 for the Constant method, 0.823 for the CAMP method and 1.05 for the RAMP method. Because VS is closest to 1 for the RAMP method, the RAMP $\lambda_2(\boldsymbol{p})$ is more accurate than the Constant or CAMP $\lambda_2(\boldsymbol{p})$. The VS for the Constant method and CAMP are less than one, meaning the Constant method and CAMP methods overestimate the CMAQ prediction error variance. This result is confirmed by the RMSS and is further quantified by the MR. The MR is the mean of the CMAQ prediction error standard deviations. The $MR(\mathcal{D})$ for RAMP indicates that the random error of CMAQ prediction has a standard deviation that is equal to 5.45 $\mu g/m^3$ on average across $\mathcal{D}$. The $MR(\mathcal{D})$ for the Constant method is 8.20 $\mu g/m^3$, indicating that the Constant method leads to a substantial overestimation of random errors by about 50% over RAMP estimates. The overestimation of random error

is attenuated with the CAMP method, which has an $MR(\mathcal{D})$ equal to $70.4\ \mu g/m^3$ corresponding to a 29% overestimation compared to the RAMP estimates.

Overall these validation results demonstrate that the RAMP method provides a $\lambda_1(\boldsymbol{p})$ value that better corrects systematic errors than other performance evaluation methods and provides a $\lambda_2(\boldsymbol{p})$ value that better estimates random errors compared to other model performance evaluation methods. We hypothesize that this is due to the RAMP method being better able to assess the spatial and temporal uncertainty of systematic and random errors compared with other model performance evaluation methods.

		CMAQ Corrected		
Statistic	**CMAQ**	Constant Correction	Non-linear/Non homoscedastic (CAMP) Correction	Non-linear/Non homoscedastic and Non-homogenous (RAMP) Correction
$ME(\mathcal{D})$ $(\mu g/m^3)$	-1.05	0.03	0.03	-0.02
$\sqrt{VE(\mathcal{D})}$ $(\mu g/m^3)$	7.77	7.18	6.58	6.34
$MSE(\mathcal{D})$ $(\mu g/m^3)^2$	61.5	51.5	43.3	40.1
$r(\mathcal{D})$ (unitless)	0.589	0.625	0.631	0.698
$VS(\mathcal{D})$ (unitless)	NA	0.766	0.823	1.05
$RMSS(\mathcal{D})$ (unitless)	NA	0.875	0.907	1.03
$MR(\mathcal{D})$ $(\mu g/m^3)$	NA	8.20	7.04	5.45

Table 2.1. Validation statistics. Statistics of the validation results of daily paired observed PM2.5 and $\lambda_1(\boldsymbol{p})$ and $\lambda_2(\boldsymbol{p})$ estimated from each of the three methods: the Constant method, CAMP and RAMP for 2001 across the continental United States. The CMAQ column are the statistics between the paired observed and CMAQ concentrations. VS is variance of the standardized errors, RMSS is square root of the mean squared standardized errors and MR is the mean of the square root of $\lambda_2(\boldsymbol{p})$.

2.3.3 Stochastic Simulation Results

The RAMP estimates of $\lambda_1(\boldsymbol{p})$ and $\lambda_2(\boldsymbol{p})$ obtained for 2001 were selected as the true mean and variance, respectively, of the observed values. The CMAQ modeled output was selected as the modeled value $\tilde{x}(\boldsymbol{p})$. We obtained a stochastic realization of $\hat{x}^*(\boldsymbol{p}) \sim N(\lambda_1(\boldsymbol{p}), \lambda_2(\boldsymbol{p}))$ for each observed space/time location. Paired $\tilde{x}(\boldsymbol{p})$ and $\hat{x}^*(\boldsymbol{p})$ values extracted for each observed space/time location. The Constant, CAMP and RAMP model performance evaluation methods are used to re-calculate $\lambda_1{}^*(\boldsymbol{p})$ and $\lambda_2{}^*(\boldsymbol{p})$ based *only* on $\tilde{x}(\boldsymbol{p})$ and $\hat{x}^*(\boldsymbol{p})$ for 2001. The re-calculated $\lambda_1{}^*(\boldsymbol{p})$ and $\lambda_2{}^*(\boldsymbol{p})$ are compared to the selected $\lambda_1(\boldsymbol{p})$ and $\lambda_2(\boldsymbol{p})$ for July 1, 2001. Detailed outputs of the results of this stochastic simulation are given in Appendix A and are summarized here.

The map of the true systematic error $\hat{x}(\boldsymbol{p}) - \lambda_1(\boldsymbol{p})$ for July 1, 2001 displays by design clear geographical trends identifying well defined regions where systematic error is large. The map of re-calculated systematic error $\hat{x}(\boldsymbol{p}) - \lambda_1{}^*(\boldsymbol{p})$ obtained using the Constant method is constant and is therefore unable to capture the spatial variability in systematic errors. The corresponding map obtained with the CAMP method is able to capture spatial variability occurring across the entire modeling domain, but unable to capture the regional and fine scale variability in systematic errors. However, the corresponding RAMP map captures spatial variability of systematic errors at a fine spatial scale. A correlation coefficient r was calculated between $\hat{x}(\boldsymbol{p}) - \lambda_1(\boldsymbol{p})$ and $\hat{x}(\boldsymbol{p}) - \lambda_1{}^*(\boldsymbol{p})$ for July 1, 2001. This correlation coefficient was 0%, 24.0% and 76.1% for the Constant, CAMP and RAMP methods, respectively. These results demonstrate that the RAMP method is better able to capture fine scale spatial variability of systematic errors.

Similar results were found when comparing the true $\lambda_2(\boldsymbol{p})$ with $\lambda_2{}^*(\boldsymbol{p})$ obtained for each model performance evaluation method, again for July 1, 2001. Qualitatively, the $\lambda_2{}^*(\boldsymbol{p})$ map obtained with the Constant method misrepresents the true $\lambda_2(\boldsymbol{p})$ map by failing to capture any of the spatial variability in random errors and overestimating the average random error. The $\lambda_2{}^*(\boldsymbol{p})$ map obtained with the CAMP method is a considerable improvement by reproducing variability at a long scale distance. However, visually, the CAMP method is unable to capture fine scale variability. The $\lambda_2{}^*(\boldsymbol{p})$ map obtained with the RAMP method provides a good visual reproduction of the true systematic error. These results are quantitatively supported by the correlation coefficient between $\lambda_2(\boldsymbol{p})$ and $\lambda_2{}^*(\boldsymbol{p})$ of 0%, 5.18% and 54.5% for the Constant, CAMP and RAMP methods, respectively.

These results demonstrate that in situations where there is regional variability in model performance, the RAMP method is better able to estimate the spatial variability of systematic errors compared to the Constant and CAMP methods. This implies the RAMP method should be considered for performance evaluation in future studies wherever it is plausible for model performance to vary spatially.

2.3.4 Evidence and Implications of Non-Linear and Non-Homoscedastic Model Performance

This work contributes novel evidence that the performance of air quality models is non-linear and non-homoscedastic. That is, λ_1 and λ_2 are a non-linear function of the modeled value $\tilde{x}_k$. This is seen through 1) the comparison of the Constant method and the CAMP method and 2) the stochastic

simulation results. The Constant method assumes that $\lambda_1 - \tilde{x}_k$ and λ_2 do not vary as a function of $\tilde{x}_k$. The CAMP method assumes that λ_1 and λ_2 are non-linear functions of $\tilde{x}_k$. The first evidence of non-linear and non-homoscedastic behavior comes from the validation results. The MSE reduces from $51.5\ (\mu g/m^3)^2$ for the Constant method to $43.3\ (\mu g/m^3)^2$ for the CAMP method, corresponding to a 16% reduction in MSE that demonstrates that model performance improves for a non-linear and non-homoscedastic model. In the stochastic simulation results, the Constant method is unable to capture the spatial variability in systematic and random errors whereas the CAMP method is able to capture domain-wide variability of these errors. Furthermore, both the validation and stochastic simulation results indicate that the Constant method significantly over predicts random errors compared to the CAMP method. Finally, the non-homoscedastic behavior in model performance is evidenced by maps of $\lambda_2(\tilde{x}_k; \mathcal{R}(\boldsymbol{p}))$ for different values fixed $\tilde{x}_k$ values (see Appendix A), showing that at a given region $\mathcal{R}(\boldsymbol{p})$, the error variance changes substantially from one value of $\tilde{x}_k$ to another.

From these results, one should be cautious when using linear and homoscedastic model performance evaluation methods to explore the spatial variability of model performance. This is the usual practice of current approaches in which models can be expressed as either $\tilde{X}(\boldsymbol{s}) = \beta_0(\boldsymbol{s}) + \beta_1(\boldsymbol{s})\hat{X}(\boldsymbol{s}) + \varepsilon(\boldsymbol{s})$ (Fuentes and Raftery, 2005) or $\hat{X}(\boldsymbol{s}, t) = \beta_0(\boldsymbol{s}, t) + \beta_1(\boldsymbol{s}, t)\tilde{x}(\boldsymbol{s}, t) + \varepsilon(\boldsymbol{s}, t)$ (Berrocal et al., 2010b). In both cases the relationship is linear and homoscedastic when assuming a constant error variance of the noise term, i.e. $\varepsilon(\boldsymbol{s}, t) \sim N(0, \sigma_\epsilon^2)$. The linearity of these models has advantages in terms of implementation, but they fail to account for the non-linear and non-homoscedastic nature of model performance. This may undermine their capacity to fully capture spatial variability in model performance. Furthermore, these methods may overestimate the error variance. By contrast the RAMP method provides a novel alternative that fully captures the space/time variability of non-linear non-homoscedastic model performance and, as a result, provides a novel description of the spatial patterns in systematic and random errors across the spatiotemporal continuum.

2.3.5 Spatial Patterns of Systematic and Random Errors

To better understand the magnitude of the systematic errors $ME^2(\boldsymbol{p})$ (Fig. 2.2a), we also show a map of $ME(\boldsymbol{p})$ (Fig. 2.3), which differentiates areas where daily PM2.5 concentrations are over predicted ($ME > 0$) versus under predicted ($ME < 0$). The picture depicted by Fig. 2.2 and Fig. 2.3 is in line with

what we know about CMAQ. That is, CMAQ generally struggles with estimating high values of PM2.5 (Yu et al., 2012, 2008). Areas shown with negative $ME(\boldsymbol{p})$ values in Fig. 2.3 (i.e. where PM2.5 is under predicted) coincide with areas shown to have high $\lambda_1(\boldsymbol{p})$ values in Fig. 2.2d (i.e. where PM2.5 levels are high).

The RAMP analysis provides a map of $ME^2(\boldsymbol{p})$ across the continuous space/time domain (as opposed to being restricted to only monitoring stations). This makes it possible to clearly delineate and identify specific regions with high $ME^2(\boldsymbol{p})$ values and quantify their geographical extent. To illustrate this capability, we identified in Fig. 2.2a six regions (labeled 1-6) defined as having relatively high systematic error (i.e. $ME^2(\boldsymbol{p}) \geq 17.4\ (\mu g/m^3)^2$). The areas of high systematic error are quantified as follows: (1) the Great Lakes ($15{,}552\ km^2$), (2) the Appalachian Mountains ($116{,}640\ km^2$), (3) the South East ($38{,}880\ km^2$), (4) Southern California ($73{,}872\ km^2$), (5) Northern California ($75{,}168\ km^2$) and (6) the Rocky Mountains ($290{,}304\ km^2$).

Some of the regions identified for their high systematic errors are corroborated in the literature. The over prediction in region 1 (the Great Lakes) is in line with an overestimation of residential wood burning in the region reported in the National Emissions Inventory (NEI) (Appel et al., 2008). Region 3 (South East) includes Atlanta where PM2.5 is over estimated and an area to its South where PM2.5 is under estimated. CMAQ is known to under predict PM in the South East. Some of this under prediction may be associated with highly uncertain SOA chemistry, particularly including chemistry from biogenic emissions (Chan et al., 2010; Morris et al., 2006). Likewise high systematic error in the mountain regions 2 and 6 (Appalachia Mountains and the Rockies) can be associated with the known difficulties in modeling air quality accurately on and near mountain ranges (Steyn et al., 2013). The causes of high systematic error identified by RAMP in other regions may not yet be documented well. For example, the identification of Northern California (region 4) and Southern California (region 5) may serve as a trigger for further investigation into the constituents and chemical pathways of PM2.5 in these regions (Motallebi et al., 2003; USEPA, 2001) so as to investigate causes that may lead to systematic errors in these areas. To our knowledge this is the first work in the model performance literature to delineate these regions and quantify their geographic extent.

The map of $VE(\boldsymbol{p})$ in Fig. 2.2b delineates areas with high random errors. It is interesting to note that areas of high systematic errors are always fully contained within areas of high random error as seen by comparing Fig. 2.2a and Fig. 2.2b. To our knowledge these are the first maps delineating regions of high random errors and finding general collocation with (and of about twice the magnitude of) systematic errors. If both systematic and random errors are caused by similar processes, then presumably reducing systematic errors could have the added benefit of also addressing collocated random errors.

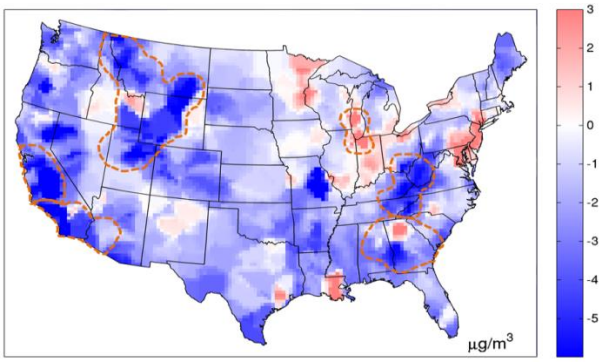


Figure 2.3. Map of RAMP mean error. Daily PM2.5 across the continental United States on 07/01/2001 displaying $ME(\boldsymbol{p})$ in $\mu g/m^3$. The 6 regions of high random error delineated in Figure 2.2b are delineated in the dashed orange line.

2.4 Conclusions

This work introduces a spatiotemporal approach that can estimate and distinguish systematic from random error of predictions made by regulatory air quality models at any location of interest. The estimation of systematic and random error is created in a manner that does not assume that the relationship between observed and modeled values is linear or homoscedastic, and estimation of errors is performed in a manner that is regionalized. By estimating errors across a continuous geographical domain for a given day of interest, this approach permits the production of maps delineating areas of high errors. These maps are useful to 1) assess model performance by quantifying systematic and random error at a fine spatial resolution across the entire space/time domain where monitoring does not exist and 2) do a model correction of systematic errors of the CMAQv4.5 estimates of PM2.5 for 2001 for individual grids. Future works include doing a data fusion of RAMP model corrected values and observations using the Bayesian Maximum Entropy (BME) method (Akita et al., 2012; Allshouse et al., 2009; de Nazelle et al., 2010; Reyes and Serre, 2014), and updating the RAMP analysis for other years. This future work will be critical for ongoing epidemiologic studies analyzing the effect of air pollution on brain aging for women

in the Women's Health Initiative-Memory Study who were exposed to air pollution between 1999 and 2006. The application of RAMP on CMAQv4.5 demonstrated that the RAMP analysis was able to successfully identify known regions of errors of this version of CMAQ. This work provides a model correction for 2001 based on the most recent of CMAQ for this year and provides a useful baseline against which future versions can be compared to explore changes in systematic and random errors.

CHAPTER 3: INCORPORATING REGIONALIZED AIR QUALITY MODEL PERFORMANCE EVALUATION IN A NATIONWDIE GEOSTATISTICAL DATA INTEGRATION OF DAILY PM2.5[2]

3.1 Introduction

The Clean Air Act of 1990 established regulatory standards for air pollutants in the United States (Boldo et al., 2006; Pope et al., 2009). Currently there are six "criteria air pollutants" regulated by the US Environmental Protection Agency (EPA) due to their detriment to human health and the environment, including Particulate Matter ≤ 2.5 micrometers (PM2.5). PM2.5 is associated with a host of adverse health outcomes including increased risk of cardiovascular and respiratory disease and mortality (Beelen et al., 2007; Krewski et al., 2009; Pope et al., 2004). To ensure PM2.5 does not exceed the regulatory standard, a nationwide monitoring network has been established that measures PM2.5 concentrations on a regular basis. However, despite the significant number of regulatory stations, there are large monitoring gaps that exist in many parts of the country. These can become problematic in both epidemiologic studies when attempting to predict exposures and in regulatory settings when establishing attainment. From an epidemiologic standpoint, modeled data such as Chemical Transport Models (CTMs) and satellite data can be a means to fill in the gaps left from observed data (Brauer et al., 2015; Tang et al., 2016; van Donkelaar et al., 2015). CTMs (e.g. Community Multiscale Air Quality (CMAQ) model) are deterministic and combine emissions, meteorology and chemistry to predict ambient pollution concentration across the entirety of a gridded modeling domain (Appel et al., 2013b; Foley et al., 2015a, 2015b).

In air quality modeling there has been a recent surge in data fusion methods. These methods combine different air quality sources together, in particular, observed data with gridded modeled data (Berrocal et al., 2010a; Crooks and Isakov, 2013; Fuentes and Raftery, 2005). Many studies that combine data sources focus on epidemiologic studies with the goal of having accurate exposure prediction that reduce misclassification (Beckerman et al., 2013). Observed data are considered highly accurate and

[2] This chapter was submitted as an article to the journal Environmental Science and Technology. Reyes, Jeanette M., Xu,Yadong, Vizuete, William, Serre, L. Marc. Incorporating Regionalized Air Quality Model Performance evaluation in a nationwide geostatistical data integration of daily PM2.5.

produce low prediction errors but are potentially sparsely measured. Corresponding prediction methods which may only use observed data (e.g. kriging) have lower spatial refinement. CMAQ data are considered less accurate than observed data but have excellent spatial and temporal coverage along with a higher level of spatial refinement.

Previous popular data fusion methods include the Downscaler method and Bayesian Melding (Berrocal et al., 2012, 2010a, 2010b). The Downscaler method takes output of a CTM model and uses it as an independent variable in the Downscaler regression model. The Bayesian Melding approach characterizes the full uncertainty in observed, modeled and the true underlying process of the air pollutant of interest. However, this method is computationally intensive and has only been applied in a spatial setting. Inherent in both of these methods are assumptions of linearity and homoscedastic behavior in the model.

This work proposes the Regionalized Air quality Model Performance (RAMP) incorporation into the Bayesian Maximum Entropy (BME) geostatistical framework (Reyes et al., 2016; Xu et al., 2016). BME is an extension of linear kriging and has the flexibility of incorporating multiple data sources together. The RAMP BME data fusion method is an extension of the Constant Air quality Model Performance (CAMP) method introduced by de Nazelle et al. (de Nazelle et al., 2010). The novel RAMP approach to model performance evaluation is able to quantify model performance metrics across the entirety of a domain and fully characterizes model performance at each space/time grid location over the fully spectrum of given modeled values. This proposed data fusion method can capture the accuracy of observed data with the spatial refinement of CMAQ data without assumptions of linearity and homoscedastic behavior.

A demonstration of the BME data fusion method developed in this work combines CMAQ modeled data with observed data to predict daily PM2.5 mass across the continental US for 2001. The BME method takes advantage of low prediction error associated with observed data along with the high spatial refinement associated with CMAQ modeled data. Results are then compared to a frequentist version of the Downscaler method. These results improve the spatial refinement of PM2.5 predictions and offer a more realistic exposure profile which can be used in epidemiologic analysis to reduce

misclassification and more clearly uncover the true association between participants' air pollution exposures and health outcomes.

3.2 Materials and Methods

3.2.1 Observed and modeled data

The daily observed PM2.5 concentration for each monitoring site/day during 2000-2002 were constructed based on raw monitoring data from monitoring stations measuring either hourly or daily PM2.5 concentrations obtained from the EPA's Air Quality Systems data base (US EPA, n.d.). Daily concentrations for PM2.5 were also constructed from hourly modeled data averaged to daily for years 2001 and 2002 using CMAQv4.5 across the contiguous United States on a 36 km grid. For more detailed information regarding the aggregation and pairing process of observed and modeled data see Appendix A.

3.2.2 BME estimation methodology

BME is a mathematically rigorous geostatistical space/time framework originally developed in a geostatistical setting by Christakos (Christakos, 2000; Christakos et al., 2001). BME can incorporate information from multiple data sources and is implemented using the *BMElib* suite of functions in MATLAB™. The buttress of BME has been detailed in other works and can be summarized as performing the following steps: 1) gathering the general knowledge base (G-KB) and site-specific knowledge base (S-KB) characterizing the Space/Time Random Field (S/TRF) $X(\boldsymbol{p})$ representing a process at space/time coordinate $\boldsymbol{p} = (\boldsymbol{s}, t)$ where $\boldsymbol{s}$ is the spatial coordinate and t is time, 2) using the *Maximum Entropy* principle of information theory to process the G-KB in the form of a prior Probability Distribution Function (PDF) f_G, 3) integrating S-KB in the form of a PDF f_S with and without measurement error using an epistemic *Bayesian* conditionalization rule on data to create a posterior PDF f_K and 4) creating space/time estimates based on the analysis. Typically the G-KB consists of the expected value and covariance of $X(\boldsymbol{p})$ denoted as $G = \{m_X(\boldsymbol{p}), c_X(\boldsymbol{p}, \boldsymbol{p}')\}$ and the S-KB consists of hard data (data measured without error) and soft data (data measured with error) denoted as $S = \{\boldsymbol{x}_h, f_s(\boldsymbol{x}_s)\}$. The BME posterior PDF f_K describing the process x_k at an estimation point of interest $\boldsymbol{p}_k$ is given by the BME equation

$$f_K(x_k) = A^{-1} \int d\boldsymbol{x}\, f_S(\boldsymbol{x}_s) f_G(\boldsymbol{x}), \qquad \text{(Equ. 3-1)}$$

where $\boldsymbol{x} = (x_k, \boldsymbol{x}_h, \boldsymbol{x}_s)$ is a realization of $\boldsymbol{X}$ at points $\boldsymbol{p} = (\boldsymbol{p}_k, \boldsymbol{p}_h, \boldsymbol{p}_s)$ and A is a normalization constant.

In this study we use an S/TRF to describe the variability of daily PM2.5 mass across the US in 2001. Our notation for an S/TRF will consist of denoting a single random variable Z in capital letters, its realization, z, in lower case and vectors and matrices in bold faces (e.g. $\boldsymbol{Z} = [Z_1, \dots, Z_n]^T$ and $\boldsymbol{z} = [z_1, \dots, z_n]^T$). Let $Z(\boldsymbol{p}) = Z(\boldsymbol{s}, t)$ be a Space/Time Random Field (S/TRF) representing daily PM2.5. The BME data fusion method incorporates both modeled and observed data. Let $\hat{Z}(\boldsymbol{p})$ be the random variable representing the observed concentration, $\hat{z}(\boldsymbol{p})$ be its known (i.e. observed) value and $\tilde{z}(\boldsymbol{p})$ be the CMAQ modeled value at location $\boldsymbol{p}$.

We define the transformation of observed PM2.5 data $\boldsymbol{z}_h$ observed at locations $\boldsymbol{p}_h$ as

$$\boldsymbol{x}_h = \boldsymbol{z}_h - o_Z(\boldsymbol{p}_h) \quad \text{(Equ. 3-2)}$$

where $o_Z(\boldsymbol{p})$ may be any deterministic offset that can be mathematically calculated without error as a function of the space/time coordinate $\boldsymbol{p}$. We then define $X(\boldsymbol{p})$ as a homogeneous/stationary S/TRF representing the variability and uncertainty associated with the transformed data $\boldsymbol{x}_h$, and we let $Z(\boldsymbol{p}) = X(\boldsymbol{p}) + o_Z(\boldsymbol{p})$ be the S/TRF representing PM2.5. We can then calculate $\ddot{z}_k$, the predicted daily PM2.5 at unmonitored location $\boldsymbol{p}_k$, by obtaining the BME estimate $\ddot{x}_k$ for the transformed S/TRF $X(\boldsymbol{p})$ at the estimation point $\boldsymbol{p}_k$ and adding $o_z(\boldsymbol{p}_k)$, the offset calculated at $\boldsymbol{p}_k$. In this work we calculate the offset using a space/time composite kernel smoothing of the data (Lee et al., 2012). The covariance model for the homogeneous/stationary S/TRF $X(\boldsymbol{p})$ is developed from the experimental covariance of the transformed data $\boldsymbol{x}_h = \boldsymbol{z}_h - o_Z(\boldsymbol{p}_h)$. The offset and the corresponding covariance in this work are chosen as having the best combination of low variance and the high autocorrelation. For detailed information regarding the calculation of the offset function and covariance model, see Appendix B.

3.2.3 Regionalized Air quality Model Performance (RAMP) soft data construction

Like stated in the BME estimation methodology section above, part of the site specific knowledge of BME comes from soft data. Soft data are constructed with paired modeled and observed data through the Regionalized Air quality Model Performance (RAMP) methodology and has been documented in previous works and is detailed in Chapter 2 (Reyes et al., 2016; Xu et al., 2016). The resulting soft data are error corrected and can be thought of as a measure of model performance. Soft data at an estimation location can be represented as $f_S(\boldsymbol{x_s})$. Soft data are created for every daily PM2.5 modeled value. At a

given estimation location, several hard and soft data go into the corresponding prediction. Thus, we can further expand $f_S(\boldsymbol{x_s})$ to the following expression:

$$f_S(\boldsymbol{x_s}) = \prod_i^{n_m} f(x_i|\tilde{x}_i, \boldsymbol{p}_i), \quad \text{(Equ. 3-3)}$$

where n_m is the number of CMAQ grids used in the calculation of the soft data, $f(x_i|\tilde{x}_i, \boldsymbol{p}_i)$ is the soft data PDF of PM2.5 concentration at the modeled data location $\boldsymbol{p}_i$ and $\tilde{x}_i$ is the modeled value of PM2.5 concentration after removing the offset.

For the sake of clarity of the soft data development, consider the non-transformed random variable Z. The soft PDF $f(z_i|\tilde{z}_i, \boldsymbol{p}_i)$ is Gaussian distributed with mean λ_1 and variance λ_2, denoted:

$$f(z_i|\tilde{z}_i, \boldsymbol{p}_i) = \Phi(z_i; \lambda_1(\boldsymbol{p}), \lambda_2(\boldsymbol{p})) \quad \text{(Equ. 3-4)}$$

The parameters λ_1 and λ_2 are dependent on the modeled value concentration around $\boldsymbol{p}$. The parameters λ_1 and λ_2 are estimated using the equations given below:(de Nazelle et al., 2010)

$$\lambda_1(\tilde{z}_k; \mathcal{R}(\boldsymbol{p})) = M\left[\hat{Z}|\tilde{z}_k; \mathcal{R}(\boldsymbol{p})\right] \approx \frac{1}{n(\tilde{z}_k; \mathcal{R}(\boldsymbol{p}))} \sum \hat{z}_i \quad \text{(Equ. 3-5)}$$

$$\lambda_2(\tilde{z}_k, \mathcal{R}(\boldsymbol{p})) = V\left[\hat{Z}|\tilde{z}_k; \mathcal{R}(\boldsymbol{p})\right] \approx \frac{1}{n(\tilde{z}_k; \mathcal{R}(\boldsymbol{p}))-1} \sum \left(\hat{z}_i - \lambda_1(\tilde{z}_k; \mathcal{R}(\boldsymbol{p}))\right)^2 \quad \text{(Equ. 3-6)}$$

where $n(\tilde{z}_k; \mathcal{R}(\boldsymbol{p}))$ is the number of paired modeled and observed points within region $\mathcal{R}(\boldsymbol{p})$ associated with space/time location $\boldsymbol{p}$ which are from the 3 closest monitoring stations within 180 days around $\tilde{z}_k$. Stated simply, λ_1 is estimated through pooling all paired modeled and observed data together in region $\mathcal{R}(\boldsymbol{p})$ associated with space/time location $\boldsymbol{p}$ and close to the modeled value $\tilde{z}_k$. The mean of all the near-by observed data are taken to calculate $\lambda_1(\tilde{z}_k; \mathcal{R}(\boldsymbol{p}))$. Similarly, the variance of all the near-by observed data are taken to calculate $\lambda_2(\tilde{z}_k, \mathcal{R}(\boldsymbol{p}))$. Modeled and observed data are paired if an observed datum lies within a given grid.

Although $\lambda_1(\tilde{z}_k; \mathcal{R}(\boldsymbol{p}))$ and $\lambda_2(\tilde{z}_k, \mathcal{R}(\boldsymbol{p}))$ can be calculated for any arbitrary $\tilde{z}_k$, in this work we set $\tilde{z}_k = \tilde{z}(\boldsymbol{p})$, where $\tilde{z}(\boldsymbol{p})$ is the CMAQ modeled value at $\boldsymbol{p}$. By replacing $\tilde{z}_k$ with $\tilde{z}(\boldsymbol{p})$ in Equ. 3-5 and Equ. 3-6, we define $\lambda_1(\tilde{z}(\boldsymbol{p}), \mathcal{R}(\boldsymbol{p})) = \lambda_1(\boldsymbol{p})$ and $\lambda_2(\tilde{z}(\boldsymbol{p}), \mathcal{R}(\boldsymbol{p})) = \lambda_2(\boldsymbol{p})$. For the sake of shorthand in this work, we further define $\lambda_1 = \lambda_1(\boldsymbol{p})$ and $\lambda_2 = \lambda_2(\boldsymbol{p})$.

3.2.4 Leave One Out Cross Validation (LOOCV) accuracy analysis

To assess the prediction accuracy of the BME data fusion method, a LOOCV accuracy analysis is performed. For each monitoring station, all data are removed one at a time and predicted through RAMP

BME for each removed datum (without recalculating the offset or the covariance model) using all the data from the remaining monitoring stations and soft data. This is repeated again for each monitoring station. However, instead of only removing data from a single monitoring station, all observed data within a given radius from the monitoring station (100 km, 200 km, 300 km, … , 900 km) are removed.

The difference between each prediction value $\ddot{z}_i$ and observed value $\hat{z}_i$ is the prediction error, $e_i = \ddot{z}_i - \hat{z}_i$. The prediction accuracy is quantified based on statistics of prediction errors, which consist of the Mean Squared Error (MSE, $(\mu g/m^3)^2$), Mean Error (ME, $\mu g/m^3$) and the Pearson's correlation coefficient (r, unitless) between observed and predicted values. BME data fusion predictions are then compared to kriging (i.e. predictions created only using observed data).

Because λ_1 and λ_2 are written in terms of an expected value and variance, respectively, performance metrics can be written in terms of these quantities. Namely, mean error and variance of error for an arbitrary $\boldsymbol{p}$ (see Chapter 2).

$$\lambda_1(\boldsymbol{p}) = \tilde{z}(\boldsymbol{p}) - ME(\boldsymbol{p}) \quad \text{(Equ. 3-7)}$$

$$\lambda_2(\boldsymbol{p}) = VE(\boldsymbol{p}) \quad \text{(Equ. 3-8)}$$

Using the relation equating mean squared error to mean error and variance of error (i.e. $MSE = ME^2 + VE$), LOOCV results can be stratified by the scaled mean error statistic,

$$SME(\boldsymbol{p}) = ME^2(\boldsymbol{p})/MSE(\boldsymbol{p}). \quad \text{(Equ. 3-9)}$$

3.2.5 Comparison to the frequentist Downscaler method

This work is compared to a frequentist implementation of the space/time Downscaler method (Berrocal et al., 2010a). A full description of this method can be found in Appendix B. In short, $Z(\boldsymbol{p}) \sim N(\mu_Z, c_Z)$ where $Z(\boldsymbol{p})$ is the pollutant of interest. $\hat{Z}(\boldsymbol{p})$ is defined as:

$$\hat{Z}(\boldsymbol{p}) = \beta_{0t} + \beta_0(\boldsymbol{p}) + \beta_{1t}\tilde{z}(\boldsymbol{p}) + \beta_1(\boldsymbol{p})\tilde{z}(\boldsymbol{p}) + \epsilon(\boldsymbol{p}), \quad \text{(Equ. 3-10)}$$

where β_{0t} is the constant additive bias, $\beta_0(\boldsymbol{p})$ is the additive bias that changes as a function of $\boldsymbol{p}$, β_{1t} is the constant multiplicative bias, $\beta_1(\boldsymbol{p})$ is the multiplicative bias that changes as a function of $\boldsymbol{p}$, $\tilde{z}(\boldsymbol{p})$ is the modeled value concentration of $\boldsymbol{p}$ and $\epsilon(\boldsymbol{p})$ is random noise. In the space/time application of the Downscaler, the additive and multiplicative biases can be treated independently across time or they can be treated in a more recursive manner. The results given in this work use the space/time Downscaler in

which both the additive and multiplicative bias are treated independently across time. The results given below are a frequentist implementation of this method (i.e. all parameters are estimated empirically).

3.3 Results and Discussion

3.3.1 PM2.5 data fusion demonstration of RAMP BME

A demonstration of the BME data fusion method was completed by combining daily observed PM2.5 with daily systematic error corrected CMAQ predictions were generated using CMAQv4.5 on a 36 km grid across the continental United States for 2001. BME predictions were created at the centroid of each CMAQ grid. Results of the BME data fusion method and kriging are displayed across the continental US on 07/01/2001 (Fig. 3.1). There is a clear spatial pattern for BME across the day as shown through the posterior mean (Fig. 3.1b). There is an area of high daily PM2.5 predictions (over $20\ \mu g/m^3$) in Southern California, which is an area known to have high PM2.5 concentrations (Fann et al., 2012). There is a distinctive band of high concentrations, also around $20\ \mu g/m^3$, in the Eastern US extending from the New England area to West Virginia ending around Illinois. Areas of low concentrations (between $0-4\ \mu g/m^3$) can mostly be found in the US states bordering Canada including Montana, North Dakota and Minnesota, which are states known to have relatively low concentrations of daily PM2.5 (Fann et al., 2012). For comparison to the BME data fusion method, kriging predictions are created for the same locations across the US on 07/01/2001 using only observed daily PM2.5 data (Fig. 3.1a,c). Generally speaking, the overarching spatial patterns of the kriging map are similar to those of BME (Fig. 3.1a). Kriging mean predictions show a PM2.5 plume encompassing a larger area over Southern California compared to BME. Likewise, the kriging map depicts the entirety of New England as having large PM2.5 concentrations. The pattern of relatively high PM2.5 concentration continues across West Virginia and Illinois, albeit at lower concentrations than are predicted for BME. For kriging, areas of the country showing the lowest concentration are the same border states. However, the lowest concentrations of PM2.5 for kriging are between $4-8\ \mu g/m^3$, while low concentrations for BME are between $1-5\ \mu g/m^3$.

The uncertainty with both the BME and kriging mean estimates is quantified by their corresponding standard deviation of estimation errors (Fig. 3.1c,d). As expected, the kriging standard deviations drop close to zero when predictions are made near observed data (Cao et al., 2014), but they reach approximately $5.5\ \mu g/m^3$ across large proportions of the Western US from Kansas to Montana,

around areas that are far from observed data. This is in stark contrast with the BME standard deviation map, which displays substantially lower standard deviations in the same areas. By design the BME framework benefits from information provided by observations as well as CMAQ data, where the latter covers the entirety of the mapping domain. It is the addition of these CMAQ data that are responsible for the sizable decrease in the standard deviation in areas suffering from sparse air quality monitoring. As a result, the average estimation error standard deviation across the US drops substantially from kriging to BME. The average standard deviation on 07/01/2001 is $5.6\ \mu g/m^3$ for kriging while that value drops to $2.0\ \mu g/m^3$ for BME, indicating a more than two fold decrease in mapping uncertainty across the US.

We are also able to visualize the differences between the BME mean and the kriging mean estimates across the US (Fig. 3.1e). The difference calculated as the BME mean minus the kriging mean is mostly positive in the Eastern US with exceptions seen from New York to Maine. In the Western US the difference is mostly negative with exceptions seen in parts of Northern California and Utah. Lastly, we are able to visualize the parameters λ_1 and λ_2 that go into the BME data fusion method (Fig. 3.1f). Areas shown in Fig. 3.1f (i.e. $\left(\lambda_1 - E_{Krig}[\boldsymbol{Z}]\right)/\sqrt{\lambda_2}$) are mostly negative.

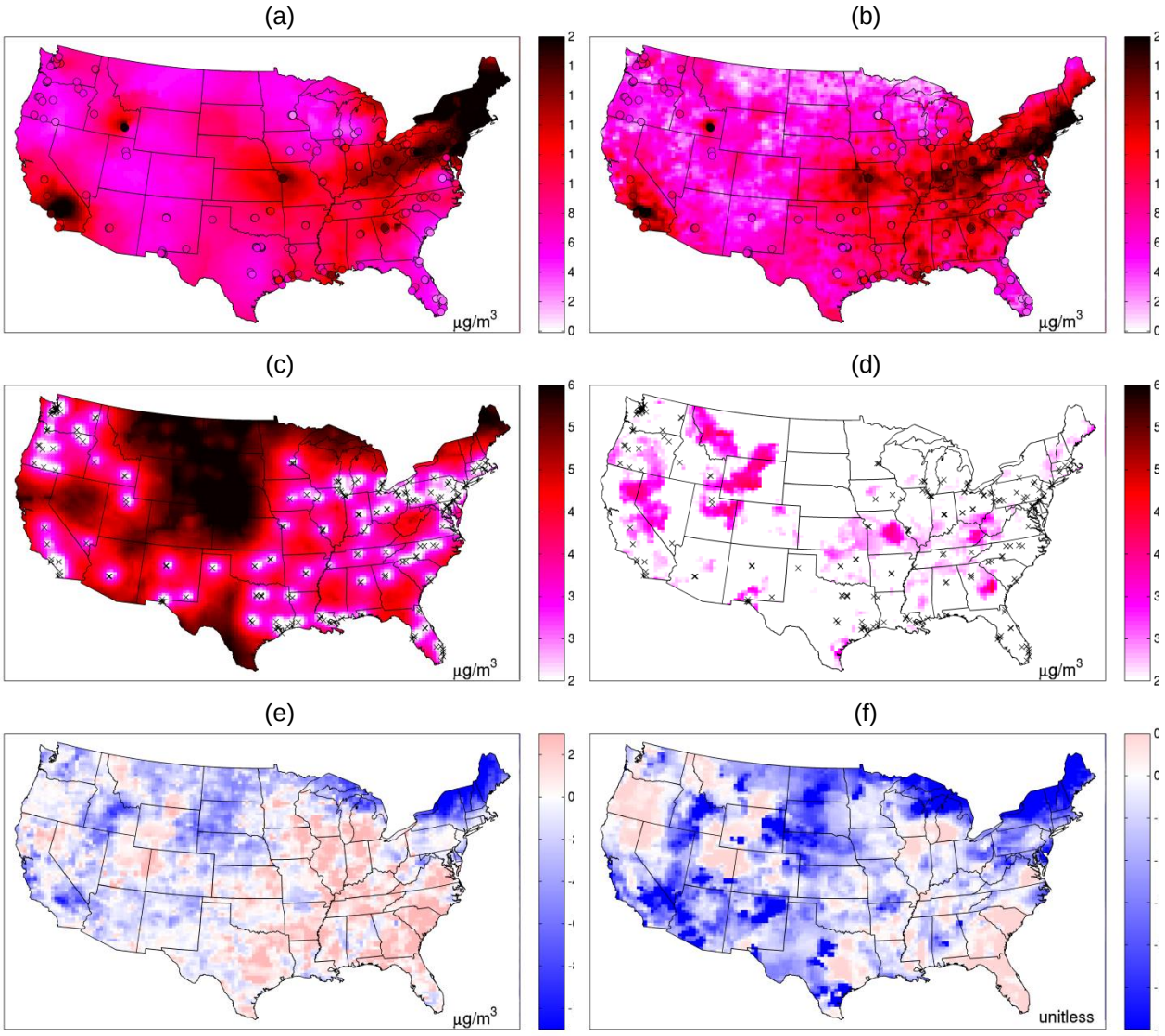


Figure 3.1. Map of kriging and BME mean and variance. Daily PM2.5 across the continental United States on 07/01/2001 displaying the (a) kriging mean estimate ($E_{Krig}[\boldsymbol{Z}]$), (b) BME mean estimate ($E_{BME}[\boldsymbol{Z}]$), (c) kriging standard deviation, (d) BME standard deviation, (e) $E_{BME}[\boldsymbol{Z}] - E_{Krig}[\boldsymbol{Z}]$ and (f) $(\lambda_1 - E_{Krig}[\boldsymbol{Z}])/\sqrt{\lambda_2}$. Plots. (a)-(e) are in $\mu g/m^3$ and (f) is unitless.

3.3.2 Validation results

An LOOCV is performed for the BME data fusion method and the kriging method on daily observed PM2.5 for 2001 for 10 different cross validation radii from 0 km to 900 km in increments of 100 km. The Percent Change in Mean Square Error (PCMSE) from kriging to BME is calculated for various cross validation radii (Fig. 3.2a,b). Negative PCMSE values indicate that BME has a lower MSE than, and therefore outperforms, kriging. The PCMSE ranges from -2.0% to -33% across cross validation radii as seen in the baseline in Fig. 3.2a, demonstrating that BME outperforms kriging across all cross validation

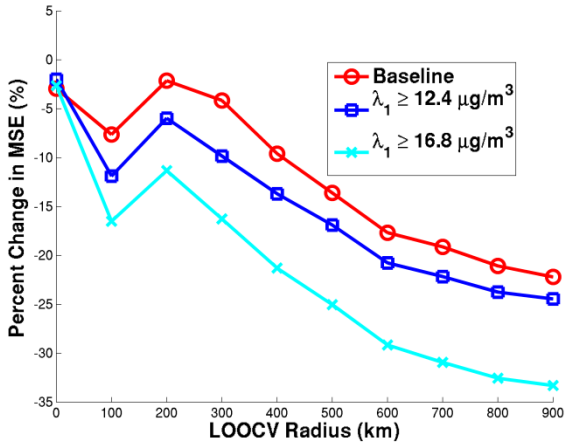
radii, and that this outperformance enhances significantly with increasing cross validation radii. As increasing distance between the prediction location and the nearest observed data increases, the kriging method suffers a significant increase in MSE, which is tampered for BME due to the information contributed by the CTM data.

Because all observations can be paired with a corresponding CMAQ value, all observed data can be paired with the corresponding error corrected CMAQ value λ_1 characterizing the expected PM2.5 concentration at that location. Hence the PCMSE can further be explored as a function of different selection criteria for λ_1. The criteria used in Fig. 3.2a correspond to the baseline case (all observations are included), $\lambda_1 \geq 12.4\,\mu g/m^3$ and $\lambda_1 \geq 16.8\,\mu g/m^3$. The PCMSE becomes more negative as λ_1 increases, and this becomes even more pronounced for larger cross validation radii.

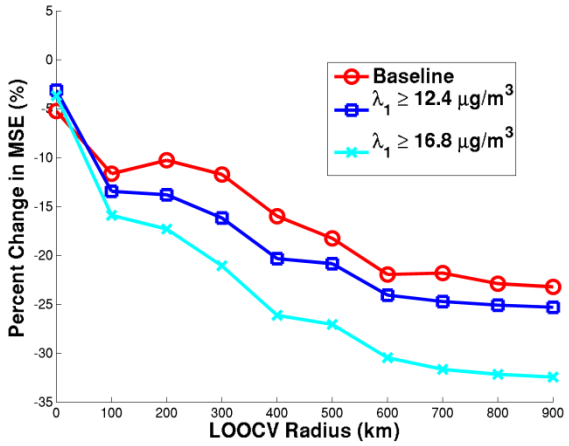
Likewise the CMAQ value collocated with each observation can be characterized by the model performance parameter $SME(\boldsymbol{p})$ at the CMAQ grid coordinate $\boldsymbol{p}$, which quantifies the proportion of systematic to total error for the CMAQ prediction.(Reyes et al., 2016) When the PCMSE is calculated using only observations for which the collocated CMAQ values are such that $SME(\boldsymbol{p}) \geq 22\%$, the reduction in MSE from kriging to BME is even more pronounced (Fig. 3.2b). The PCMSE ranges from -3.1% to -32% depending on λ_1 and cross validation radii. The PCMSE falls more quickly for increasing radii as the $SME(\boldsymbol{p}) \geq 22\%$ criteria is added, especially for radii 200-400 km. For the baseline case, previous to the application of the proportional systematic error criteria, the PCMSE is -2.1% for 200 km, -4.2% for 300 km and -9.6% for 400 km. After the proportional systematic error criterion is applied the PCMSE is -10% for 200 km, -12% for 300 km and -16% for 400 km.

Within each cross validation radii, the increasing exclusion criteria makes the PCMSE more pronounced (Table 3.1). Within in a 0 km cross validation radius, the PCMSE ranges from -2.9% to -3.7%. At 400 km, the PCMSE ranges from -9.6% to -26.1%. At 800 km, the PCMSE ranges from -19.1% to -31.6%.

(a)



(b)



(c)

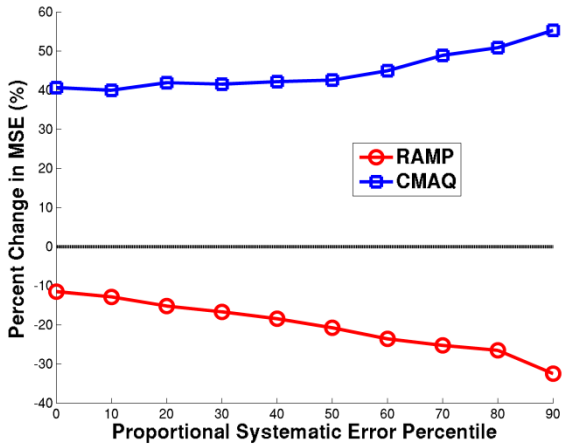


Figure 3.2. Cross validation comparing kriging with BME and the frequentist Downscaler. (a) Percent change in MSE from Kriging to BME calculated for $SME(\boldsymbol{p}) \geq 0\%$ with increasing LOOCV radii and increasing λ_1. A negative percent change indicates that BME has lower error. (b) Percent change in MSE from Kriging to BME calculated for $SME(\boldsymbol{p}) \geq 22\%$ with increasing LOOCV radii and increasing λ_1. (c) Percent change in MSE from the frequentist Downscaler to both BME and CMAQ calculated for

increasing percentiles of $SME(\boldsymbol{p})$ for the 200 km LOOCV radius. In (a) Baseline comprises 174,531 (100%) points, $\lambda_1 \geq 12.4$ comprises $69{,}814$ (40%) points, $\lambda_1 \geq 16.8$ comprises $34{,}906$ (20%) points. In (b) Baseline comprises $52{,}360$ (30%) points, $\lambda_1 \geq 12.4$ comprises $21{,}636$ (12%) of points, $\lambda_1 \geq 16.8$ comprises $11{,}386$ (6.5%) points.

3.3.3 Non-homogenous behavior of BME data fusion

The model performance of CMAQ has been shown to be non-linear and non-homoscedastic with respect to modeled value, and non-homogenous across space/time (Reyes et al., 2016). The RAMP method fully captures the non-linear, non-homoscedastic and non-homogenous behavior of model performance through its $\lambda_1(\boldsymbol{p})$ and $\lambda_2(\boldsymbol{p})$ parameters. The $\lambda_1(\boldsymbol{p})$ parameter is calculated through paired observed and modeled data, while $E_{Krig}[\boldsymbol{Z}]$ is the kriging prediction based on observations alone. Therefore, the difference $\lambda_1(\boldsymbol{p}) - E_{Krig}[\boldsymbol{Z}]$ is the expected correction when comparing the kriging estimate based only on observations, to an estimate based on data fusion incorporating both observations and CMAQ data. The parameter $\lambda_2(\boldsymbol{p})$ characterizes the uncertainty associated with the CMAQ data. Looking at the standardized metric $\frac{\lambda_1 - E_{Krig}[\boldsymbol{Z}]}{\sqrt{\lambda_2}}$ is a means to see apriori regions of the country in which the data fusion method will be most influential (Fig. 3.1f). The locations in which that quantity is largest in magnitude correspond to where the data fusion predictions differ the most compared to kriging predictions. We found that this is indeed the case by comparing Fig. 3.1e with Fig. 3.1f. This demonstrates that the BME data fusion properly incorporated the soft data generated by the RAMP analysis. The RAMP BME data fusion method is in contrast with current data fusion approaches that assume linearity and a homoscedastic relationship between observed and gridded modeled data. These approaches have been expressed through Bayesian Melding as $\tilde{X}(\boldsymbol{s}) = \beta_0(\boldsymbol{s}) + \beta_1(\boldsymbol{s})\hat{X}(\boldsymbol{s}) + \varepsilon(\boldsymbol{s})$(Fuentes and Raftery, 2005) and through the Downscaler as $\hat{X}(\boldsymbol{s}, t) = \beta_0(\boldsymbol{s}, t) + \beta_1(\boldsymbol{s}, t)\tilde{x}(\boldsymbol{s}, t) + \varepsilon(\boldsymbol{s}, t)$ (Berrocal et al., 2010b). The linear nature and homoscedastic assumption of these models has advantages in terms of implementation, but they fail to account for the non-linear and non-homoscedastic nature of model performance (Reyes et al., 2016; Xu et al., 2016). The RAMP BME framework presented here provides an attractive alternative for data fusion because it explicitly accounts for the non-linear, non-homoscedastic and non-homogenous nature of model performance.

To investigate this, we compare the data fusion BME method with a frequentist implementation of the Downscaler equation (Berrocal et al., 2010b) that assumes model performance is linear and

homoscedastic. The statistical underpinnings of the steps taken to implement the frequentist Downscaler can be found in Appendix B. Through a LOOCV analysis, the PCMSE was calculated between the frequentist Downscaler and the raw CMAQ data as well as the BME data fusion method for the 200 km cross validation radius (Fig. 3.2c). Results are stratified by percentile of the proportion of systematic error to total error $SME(\boldsymbol{p})$. An increase in proportional systematic error implies a decrease in the proportion of random error. Decreased proportion of random error in the model performance metric translates into less uncertainty in the CMAQ data. As a result, the PCMSE between the frequentist Downscaler and the CMAQ data increases from 40% to 55% (Fig. 3.2c), indicating a clear trend of improved performance as a function of increasing proportional systematic error. This shows that the frequentist Downscaler is a successful data fusion method. However the PCMSE between RAMP BME and the frequentist Downscaler ranges from -11% to -32% (Fig. 3.2c), meaning the MSE is 11 to 32% lower for BME compared to the frequentist Downscaler approach. This substantial improvement demonstrates that BME has higher estimation accuracy.

Table 3.1. Cross validation statistics across radii. Cross validation results for daily PM2.5 across the contiguous US for 2001 from three LOOCV radii: 1) 0 km, 2) 400 km, and 800 km displaying results of the statistics of Mean Squared Error, MSE ($(\mu g/m^3)^2$), correlation coefficient, r (unitless) and Percent Change in MSE, PCMSE (%). Statistics are displayed comparing observed values with estimates calculated through kriging and BME. Within each LOOCV radius there are three exclusion criteria implemented: 1) a "Baseline" including all data, 2) $SME(\boldsymbol{p}) \geq 22\%$, including only observed space/time locations associated with proportional systematic error model performance greater than or equal to 22% and 3) adding an additional exclusion criteria $\lambda_1 \geq 16.8$, including only observed space/time locations associated with the λ_1 parameter greater than or equal to 16.8.

		0 km Radius			**400 km Radius**			**800 km Radius**		
Statistic	**Method**	Baseline	$SME \geq 22\%$	$SME \geq 22\%$, $\lambda_1 \geq 16.8$	Baseline	$SME \geq 22\%$	$SME \geq 22\%$, $\lambda_1 \geq 16.8$	Baseline	$SME \geq 22\%$	$SME \geq 22\%$, $\lambda_1 \geq 16.8$
MSE	Kriging	20.5	23.7	40.6	47.9	60.1	159.3	56.8	66.4	173.8
	BME	19.9	22.4	39.1	43.3	50.5	117.7	45.9	51.9	118.8
r	Kriging	0.863	0.859	0.841	0.629	0.579	0.382	0.539	0.517	0.378
	BME	0.866	0.865	0.846	0.673	0.660	0.459	0.649	0.647	0.457
PCMSE	---	-2.9	-5.2	-3.7	-9.6	-16.0	-26.1	-19.1	-21.8	-31.6

3.3.4 Stratification of BME data fusion performance

As discussed above with the cross-validation results, percent change in MSE comparing kriging to BME improves with increasing cross-validation radii, increasing λ_1 and increasing systematic error. In the RAMP methodology λ_1 is a CMAQ prediction that has been corrected for systematic errors, hence large λ_1 values are associated with high PM2.5 concentrations.

The cross validation MSE increases for both kriging and the BME data fusion method as the exclusion criteria becomes increasingly more stringent (Table 3.1). Increasing MSE is expected with increasing λ_1 due to the higher errors being associated with high observed concentrations. Performance of BME over kriging accentuates with increasing λ_1 values (Fig. 3.2a). This differentiation leads to an improvement in MSE when comparing BME with kriging. For BME we hypothesize incorporation of λ_1s improve prediction due to high λ_1s being surrogates for high PM2.5 concentration. When looking at the data fusion method stratifying by increasing λ_1 will, by definition, include larger λ_1 values into the prediction. This will increases the magnitude of the prediction and create an improved ME when compared with kriging.

Data fusion performance is differentiated even further when data are subset to high levels of proportional systematic error. As proportional systematic error increases, random error decreases. As has been shown in previous work (Reyes et al., 2016), there is a correspondence between random error and λ_2. By stratifying performance by both high PM2.5 concentrations and low random error, predictions have a low uncertainty and soft data contribute more to predictions at unmonitored locations.

Investigating model performance metrics provide insight into locations where the BME data fusion method would be most beneficial and show the largest improvement in prediction compared with using an observational based method, like kriging. Increased performance in the data fusion method with increasing cross validation radii demonstrates that the data fusion method can increase performance far from monitors. This has implications in epidemiologic studies when assigning exposure to participants who live in rural areas or areas located far from regulatory monitoring stations.

3.3.5 Data fusion corrects the bias of observation based predictions

Looking at the difference between the BME and kriging mean predicted across all CMAQ grid locations for a given day (Fig. 3.1e), the difference between BME mean and kriging mean is mostly

negative. Of the 16,576 prediction grids shown for 07/01/2001, 12,310 (74.26%) are negative and 4,266 (25.74%) are positive. For this given day there is nearly a 2.9:1 odds that the kriging prediction is larger than the BME prediction. This is due to the coarseness of the inputs in kriging. Kriging predictions are created only using observed data from regulatory monitors. That is, kriging is limited to prediction at the scale in which observed data exist. In between monitors, particularly where there is sparse monitoring, there is a failure to see fine scale variation. Most regulatory monitors are located in urban areas with denser monitoring. With kriging, urban spatial gradients are then incorrectly applied to rural areas. Rural areas with lower daily PM2.5 concentrations typically have sparse monitoring. To provide a prediction in rural areas, kriging must utilize observed data large distances from the prediction locations, leading kriging predictions to over predict low concentrations. Along with low concentrations seen largely in rural areas, kriging has difficulty estimating the highest PM2.5 concentrations. Typically, the highest daily PM2.5 concentrations typically are observed in isolation, with the surrounding stations observing relatively lower concentrations. In an LOOCV, an interpolation of surrounding stations is unable to predict at the highest concentrations.

The soft data incorporated into BME have been removed of systematic error. Therefore, on average, the mean error between observed and the soft data is zero. Not only does the soft data match closely with the observed data where the observed data exist, but the soft data are then able to pick up on small scale variation in between monitors. It is this small scale variation that is incorporated into BME.

This work implies that modelers should be cautious when using observation only prediction methods far from observed data or when PM2.5 studies include the highest PM2.5 concentrations. There may be a tendency to pick up on trends seen in areas with denser monitoring. These same trends may not apply in more rural areas. From an epidemiologic perspective, an observation only geostatistical approach can be problematic for rural areas.

3.3.6 Data fusion captures fine scale variability of PM2.5

The BME data fusion method is able to capture fine scale variability of daily PM2.5 better than observation based only prediction methods. Three areas of the country were explored to quantify the magnitude of refinement. The covariance ranges of the BME and kriging posterior means on 07/01/2001 were calculated from three distinct areas of the country: 1) Southern California, 2) the Mid-East and 3)

Missouri (Fig. B.6). To quantify the spatial refinement, covariance models were calculated on the posterior means in each of these three areas (Fig. B.7). The kriging covariances fit a one-structured model while the BME covariances fit a two-structured model. The shortest ranges for kriging are 422 km, 296 km and 386 km for Southern California, the Mid-East and Missouri, respectively. The shortest ranges for BME are 66 km, 39 km and 58 km for Southern California, the Mid-East and Missouri, respectively. The ratios between the shortest kriging ranges to the shortest BME range are 6.4, 7.6 and 6.7 for Southern California, the Mid-East and Missouri, respectively. That is, BME allows for over 6 times the amount of spatial refinement compared with kriging. This can we seen visually through maps (Fig. 3.1a,b). In Southern California the BME mean map is able to refine spatial variability by only having high PM2.5 levels concentrated in and around Los Angeles, California, while the kriging map encompasses a much larger proportion of California. The Mid-East area is much further refined by having a more pronounced gradient between high and low concentrations. By contrast, the kriging mean has the high concentrations in the Mid-East as one contiguous front.

This finding has major implications in epidemiologic studies. When exploring long-term exposure of ambient concentrations to air pollutants at a coarse scale, misclassification can be highly problematic. Exposure predictions benefit both from the accuracy of observed data and the coverage and spatial refinement of modeled data. The BME data fusion method has both of these characteristics. With high levels of potential misclassification, links between disease and exposure can be understated or missed.

Exposure predictions from the BME data fusion method have been used to investigate associations between long term exposure to PM2.5 and brain mass in elder women in a cohort from the Women's Health Initiative (WHI) focusing on memory studies (Chen et al., 2015). Exposure prediction profiles for individual participants are considered more realistic when variation within a given participant is high (Chen et al., 2015). Variation for the BME data fusion method remains higher than kriging when investigating posterior means across the US even when the spatial domain is subset (Fig. B.8). The average variance within increasingly smaller spatial subdomains for BME is up to 15 times that of kriging. With larger spatial variation, exposure profiles in an epidemiologic setting because increasingly more realistic.

3.3.7 Overall contributions and future works

This work incorporated the RAMP error corrected CMAQ data with observed data into the BME data fusion method. The soft data added in this work takes into account the non-linear, non-homoscedastic and non-homogeneous model performance seen in CMAQ. This is the first work to incorporate observed and CMAQ modeled data together in the BME framework to predict daily PM2.5. The BME data fusion method improves spatial refinement of predictions and captures fine scale variability. Limitations in this work include using an older version of CMAQ. CMAQv4.5 is the most current version of CMAQ predicting across the US for the year 2001, which was needed for the Women's Health Initiative-Memory Study (WHI-MS) associated with this work (Chen et al., 2015). However, any epidemiologic study investigating 2001 can benefit from this work. Using a CMAQ model version with known deficiencies (Foley et al., 2010) provides an ideal case study for exploring a data fusion method. From an epidemiologic perspective, having fine scale predictions from this time period can be helpful when assessing long term exposure to an air pollutant (Chen et al., 2015). Future work includes reevaluating the BME data fusion method with an updated version of CMAQ and reanalyzing the resulting improvement in predictions.

CHAPTER 4: INCORPORATING MASS FRACTION OF POLYCYCLIC AROMATIC HYDROCARBONS INTO THE BAYESIAN MAXIMUM ENTROPY FRAMEWORK ACROSS NORTH CAROLINA[3]

4.1 Introduction

Polycyclic Aromatic Hydrocarbons (PAHs) are a class of organic compounds containing 2 or more fused aromatic rings created by incomplete fuel combustion from a variety of sources including biofuel burning, wildfires, coal production, etc. (Di-Toro et al., 2000; Zhang and Tao, 2009). Several species of PAHs have been designated by the US Environmental Protection Agency (EPA) as being probably human carcinogens (Bocskay et al., 2005; Menzie et al., 1992; Wolff et al., 2005). Currently the EPA only has PAH regulatory standards for drinking water and National Institute for Occupational Safety and Health (NIOSH) has established occupational exposure limits to coal tar pitch volatiles (Kim et al., 2013). International organizations and other countries have established ambient concentration guidelines for one of the more toxic PAHs, benz(a)pyrene (Ravindra et al., 2008). However, currently in the US there are no regulatory standards for ambient concentrations of PAHs. PAHs can be costly to measure (Pleil et al., 2004). Compared to regulated ambient air pollutants, there are few epidemiologic studies that have utilized observed data or explored ambient exposures to different PAHs (Abdel-Shafy and Mansour, 2015). From a geostatistical perspective, limited ambient observed data have resulted in few studies creating maps of PAHs concentrations across space/time (Allshouse et al., 2009; Augusto et al., 2009; Lee et al., 2016; Ribeiro et al., 2015). Others have used Chemical Transport Models (CTMs) to predict PAH concentrations (Guerreiro et al., 2016; Ravindra et al., 2008). However, these studies are limited in number. As a result, there is a gap in the literature exploring ambient PAH exposures and their associations with various health endpoints.

[3] This chapter is planned to be submitted as an article to the Journal of Exposure Science and Environmental Epidemiology. Reyes, Jeanette M., Hubbard, Heidi, Stiegel, Matthew A., Pleil, Joachim D., Serre, L. Marc. Incorporating Mass Fraction of Polycyclic Aromatic Hydrocarbons into the Bayesian Maximum Entropy Framework across North Carolina.

There is a lack of consistent PAH monitoring outside of monitoring campaigns conducted for specific studies. In contrast to the data poor environment of PAH monitoring, Particulate Matter ≤ 2.5 micrometers (PM2.5) exists in a data rich environment with a vast, consistent, historical monitoring network across the US.(US EPA, n.d.) Currently there are 16 EPA designated priority PAHs, 9 of which are particle-bound (Allshouse et al., 2009). Thus, a portion of PM2.5 is particle-bound PAH. The US state of North Carolina currently has no maps displaying PAH concentration. For this work 84 PM2.5 filters were collected from North Carolina in 2005 and analyzed for 9 particle-bound species of PAH. The relationship between collocated PAH and PM2.5 data is developed and applied elsewhere where PM2.5 is known. The goals of this work is the provide maps and present a novel method of modeling sparse data, supported through a data rich environment, which can be applied elsewhere through the modern geostatistical Bayesian Maximum Entropy (BME) framework (Christakos, 2000; Christakos et al., 2001).

This study explores the relationship between PM2.5 and PAH and is an extension of previous work done by Allshouse et al. (Allshouse et al., 2009) that investigated PAH near the World Trade Center after September 11th. An empirical approach is taken to find the optimal neighborhood size for each of the 9 PAHs, after which, two different methods are used to relate PAH to PM2.5: 1) a simple Linear Regression (LR) approach and 2) a Mass Fraction (MF) approach. The MF method assumes that the ratio of PAH/PM2.5 is constant within an estimation neighborhood. PAH is then estimated across North Carolina at PM2.5 monitoring locations using both approaches. These estimated PAHs assume a Gaussian distribution and are incorporated into to BME framework to estimate PAH at unmonitored space/time locations across North Carolina. The MF BME method was developed by Allshouse et al. However, this is the first work to compare MF BME with LR BME.

This work implements the LR BME and MF BME method to predict PAH concentration at unmonitored locations creating the first maps of PAH across the US state of North Carolina for 2005. The MF BME prediction method is compared to more traditional geostatistical methods and is evaluated through cross validation. Predictive maps also allow to visualize probability of exceeding PAH cutoff concentrations. Lastly, a comparison is performed between the MF BME and other methods showing how the relationship between PAH concentrations near fires change across prediction methods. These results

provide a method for which a data poor environment can be exploited in an efficient manner in conjunction with a data rich environment, where the relationship between the two can be applied elsewhere in a given study area. This cost-effective method can be applied to other air pollution parameters that have not been previously mapped. This methodology opens to doors for greater epidemiologic studies exploring the association between ambient concentrations of PAHs and various health endpoints.

4.2. Materials and Methods

4.2.1 Observed PM2.5 and PAH data

Daily PM2.5 filters for each space/time location during 2004-2005 in North Carolina were collected from the EPA's Air Quality Systems (AQS) data base (US EPA, n.d.). Of the PM2.5 filters collected during this time period, 84 filters from 2005 were analyzed by the US EPA for the following 9 species of PAHs: benz(a)anthracene, chrysene, benzo(b)fluoranthrene, benzo(k)fluoranthrene, benzo(e)pyrene, benzo(a)pyrene, indeno(1,2,3-cd)pyrene, benzo(g,h,i)perylene, dibenzo(a,h)anthracene, and the summation of the 9 PAH species called Total PAH. PM2.5 has units of $\mu g/m^3$ and PAH has units of ng/m^3.

4.2.2 The Mass Fraction (MF) and Linear Regression (LR) method

There are approximately $8{,}000$ space/time locations where PM2.5 is observed and PAH is estimated. PAH is estimated using surrounding PAH and PM2.5 information. There are two different PAH estimation methods: 1) a Linear Regression (LR) method where a regression is created from paired PM2.5 and PAH in an estimation neighborhood, and PAH is then predicted at locations where PM2.5 is known and 2) a Mass Fraction (MF) method where it is assumed that the ratio of PAH/PM2.5 is constant within an estimation neighborhood, and PAH is then predicted by applying the ratio at locations where PM2.5 is known.

The MF method introduced here builds on previous work and is used to estimate PAH at unmonitored locations (Allshouse et al., 2009). The log-mas fraction (log-MF), is calculated at locations for paired PAH and PM2.5 data and is defined as

$$MF_{hard,i} = ln\left(\frac{PAH_{hard,i}}{PM2.5_{hard,i}}\right) \quad \text{(Equ. 4-1)}$$

The MF is then estimated at PM2.5 space/time locations without PAH observed data with mean $\mu_{MF,j}$ and variance $\sigma^2_{MF,j}$ defined below

$$\mu_{MF,j} = \sum_{i=1}^{N_{MF}(p_j)} MF_{hard,i}/N_{MF}(\boldsymbol{p}_j) \quad \text{(Equ. 4-2)}$$

$$\sigma^2_{MF,j} = \sum_{i=1}^{N_{MF}(p_j)} \left(MF_{hard,i} - E\left[MF_{soft,j}\right]^2\right)/\left(N_{MF}(\boldsymbol{p}_j) - 1\right). \quad \text{(Equ. 4-3)}$$

$N_{MF}(\boldsymbol{p}_j)$ is the number of $MF_{hard,i}$ closest to the space/time location $\boldsymbol{p}_j = (\boldsymbol{s}_j, t_j)$ used in the calculation. This number is optimized and is described in section 2.3. The terms in Equ. 4-1 can be rearranged to calculate $PAH_{hard,i}$ as follows

$$ln(PAH_{hard,i}) = MF_{hard,i} + ln(PM2.5_{hard,i}) \quad \text{(Equ. 4-4)}$$

The relationship in Equ. 4-4 can then be used to estimate PAH at PM2.5 locations with the following distribution

$$ln\left(PAH_{soft,j}\right) \sim N\left(\mu_{MF,j} + ln\left(PM2.5_{hard,j}\right), \sigma^2_{MF,j}\right) \quad \text{(Equ. 4-5)}$$

Equ. 4-5 becomes the soft data in the BME estimation methodology described in section 2.4.

The LR method is also used to estimate PAH at PM2.5 space/time locations where PAH was not directly measured. The LR method is a simple linear regression equation, like the MF, calculated at paired PAH and PM2.5 locations given in the equation below.

$$ln\left(PAH_{hard,i}\right) = \beta_0 + \beta_1\, ln\left(PM2.5_{hard,i}\right) \quad \text{(Equ. 4-6)}$$

The number of points used to estimate the parameters β_0 and β_1, $N_{LR}(\boldsymbol{p}_j)$ is described in section 2.3. The relationship in Equ. 4-6 can then be used to estimate PAH at PM2.5 locations with the following distribution

$$ln\left(PAH_{soft,j}\right) \sim N\left(\hat{\beta}_0 + \hat{\beta}_1\, ln\left(PM2.5_{hard,j}\right), \sigma^2_{LR,j}\right) \quad \text{(Equ. 4-7)}$$

where $\sigma^2_{LR,j}$ is the linear regression prediction variance. In the limiting case, the MF and LR method are equivalent when $\beta_0 = MF_{hard,j}$ and $\beta_1 = 1$.

4.2.3 Soft data neighborhood validation optimization

As discussed in the section above, parameters $N_{MF}(\boldsymbol{p}_j)$ and $N_{LR}(\boldsymbol{p}_j)$ are optimized. An exhaustive validation approach was used for each of the 9 PAHs (and Total PAH) at the 84 observed PAH space/time locations. The neighborhood is created in such a way that points included are informative but can also characterize a non-trivial area. For each of the 84 locations, an exhaustive

combination of the n closest pairs, as determined by a space/time metric, is collected and PAH is estimated using either the MF method (Equ. 4-4) or the LR method (Equ. 4-6), excluding the space/time location of interest. From this, a Mean Squared Error (MSE) is calculated from the 84 errors. A MSE is calculated for $75{,}600$ different combinations of n and the space/time metric for each PAH and method. For each PAH, the parameters $N_{MF}(\boldsymbol{p}_j)$ and $N_{LR}(\boldsymbol{p}_j)$ are selected from the n and space/time metric that resulted in the lowest MSE. Due to the parsimony of the MF and LR methods, $N_{MF}(\boldsymbol{p}_j) \geq 1$ while $N_{LR}(\boldsymbol{p}_j) \geq 2$.

The values found for $N_{MF}(\boldsymbol{p}_j)$ and $N_{LR}(\boldsymbol{p}_j)$ for each PAH are then applied to locations where observed PM2.5 exists to estimate PAH using Equ. 4-5 and Equ. 4-7. These PAH estimates become the soft data in the BME estimation framework described next.

4.2.4 Bayesian Maximum Entropy (BME) estimation methodology

BME is a mathematically rigorous geostatistical space/time framework originally developed by Christakos (Christakos, 2000; Christakos et al., 2001). BME can incorporate information from multiple data sources and is implemented using the *BMElib* suite of functions in MATLAB™. The buttress of BME has been detailed in other works, and can be summarized as performing the following steps: 1) gathering the general knowledge base (G-KB) and site-specific knowledge base (S-KB) characterizing the Space/Time Random Field (S/TRF) $X(\boldsymbol{p})$ representing a process at space/time coordinate $\boldsymbol{p} = (\boldsymbol{s}, t)$ where $\boldsymbol{s}$ is the spatial coordinate and t is time, 2) using the *Maximum Entropy* principle of information theory to process the G-KB in the form of a prior Probability Distribution Function (PDF) f_G, 3) integrating S-KB in the form of a PDF f_S with and without measurement error using an epistemic *Bayesian* conditionalization rule on data to create a posterior PDF f_K and 4) creating space/time estimates based on the analysis. Typically, the G-KB consists of a mean trend and covariance of $X(\boldsymbol{p})$ denoted as $G = \{m_X(\boldsymbol{p}), c_X(\boldsymbol{p}, \boldsymbol{p}')\}$ and the S-KB consists of hard data (data measured without error) and soft data (data measured with error) denoted as $S = \{\boldsymbol{x}_h, f_s(\boldsymbol{x}_s)\}$. The BME posterior PDF f_K describing the process x_k at an estimation point of interest $\boldsymbol{p}_k$ is given by the BME equation

$$f_K(x_k) = A^{-1} \int d\boldsymbol{x}\, f_S(\boldsymbol{x}_s) f_G(\boldsymbol{x}) \qquad \text{(Equ. 4-8)}$$

where $\boldsymbol{x} = (x_k, \boldsymbol{x}_h, \boldsymbol{x}_s)$ is a realization of $\boldsymbol{X}$ at points $\boldsymbol{p} = (\boldsymbol{p}_k, \boldsymbol{p}_h, \boldsymbol{p}_s)$ and A is a normalization constant.

In this study we use an S/TRF to describe the variability of PAH across North Carolina in 2005. In this work $\boldsymbol{x}_h$ are the observed PAH data and $f_S(\boldsymbol{x}_s)$ is estimated through either the LR or MF method. Our notation for an S/TRF will consist of denoting a single random variable X in capital letters, its realization, x, in lower case and vectors and matrices in bold faces (e.g. $\boldsymbol{X} = [X_1, \ldots, X_n]^T$ and $\boldsymbol{x} = [x_1, \ldots, x_n]^T$). Let $X(\boldsymbol{p}) = X(\boldsymbol{s}, t)$ be a Space/Time Random Field (S/TRF) representing daily PAH. We can then calculate x_k, the predicted daily PAH at the unmonitored location $\boldsymbol{p}_k$. In this work, $m_X(\boldsymbol{p})$ is assumed to be constant. The covariance model for the homogeneous/stationary S/TRF $X(\boldsymbol{p})$ is developed from the experimental covariance of the data fit through least squares empirical fitting. For each PAH, the 84 observed data were used to fit a single-structured space/time exponential covariance model given by the equation

$$c_X(r, \tau) = C_0 \exp\left(\frac{-3r}{a_r}\right) \exp\left(\frac{-3\tau}{a_t}\right) \quad \text{(Equ. 4-9)}$$

where r is the spatial distance (km), τ is the temporal distance (days), C_0 is the variance, a_r is the spatial range (km) and a_t is the temporal range (days).

4.2.5 Leave One Out Cross Validation (LOOCV) accuracy analysis

To assess the prediction accuracy of the MF and LR methods, a LOOCV accuracy analysis is performed. For each monitoring station where observed PAH data exist, all observed data from a given station are removed one at a time and a BME prediction was conducted (without recalculating the mean trend or the covariance model) to obtain the BME predictions at that station using all the remaining observed data and soft data as estimated by the MF and LR method.

The difference between each prediction value $\tilde{x}_i$ and observed value $\hat{x}_i$ is the prediction error, $e_i = \tilde{x}_i - \hat{x}_i$. The prediction accuracy is quantified based on prediction error statistics, which consist of the Mean Error (ME, ng/m^3), Variance of Errors (VE, $(ng/m^3)^2$), Root Mean Squared Error (RMSE, ng/m^3), Mean Squared Error (MSE, $(ng/m^3)^2$), and the squared of the Pearson's correlation coefficient (r^2, unitless) calculated between observed and predicted values. MF BME and LR BME predictions are then compared to kriging (i.e. predictions created only using observed data) and cokriging (i.e. predictions created using paired PAH and PM2.5 observed data).

4.2.6 Fire comparisons

The difference in PAH concentrations near known fire locations were estimated. PAH is estimated on a fine grid across North Carolina on days with observed PAH data. PAH is estimated using 4 different prediction methods: 1) kriging, 2) cokriging, 3) LR BME and 4) MF BME. Fire data are obtained from the Federal Wildfire Fire Occurrence Website (United States Geological Survey, 2016). All fires greater than or equal to one acre in North Carolina, Virginia, Tennessee and South Carolina were collected in 2005 on days for which PAH observed data were measured where the start and control date of the fires are known. A two-tailed two-sampled t-test (assuming unequal variances) is calculated on the PAH predictions on a fine grid at a 5% significance level. The significance test is performed on all fine grid predictions within 100 km of known fire locations and all fine grid predictions outside of 100 km. We explored the statistically significant difference in the PAH predictions near versus far from the known fire locations across all 4 prediction methods.

Table 4.1. Soft data neighborhood optimization. Optimized n closest observed data locations (as determined by the space/time metric) corresponding to the minimized mean squared error validation statistic calculated through the linear regression and mass fractions methods across the 9 PAHs, with Total PAH being the summation. Bolded numbers indicate the lowest MSE across PAHs.

	Linear Regression			Mass Fraction		
PAH	n	S/T Metric (km/days)	MSE $(ng/m^3)^2$	n	S/T Metric (km/days)	MSE $(ng/m^3)^2$
benz(a)anthracene	14	0.891	1.128	5	0.839	**0.908**
chrysene	7	0.600	0.979	5	0.839	**0.799**
benzo(b)fluoranthrene	7	0.863	1.358	5	0.899	**1.180**
benzo(k)fluoranthrene	14	0.895	1.375	5	0.842	**1.046**
benzo(e)pyrene	14	0.895	1.006	2	0.868	**0.726**
benzo(a)pyrene	14	0.895	**1.332**	5	0.899	1.417
indeno(1,2,3-c,d)pyrene	14	0.891	0.892	2	0.868	**0.702**
benzo(g,h,i)perylene	14	0.895	0.757	2	0.777	**0.742**
dibenzo(a,h)anthracene	14	0.772	1.532	3	0.820	**1.115**
Total PAH	14	0.895	0.890	3	0.820	**0.675**

4.3. Results and Discussion

4.3.1 Neighborhood optimization

Due to the skewed nature of PM2.5 and PAH, a log-transformation of both the data sets were taken. Due to the limited number of observed data, initial soft data neighborhood optimization was important. The soft data for each PAH needed to be optimized in such a manner that the few observed

data were being utilized in the most efficacious manner possible through an exhaustive search of the n closest observed data. A validation of an exhaustive combination of n closest stations using several different space/time metrics of the 84 collocated PAH/PM2.5 space/time locations were taken to optimize the soft data neighborhood needed to construct estimated PAH and the remaining PM2.5 space/time locations. For each PAH and soft data estimation method (i.e. LR and MF), the optimized n and space/time metric was selected such that it minimized the MSE. The neighborhood optimization was done to ensure that the neighborhood selected for each PAH and soft data estimation method would be as representative as possible considering the limited collocated values. Across each PAH the n closest stations that optimized the soft data neighborhood was always smaller for the MF method compared to the LR method (Table 4.1). Across most of the PAHs, the space/time metrics were similar indicating that the differences in the minimized MSEs were being driven by the number of collocated values and the estimation method and less so by the choice of the space/time metric. By definition the MF method requires less collocated PAH and PM2.5 to calculate a PAH estimate, due to the MF method being more parsimonious (i.e. having less parameters to estimate). The parameter n ranges from 2-5 for the MF method and n ranges from 7-14 for the LR method. Benzo(g,h,i)perylene, indeno(1,2,3-c,d)pyrene and benzo(e)pyrene require $n = 2$ from the MF method, requiring the least amount of points across all PAHs. Seven out of 9 PAHs in the LR method require $n = 14$. Across each PAH, the minimized MSE is consistently lower for the MF method than the LR method. The only exceptions are benzo(a)pyrene. The MSE for the MF method is smaller than the LR method for Total PAH. With these optimized neighborhoods, soft data are created by each method and predicted across North Carolina using BME.

The PAH estimation neighborhood for the MF BME method is smaller than the LR BME method. Out of the previously mentioned studies, very little use observed PAH data and of those studies that do, most observed data come from short-lived monitoring campaigns. The results presented in this work utilize long-term, established regulatory monitoring sites (i.e. PM2.5 sites). PM2.5 data is comparatively plentiful. By developing a relationship between a few PAH observations and PM2.5, the door opens to applying this relationship to a network with a large amount of publically available data. Data poor environments (e.g. PAH) can benefit from data rich environments (e.g. PM2.5). However, for this relationship to be fully exploited, it must be constructed in a way that best utilizes the limited data set.

That is, the relationship between PAH and PM2.5 must be parsimonious. The MF method only has one parameter to be estimated, namely, $MF_{hard,i}$ (Equ. 4-4). The minimum number of observed data needed to construct a PAH estimation is low with $N_{MF}(\boldsymbol{p}_i) \geq 1$. The soft data created from the MF method required less observed data than the LR method. This resulted in a superior cross validation model performance (Table 4.2). The smaller neighborhood also has a physical interpretation. We hypothesize that the smaller number of observed data made the estimation more localized and more relevant to the air shed being predicted.

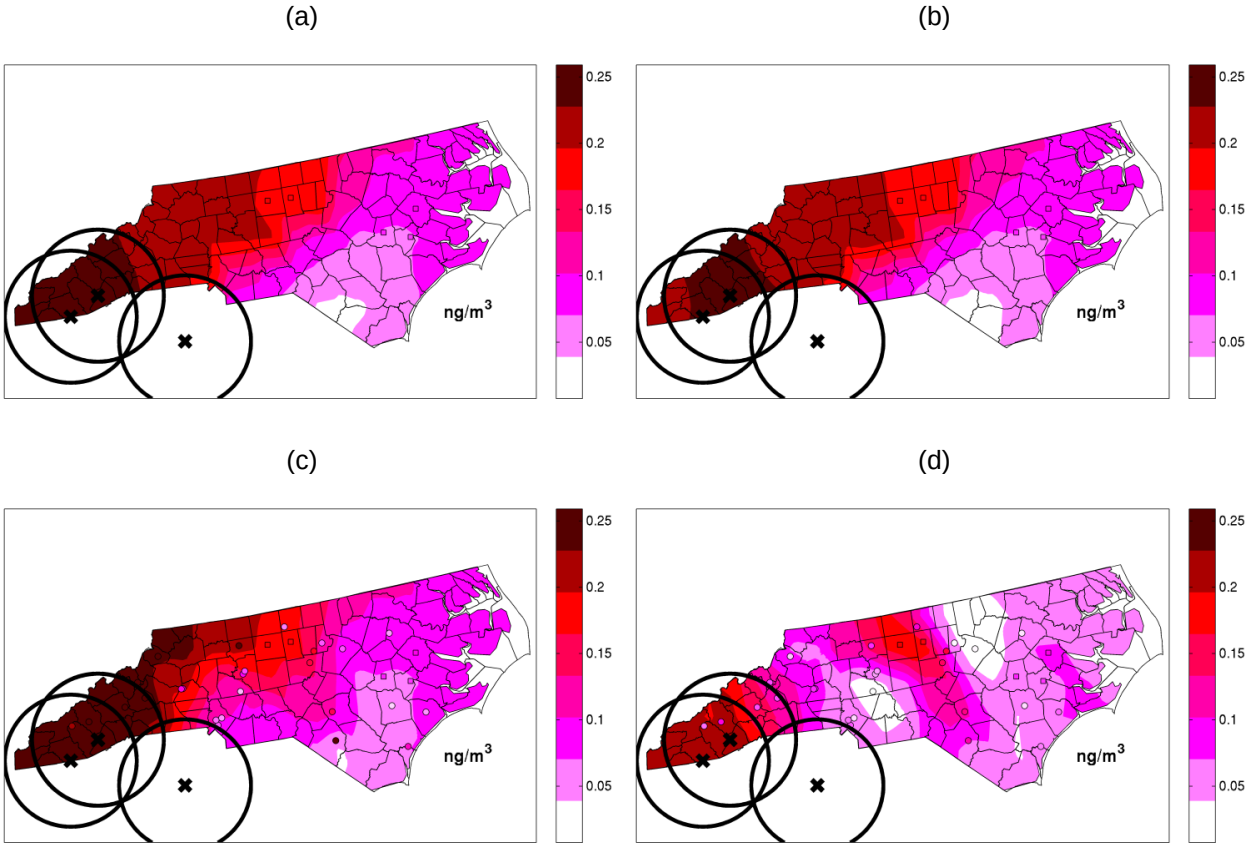


Figure 4.1. Map of benzo(g,h,i)perylene. Maps of mean benzo(g,h,i)perylene concentration for North Carolina on March 11, 2005 across the 4 prediction methods: (a) kriging, (b) cokriging, (c) BME linear regression, (d) BME mass fraction. Square markers indicate observed data, circle markers indicate soft data, X's mark known fires for that day with a $100\ km$ buffer. Units are in ng/m^3.

4.3.2 PAH prediction maps

This work created the first maps of predicted PAH in space/time across the US state of North Carolina for 2005. With only a handful of observed PAH data taken throughout the year, the LR BME and MF BME method were able to create soft data with a corresponding uncertainty that was incorporated

into the BME framework. Incorporating soft data allowed for increased spatial variation. Both BME methods outperformed kriging and cokriging. Of the BME methods, MF BME was superior in terms of visually distinguishing spatial variations of PAH.

Each PAH was predicted on a fine grid across the US state of North Carolina every day observed PAH data were collected (41 days) across 2005 for the four prediction methods: kriging, cokriging, LR BME and MF BME. Covariance parameters can be found in Appendix C (Table C.1 and Table C.2). These four mean prediction maps are displayed across North Carolina on March 11, 2005 for the PAH benzo(g,h,i)perylene with observed and soft data pictured (Figure 4.1). The kriging map consistently predicts the highest PAH concentrations at unmonitored locations with the least realistic gradient. Kriging has difficulty distinguishing between multiple PAH fronts and plumes. The minimal gradation is influenced by the sparse data set of only 84 observed PAH values. The cokriging map is visually similar to the kriging map. There is a slight reduction of PAH concentration notably in Western North Carolina along with a reduction in concentration near the South Carolina border. Taking into account the relationship between PAH and PM2.5 contributed little to the cokriging predictions. The prediction map becomes visibly different for the LR BME method. The area of highest concentration increases along the Appalachian Mountains. The gradient for the LR BME method falls more in line with a geographical pattern across the state. There is more of a gradual decrease in concentration across the state. Overall the LR method has a larger region of the state with relatively lower concentrations. Overall the MF BME method has the lowest concentrations across the state. Across methods, the relatively highest concentrations were found in Western North Carolina and concentrations become increasingly more refined across methods. The MF map is the only map to show two different fronts: one in the western part of the state and another separate front in the central part of the state. The lowest concentrations occur between the two fronts.

Very few other studies have created maps of ambient PAH concentrations across a given area using a geostatistical method. These limited studies are due in part to the lack of observed data, much like the mapping scenario presented in this work and previous works (Allshouse et al., 2009). One previous study fit a temporal trend comparing a few long-running PAH stations from the Great Lakes region and a few stations across Europe. However, only a temporal trend was fit through a regression

and a spatial interpolation was not conducted (Liu et al., 2013). One of the few studies that create maps over a large area, displayed benz(a)pyrene across Europe for 1990, 2001 and 2005 using a transport model (Ravindra et al., 2008). Another study creating maps of PAH across Europe utilized kriging to estimated benz(a)pyrene for 2012 using two different chemical transport models as data (Guerreiro et al., 2016). A study in Portugal used observed PAH data extracted from lichen and interpolated to create maps using kriging (Augusto et al., 2009; Ribeiro et al., 2015). Land use regression models have also been used to estimate PAH (Jedynska et al., 2014; Noth et al., 2011). The closest study to the LR BME method presented in this work used a monitoring campaign along with personal monitors to analyze PAH from PM2.5 in which predictions were made at unmonitored locations using kriging in Kaohsiung city, Taiwan (Lee et al., 2016). A regression model with a variety of explanatory variables was then applied to PM2.5 data to predict PAH. However, the explanatory variables used in their work (i.e. PM10, NOx, CO and temperature) are less relevant than the explanatory variable used in the LR BME method (i.e. using PM2.5 directly).

Table 4.2. Cross validation statistics. Leave-One-Out-Cross-Validation statistics for Total PAH (summation of the 9 PAHs) comparing observed and predicted concentrations across the 4 prediction methods for North Carolina in 2005. ME is Mean Error, VE is Variance of Error, RMSE is Root Mean Squared Error, MSE is Mean Squared Error and r^2 is the Pearson's correlation coefficient squared.

Statistic	Kriging	Cokriging	Linear Regression	Mass Fraction
ME (ng/m^3)	-0.145	-0.137	-0.102	-0.042
VE $(ng/m^3)^2$	0.806	0.782	0.764	0.591
RMSE (ng/m^3)	0.904	0.890	0.875	0.765
MSE $(ng/m^3)^2$	0.818	0.792	0.766	0.586
r^2 (unitless)	0.747	0.752	0.744	0.821

4.3.3 Cross-validation

A Leave-One-Out Cross Validation (LOOCV) method (where one monitoring station at a time is left out) was calculated across 2005. The "left out" observed space/time locations are then re-estimated using the four prediction methods. Statistics were calculated showing performance for Total PAH (Table 4.2). Cross validations statistics for all 9 PAHs can be found in Appendix C (Table C.3). ME decreases as the method become increasingly more complex. ME is negative across each prediction method meaning that overall, the methods under-predict observed Total PAH concentrations. ME magnitude is highest for kriging and closest to zero for MF BME. There is an over 58% reduction in ME from LR BME to MF BME.

There is less variation in error across methods. There is an over 36% reduction in VE from kriging to MF BME. There is a consistent reduction in MSE across methods. There is an over 39% reduction in MSE from kriging to MF BME. The correlation coefficient increases across methods. There is an over 10% increase in r^2 from LR BME to MF BME. The performance statistics from kriging are similar to cokriging. This echoes the results of the prediction maps. Traditional incorporation of co-pollutants through cokriging adds little to the predictive captivity of PAH. Incorporating the soft data showed improvement in statistics, while incorporating of soft data through the MF method showed the best performance across statistics.

The MF BME method consistently outperformed all other comparison methods as seen visually through maps and through the LOOCV statistics. Of the four prediction methods, kriging performed the worst. Kriging predictions were driven exclusively by the observed data. Predictions made far from observed data therefore had a large associated variance. The sparse data was only able to pick up the coarsest of PAH gradients. Cokriging performed similarly to kriging. Cokriging is an intuitive choice for collocated, ambient, environmental parameters in a geostatistical setting. In the literature, to the best of our knowledge, cokriging has not been used to prediction ambient PAH concentrations, making it an ideal candidate method to explore. In this work the cokriging cross-covariance is able to develop the relationship between PAH and PM2.5. However, as seen through predictive maps and through cross validation, the cokriging incorporation of PM2.5 contributes little in terms of predictive capacity. Linear regression is another intuitive choice with collocated data. The LR BME method shows a marked improvement visually and through estimation accuracy. The LR method is able to estimate PAH at PM2.5 space/time locations using an optimized neighborhood customized for each PAH. However, LR performed consistently worse than the MF method. The LR method requires the estimation of 2 parameters (i.e. β_0 and β_1). We hypothesize that this increase in the number of parameters makes the LR model less parsimonious, requiring more paired PAH and PM2.5 to optimize the estimation neighborhood. The PAH paired data is then outside of the relevant air shed of estimation. The LR method also assumes that PM2.5 and PAH follows a simple linear regression relationship. However, the more direct MF approach may be closer to the true relationship between paired PAH and PM2.5 data.

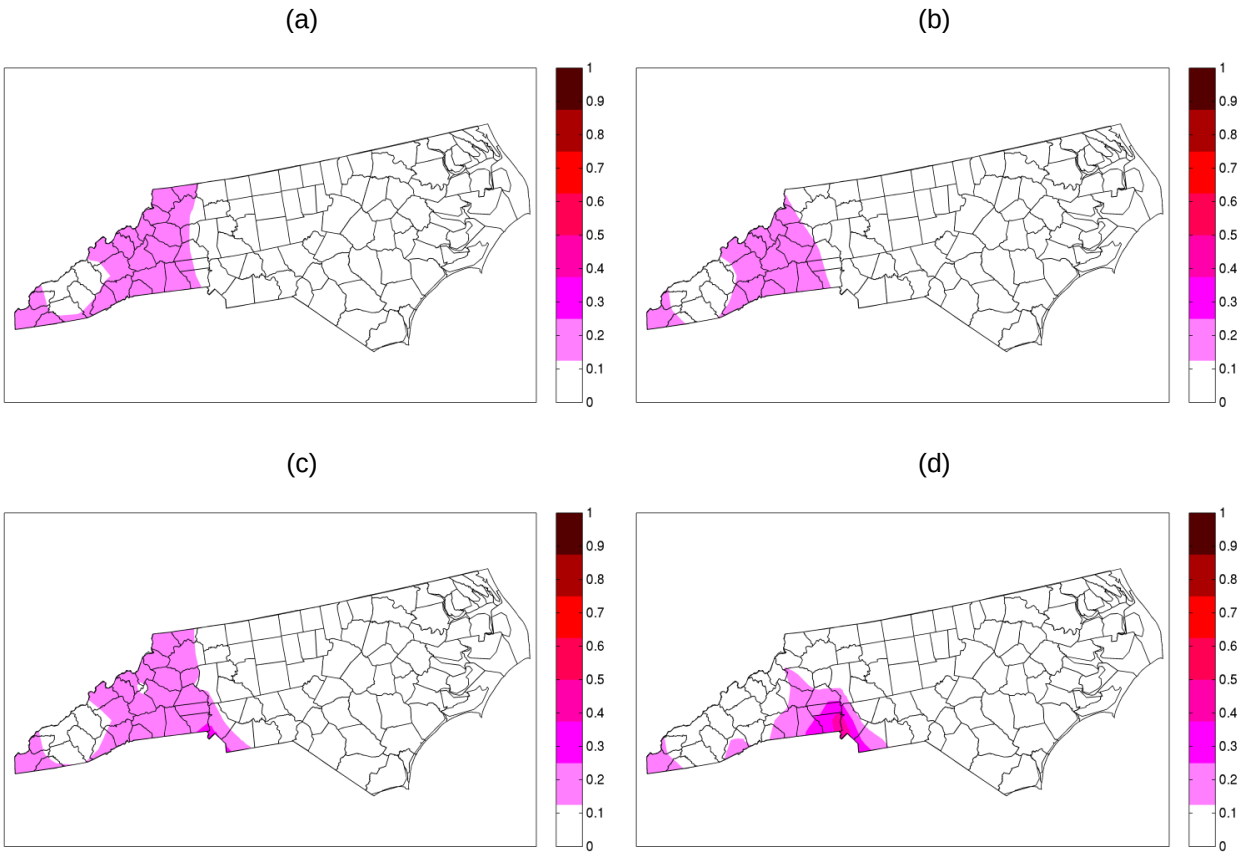


Figure 4.2. Probability of exceedance. Probability of annual benz(a)pyrene exceeding $0.25\ ng/m^3$ across North Carolina in 2005 as predicted by (a) kriging, (b) cokriging, (c) BME linear regression and (d) BME mass fraction.

4.3.4 Probability of exceedance

In a geostatistical framework, predictions come in the form of a Probability Distribution Function (PDF) with a corresponding mean and variance. With this PDF, the probability of exceeding a given value can be calculated. An annual benz(a)pyrene concentration of $0.25\ ng/m^3$ has been suggested in the United Kingdom (Ravindra et al., 2008). With this standard in mind, the probability of exceeding this cutoff was calculated on annual benz(a)pyrene concentrations in North Carolina in 2005 for each prediction method (Fig. 4.2). As seen from Fig. 4.1, overall PAH concentration decrease across methods, thus the probability of exceeding the $0.25\ ng/m^3$ cutoff in turn decreases across methods. Across methods, the region of the state with the relatively highest probability of exceedance is maintained as Western North Carolina as well as the border with South Carolina. Interestingly, these same regions correspond to high levels of benzo(g,h,i)perylene on March 11, 2005 (Fig. 4.1). Across all prediction methods, the probability of exceedance remains low with the maximum probability remaining below 0.50. The distribution of the

probability of exceedance becomes increasingly skewed across methods. The prediction locations for kriging have the largest median probability of exceedance (0.074) across the methods. The cokriging method has the lowest maximum probability of exceedance (0.16) across the annual prediction locations and has the lowest proportions of probabilities under 0.1 (61%). While the median exceedance probability drops for the LR BME method (0.067) and MF BME method (0.059), their maximum probabilities increase (0.33 and 0.44, for LR and MF, respectively). The BME methods also contain the highest proportional of probabilities above 0.2 (10% and 8% for LR and MF, respectively). Thus, we see the BME methods are better able to differentiate areas of high and low probabilities of exceedance. The MF BME was best able to distinguish the maximum probability of exceedance. Through having more realistic ambient predictive gradients, the MF BME method becomes an effective tool to identify areas of exceedance of different PAH concentrations. The border between North and South Carolina showed the highest probability of exceedance of benz(a)pyrene. This calls for more monitoring in that area as well as a further investigation of associated health effects (e.g. lung cancer) in the region.

Table 4.3. Mean difference in PAH near versus far from fires. 95% confidence intervals comparing the mean difference in predicted PAH near (within $100\ km$) versus far ($> 100\ km$) from fires for each of the 9 PAH and Total PAH across the 4 prediction methods. Units are in ng/m^3. *mean difference is statistically significant (p-value≤ 0.05), $^{\#}$mean difference > 0.

PAH	Kriging	Cokriging	Linear Regression	Mass Fraction
benz(a)anthracene	(-0.004944,-0.002174)*	(-0.002610,-0.000011)*	(-0.001264,0.001066)	(0.001575,0.004379)*,#
chrysene	(-0.006666,-0.003529)*	(-0.003737,-0.000828)*	(-0.000940,0.001671)	(0.002070,0.005392)*,#
benzo(b)fluoranthrene	(0.003979,0.011092)*,#	(0.003779,0.011006)*,#	(0.010934,0.022334)*,#	(0.023594,0.030234)*,#
benzo(k)fluoranthrene	(0.003136,0.006467)*,#	(0.002319,0.005008)*,#	(0.005271,0.007937)*,#	(0.007804,0.010768)*,#
benzo(e)pyrene	(-0.002922,0.002229)	(-0.003171,0.001705)	(0.005219,0.009804)*,#	(0.018337,0.024890)*,#
benzo(a)pyrene	(-0.003827,0.001843)	(-0.006237,-0.000842)*	(0.002226,0.013734)*,#	(0.005135,0.010157)*,#
indeno(1,2,3-c,d)pyrene	(0.018667,0.030507)*,#	(0.017301,0.028689)*,#	(0.020396,0.032370)*,#	(0.047856,0.061099)*,#
benzo(g,h,i)perylene	(0.030368,0.042666)*,#	(0.025449,0.036588)*,#	(0.027206,0.040613)*,#	(0.031329,0.040613)*,#
dibenzo(a,h)anthracene	(-0.020700,-0.013553)*	(-0.017140,-0.010665)*	(-0.004795,0.002157)	(0.001898,0.008765)*,#
Total PAH	(0.022819,0.067469)*,#	(0.021261,0.065714)*,#	(0.062870,0.103406)*,#	(0.171993,0.230460)*,#

4.3.5 Association with fires

The mean difference in PAH predictions (for the 9 PAHs and Total PAH) as calculated through the 4 prediction methods was found through a two-sampled t-test comparing areas near ($\leq 100\ km$) and far ($> 100\ km$) from known fire locations (Table 4.3). For the MF method, all 9 PAHs and Total PAH showed a statistically significantly difference between prediction near versus far from fires, for the MF method 6 PAHs and Total PAH showed a significant difference and for both kriging and cokriging 4 PAHs and Total PAH showed a significant difference greater than zero. Of those PAHs that showed a significant difference and whose differences were positive, the MF method had the largest differences across 8 PAHs (benzo(g,h,i)perylene being the exception) and Total PAH. Known fire locations for March 11, 2005 are marked along with a $100\ km$ radial buffer surrounding each location (Fig. 4.1). Across prediction methods, PAH concentrations are higher within/near these buffers. Indeed, benzo(g,h,i)perylene (depicted in Fig. 4.1) was one of the 4 PAHs (along with Total PAH) that showed both a significant, positive difference across all prediction methods.

This work investigates ambient concentrations of a particular set of particle-bound PAHs. Concentrations alone cannot distinguish sources. However, there are PAH ratios associated with certain sources. The diagnostic ratio of indeno(1,2,3-c,d)pyrene /(indeno(1,2,3-c,d)pyrene + benzo(g,h,i)perylene)=0.62 is associated with wood burning.(Ravindra et al., 2008) This ratio was calculated across March 11, 2005 across all 4 prediction methods (Fig. 4.3). The ratio for kriging and cokriging remained under 0.62 across all prediction locations of the day. There is little variation of this ratio across the day for kriging and cokriging. This ratio increases and approaches 0.62 for the BME methods. There is more variation of this ratio for the LR BME method, possibly implying better differentiation between PAH sources. MF BME has the largest variation of the PAH diagnostic ratio, with the largest number of predictions near the 0.62 value. Both kriging and cokriging have $\leq 1\%$ of prediction ratios for the day around 0.62 (0.62 ± 0.05), LR BME has 3% of prediction ratios around 0.62 and MF BME has 10% of prediction ratios around 0.62.

The MF BME method was better able to distinguish higher significant differences in PAH concentrations near known fire locations compared with other prediction methods. Of the four prediction methods the MF was the only method that showed statistically significant differences that were positive

around areas with fires across all 9 PAHs and Total PAH. Although the same buffer size was used for all the fires, the significance implies an association. Depending on the acreage burned from a fire, the type of vegetation burned and the duration of the fire, the smoke produced may be long lasting and may have long range transport. With this in mind, it is also important to take into consideration nearby fires in days previous to observed PAH concentrations. A diagnostic ratio can be used in conjunction with known sources. Diagnostic ratios should not be used in isolation. However, when used along known fire locations, it can strengthen the association between PAH concentration and its known sources. Gathering information about wildfire smoke has become increasingly more important as the number of large wildfires have increased in recent years (Dennison et al., 2014).

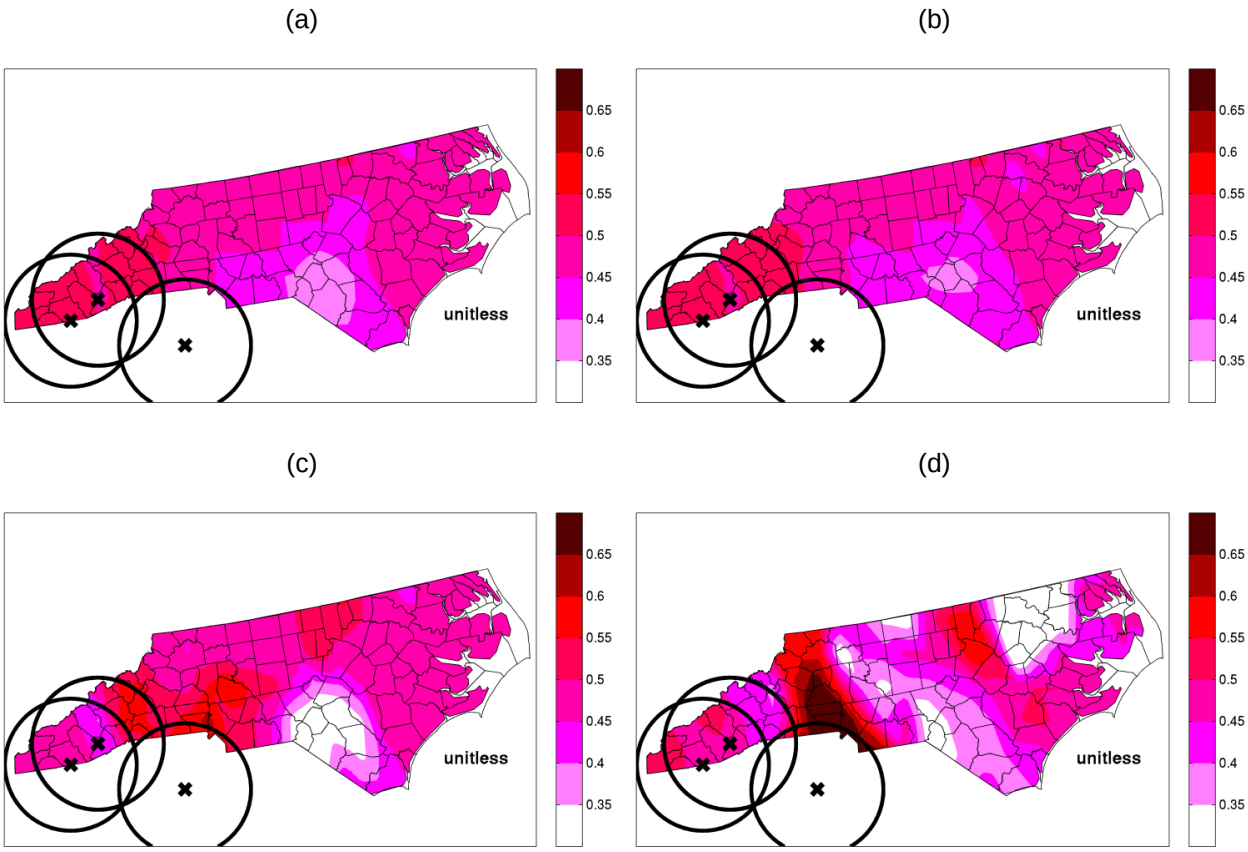


Figure 4.3. PAH ratios. Ratio of indeno(1,2,3-c,d)pyrene/(indeno(1,2,3-c,d)pyrene+benzo(g,h,i)perylene) on March 11, 2005 in North Carolina across the 4 prediction methods: (a) kriging, (b) cokriging, (c) BME linear regression, (d) BME mass fraction. X's mark known fires for that day with a $100\ km$ buffer.

4.3.6 Overall contributes and concluding statements

The MF BME method allows for straightforward predictions of PAHs to be used for exposure assessments. There are a plethora of studies exploring the association between ambient PM2.5 and

various health endpoints (Beelen et al., 2007; Krewski et al., 2009; Pope et al., 2004). However, there are far less studies that explore ambient PAH exposures and associated health effects. Occupational inhalation exposures have been more thoroughly investigated with health outcomes including lung cancer (Kim et al., 2013). Few studies have investigated chronic ambient concentrations of PAHs. Many of the epidemiologic studies that have been explored investigate respiratory illnesses such as lung cancer and pulmonary function (Guerreiro et al., 2016; Kim et al., 2013; Padula et al., 2015). However, these studies are small in number. The current state of the literature is lacking in epidemiologic studies. The lack of long-term ambient concentrations to PAHs may be due to inadequate exposure data. Analyzing PM filters for specific PAHs can be very costly, making it difficult of obtain larger amounts of observed data needed for exposure assessment (Pleil et al., 2004). The MF BME method allows for an efficient and cost efficient way to utilized minimal PAHs observed data. The MF BME method can be easily utilized to fill in this clear gap in the literature. Tied with corresponding health data, ambient predictions calculated through the MF BME method could be used to assign exposure. Health metrics can then be calculated from the exposures. This opens to door to investigate possible health endpoints as well as assigning risk.

In conclusion, this work creates the first maps of ambient PAH concentration across the US state of North Carolina through the creation of the MF BME geostatistical method. This method developed a relationship between paired PAH and PM2.5 data in a manner that is a parsimonious and cost-effective that can be utilized in a data poor environment. The MF BME method outperforms more traditionally used geostatistical methods and has the ability to elucidate a significant association between PAH predictions and known fire locations. The MF BME method has the potential to be used to assign exposure in epidemiologic analyses to fill in the significant knowledge gap currently existing in the literature between PAH exposures and potential health outcomes.

CHAPTER 5: CONCLUDING REMARKS

The goal of this work was to develop a data fusion methodology combining error-corrected gridded modeled with observed ambient air pollution data to be used in a geostatistical context. Combining of different data sources was performed in a data rich environment with daily PM2.5 and a data poor environment with PAH. This was achieved in a data rich environment through the development of the Regionalized Air quality Model Performance (RAMP) method and achieved in a data poor environment using the Mass Fraction (MF) method. Both were implemented in the Bayesian Maximum Entropy (BME) geostatistical framework. Generally speaking, the data fusion methods were able to improve upon more typically used and more well-known methods such as kriging and an implementation of the Downscaler method. In a data rich environment, the data fusion method was able to have the accuracy of observed data with the spatial refinement of modeled data. In the data poor environment, the data enrichment method was able to incorporate information about the joint relationship between PAH and PM2.5 data.

Chapters 2 developed the RAMP method for the model performance evaluation and subsequent error-correction of CMAQ. Validation results demonstrated that the RAMP parameters were better able to predict observed PM2.5 concentrations than other methods with up to a 22.1% reduction in MSE. A Simulation of the RAMP method shows that the RAMP method is better able to capture the homogeneity of simulated model performance compared with other methods. The RAMP method allows to a more flexible relationship between modeled and observed data compared to other data fusion method by not assuming a linear or homoscedastic relationship. The RAMP method was able to identify six regions of the country with high error and was able to delineate and quantify the sizes of these regions. Future research includes investigating more recent versions of CMAQ for the BME RAMP method and visualizing geographical changes in error. In Chapter 3, the RAMP corrected CMAQ data were incorporated with observed data in the BME framework. The BME RAMP method was compared with kriging and an implementation of the Downscaler method. In a LOOCV, the BME RAMP method lowered

the MSE between 2%-33% compared with kriging. Improvements were even more pronounced when performing the cross validation far from monitoring stations and when results were stratified by RAMP metrics. The RAMP method was able to have 6-7 times the level of spatial refinement compared with kriging in a few different areas of the county and reduce the prediction variance compared with kriging.

Chapter 4 developed the MF method across North Carolina in 2004 to predict PAH from only a handful of paired PAH and PM2.5 values. The MF BME method was compared with kriging, cokriging and a LR BME method. A soft data validation was performed to optimize with number of paired PAH and PM2.5 used to estimate the soft data. In the MF soft data validation, the number of paired points was smaller than the LR, indicating that the MF method is more specific to the estimation location of interest and is therefore more geographical relevant to the air shed in question. In a LOOCV, the MF method produced in the lowest MSE of 7 out of 9 PAHs investigated. The MF method is also able to show a statistically significant difference in PAH concentrations near known fire locations versus far from known fire locations. Other methods were not able to distinguish differences in PAH concentration by fire information. To the best of our knowledge, this is currently the most comprehensive map of PAH concentration for North Carolina during 2004.

Overall, this work was able to merge multiple data sets together to estimate PM2.5 in a data rich environment and PAH in a data poor environment into the BME geostatistical framework. Both pollutants achieved notable increases in spatial refinement and sizable increases in accuracy when comparing cross-validation statistics. Both PM2.5 and PAH were able to achieve high quality prediction maps and corresponding prediction uncertainties. This work has many implications for several diverse arrays of disciplines. Understanding model performance of regulatory models, especially in areas without monitoring, is important for regulatory agencies as well as model developers. Improvements in prediction accuracy are important to epidemiologists for reducing misclassification and revealing the underlying health measures that tie air pollution exposures to health endpoints. Geostatistical estimation of air pollution casts a wide net under which resides public health officials, epidemiologist, atmospheric chemists and the general public.

APPENDIX A: SUPPORTING INFORMATION FOR REGIONALIZED PM2.5 COMMUNITY MULTISCALE AIR QUALITY MODEL PERFORMANCE EVALUATION ACROSS A CONTINUOUS SPATIOTEMPORAL DOMAIN[4]

A.1 Model Performance Metrics

In the model performance literature typical definitions exist regarding nomenclature of certain metrics, namely, "bias" and "error". In the Statistics realm, metrics such as "bias" and "error" are defined differently. This work utilizes naming schemes of metrics that is consistent with Statistical language. In Table A.1 $\tilde{x}_i$ denotes a modeled value as some space/time point $\boldsymbol{p}_i$, $\hat{x}_i$ is its paired observed value (i.e. observed at the same space/time location) and $e_i = \tilde{x}_i - \hat{x}_i$ is the corresponding error. In Table A.2 $\hat{\sigma}_i$ denotes an estimate of the error standard deviation, which in our work is obtained using $\sqrt{\lambda_2(\boldsymbol{p})}$.

Table A.1. Table of commonly used model performance evaluation statistics used in the CMAQ literature. The left column displays the typical nomenclature used and the right column displays the nomenclature used in this work. The third column states whether the metrics quantifies systematic or random error. The second half of the table displays metrics less commonly used in the literature.

Metric name used in the CMAQ literature	Definition	Systematic / Random	Metric Name used in this Work
Regulatory Performance Metrics Used in Air Quality Modeling			
# of data pairs	n	NA	**# of data pairs**
Mean observation value	$\frac{1}{n}\sum_{i=1}^{n}(\hat{x}_i) = \bar{\hat{x}}$	NA	**Mean observation value**
Mean simulation value	$\frac{1}{n}\sum_{i=1}^{n}(\tilde{x}_i) = \bar{\tilde{x}}$	NA	**Mean modeled value**
Mean bias	$\frac{1}{n}\sum_{i=1}^{n}(\tilde{x}_i - \hat{x}_i)$	Systematic	**Mean error**
Normalized bias	$100\% \times \frac{1}{n}\sum_{i=1}^{n}\left(\frac{\tilde{x}_i - \hat{x}_i}{\hat{x}_i}\right)$	Systematic	**Mean normalized Error**
Normalized mean bias	$100\% \times \frac{\sum_{i=1}^{n}(\tilde{x}_i - \hat{x}_i)}{\sum_{i=1}^{n}\hat{x}_i}$	Systematic	**Normalized mean Error**
Fractional bias	$100\% \times \frac{1}{n}\sum_{i=1}^{n}\left(\frac{\tilde{x}_i - \hat{x}_i}{0.5 \times (\tilde{x}_i + \hat{x}_i)}\right)$	Systematic	**Fractional error**
Mean Error	$\frac{1}{n}\sum_{i=1}^{n}\lvert\tilde{x}_i - \hat{x}_i\rvert$	Systematic / Random	**Mean absolute error**

[4] This appendix was submitted as the supporting information of an article to the journal Atmospheric Environment. Reyes, Jeanette M., Xu,Yadong, Vizuete, William, Serre, L. Marc. Regionalized PM2.5 Community Multiscale Air Quality model performance evaluation across a continuous spatiotemporal domain.

Normalized Error	$100\% \times \frac{1}{n}\sum_{i=1}^{n}\left(\frac{\lvert\tilde{x}_i - \hat{x}_i\rvert}{\hat{x}_i}\right)$	Systematic / Random	**Mean normalized absolute error**
Normalized Mean Error	$100\% \times \frac{\sum_{i=1}^{n}\lvert\tilde{x}_i - \hat{x}_i\rvert}{\sum_{i=1}^{n}\hat{x}_i}$	Systematic / Random	**Normalized mean absolute error**
Fractional error	$100\% \times \frac{1}{n}\sum_{i=1}^{n}\left(\frac{\lvert\tilde{x}_i - \hat{x}_i\rvert}{0.5 \times (\tilde{x}_i + \hat{x}_i)}\right)$	Systematic / Random	**Fractional absolute error**
Correlation	$\frac{\sum_{i=1}^{N}(\tilde{x}_i - \bar{\tilde{x}})(\hat{x}_i - \bar{\hat{x}})}{\sqrt{\sum_{i=1}^{N}(\tilde{x}_i - \bar{\tilde{x}})^2}\sqrt{\sum_{i=1}^{N}(\hat{x}_i - \bar{\hat{x}})^2}}$	Random	**Correlation**
Less Commonly Used Regulatory Performance Metrics			
Correlation squared	r^2	Random	**Correlation squared**
Standard bias	$\sqrt{V[\tilde{x}_i - \hat{x}_i]}$	Random	**Standard error**
Mean squared bias	$\frac{1}{n}\sum_{i=1}^{n}(\tilde{x}_i - \hat{x}_i)^2$	Systematic / Random	**Mean squared error**
Root mean squared bias	$\sqrt{\frac{1}{n}\sum_{i=1}^{n}(\tilde{x}_i - \hat{x}_i)^2}$	Systematic / Random	**Root mean squared error**
Normalized root mean squared bias	$\frac{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(\tilde{x}_i - \hat{x}_i)^2}}{\frac{1}{n}\sum_{i=1}^{n}\hat{x}_i}$	Systematic / Random	**Normalized root mean squared error**
Mean bias/standard bias	$\frac{1}{n}\sum_{i=1}^{n}(\tilde{x}_i - \hat{x}_i) / \sqrt{V[\tilde{x}_i - \hat{x}_i]}$	Systematic	**Mean error/standard error**
Mean bias squared/mean squared bias	$\frac{(\sum_{i=1}^{n}(\tilde{x}_i - \hat{x}_i))^2}{\sum_{i=1}^{n}(\tilde{x}_i - \hat{x}_i)^2} = ME^2/MSE$	Proportion of Systematic	**Mean error squared/mean squared error**
Variance of bias/mean squared bias	$\frac{V[\tilde{x}_i - \hat{x}_i]}{\frac{1}{n}\sum_{i=1}^{n}(\tilde{x}_i - \hat{x}_i)^2}$	Proportion of Random	**Variance of errors/mean squared error**

Table A.2. Table of model performance evaluation statistics used when the estimate $\tilde{x}_i$ has a corresponding variance $\hat{\sigma}_i^2$. These metrics are used in the validation statistics**.**

BME Metric	Definition
Variance standardized error	$V\left[\frac{\tilde{x}_i - \hat{x}_i}{\hat{\sigma}_i}\right]$
Root mean squared standardized error	$\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\frac{\tilde{x}_i - \hat{x}_i}{\hat{\sigma}_i}\right)^2}$
Mean root variance	$\frac{1}{n}\sum_{i=1}^{n}\left(\sqrt{\hat{\sigma}_i}\right)$

A.2 Data

A.2.1 Observed Data

The daily PM2.5 concentration for each monitoring site/day during 2000-2002 were constructed based on raw monitoring data from monitoring stations measuring either hourly or daily PM2.5 concentrations using the procedure described below.

PM2.5 monitoring data (raw data) sampled during the study period (2000-2002) were obtained from the Air Quality Systems (AQS) database maintained by the EPA, a repository of the monitoring data collected across various monitoring networks. PM2.5 data are available in a few data files on AQS depending on the source of data. These files are described in AQS as follows: 1) daily PM2.5 local conditions, 2) daily PM fine speciation from the Chemical Speciation Network (CSN) monitoring network and 3) daily PM fine speciation from the Interagency Monitoring of Protected Visual Environments (IMPROVE) monitoring network. Within each data file, the methodologies used to measure PM2.5 are defined using a parameter code which takes the following values: 1) 88101 for daily and hourly PM2.5 concentrations measured using a Federally Referenced Method (FRM) and 2) 88502 for PM2.5 Air Quality Index (AQI) values that provide acceptable measurements of PM2.5 concentrations in that they are comparable to FRM measurements. Data from the parameter 88502 are also known as Tapered Element Oscillating Microbalance (TEOM) data.

Hourly PM2.5 data were averaged into daily PM2.5 if at least 18 out of 24 hours were measured for a given day/monitor. Otherwise, a daily average was not constructed. More than 99.9% of hourly records were reported every hour on the hour. However, there were several records not reported on the hour. These hourly records were removed before constructing daily concentrations. All observations sampled at monitors whose measurement scale was “Microscale” were removed.

At each monitoring site with multiple monitors, the collocated daily concentrations recorded at any given day were combined using the following procedure to produce a constructed daily concentration for that site/day. First, priority rank scores were assigned to each collocated daily concentration based on its data source and type as follows:

Table A.3. Ranking scores used for averaging collocated PM2.5 values for a given site/day

Rank 1	FRM daily PM2.5
Rank 2	TEOM daily PM2.5 from CSN
Rank 3	TEOM daily PM2.5 from IMPROVE
Rank 4	TEOM hourly PM2.5

If the collocated concentrations for a given site/day had varying priority ranks, then only the concentration with the highest rank (i.e. the smallest priority score) was retained. If there were more than one collocated daily concentrations with the highest priority rank, then these daily concentrations were averaged to produce a single daily concentration at that site/day.

A.2.2 Modeled Data

Daily concentrations for PM2.5 were also constructed from modeled CMAQ data. CMAQ inputs emissions and meteorological data which are then translated into complex chemical processes to estimate ambient air pollution over gridded geographical boundaries for different time steps. The modeled data used for this work were available at a 36km resolution every hour for the years 2001 and 2002 across the continental US. Data are projected using a Lambert Conic Conformal (LCC) projection.

Table A.4. Description of available CMAQ modeling data

Year	Model	Domain	Resolution	Source	Received date
2001	CMAQ v4.5	The contiguous US	36km	EPA	08-26-2011
2002	CMAQ v4.5	The contiguous US	36km	CENRAP	08-02-2011

The CMAQ data have full spatial and temporal coverage for the continental US (Table A.4). All modeled runs were done using hindcasting. Daily modeled values were constructed by averaging the 24 hourly modeled values for a given grid location/day. To reconcile the spatial misalignment of defining the modeled concentration over an area (i.e. the modeled concentration over a grid), the location of modeled values are defined by the centroid of each grid.

A.3 Choice of S-Curve Parameters for the RAMP analysis

We conducted a visual analysis to select the parameters used to construct the λ_1 and λ_2 S-curves in the RAMP analysis. In the CAMP analysis λ_1 and λ_2 are calculated across the domain $\mathcal{D}$. The main limitation of the CAMP analysis is that the S-curves do not capture the variability of model performance at a fine spatial scale. To address this issue the RAMP analysis regionalizes the calculation to a space/time

region $\mathcal{R}(\boldsymbol{p})$ that consists of paired modeled/observed values at the n closest stations within T days of $\boldsymbol{p}$. Our aim is to capture fine scale spatial variability in the S-curves. In order to achieve the finest spatial resolution possible we choose $n = 3$. Using 3 proximal stations define the smallest region $\mathcal{R}(\boldsymbol{p})$ possible near any space/time location $\boldsymbol{p}$ of interest, which provides a description of model performance at the finest spatial resolution possible for any given location of interest. The time window was thus increased to ± 180 days to allow approximately 150 paired values needed in the creation of the S-curve. To test whether $T = \pm 180$ days and $n = 3$ leads to a stable analysis we performed a validation and stochastic simulation analysis that demonstrates that RAMP produces λ_1 and λ_2 values that outperforms CAMP. We conducted a visual sensitivity analysis by increasing n and inspecting whether the λ_1 and λ_2 maps change appreciatively as n increases (Fig. A.2). As seen in this figure, the λ_1 and λ_2 maps do not change appreciatively as n increases from 3 to 6. The same result is obtained using other values of n between 3 and 6 (results not shown). This demonstrates that the parameters chosen ($T = \pm 180$ days and $n = 3$) are as spatially specific as possible while still maintaining a stable estimation of λ_1 and λ_2.

Another choice for the number of station n would be to use a fixed radius r, and select all stations within r of the location $\boldsymbol{p}$ of interest. We found that in order to achieve a stable estimate of λ_1 and λ_2, the radius has to be set to a long distance r such that at least 3 stations are included in the most sparsely monitored area of the continental US. When moving to densely monitored areas, the number of station n within the fixed r radius becomes so large that essentially the RAMP method becomes equivalent to the CAMP method in these areas, and as a result this approach fails to assess model performance at fine spatial resolution.

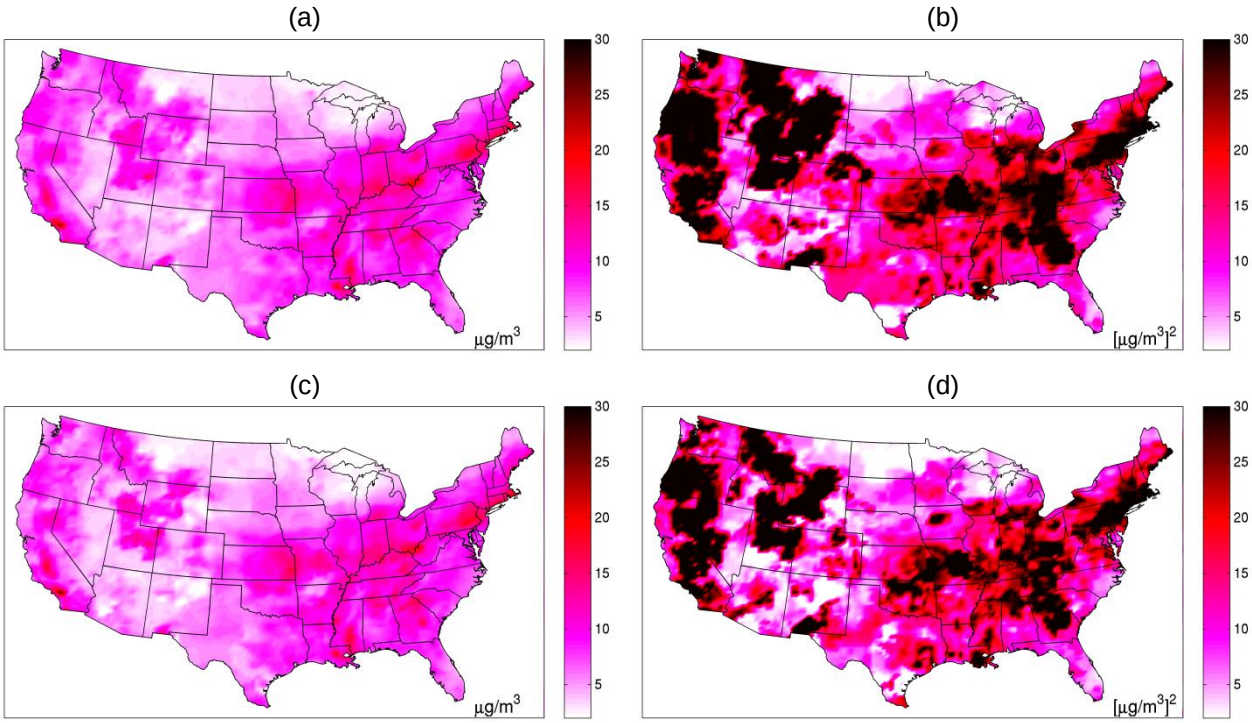


Figure A.1. Maps of $\lambda_1(p)$ and $\lambda_2(p)$ across the US on July 1, 2001 calculated using the RAMP method with two sets of S-curve parameters. $\lambda_1(p)$ is displayed in (a) and (c). $\lambda_2(p)$ is displayed in (b) and (d). (a) and (b) are obtained using the 6 closest stations within 180 days of p, while (c) and (d) are obtained using the 3 closest stations within 180 days of p. No appreciable difference can be seen by comparing (a) and (b) against (c) and (d).

A.4 Model Performance Metrics for Different Fixed Modeled Values

The RAMP method allows for construction of model performance metrics as a function of both space/time region and arbitrary modeled values. Through the equation $ME(\tilde{x}_k; \mathcal{R}(\boldsymbol{p})) = \tilde{x}_k - \lambda_1(\tilde{x}_k; \mathcal{R}(\boldsymbol{p}))$ and an analogous equation for $VE(\tilde{x}_k; \mathcal{R}(\boldsymbol{p}))$, one can visualize how mean error and variance of error changes across the United States for a given day for different increasing modeled values. Mean error for PM2.5 decreases consistently across the United States from $5\ \mu g/m^3$ to $10\ \mu g/m^3$. CMAQ has consistent problems estimating PM2.5 at smaller concentrations and high concentrations. CMAQ performs best at mid-range values of PM2.5 (Fig. A.3). Standard error increases as a function of modeled value (Fig. A.4). Change in the standard deviation of errors demonstrates that model performance is non-homoscedastic. The RAMP method allows the flexibility of allowing model performance to be both non-linear and non-homoscedastic. Non-homoscedastic behavior is most clearly observed in the Great Lakes region of the county.

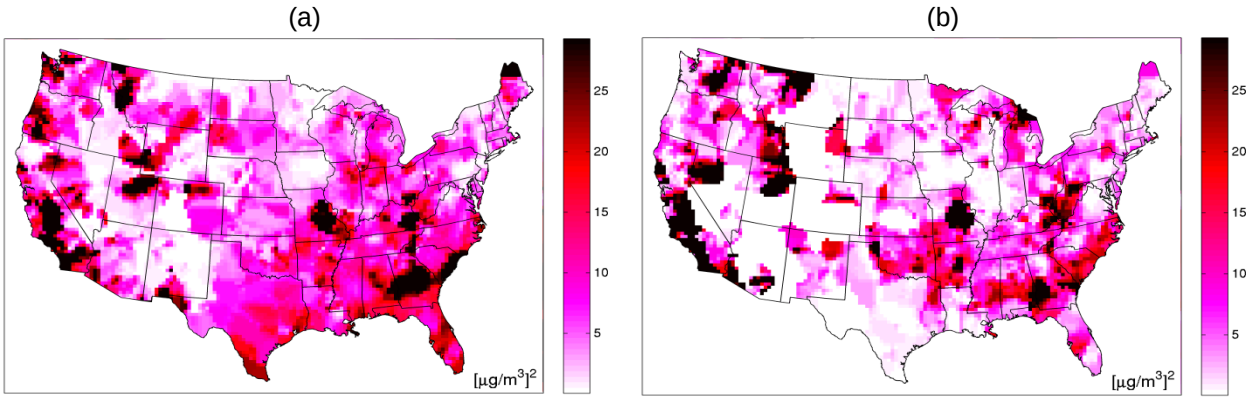


Figure A.2. Systematic error ($ME^2(\tilde{x}_k; \mathcal{R}(p))$) for fixed modeled values for (a) $5\ \mu g/m^3$ and (b) $10\ \mu g/m^3$ across the US on July 1, 2001 using the RAMP method as interpolated by S-Curves. Areas where $ME^2(\tilde{x}_k; \mathcal{R}(p))$ are not displayed are outside the range of modeled values for the corresponding S-Curve and are therefore not interpolated.

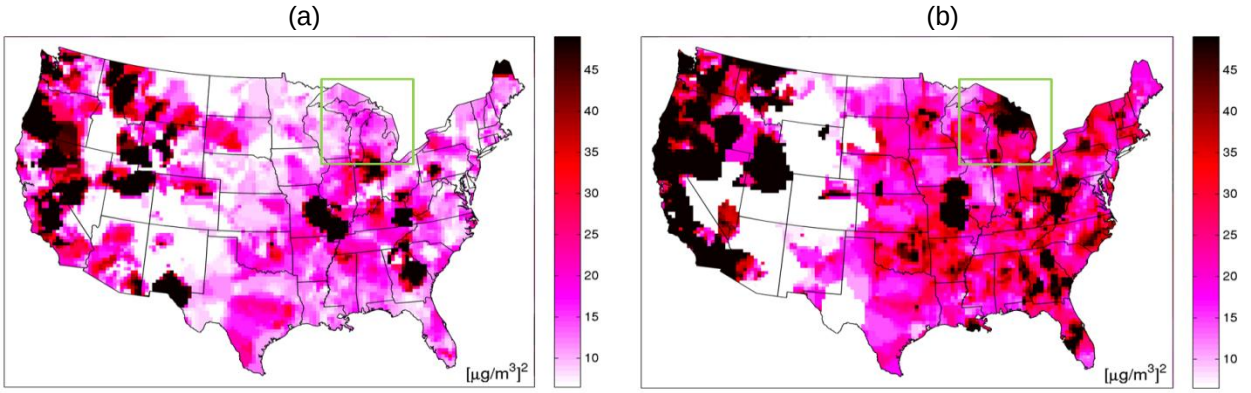


Figure A.3. Random error ($VE(\tilde{x}_k; \mathcal{R}(p))$) for fixed modeled values $5\ \mu g/m^3$ (a) and $10\ \mu g/m^3$ (b) across the US on July 1, 2001 using the RAMP method as interpolated by S-Curves. Areas where $VE(\tilde{x}_k; \mathcal{R}(p))$ are not displayed are outside the range of modeled values for the corresponding S-Curve and are therefore not interpolated. The boxed area in green corresponds to the Great Lakes region.

A.5 Maps of Other Model Performance Metrics

All metrics can be visualized for each CMAQ grid cell. When looking at systematic error and random error as a proportion of total error, most error coming from CMAQ can be defined as random (Fig. A.5a-b). Because such a large proportion of total error is coming from random error, visually, maps of random error and total error look similar (Fig. A.5c). MNAE (Fig. A.5e) is lowest in the Eastern part of the US where overall performance of CMAQ is known to be better. MNE has a large range illustrating the large potential normalized errors when observed values are small (Fig. A.5d). Visually, the maps of MNE look similar to $VE(\boldsymbol{p})/MSE(\boldsymbol{p})$.

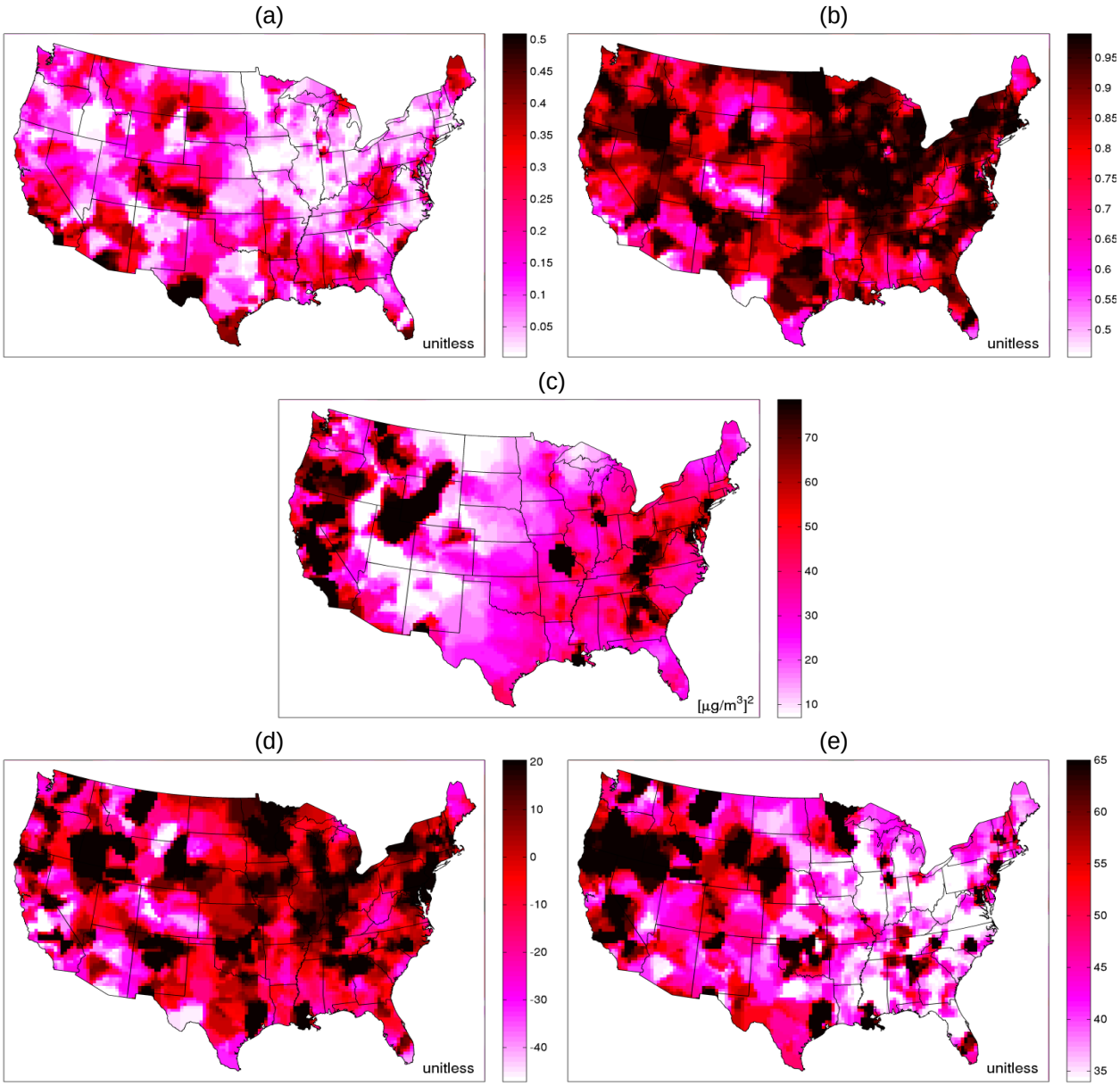


Figure A.4. Maps of various metrics across the continental United States on 07/01/2001 displaying (a) $ME^2(p)/MSE(p)$, (b) $VE(p)/MSE(p)$, (c) $MSE(p)$, (d) $MNE(p)$ and (e) $MNAE(p)$.

A.6 $\lambda_1(\tilde{x}_k;\mathcal{R}(\boldsymbol{p}))$ and $\lambda_2(\tilde{x}_k;\mathcal{R}(\boldsymbol{p}))$ for Different Fixed Modeled Values

Patterns of $\lambda_1(\tilde{x}_k;\mathcal{R}(\boldsymbol{p}))$ for increasing modeled values follow along similar lines of $ME(\tilde{x}_k;\mathcal{R}(\boldsymbol{p}))$. That is, performance is most consistent with mid-levels of PM2.5. Most systematic error is seen with low and high concentrations. The corresponding error correction of $\lambda_1(\tilde{x}_k;\mathcal{R}(\boldsymbol{p}))$ shows the lowest levels of variation for $5\ \mu g/m^3$, $10\ \mu g/m^3$ and $15\ \mu g/m^3$. Areas that show the highest levels of error correction in the US start in the Eastern US and move predominately to the Appalachian Mountain region of the country. Patterns of $\lambda_2(\tilde{x}_k;\mathcal{R}(\boldsymbol{p}))$ for increasing modeled values follow along similar lines of $VE(\tilde{x}_k;\mathcal{R}(\boldsymbol{p}))$.

That is, performance is most homogenous for lower levels of PM2.5. Most random error is seen with mid to high level concentrations.

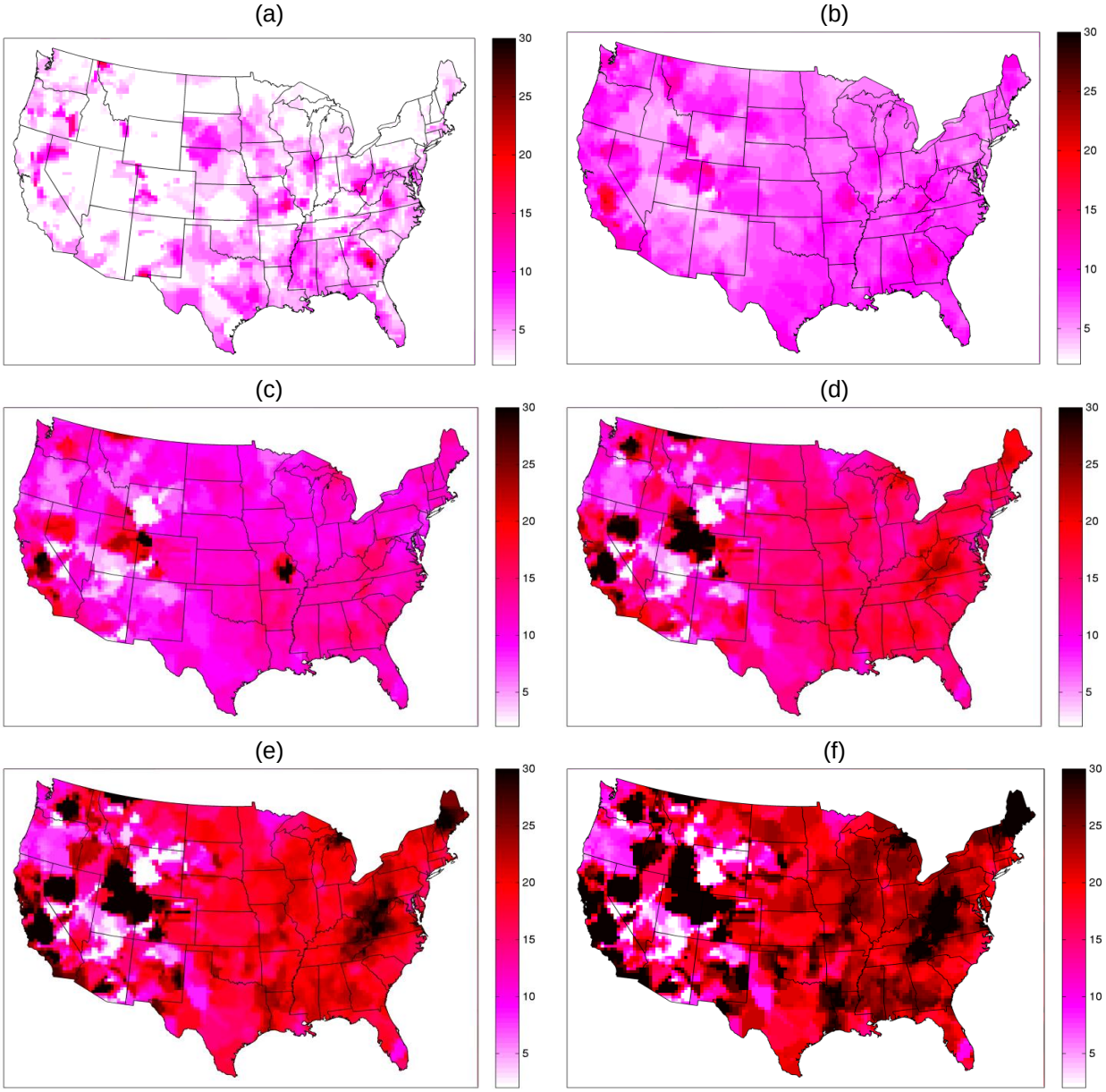


Figure A.5. Maps of $\lambda_1(\tilde{x}_k; \mathcal{R}(p))$ PM2.5 concentrations in increments of 5 $\mu g/m^3$ across the continental United States on 07/01/2001 with (a) being 0 $\mu g/m^3$ and (f) being 25 $\mu g/m^3$. Units for all figures are $\mu g/m^3$.

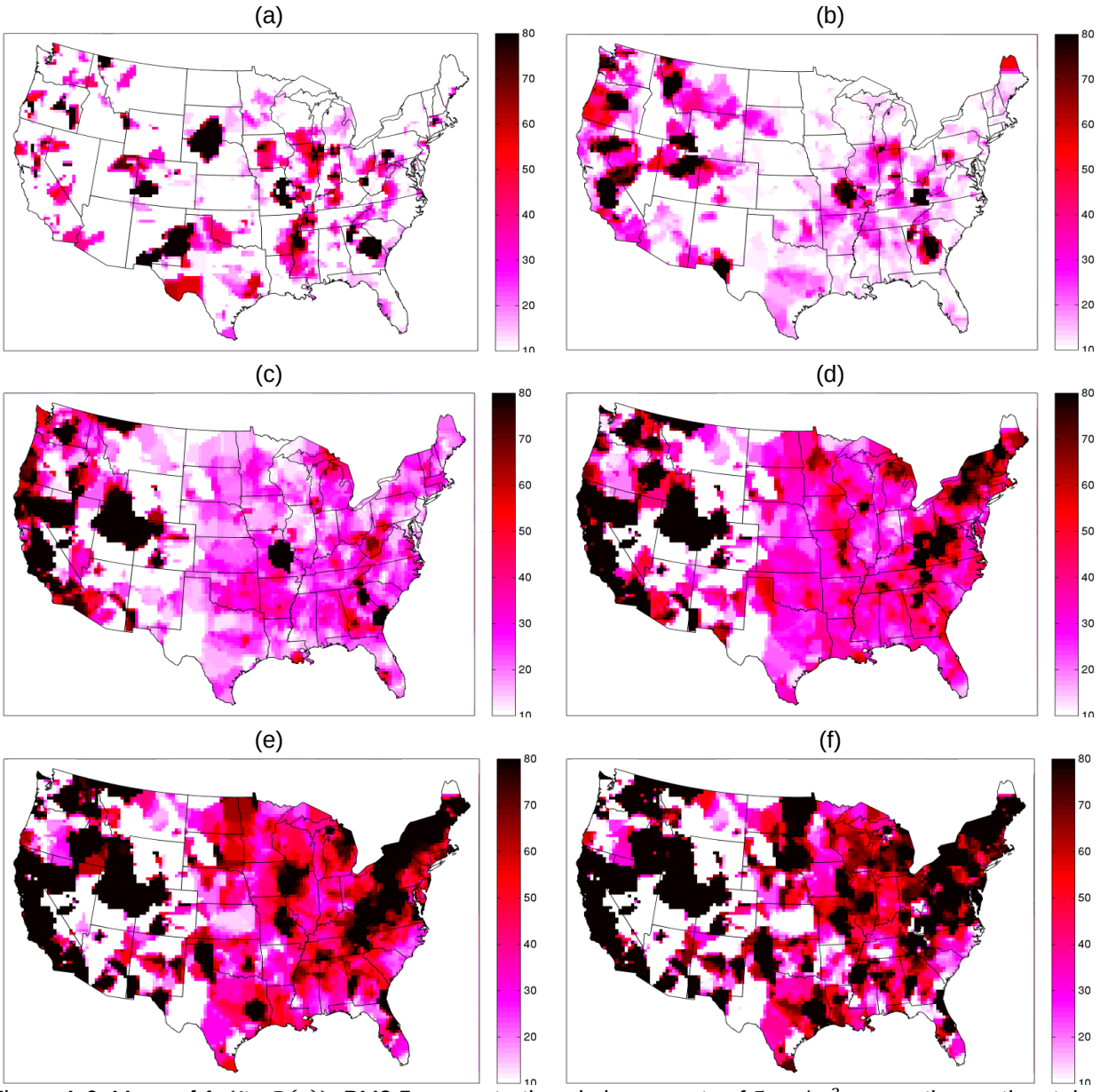


Figure A.6. Maps of $\lambda_2(\tilde{x}_k; \mathcal{R}(p))$ PM2.5 concentrations in increments of $5\ \mu g/m^3$ across the continental United States on 07/01/2001 with (a) being $0\ \mu g/m^3$ and (f) being $25\ \mu g/m^3$. Units for all figures are $(\mu g/m^3)^2$.

A.7 RAMP Stochastic Simulation

We do not know the true values of $\lambda_1(\boldsymbol{p})$ and $\lambda_2(\boldsymbol{p})$ because they are not directly measured. As a result it is not possible to validate the RAMP method based on true or measured $\lambda_1(\boldsymbol{p})$ and $\lambda_2(\boldsymbol{p})$ values. However we can use stochastic simulation to create a set of simulated values for $\lambda_1(\boldsymbol{p})$, $\lambda_2(\boldsymbol{p})$, $\tilde{x}(\boldsymbol{p})$ and $\hat{x}(\boldsymbol{p})$ that have the same statistical properties as the true values. These simulated values are taken as the simulated truth, which can then be used to validate the RAMP method.

To do this, $\lambda_1(\tilde{x}_k;\mathcal{R}(\boldsymbol{p}))$ and $\lambda_2(\tilde{x}_k;\mathcal{R}(\boldsymbol{p}))$ that were obtained in this work were selected as the "true" values. From the selected $\lambda_1(\tilde{x}_k;\mathcal{R}(\boldsymbol{p}))$ and $\lambda_2(\tilde{x}_k;\mathcal{R}(\boldsymbol{p}))$ and the CMAQ modeled values $\tilde{x}(\boldsymbol{p})$ used in this work, $\lambda_1(\boldsymbol{p})$ and $\lambda_2(\boldsymbol{p})$ were obtained by substituting $\tilde{x}_k$ with $\tilde{x}(\boldsymbol{p})$ in $\lambda_1(\tilde{x}_k;\mathcal{R}(\boldsymbol{p}))$ and $\lambda_2(\tilde{x}_k;\mathcal{R}(\boldsymbol{p}))$, respectively. Observed data were statistically simulated by randomly generating a stochastic realization $\hat{x}^*(\boldsymbol{p})\sim N(\lambda_1(\boldsymbol{p}),\lambda_2(\boldsymbol{p}))$. The set of $\lambda_1(\boldsymbol{p})$, $\lambda_2(\boldsymbol{p})$, $\tilde{x}(\boldsymbol{p})$ and $\hat{x}^*(\boldsymbol{p})$ values represent the simulated truth. The validation then consists of using the Constant, CAMP and RAMP methods to obtain ${\lambda_1}^*(\boldsymbol{p})$ and ${\lambda_2}^*(\boldsymbol{p})$ based on a re-estimation that uses *only* the $\tilde{x}(\boldsymbol{p})$ and $\hat{x}^*(\boldsymbol{p})$. The re-estimates ${\lambda_1}^*(\boldsymbol{p})$ and ${\lambda_2}^*(\boldsymbol{p})$ can be compared with the selected "true" $\lambda_1(\boldsymbol{p})$ and $\lambda_2(\boldsymbol{p})$ to evaluate the estimation accuracy of each method. Fig. A.8a-b show $\lambda_1(\boldsymbol{p})$ and $\lambda_2(\boldsymbol{p})$, respectively. Fig. A.8c-d show $\tilde{x}(\boldsymbol{p})$ and $\hat{x}^*(\boldsymbol{p})$, respectively, for the day of interest. The results of the re-estimation for the RAMP methods are shown in Fig. A.8e-f. As can been seen in the figure, the maps of ${\lambda_1}^*(\boldsymbol{p})$ and ${\lambda_2}^*(\boldsymbol{p})$ are very similar to those of $\lambda_1(\boldsymbol{p})$ and $\lambda_2(\boldsymbol{p})$, respectively, which demonstrates that the RAMP method is able to correctly estimate $\lambda_1(\boldsymbol{p})$ and $\lambda_2(\boldsymbol{p})$ across the space/time study domain based only on paired $\tilde{x}(\boldsymbol{p})$ and $\hat{x}^*(\boldsymbol{p})$ values.

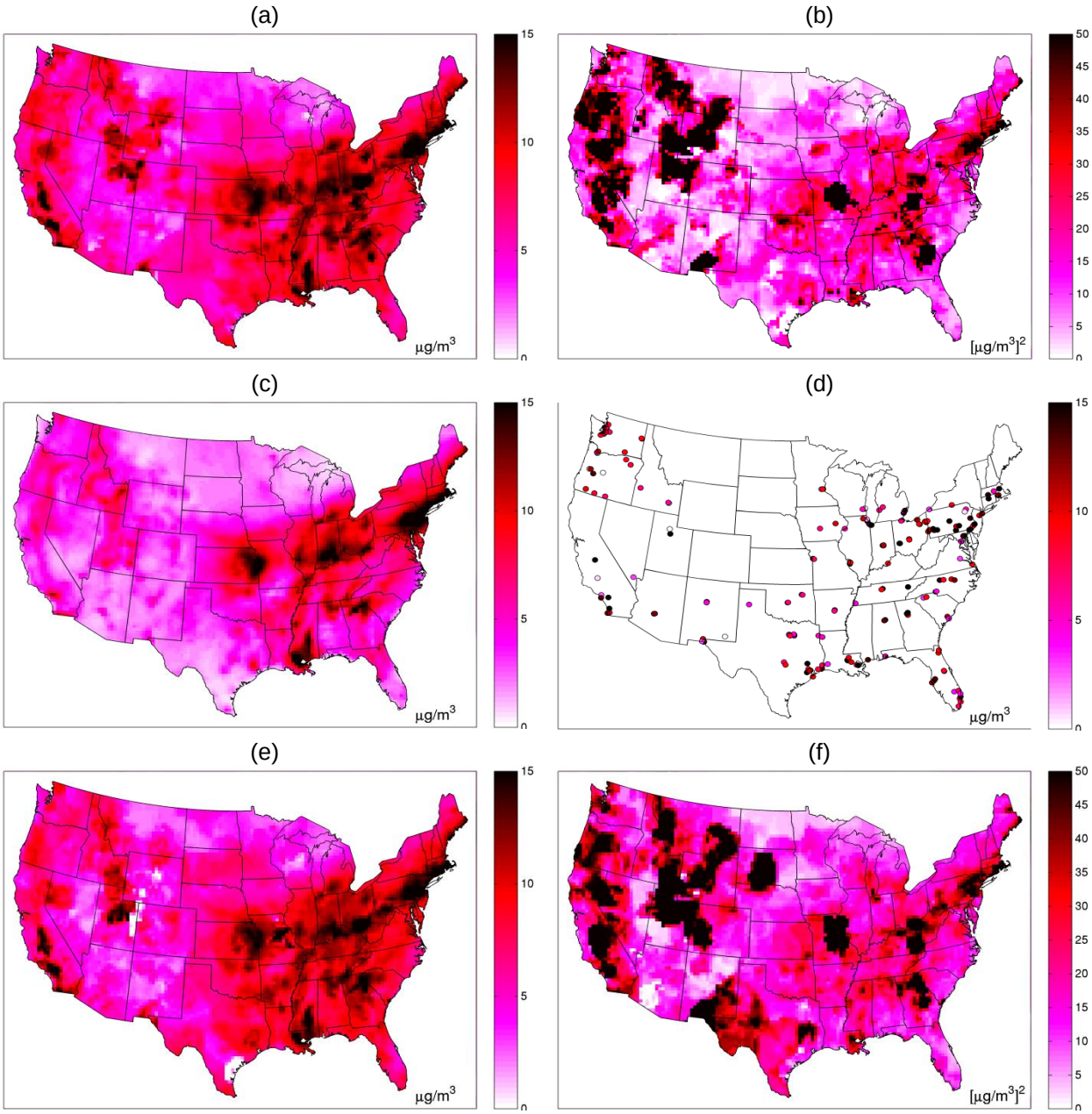


Figure A.7. Maps of stochastic simulation of daily PM2.5 across the continental United States on 07/01/2001. (a) Selected true $\lambda_1(p)$, (b) selected true $\lambda_2(p)$, (c) CMAQ modeled values $\tilde{x}(p)$ for that day, (d) stochastic realization $\hat{x}^*(p) \sim N(\lambda_1(p), \lambda_2(p))$, (e) $\lambda^*{}_1(p)$ and (f) $\lambda^*{}_2(p)$ re-estimated by the RAMP method based only on $\tilde{x}(p)$ and $\hat{x}^*(p)$.

In order to assess how well the RAMP method evaluates systematic errors compared to other methods, we show in Fig. A.9a $ME(\boldsymbol{p}) = \hat{x}(\boldsymbol{p}) - \lambda_{\boldsymbol{1}}(\boldsymbol{p})$ that was selected as the "truth" across the continental United States on 07/01/2001. We show in Fig. A.9b-d $ME^*(\boldsymbol{p}) = \hat{x}(\boldsymbol{p}) - \lambda_{\boldsymbol{1}}{}^*(\boldsymbol{p})$ values that were obtained using the Constant, CAMP and RAMP methods. The Pearson Correlation coefficient between $ME(\boldsymbol{p})$ and $ME^*(\boldsymbol{p})$ is 0.00% for the Constant, 24.0% for the CAMP and 76.1% for the RAMP

methods. These correlation values and the corresponding figures demonstrate that the Constant method is not able to capture any of the spatial variability in systematic errors and the CAMP method captures only some of the spatial variability of systematic errors. By comparison the RAMP method captures the spatial variability of systematic errors well.

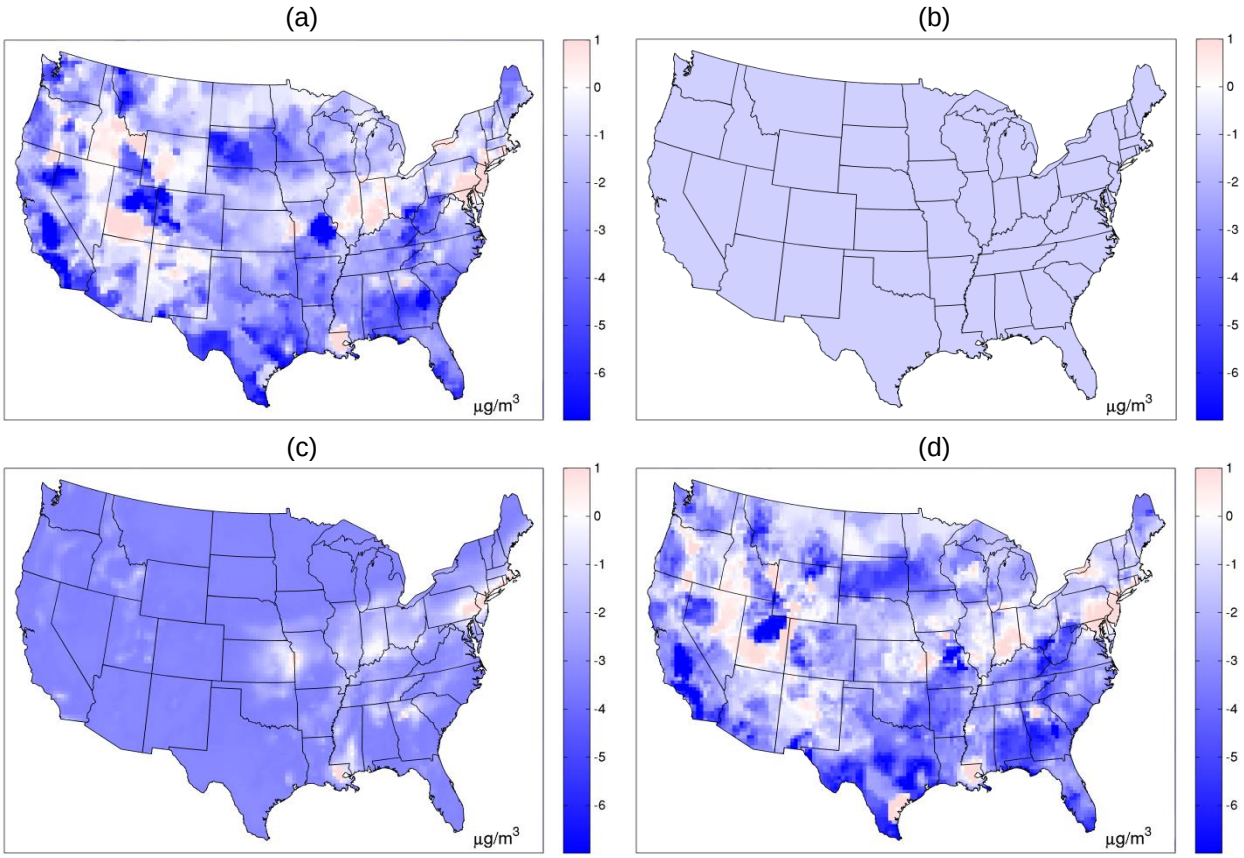


Figure A.8. (a) Map of the selected true $ME(p) = \hat{x}(p) - \lambda_1(p)$ for daily PM2.5 across the continental United States on 07/01/2001, and maps of the corresponding re-estimated $ME^*(p) = \hat{x}(p) - \lambda_1{}^*(p)$ obtained using the (b) the Constant method, (c) the CAMP method and (d) the RAMP method.

In order to assess how well the RAMP method is able to assess systematic errors compared to other methods, we show in Fig. A.10a $VE(\boldsymbol{p}) = \lambda_2(\boldsymbol{p})$ that was selected as the "truth" across the continental United States on 07/01/2001, and we show in Fig. A.10b-d $VE^*(\boldsymbol{p}) = \lambda_2{}^*(\boldsymbol{p})$ values that were obtained using the Constant, CAMP and RAMP methods. These maps indicates that the Constant method is unable to capture the spatial variability in random errors, the CAMP method captures only some of the spatial variability of random errors and, by comparison, the RAMP method is able to capture areas high and low random errors well.

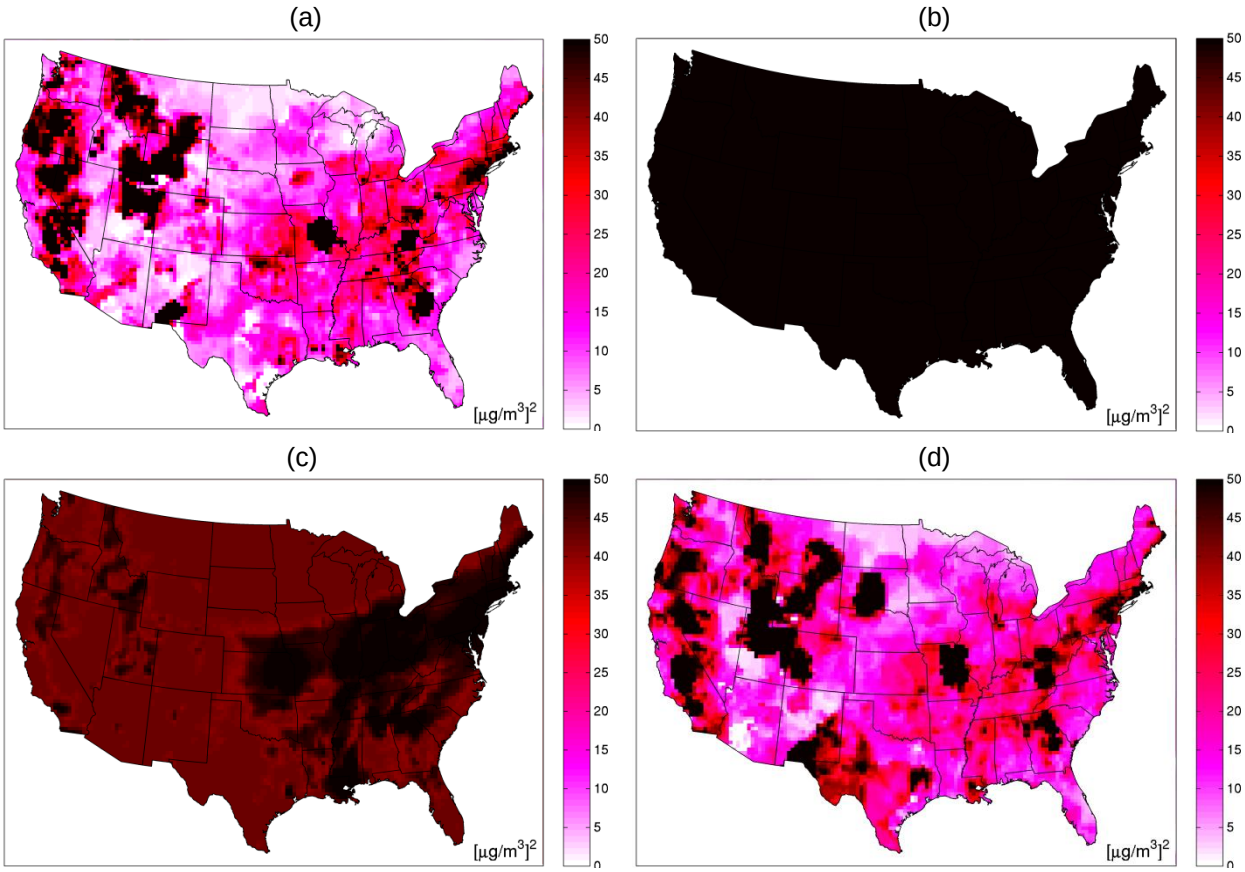


Figure A.9. (a) Map of the selected true $VE(p) = \lambda_2(p)$ for daily PM2.5 across the continental United States on 07/01/2001, and maps of the corresponding re-estimated $VE^*(p) = \lambda_2{}^*(p)$ obtained using the (b) the Constant method, (c) the CAMP method and (d) the RAMP method.

It should be noted that the $ME^*(\boldsymbol{p})$ and $VE^*(\boldsymbol{p})$ values were obtained using *only* the paired modeled and randomly generated observed values and yet the RAMP method is able to capture areas of high and low systematic and random errors across the continuous mapping domain. This demonstrates that the RAMP method is able to assess model performance at unsampled locations, i.e. at locations where observations are *not* available.

APPENDIX B: SUPPORTING INFORMATION FOR INCORPORATING REGIONALIZED AIR QUALITY MODEL PERFORMANCE EVALUATION IN A NATIONWDIE GEOSTATISTICAL DATA INTEGRATION OF DAILY PM2.5[5]

B.1 Offset and Covariance Optimization

The offset is considered a deterministic function of space and time that can be mathematically calculated for any space/time point $\boldsymbol{p}$ without error. The offset is used to transform the PM2.5 data into residual offset-removed data. The BME analysis is performed on these transformed data. Ideally an offset is created to ensure that the transformed data have low variance to ensure accuracy of the estimation and high autocorrelation to ensure that neighboring data locations are informative to the estimation location. In this study we elect to create several offset functions that capture variability of PM2.5 at varying spatial and temporal scales and pick one which meets the above criteria most closely.

We calculate the offset using a space/time composite kernel smoothing of the data. The equation to calculate the offset $o_Z(\boldsymbol{p}_j)$ at location $\boldsymbol{p}_j = (\boldsymbol{s}_j, t_j)$ is given by.

$$o_z(\boldsymbol{s}_j, t_j) = \sum_{i=1}^{N} w_i Z(\boldsymbol{s}_i, t_i) / \sum_{i=1}^{N} w_i \qquad \text{(Equ. B-1)}$$

where $Z(\boldsymbol{s}_i, t_i)$ is the measured value at space/time point $\boldsymbol{p}_i = (\boldsymbol{s}_i, t_i)$ within the neighborhood, $w_i = \exp(-\frac{\|\boldsymbol{s_i}-\boldsymbol{s_j}\|}{a_r} - \frac{|t_i-t_j|}{a_t})$ and a_r is the spatial smoothing range and a_t is the temporal smoothing range. Intuitively there is an inverse relationship between the amount of variability in the offset and the remaining variability of the transform. If the offset describes short range space/time variability (i.e. a_r and a_t are short), then the offset has large variability and there is little remaining variability in autocorrelation of the transform. Conversely if the offset only describes long range variability (i.e. a_r and a_t are large), then the resulting transform retains much of the variability of the original data and thus has large variance and autocorrelation.

The covariance model for the homogeneous/stationary S/TRF $X(\boldsymbol{p})$ is developed from the experimental covariance of the transformed data $\boldsymbol{x}_h = \boldsymbol{z}_h - o_Z(\boldsymbol{p}_h)$. The experimental covariance value for a spatial lag r and a temporal lag τ is calculated as

[5] This appendix was submitted as the supporting information of an article to the journal Environmental Science and Technology. Reyes, Jeanette M., Xu,Yadong, Vizuete, William, Serre, L. Marc. Incorporating Regionalized Air Quality Model Performance evaluation in a nationwide geostatistical data integration of daily PM2.5.

$$\hat{c}_X(r,\tau) = \frac{1}{N(r,\tau)} \sum_{i=1}^{N(r,\tau)} X_{head,i} X_{tail,i} - {m_X}^2 \quad \text{(Equ. B-2)}$$

where $N(r,\tau)$ is the number of pairs of values $(X_{head,i}X_{tail,i})$ separated by a spatial lag of r and time lag of τ and m_X is the mean of the $\boldsymbol{x}_h$ data. In practice $\hat{c}_X(r,0)$ and $\hat{c}_X(0,\tau)$ are plotted separately to facilitate the visualization of the space/time covariance models.

In order to investigate the effect of the variance and autocorrelation on the transformed data, we constructed 4 offsets, describing variability at short, intermediate, long and very long size scales, respectively, with a_r and a_t values (Table B.1).

Table B.1. Offset parameter values and namings used to smooth PM2.5 in space/time

Offset name	a_r (km)	a_t (days)
short	20	10
intermediate	50	20
long	300	50
very long	1,000	200

Each offset can be assessed visually through maps (Fig. B.1) and time series (Fig. B.2). The following four maps show the short, intermediate, long and very long offsets, respectively of PM2.5 concentration on July 30, 2001. As can be seen from these figures, the short offset describes variability at a fine scale, while the very long offset smoothed out the data.

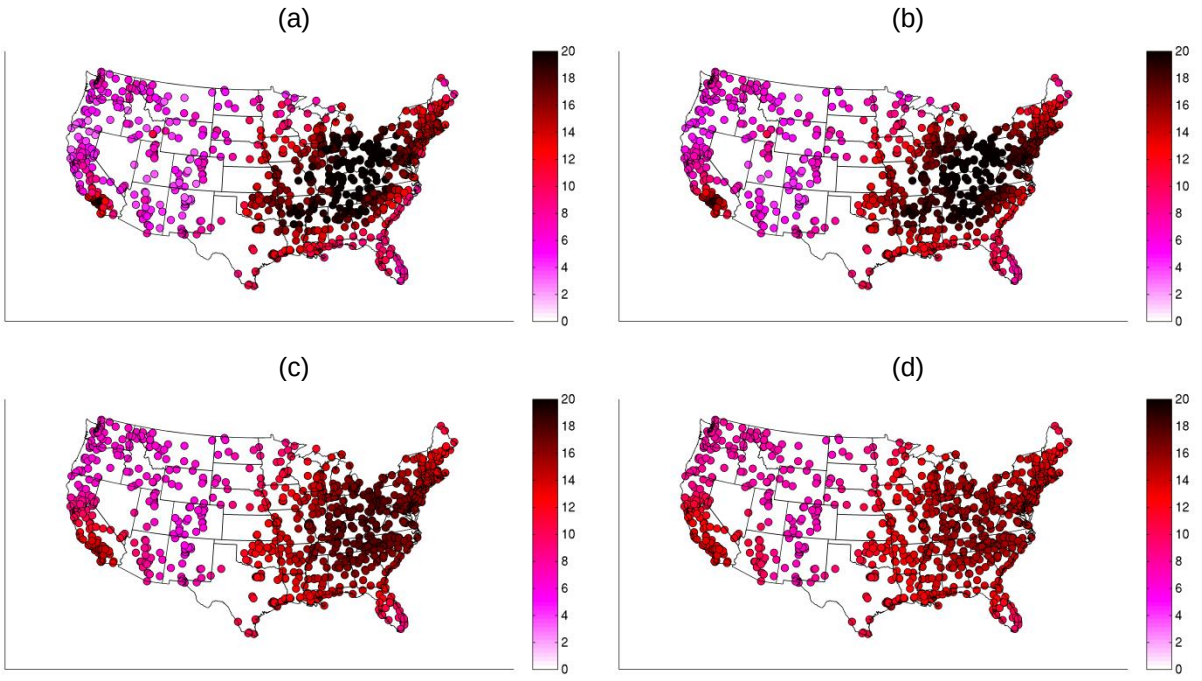


Figure B.1. PM2.5 concentration across the continental US on July 30, 2001 after smoothing the data using the (a) short, (b) intermediate, (c) long and (d) very long offset smoothing parameters described in

Table B.1. Colored circles are the concentration of PM2.5 of monitoring station locations on that day after smoothing by the parameters. Units are in $\mu g/m^3$.

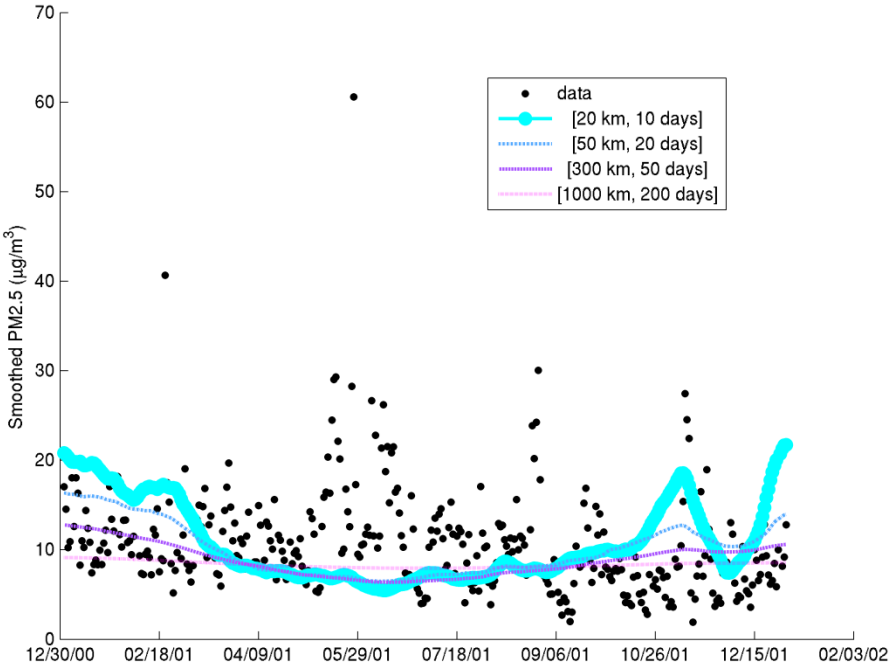


Figure B.2. Time series of PM2.5 concentration across an arbitrary PM2.5 monitoring station across 2001 using the short, intermediate, long, and very long offset smoothing parameters described in Table B.1. Black dots are observed PM2.5 values across the station for 2001.

The offsets with the shortest parameter values smooth the data the least. As the offset smoothing parameters increase, the smoothing increases as well (Fig. B.1, Fig. B.2).

(a) (b)

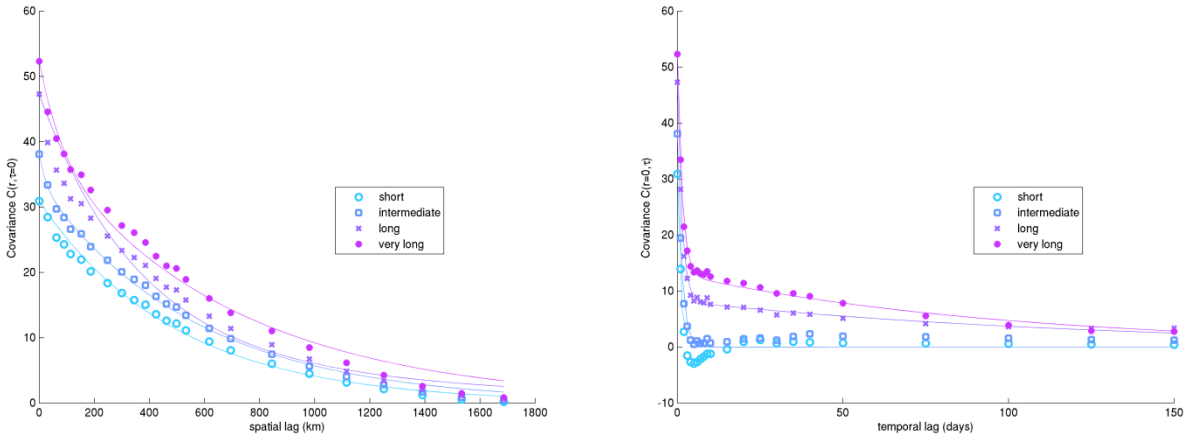


Figure B.3. Experimental and modeled covariance of the transform of the short, intermediate, long and very long offset in (a) space and (b) time for PM2.5.

Several covariance models were evaluated (plots not shows), and a two-structured exponential model was selected because it provided the best overall performance with respect to least squared fit, model interpretability, and consistency with models used in previous works on PM. The equation of the two-structured exponential covariance model is given by

$$c_X(r,\tau) = C_0[\alpha \exp\left(\frac{-3r}{ar_1}\right)\exp\left(\frac{-3\tau}{at_1}\right) + (1-\alpha)\exp\left(\frac{-3r}{ar_2}\right)\exp\left(\frac{-3\tau}{at_2}\right)] \quad \text{(Equ. B-3)}$$

where C_0 is the sill (variance), ar_1 and at_1 are the spatial and temporal ranges of the first covariance structure, ar_2 and at_2 are the spatial and temporal ranges of the second covariance structure, and α is the proportion of variability explained by the first covariance structure. The parameters α, ar_1, ar_2, at, at_2 obtained by joint least square fitting for each offset (Table B.2).

Table B.2. Covariance model and parameter values for each offset calculated through least squares fitting

Offset	α	ar_1 (km)	ar_2 (km)	at_1 (days)	at_2 (days)
short	0.03	872.25	1505.48	2.91	2.91
intermediate	0.89	1684.71	68.73	4.04	4.04
long	0.83	1057.08	3868.38	4.08	379.14
very long	0.25	237.36	2068.51	296.88	4.28

In order to assess which of the 4 offsets should be selected in the BME estimation, dominance plots are created (Fig. B.4). Like stated above, the offset selected will have a combination of the lowest variance and the highest autocorrelation.

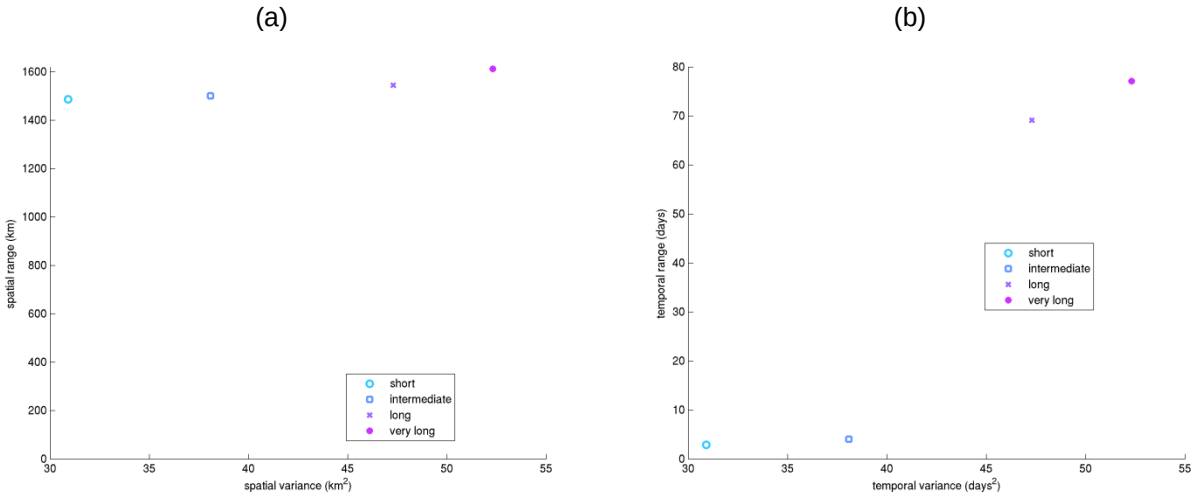


Figure B.4. Dominance plots in (a) space and (b) time for the transforms of the short, intermediate, long and very long offset displaying covariance ranges as a function of variance.

Based on these plots, the long offset was selected and BME analysis was completed on the corresponding transform. As can be seen from these figures the variance increases at regular increments from the remaining offsets. The temporal covariance range, however, changes drastically from the intermediate to the long offset. Therefore the long offset is selected because it produces much larger autocorrelation than the intermediate offset, while only sacrificing approximately a quarter of the variance. Thus the long offset offers the best tradeoff of lowering variance while maintaining autocorrelation in the transformed data. This offset is used in the subsequent BME analysis.

B.2 Quantification of Spatial Refinement

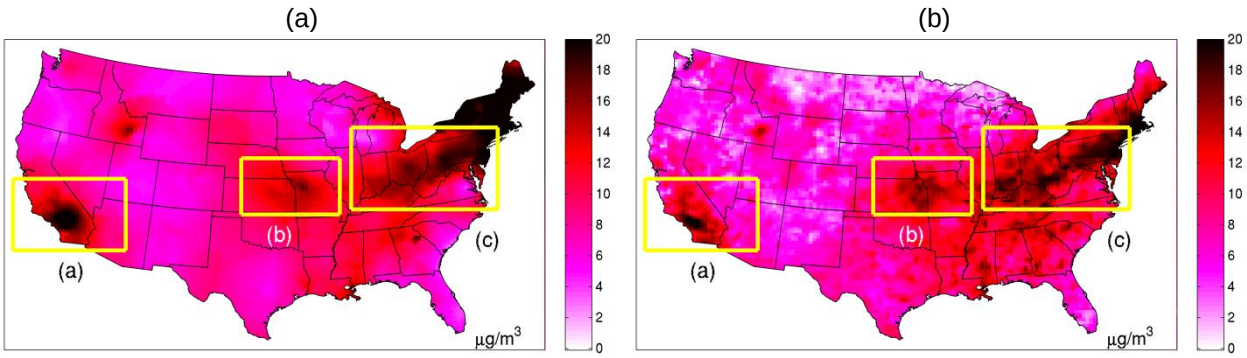
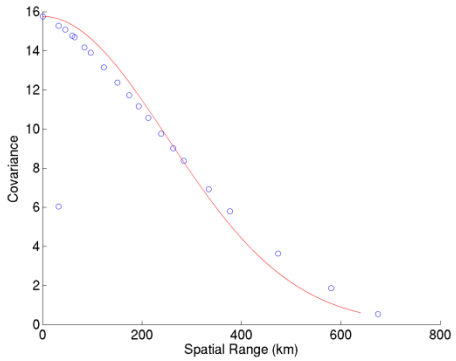
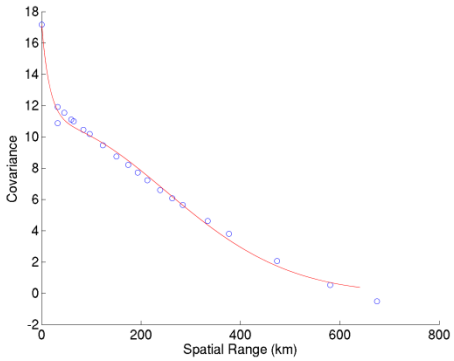


Figure B.5. Posterior mean of PM2.5 across the contiguous US on July 1, 2001 as estimated by (a) kriging and (b) the BME data fusion method. In both (a) and (b) the same three regions are boxed in: California (a), Missouri (b) and the Mid-East (c). Units are in $\mu g/m^3$.

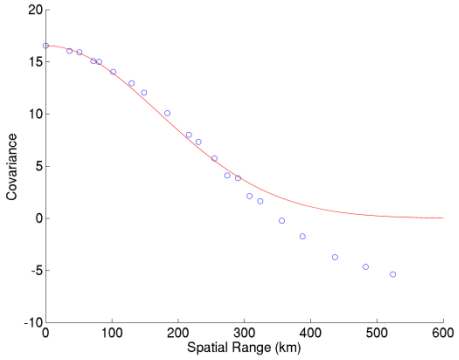
(a)


Covariance
Spatial Range (km)

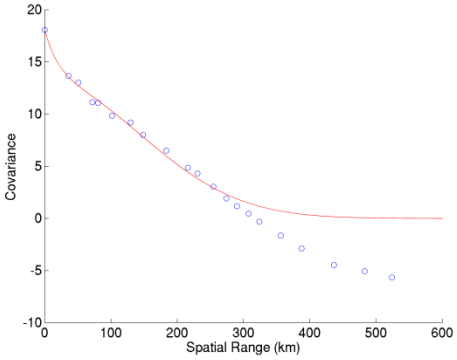
(b)


Covariance
Spatial Range (km)

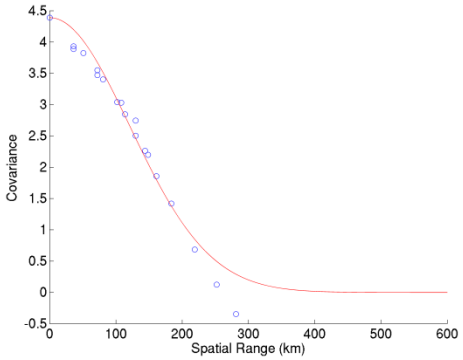
(c)


Covariance
Spatial Range (km)

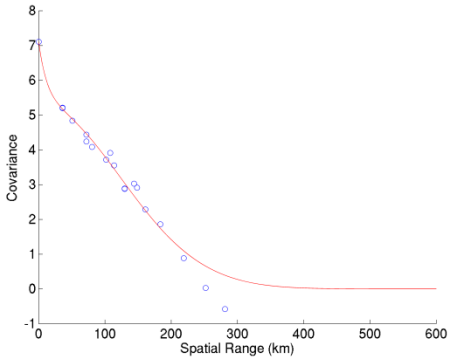
(d)


Covariance
Spatial Range (km)

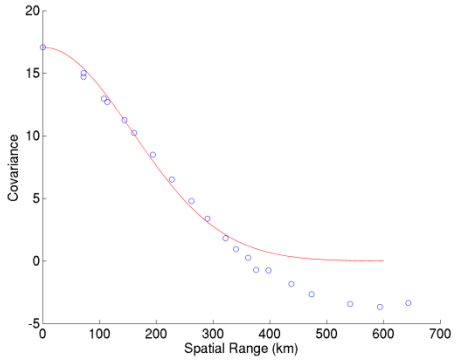
(e)


Covariance
Spatial Range (km)

(f)


Covariance
Spatial Range (km)

(g)


Covariance
Spatial Range (km)

(h)

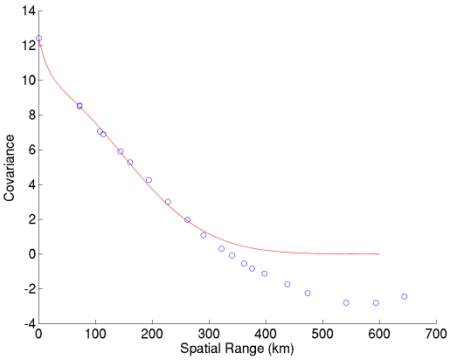

Covariance
Spatial Range (km)

Figure B.6. The experimental covariance and covariance models for the posterior mean estimates for kriging and BME within the boxed areas displayed in Fig. B.6. The first column (a, c, e, g) are all the kriging models and the second column (b, d, f, h) are all the BME models. The first row (a, b) are models developed from all posterior mean across the US, the second row (c, d) are models developed from the boxed region of California, the third row (e, f) are models developed from the boxed region of Missouri, the fourth row (g, h) are models developed from the boxed region of the Mid-East.

Table B.3. Spatial covariance ranges of the posterior means of the boxed regions presented in Fig. B.6 for kriging and BME. The kriging covariance is a one structured Gaussian model and the BME covariance is a two structured Gaussian and exponential model. The last column displays the ratio of the shortest kriging range (i.e. the only kriging range) with the shortest BME range.

Region	Kriging Range (km)	BME Range (km)		Short Range Ratio
		Long	Short	
California	422	362	66	6.4
Missouri	296	302	39	7.7
Mid-East	386	358	58	6.6

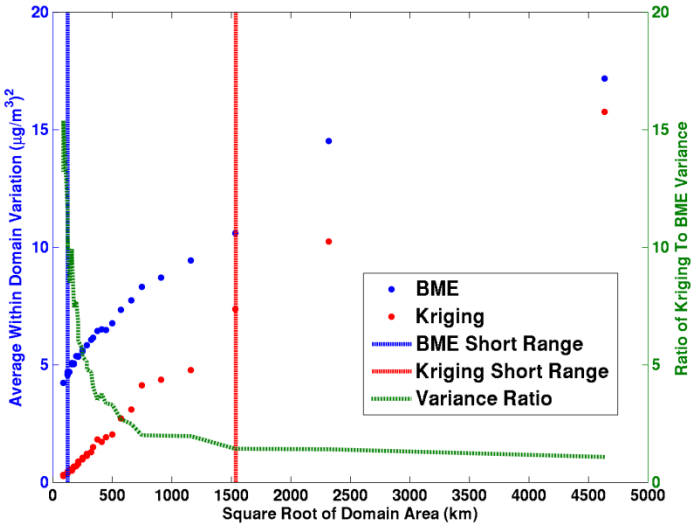


Figure B.7. The average variation of BME and kriging within a subdomain of the US with an increasingly smaller area. Posterior mean estimates for daily PM2.5 were taken across the contiguous US as estimated by both kriging and BME. The entirety of the domain was then broken up into subdomains of equal size as displayed on the independent axis. The average variance of the posterior mean estimates of all subdomains were taken and plotted on the left dependent axis. The horizontal red line is the shortest covariance spatial range calculated from all posterior kriging mean estimates across the contiguous US on July 1, 2001. The horizontal blue line is the shortest covariance spatial range calculated from all posterior BME mean estimates across the contiguous US on July 1, 2001. The dotted green line is the ratio between the average BME to kriging variation as measured on the right dependent axis.

B.3 Implementation of the Frequentist Downscaler Method

B.3.1 Equations

$$Y(s,t) = \beta_{0t} + \beta_0(s,t) + \beta_{1t}x(B,t) + \beta_1(s,t)x(B,t) + \epsilon(s,t) \quad \text{(Equ. B-4)}$$

There are several options to extend of the downscaler model to include a temporal component. This work looks at the additive bias in a dynamic manner and the multiplicative bias in an independent manner.

$$Y(s,t) = \rho_0\beta_{0t-1} + \eta_{0t} + \rho_1\beta_{1t-1}x(B,t) + \eta_{1t}x(B,t) + (A_{11} + A_{12}x(B,t))w_{0t}(s,t) + (A_{12} + A_{22}x(B,t))w_{1t}(s,t) + \epsilon(s,t) \quad \text{(Equ. B-5)}$$

$$Y(s,t) \sim GRF(m_Y = \rho_0\beta_{0t-1} + \eta_{0t} + \rho_1\beta_{1t-1}x(B,t) + \eta_{1t}x(B,t), c_Y = \left(A_{11} + A_{21}x(B,t)\right)^2 c_{w_{0t}} + \left(A_{12} + A_{22}x(B,t)\right)^2 c_{w_{1t}} + \tau^2 c_\epsilon \quad \text{(Equ. B-6)}$$

The covariance model is not only a function of the random fields $w_{0t}(s)$, $w_{0t}(s)$, and $\epsilon(s,t)$, but it is a function of the variable (known everywhere) $x(B,t)$. Thus, c_Y can be more explicitly defined as:

$$c_Y|_{x(B)} = cov\left(Y(s,t), Y(s',t')\right) = (A_{11} + A_{12}x(B,t))(A_{11} + A_{12}x(B',t'))\exp(-\phi_0|s-s'|)\delta(|t-t'|) + (A_{12} + A_{22}x(B,t))(A_{12} + A_{22}x(B',t'))\exp(-\phi_1|s-s'|)\delta(|t-t'|) + \tau^2\delta(|s-s'|,|t-t'|) \quad \text{(Equ. B-7)}$$

B.3.2 Empirical Estimation of Parameters

To estimate the parameters in a non-Bayesian way, an empirical approach is taken.

Estimation of ρ_0, β_{0t-1}, η_{0t}, ρ_1, β_{1t-1}, and η_{1t}

For every day, all paired modeled and observed data are collected. For each day, the parameters are fit to the following optimization function.

$$min\left\{\sum_{t=1}^{T}\left(\hat{m}_{Y,t} - Y(s,t)\right)^2\right\} \quad \text{(Equ. B-8)}$$

where, $\hat{m}_{Y,t} = \hat{\rho}_0\hat{\beta}_{0t-1} + \hat{\eta}_{0t} + \hat{\rho}_1\hat{\beta}_{1t-1}x(B,t) + \hat{\eta}_{1t}x(B,t)$. For across all t days the empirical covariance model is fit through least squares.

Estimation of τ^2

The parameter τ^2 is the variance of the random white noise of $Y(s,t)$. To minimized the skewness of the residuals of $Y(s,t)$, variance of the 25% and 75% percentiles were taken.

$$\hat{\tau}^2 = V[\hat{m}_Y - Y(s,t)] \quad \text{(Equ. B-9)}$$

Estimation of A_{11}, A_{12},A_{22}, ϕ_0, and ϕ_1

These parameters are constant across all days on the mean trend-removed data. For every day, all paired modeled and observed data are collected. From this set, modeled data are partitioned into deciles. The experimental covariance is calculated between every combination of each decile and for $B = B'$ (i.e. $\binom{10}{2} + 10 = 55$ combinations).

For each combination/day, the parameters are fit to the following optimization function.

$$min\left\{\sum_{t=1}^{T}\sum_{i=1}^{10}\sum_{j=i}^{10}\left(\hat{C}_{Y|x(B_i,t),x(B_j,t)} - C_{Y|x(B_i,t),x(B_j,t)}\right)^2\right\} \quad \text{(Equ. B-10)}$$

For every (i, j) combination for all t days the empirical covariance model is fit through least squares.

$$c_Y = \left(A_{11} + A_{12}x(B_i,t)\right)\left(A_{11} + A_{12}x\left(B_j,t\right)\right)\exp(-\phi_0|s-s'|)\delta(|t-t'|) + \left(A_{12} + A_{22}x(B_i,t)\right)\Big(A_{12} + A_{22}x\left(B_j,t\right)\Big)\exp(-\phi_1|s-s'|)\delta(|t-t'|) + \tau^2\delta(|s-s'|)\delta(|t-t'|) \quad \text{(Equ. B-11)}$$

Keep in mind that the variance of each covariance model can only be evaluated when $B_i = B_j$ for each day.

B.3.3 Development of the Predictive Distribution

$$Y_{k|Y_d,\theta} = m_{k|Y_d} + C_{Y_k,Y_d,\theta}C^{-1}{}_{Y_d,Y_d,\theta}\left(Y_d - m_{d,\theta}\right) \quad \text{(Equ. B-12)}$$

$$C_{k|Y_d,\theta} = C_{Y_k,Y_k,\theta} - 2C_{Y_k,Y_d,\theta}C^{-1}{}_{Y_d,Y_d,\theta}C_{Y_d,Y_k,\theta} \quad \text{(Equ. B-13)}$$

Let $x(B_k, t_k)$ be the modeled concentration of the grid of the estimation location.

$$m_{k|Y_d} = \rho_0\beta_{0t-1} + \eta_{0t} + \rho_1\beta_{1t-1}x(B_k,t_k) + \eta_{1t}x(B_k,t_k) \quad \text{(Equ. B-14)}$$

$$C_{Y_k,Y_d,\theta} = \sum_{i=1}^{10}\boldsymbol{I}_{\boldsymbol{x(B_i,t_i)}}\Big((A_{11} + A_{12}x(B_k,t_k))\big(A_{11} + A_{12}x(\boldsymbol{B}_i,\boldsymbol{t})\big)\exp(-\phi_0|s_k - \boldsymbol{s}_i|)\delta(|t_k - \boldsymbol{t}|) + (A_{12} + A_{22}x(B_k,t_k))\big(A_{12} + A_{22}x(\boldsymbol{B}_i,\boldsymbol{t})\big)\exp(-\phi_1|s_k - \boldsymbol{s}_i|)\delta(|t_k - \boldsymbol{t}|) + \tau^2\delta(|s_k - \boldsymbol{s}_i|)\delta(|t_k - \boldsymbol{t}|)\Big) \quad \text{(Equ. B-15)}$$

$$C_{Y_d,Y_d,\theta} = \sum_{i=1}^{10}\sum_{j=i}^{10}\boldsymbol{I}_{\boldsymbol{x(B_i,t_i)},\boldsymbol{x(B_j,t_i)}}\Big(\left(A_{11} + A_{12}x(\boldsymbol{B}_i,\boldsymbol{t})\right)\left(A_{11} + A_{12}x\left(\boldsymbol{B}_j,\boldsymbol{t}\right)\right)\exp\left(-\phi_0\left|\boldsymbol{s}_i - \boldsymbol{s}_j\right|\right)\delta(\left|\boldsymbol{t}_i - \boldsymbol{t}_j\right|) + \left(A_{12} + A_{22}x(\boldsymbol{B}_i,\boldsymbol{t})\right)\left(A_{12} + A_{22}x\left(\boldsymbol{B}_j,\boldsymbol{t}\right)\right)\exp\left(-\phi_1\left|\boldsymbol{s}_i - \boldsymbol{s}_j\right|\right)\delta(\left|\boldsymbol{t}_i - \boldsymbol{t}_j\right|) + \tau^2\delta(\left|\boldsymbol{s}_i - \boldsymbol{s}_j\right|)\delta(\left|\boldsymbol{t}_i - \boldsymbol{t}_j\right|)\Big) \quad \text{(Equ. B-16)}$$

$\boldsymbol{I}_{\boldsymbol{x(B_i,t_i)}}$ is an indicator function where $\boldsymbol{I}_{\boldsymbol{x(B_i,t_i)}} = \begin{cases}1, when\ B_i \in decile_i\\ 0, when\ B_i \notin decile_i\end{cases}$

$\boldsymbol{I}_{\boldsymbol{x(B_i,t_i)},\boldsymbol{x(B_j,t_i)}}$ is an indicator function where $\boldsymbol{I}_{\boldsymbol{x(B_i,t_i)},\boldsymbol{x(B_j,t_i)}} = \begin{cases}1, when\ B_i \in decile_i\ and\ B_j \in decile_j\\ 0,\ B_i \notin decile_i\ and/or\ B_j \notin decile_j\end{cases}$

The size of $C_{Y_k,Y_d,\theta}$ is maintained throughout its summation, meaning $C_{Y_k,Y_d,\theta}$ is a $k \times d$ matrix and each term in the summation is also a $k \times d$ matrix. The same is true for $C_{Y_d,Y_d,\theta}$ and $C_{Y_k,Y_k,\theta}$.

B.3.4 Development of the Distribution of the bias (additive and multiplicative)

Recall $\tilde{\beta}_{0t}(s,t) = \rho_0\beta_{0t-1} + \eta_{0t} + \beta_0(s,t) = \beta_{0t} + A_{11}w_{0t}(s,t) + A_{\mathbf{12}}w_{1t}(s,t)$

$\tilde{\beta}_{0t}(s,t) \sim GRF(m_{\tilde{\beta}_{0t}} = \rho_0\beta_{0t-1} + \eta_{0t}, c_{\tilde{\beta}_{0t}} = {A_{11}}^2 c_{w_{0t}} + {A_{12}}^2 c_{w_{1t}})$ (Equ.B-17)

Because $\tilde{\beta}_{0t}(s,t)$ is not calculated directly, the predictive distribution is slightly different than $Y(s,t)$.

$Y_{\tilde{\beta}_{0t_k}|Y_d,\theta} = m_{\tilde{\beta}_{0t_k}|Y_d} + C_{\tilde{\beta}_{0t_k},Y_d,\theta}{C^{-1}}_{Y_d,Y_d,\theta}\left(Y_d - m_{d,\theta}\right)$ (Equ. B-18)

$C_{\tilde{\beta}_{0t_k}|Y_d,\theta} = C_{\tilde{\beta}_{0t_k},\tilde{\beta}_{0t_k},\theta} - 2C_{\tilde{\beta}_{0t_k},Y_d,\theta}{C^{-1}}_{Y_d,Y_d,\theta}C_{Y_d,\tilde{\beta}_{0t_k},\theta}$ (Equ. B-20)

$C_{\tilde{\beta}_{0t_k},Y_d,\theta} = cov\left(\tilde{\beta}_{0t}(s,t), Y(s',t')\right) = cov\big(\rho_0\beta_{0t-1} + \eta_{0t} + A_{11}w_{0t}(s,t) + A_{\mathbf{12}}w_{1t}(s,t), \beta_{0t} + \beta_{1t}x(B',t') +$

$(A_{11} + A_{12}x(B',t'))w_{0t}(s',t') + (A_{12} + A_{22}x(B',t'))w_{1t}(s',t') + \epsilon(s',t')\big) = A_{11}\big(A_{11} + A_{21}x(B',t')\big)c_{w_{0t}} +$

$A_{12}\big(A_{12} + A_{22}x(B',t')\big)c_{w_{1t}}$ (Equ. B-21)

$C_{\tilde{\beta}_{0t_k},\tilde{\beta}_{0t_k},\theta} = cov\left(\tilde{\beta}_0(s,t), \tilde{\beta}_0(s',t')\right) = cov\big(\rho_0\beta_{0t-1} + \eta_{0t} + A_{11}w_{0t}(s,t) + A_{\mathbf{12}}w_{1t}(s,t), \rho_0\beta_{0t-1} + \eta_{0t} +$

$A_{11}w_{0t}(s',t') + A_{\mathbf{12}}w_{1t}(s',t')\big) = {A_{11}}^2 c_{w_{0t}} + {A_{12}}^2 c_{w_{1t}}$ (Equ. B-22)

$\tilde{\beta}_1(s,t)$ follows similarly.

$\tilde{\beta}_1(s,t) = \rho_0\beta_{0t-1} + \eta_{0t} + \beta_1(s,t) = \rho_0\beta_{0t-1} + \eta_{0t} + A_{12}w_{0t}(s,t) + A_{\mathbf{22}}w_{1t}(s,t)$ (Equ. B-23)

$\tilde{\beta}_{1t}(s,t) \sim GRF(m_{\tilde{\beta}_{1t}} = \rho_1\beta_{1t-1} + \eta_{0t}, c_{\tilde{\beta}_{1t}} = {A_{12}}^2 c_{w_{0t}} + {A_{22}}^2 c_{w_{1t}})$ (Equ B-24)

Because $\tilde{\beta}_{1t}(s,t)$ is not calculated directly, the predictive distribution is slightly different than $Y(s,t)$.

$Y_{\tilde{\beta}_{1t_k}|Y_d,\theta} = m_{\tilde{\beta}_{1t_k}|Y_d} + C_{\tilde{\beta}_{1t_k},Y_d,\theta}{C^{-1}}_{Y_d,Y_d,\theta}\left(Y_d - m_{d,\theta}\right)$ (Equ. B-25)

$C_{\tilde{\beta}_{1t_k}|Y_d,\theta} = C_{\tilde{\beta}_{1t_k},\tilde{\beta}_{1t_k},\theta} - 2C_{\tilde{\beta}_{1t_k},Y_d,\theta}{C^{-1}}_{Y_d,Y_d,\theta}C_{Y_d,\tilde{\beta}_{1t_k},\theta}$ (Equ. B-26)

$C_{\tilde{\beta}_{1t_k},Y_d,\theta} = cov\left(\tilde{\beta}_{1t}(s,t), Y(s',t')\right) = cov\big(\rho_1\beta_{1t-1} + \eta_{1t} + A_{12}w_{0t}(s,t) + A_{\mathbf{22}}w_{1t}(s,t), \beta_{0t} + \beta_{1t}x(B',t') +$

$(A_{11} + A_{12}x(B',t'))w_{0t}(s',t') + (A_{12} + A_{22}x(B',t'))w_{1t}(s',t') + \epsilon(s',t')\big) = A_{12}\big(A_{11} + A_{12}x(B',t')\big)c_{w_{0t}} +$

$A_{22}\big(A_{12} + A_{22}x(B',t')\big)c_{w_{1t}}$ (Equ. B-27)

$C_{\tilde{\beta}_{1t_k},\tilde{\beta}_{1t_k},\theta} = cov\left(\tilde{\beta}_{1t}(s,t), \tilde{\beta}_{1t}(s',t')\right) = cov\big(\rho_1\beta_{1t-1} + \eta_{1t} + A_{12}w_{0t}(s,t) + A_{\mathbf{22}}w_{1t}(s,t), \beta_{1t} +$

$A_{12}w_{0t}(s',t') + A_{\mathbf{22}}w_{1t}(s',t')\big) = {A_{12}}^2 c_{w_{0t}} + {A_{22}}^2 c_{w_{1t}}$ (Equ. B-28)

APPENDIX C: SUPPORTING INFORMATION FOR INCORPORATING MASS FRACTION OF POLYCYCLIC AROMATIC HYDROCARBONS INTO THE BAYESIAN MAXIMUM ENTROPY FRAMEWORK ACROSS NORTH CAROLINA[6]

Table C.1. Covariance model parameters for observed PAH data.

PAH	C_0 $((ng/m^3)^2)$	a_r (km)	a_t (days)
benz(a)anthracene	2.26	518	142
chrysene	2.09	521	147
benzo(b)fluoranthrene	2.38	914	109
benzo(k)fluoranthrene	2.11	362	155
benzo(e)pyrene	2.20	430	119
benzo(a)pyrene	2.13	247	146
indeno(1,2,3-c,d)pyrene	1.84	507	123
benzo(g,h,i)perylene	1.43	750	129
dibenzo(a,h)anthracene	3.01	312	123
Total PAH	1.68	356	123

The covariance model for the homogeneous/stationary S/TRF $X(\boldsymbol{p})$ is developed from the PAH experimental covariance. The experimental covariance value for a spatial lag r and a temporal lag τ is calculated as

$$\hat{c}_X(r,\tau) = \frac{1}{N(r,\tau)} \sum_{i=1}^{N(r,\tau)} X_{head,i} X_{tail,i} - {m_X}^2 \quad \text{(Equ. C-1)}$$

where $N(r,\tau)$ is the number of pairs of values ($X_{head,i}X_{tail,i}$) separated by a spatial lag of r and time lag of τ and m_X is the mean of the $\boldsymbol{x}_h$ data. A one-structured exponential model was selected because it provided the best overall performance with respect to least squared fit. The equation of the one-structured exponential covariance model is given by

$$c_X(r,\tau) = C_0 \exp\left(\frac{-3r}{a_r}\right) \exp\left(\frac{-3\tau}{a_t}\right) \quad \text{(Equ. C-2)}$$

where C_0 is the sill (variance) and a_r and a_t are the spatial and temporal ranges, respectively. The parameters a_r and a_t obtained from observed PAH data are given in Table C.1.

[6] This appendix is planned to be submitted as the supporting information of an article to the Journal of Exposure Science and Environmental Epidemiology. Reyes, Jeanette M., Hubbard, Heidi, Stiegel, Matthew A., Pleil, Joachim D., Serre, L. Marc. Incorporating Mass Fraction of Polycyclic Aromatic Hydrocarbons into the Bayesian Maximum Entropy Framework across North Carolina.

Table C.2. Cokriging covariance model parameters for observed PAH and PM2.5 data. C_{PAH} is in in $(ng/m^3)^2$, C_{PM} is in $(\mu g/m^3)^2$and $C_{PAH,PM}$ is in $(ng/m^3) * (\mu g/m^3)$.

PAH	C_{PAH}	C_{PM}	$C_{PAH,PM}$	a_r (km)	a_t (days)
benz(a)anthracene	2.257	0.365	0.157	188	118
chrysene	2.091	0.365	0.135	206	119
benzo(b)fluoranthrene	2.381	0.365	-0.031	1176	78
benzo(k)fluoranthrene	2.112	0.365	0.069	223	102
benzo(e)pyrene	2.202	0.365	0.012	330	96
benzo(a)pyrene	2.131	0.365	-0.093	206	114
indeno(1,2,3-c,d)pyrene	1.838	0.365	-0.002	437	103
benzo(g,h,i)perylene	1.429	0.365	-0.054	650	101
dibenzo(a,h)anthracene	3.005	0.365	0.081	128	87
Total PAH	1.683	0.365	0.037	322	96

Table C.3. Cross validation statistics for all 9 PAHs and Total PAH.

PAH	statistic	kriging	cokriging	soft LR	soft MF
benz(a)anthracene	ME	-0.210	-0.246	-0.228	-0.049
	VE	1.037	1.017	0.923	0.716
	MSE	1.069	1.065	0.964	0.710
	R2	0.747	0.750	0.773	0.830
chrysene	ME	-0.202	-0.238	-0.200	-0.020
	VE	1.092	0.999	1.032	0.807
	MSE	1.119	1.044	1.060	0.798
	R2	0.712	0.735	0.717	0.789
benzo(b)fluoranthrene	ME	0.015	0.029	0.157	0.018
	VE	1.215	1.329	1.724	1.019
	MSE	1.200	1.314	1.728	1.007
	R2	0.723	0.701	0.637	0.762
benzo(k)fluoranthrene	ME	-0.219	-0.239	-0.179	-0.130
	VE	1.416	1.369	1.276	1.347
	MSE	1.447	1.410	1.292	1.348
	R2	0.611	0.621	0.636	0.611
benzo(e)pyrene	ME	-0.137	-0.147	-0.068	0.019
	VE	0.948	0.933	0.871	0.646
	MSE	0.955	0.944	0.865	0.639
	R2	0.776	0.779	0.781	0.848
benzo(a)pyrene	ME	-0.169	-0.174	0.069	-0.006
	VE	0.998	0.969	0.895	1.159
	MSE	1.015	0.988	0.889	1.145
	R2	0.739	0.748	0.767	0.698
indeno(1,2,3-c,d)pyrene	ME	-0.078	-0.083	-0.017	0.020
	VE	0.842	0.831	0.896	0.686
	MSE	0.839	0.828	0.885	0.678
	R2	0.758	0.761	0.755	0.804
benzo(g,h,i)perylene	ME	0.030	0.023	0.095	0.076
	VE	0.921	0.923	0.940	0.885
	MSE	0.911	0.912	0.938	0.880
	R2	0.650	0.650	0.650	0.646
dibenzo(a,h)anthracene	ME	-0.205	-0.249	-0.031	-0.019
	VE	1.028	1.205	0.897	0.885
	MSE	1.058	1.253	0.888	0.874
	R2	0.818	0.781	0.848	0.845
Total PAH	ME	-0.145	-0.137	-0.102	-0.042
	VE	0.806	0.782	0.764	0.591
	MSE	0.818	0.792	0.766	0.586
	R2	0.747	0.752	0.744	0.821

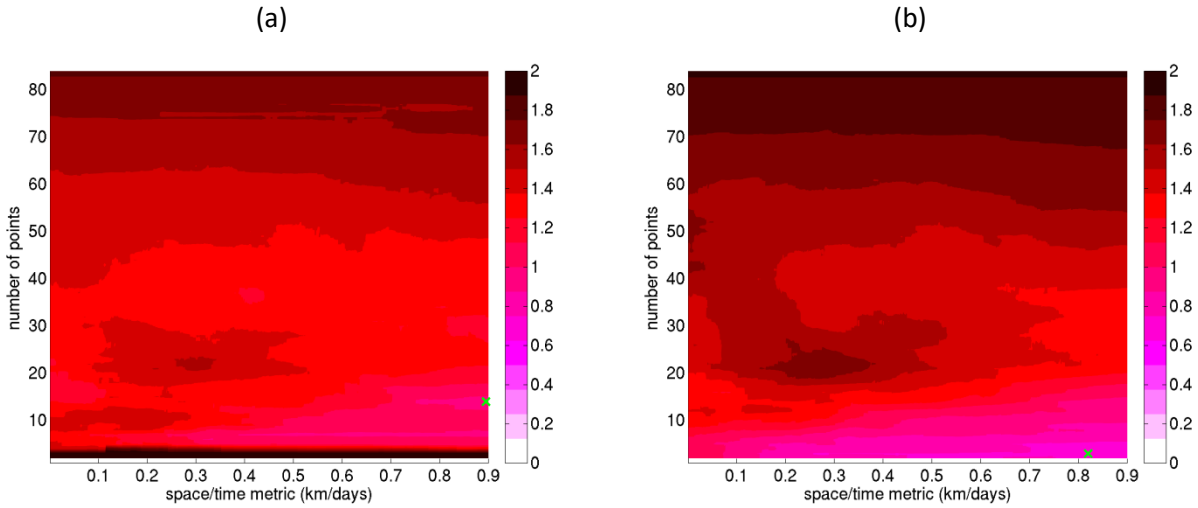


Figure C.1. Exhaustive validation search of optimal soft data neighborhood for the (a) the linear regression method and (b) the mass fraction method for Total PAH displaying the MSE. The green “X” marks the lowest MSE.

APPENDIX D: WHIMS CODE DOCUMENTATION AND QUALITY ASSURANCE FOR THE ESTIMATION OF PM2.5 AFTER 1999 USING OBSERVATION AND CTM

D.1 Introduction

In the context of an epidemiological study there is need to obtain estimates of the PM2.5 ambient concentration for which WHIMS study participants are exposed. This report describes how the BME estimation method is used to perform an interpolation of observed daily PM2.5 concentrations and obtain at each participant location an estimate of the daily PM2.5 concentration for each day from 1999 to 2010. The daily concentrations were downloaded from the Air Quality System (AQS) maintained by the U.S. Environmental Protection Agency (EPA). The AQS raw data were processed by the BME method as follows: The AQS raw data for PM2.5 were used to obtain a database of observed daily concentrations. A transformation of these data was used, which consisted of removing from the data an offset obtained using an exponential kernel smoothing (Lee et al., 2013). The exponential kernel smoothing was set so that the offset captured the spatial variability of the data over long spatial distances and long time scales. An exponential space/time covariance model was used to characterize the space/time autocorrelation in the offset removed data. The BME method was then used to estimate daily PM2.5 at unsampled locations using the offset removed daily observations treated as hard data. Since the observations are treated as hard data, the BME method reduces to the space/time Simple Kriging (SK) method of linear geostatistics, in which case BME is also referred to as space/time Simple Kriging (SK).

D.2 Materials

D.2.1 PM2.5 daily data

The daily PM2.5 concentration for each monitoring site/day during the study period (1999-2010) were constructed based raw monitoring data from monitoring stations measuring either hourly or daily PM2.5 concentrations using the procedure described here.

We obtained PM2.5 monitoring data (raw data) sampled during the study period (1999-2010) from the Air Quality Systems (AQS) database maintained by the U.S. Environmental Protection Agency (EPA), which is a repository of the monitoring data collected across various monitoring networks. The PM2.5 data are available in a few data files on AQS depending on the source of data. These files are described in the AQS as follow: 1) daily PM2.5 local conditions, 2) daily PM fine speciation from the Chemical Speciation Network (CSN) monitoring network, 3) daily PM fine speciation from the Interagency

Monitoring of Protected Visual Environments (IMPROVE) monitoring network, and 4) PM2.5 non-referenced method hourly. Within each data file, the methodologies used to measure PM2.5 are defined using a parameter code which takes the following values: 1) 88101 for daily and hourly PM2.5 concentrations measured using a federally referenced method (FRM), 2) 88501 for raw PM2.5 concentrations measured using methods that are not federally acceptable, and 3) 88502 for PM2.5 air quality index (AQI) values that provide acceptable measurements of PM2.5 concentrations in that they are comparable to FRM measurements. In this work only the 88101 and 88502 data were treated as reliable data for the purpose of constructing the database of daily PM2.5 concentrations. Data from the parameter 88502 are also known as Tapered element oscillating microbalance (TEOM) data.

All observations sampled at monitors whose measurement scale is “Microscale” were removed. Stations located in Canada were excluded because they did not have a numeric code.

Hourly $PM_{2.5}$ data were averaged into daily $PM_{2.5}$ if at least 18 hours out of 24 were measured for a given day/monitor. Otherwise, a daily average was not constructed. More than 99.9% of hourly records were reported every hour on the hour. However, there were several records not reported on the hour. These hourly records were removed, before constructing the daily concentrations. Since the hourly data were sampled using continuous monitors, we assigned the sampling frequency code 1 (daily) to the daily concentrations constructed from hourly data.

At each monitoring site with multiple monitors, the collocated daily concentrations recorded at any given day were combined using the following procedure to produce a constructed daily concentration for that site/day. First, priority rank scores were assigned to each collocated daily concentrations based on its data source and type as follow.

Rank 1: FRM daily PM2.5
Rank 2: FRM daily PM2.5 from CSN
Rank 3: FRM hourly PM2.5
Rank 4: TEOM daily PM2.5 from CSN
Rank 5: TEOM daily PM2.5 from IMPROVE
Rank 6: TEOM hourly PM2.5

If the collocated concentrations for a given site/day had varying priority ranks, then only the concentration with the highest rank (i.e. the smallest priority score) was retained. For example, if there were concentrations with rank scores 1 and 4 at a given site/day, then only the concentration with priority rank 1 was used at that site/day. If there were more than one collocated daily concentrations with the

highest priority rank, then these daily concentrations were averaged to produce a single daily concentration at that site/day.

Finally, the constructed daily PM2.5 concentrations were joined with the geographic coordinates of the sites. If the longitude and latitude of a site were not defined in the original raw data file, then the individual site information file was searched. If the longitude and latitude were found in that file, they were added to the data file of constructed daily concentrations. By searching through the individual site information file, all locations were accounted for and no location was missing a longitude and latitude after searching.

D.2.2 PM2.5 Modeled Data

Daily concentrations for PM2.5 were also constructed from modeled data from both the Community Multiscale Air Quality (CMAQ) and Comprehensive Air Quality Model with Extensions (CAMx) models. CMAQ and CAMx are Chemical Transport Models (CTMs). They use as input emissions information as well as meteorological data which is then translated into complex chemical processes to estimate ambient air pollution over gridded geographical boundaries for different time steps. Modeled data are available at a 36km resolution every hour for the years 2001, 2002, 2005, and 2007. Modeled data are available every hour for the western part of the United States for part of the spring and summer for 2006 at a 12km resolution. Data are projected using a Lambert Conic Conformal (LCC) projection.

Daily modeled values were constructed by averaging the 24 hourly modeled values for a given grid location/day. To reconcile the spatial misalignment of defining the modeled concentration over an area (i.e. the modeled concentration over a grid), the location of modeled values are defined by the centroid of each grid.

Location of the study participants

In order to protect the confidentiality of the location of WHIMS participants, Dr. Whitsel provided to the Serre lab a large set of locations (N=17,461) that included within it the WHIMS participant locations (n+7479). The Serre lab was not given knowledge of which of the 17,461 locations were actual WHIMS participants locations. This in effect “hides” the participants amongst the large set of locations, which provides an added level of data protection. The locations were saved in a file named partdata.mat containing three columns: “partid”, “partx”, and “party”. In order to protect the confidentiality of WHIMS

participants' data, the "partid" field contains a randomized id that is generated solely for the purpose of this study and does not correspond to the actual WHIMS participant id. The "partx", and "party" fields provide the spatial coordinates of each location record.

In order to protect the confidentiality of the location data, we present here examples that are based on 500 simulated (fake) case locations randomly located across the contiguous US. An example of these simulated (fake) case locations is shown below in the figure below.

partid, partx, party
1, -1132770.34, -659383.01
2, -1789928.05, 257671.13
3, 625256.73, -669169.08
4, 626906.89, -847002.75
5, 1603181.17, 127434.18
6, -503791.48, -724386.12
7, 15752.19, 582908.44
8, 1206674.01, -154597.59
9, -1837605.24, 62486.18
10, -1765398.46, 527184.83

D.3 Methods

D.3.1 Estimation of Daily PM2.5 Concentration

BME estimation

The BME estimation method is used to perform an interpolation of observed daily PM2.5 concentrations and obtain at each participant location an estimate of the daily PM2.5 concentration for each day from 1999 to 2010. The BME method was then used to estimate daily PM2.5 at unsampled locations using the offset removed daily observations treated as hard data. The BME (kriging with measurement error) mean estimates are in good agreement with the observed data, and the BME (kriging with measurement error) variance show that the estimation is least accurate in areas where monitoring stations are sparse.

Estimation accuracy

In order to assess the estimation accuracy of the BME estimates, a 10-fold estimation was performed. For each fold, a BME estimation was conducted (without recalculating the offest or the covariance model) to obtain the BME estimates of the data in that fold using only the data from the remaining 90% of the monitoring stations. The r^2 is 0.702 for the long offset with soft data.

Quality Assurance

In order to ensure that the estimation of PM2.5 concentrations were performed correctly, quality assurance plots were created showing the concentration at randomly selected estimation locations along with the concentration of the closest 5 monitoring stations. The estimation and the randomly selected locations matches well with the estimation are the 5 surrounding monitors implying that the estimation was performed correctly.

D.4 Numerical implementation

D.4.1 Data and analysis folders

All data and analysis are housed in the folder “C:\AirCTMneuro\PM2p5est_a99”. All steps of this analysis, background information, and results have been documented in “00_DataDocumentationAndQualityAssurance”. All data files (i.e. observed data, modeled data, and paired modeled and observed data are stored in “datafiles”. The collection of functions needed to implement the BME analysis is housed in the “BMELIB2.0b” folder. All data sources have to be inputted into MATLAB and saved in a file format (i.e. .mat) that allows for easy access to all data. These data sources include all observed data, all modeled data, and all paired modeled and observed data. The .m files needed are in the “01_mfiles_prepdata” folder. All data sources are converted to the same projection using the .m files located in the “09_mfiles_projections” folder. An exploratory analysis was performed on the observed data in order to optimize the parameter for the offset. The .m files needed are in the “02_mfiles_offset” folder. The offset is calculated in space and time jointly with the .m files located in “10_mfiles_newmeantrend”. All figures created are saved in the “plots_meanTrend” folder. An exploratory analysis was performed on the observed data in order to find the best covariance model for the offset-removed data. The .m files needed are in the “03_mfiles_covariance” folder. All figures created are saved in the “plots_covModel” folder. An exploratory analysis was performed on the paired observed and modeled data to find the optimized soft data parameters. These parameters were used to calculate the bias-correct mean and variance for every CTM grid. The .m files needed are in the “04_mfiles_softdata” folder. All files created are saved in the .mat format and saved in the “matfiles” folder. In order to compare the performance of each method (e.g. using only observed data versus using observed and modeled data), cross validation was performed. The .m files needed are in the “05_mfiles_crossvalidation” folder.

BME estimates are calculated every day from 1999-2010 at the locations of the WHIMS participants. The .m files needed are in the “06_mfiles_estimation” folder. All files created are saved in .mat format in the “matfiles_est” folder. Figures created are saved in the “plots_est” folder. In order to visualize the results of the estimation, maps of the BME mean and BME variance are created across the US for select days during the time period. The .m files needed are in the “07_mfiles_map” directory. Maps are saved in the “plots” folder. In order to check to see if the estimation of PM2.5 after 1999 was performed correctly, a series of QAQC checks were performed. The .m files needed are in the “08_mfiles_QAQC” folder. All files created are saved in .mat format in the “matfiles_QAQC” folder. Figures created are saved in the “plots_QAQC” folder. All subfolder are as follows.

Table D.1. Folder Directory for WHIMS.

Folder
00_DataDocumentationAndQualityAssurance
01_mfiles_prepdata
02_mfiles_offset
03_mfiles_covariance
04_mfiles_softdata
05_mfiles_crossvalidation
06_mfiles_estimation
07_mfiles_map
08_mfiles_QAQC
09_mfiles_projections
10_mfiles_newmeantrend
BMELIB2.0b
datafiles
matfiles
matfiles_est
matfiles_QAQC
plots
plots_covModel
plots_est
plots_meanTrend
plots_QAQC

D.4.2 Instructions to estimate PM2.5 concentration after to 1999

In order to reduce the computational time, several MATLAB codes need to be executed in parallel on the Linux cluster. Shell scripts were prepared for submitting multiple jobs at a time. In order to run the shell script, use the following command.

```
sh (shell script name)
```

Shell scripts need to be executed in the following sequential order. Note that the next shell script cannot be executed until the previous one finishes.

Table D.2. Shell scripts to run for each folder.

Folder	Name of Shell script
01_mfiles_prepdata	runall_Cluster_01.sh
02_mfiles_offset	runall_Cluster_02a.sh runall_Cluster_02b.sh
03_mfiles_covariance	runall_Cluster_03.sh
04_mfiles_softdata	runall_Cluster_04a.sh runall_Cluster_04b.sh runall_Cluster_04c.sh runall_Cluster_04d.sh
05_mfiles_crossvalidation	runall_Cluster_05a.sh runall_Cluster_05b.sh runall_Cluster_05c.sh runall_Cluster_05d.sh runall_Cluster_05e.sh
06_mfiles_estimation	runall_Cluster_06a.sh runall_Cluster_06b.sh
07_mfiles_map	runall_Cluster_07.sh
08_mfiles_QAQC	runall_Cluster_08a.sh runall_Cluster_08b.sh runall_Cluster_08c.sh

D.5 Results

Each record in the data file case_PM25d_a99_CTM_YYYY.csv has the data fields described below, where YYYY is the 4 digit year ranging from 1999-2010.

Table D.3. Format of WHIMS prediction file.

Field Name	Description
id	participants' identification number
PM2.5m_YYYYMMDD	BME mean estimate of daily PM2.5 concentration on YYYY (4 digit year) MM (2 digit month) DD (2 digit day). This BME estimate is obtained using AQS observations treated as hard data and bias-corrected CTM treated as soft data, in which case BME is also referred to as space/time kriging with measurement error. The date ranges from YYYY0101 to YYYY1231. Each column corresponds to a particular day.
PM2.5sd_YYYYMMDD	corresponding BME standard deviation of daily PM2.5 concentration on YYYY (4 digit year) MM (2 digit month) DD (2 digit day). Since BME uses observed values treated as hard data and bias-corrected CTM values treaded as soft data, then the BME variance is the variance of space/time kriging with measurement error. Each column corresponds to a particular day. Days range from YYYY0101 to YYYY1231.

D.6 QAQC

In order to ensure that the estimation of PM2.5 concentrations were performed correctly, quality assurance plots were created. Below are plots showing the concentration at randomly selected estimation locations along with the concentration of the closest 5 monitoring stations and 2 closest soft data

locations. Intuitively, BME estimations should be close to the values of its surrounding stations. In each figure, the estimation and the randomly selected locations matches well with the estimation are the 5 surrounding monitors implying that the estimation was performed correctly. The time series shows how the soft data influences the BME estimate.

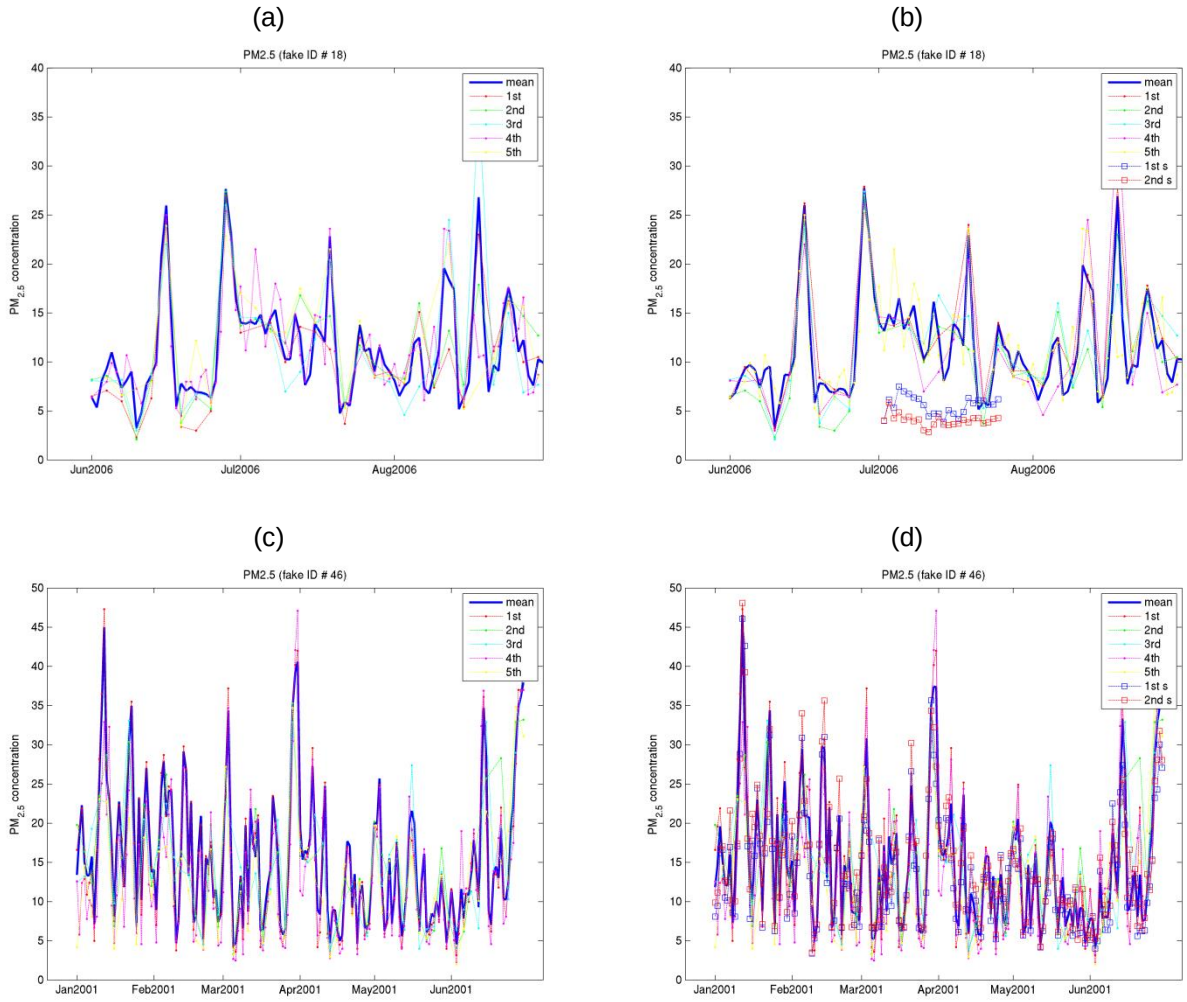


Figure D.1. Time series of random locations with and with modeled data.

D.7 Date and version number

We obtained the study participants' location data from Dr. Eric Whitsel in a file named will_marc_10_21_2013.sas7bdat on October 25, 2013. The estimation of daily PM2.5 concentrations for these locations was completed by Jeanette Reyes in June 2013. Results of estimation were copied to a file named case_PM25d_a99_CTM_YYYY.csv, which was delivered to Eric Whitsel in June 2013 as version 1.0.

APPENDIX E: GITHUB URL

All MATLAB scripts used in the creation of this dissertation can be found at https://github.com/reyesjmUNC/Dissertation-Scripts.

REFERENCES

Abdel-Shafy, H.I., Mansour, M.S.M., 2015. A review on polycyclic aromatic hydrocarbons: Source, environmental impact, effect on human health and remediation. Egypt. J. Pet. 25, 107–123. doi:10.1016/j.ejpe.2015.03.011

Akita, Y., Carter, G., Serre, M.L., 2007. Spatiotemporal Nonattainment Assessment of Surface Water Tetrachloroethylene in New Jersey. J. Environ. Qual. 36, 508–520. doi:10.2134/jeq2005.0426

Akita, Y., Chen, J.-C., Serre, M.L., 2012. The moving-window Bayesian maximum entropy framework: estimation of PM2.5 yearly average concentration across the contiguous United States. J. Expo. Sci. Environ. Epidemiol. 22, 496–501. doi:10.1038/jes.2012.57

Allshouse, W.B., Fitch, M., Hampton, K., 2011. Geomasking sensitive health data and privacy protection: an evaluation using an E911 database. Geocarto Int. 25, 443–452. doi:10.1080/10106049.2010.496496.Geomasking

Allshouse, W.B., Pleil, J.D., Rappaport, S.M., Serre, M.L., 2009. Mass fraction spatiotemporal geostatistics and its application to map atmospheric polycyclic aromatic hydrocarbons after 9/11. Stoch. Environ. Res. Risk Assess. 23, 1213–1223. doi:10.1007/s00477-009-0326-y

Appel, K.W., Bhave, P. V., Gilliland, A.B., Sarwar, G., Roselle, S.J., 2008. Evaluation of the community multiscale air quality (CMAQ) model version 4.5: Sensitivities impacting model performance; Part II—particulate matter. Atmos. Environ. 42, 6057–6066. doi:10.1016/j.atmosenv.2008.03.036

Appel, K.W., Chemel, C., Roselle, S.J., Francis, X. V., Hu, R.M., Sokhi, R.S., Rao, S.T., Galmarini, S., 2012. Examination of the Community Multiscale Air Quality (CMAQ) model performance over the North American and European domains. Atmos. Environ. 53, 142–155. doi:10.1016/j.atmosenv.2011.11.016

Appel, K.W., Pouliot, G.A., Simon, H., Sarwar, G., Pye, H.O.T., Napelenok, S.L., Akhtar, F., Roselle, S.J., 2013a. Evaluation of dust and trace metal estimates from the Community Multiscale Air Quality (CMAQ) model version 5.0. Geosci. Model Dev. Discuss. 6, 1859–1899. doi:10.5194/gmdd-6-1859-2013

Appel, K.W., Pouliot, G.A., Simon, H., Sarwar, G., Pye, H.O.T., Napelenok, S.L., Akhtar, F., Roselle, S.J., 2013b. Evaluation of dust and trace metal estimates from the Community Multiscale Air Quality (CMAQ) model version 5.0. Geosci. Model Dev. 6, 883–899. doi:10.5194/gmd-6-883-2013

Augusto, S., Máguas, C., Matos, J., Pereira, M.J., Soares, A., Branquinho, C., 2009. Spatial modeling of PAHs in lichens for fingerprinting of multisource atmospheric pollution. Environ. Sci. Technol. 43, 7762–7769. doi:10.1021/es901024w

Beckerman, B., Jerrett, M., Martin, R. V, Lee, S., Donkelaar, A. Van, Ross, Z., Su, J., Burnett, R., 2013. A Hybrid Approach to Estimating National Scale Spatiotemporal Variability of PM 2.5 in the Contiguous United States. Environ. Sci. Technol. 47, 7233–41. doi:10.1021/es400039u

Beelen, R., Hoek, G., Fischer, P., van den Brandt, P.A., Brunekreef, B., 2007. Estimated long-term outdoor air pollution concentrations in a cohort study. Atmos. Environ. 41, 1343–1358. doi:10.1016/j.atmosenv.2006.10.020

Berrocal, V.J., Gelfand, A.E., Holland, D.M., 2012. Space-time data fusion under error in computer model output: an application to modeling air quality. Biometrics 68, 837–48. doi:10.1111/j.1541-0420.2011.01725.x

Berrocal, V.J., Gelfand, A.E., Holland, D.M., 2010a. A Spatio-Temporal Downscaler for Output From Numerical Models. J. Agric. Biol. Environ. Stat. 15, 176–197. doi:10.1007/s13253-009-0004-z

Berrocal, V.J., Gelfand, A.E., Holland, D.M., 2010b. A Bivariate Space-Time Downscaler Under Space

and Time Misalignment. Annu. Appl. Stat. 1–41.

Bocskay, K.A., Tang, D., Orjuela, M.A., Liu, X., Warburton, D.P., Perera, F.P., 2005. Chromosomal Aberrations in Cord Blood Are Associated with Prenatal Exposure to Carcinogenic Polycyclic Aromatic Hydrocarbons. Caner Epidemiol. Biomarkers Prev. 14, 506–511.

Boldo, E., Medina, S., LeTertre, A., Hurley, F., Mücke, H.G., Ballester, F., Aguilera, I., Eilstein, D., 2006. Apheis: Health impact assessment of long-term exposure to PM2.5 in 23 European cities. Eur. J. Epidemiol. 21, 449–58. doi:10.1007/s10654-006-9014-0

Brauer, M., Freedman, G., Frostad, J., van Donkelaar, A., Martin, R. V, Dentener, F., van Dingenen, R., Estep, K., Amini, H., Apte, J.S., Balakrishnan, K., Barregard, L., Broday, D.M., Feigin, V., Ghosh, S., Hopke, P.K., Knibbs, L.D., Kokubo, Y., Liu, Y., Ma, S., Morawska, L., Sangrador, J.L.T., Shaddick, G., Anderson, H.R., Vos, T., Forouzanfar, M.H., Burnett, R.T., Cohen, A., 2015. Ambient Air Pollution Exposure Estimation for the Global Burden of Disease 2013. Environ. Sci. Technol. acs.est.5b03709. doi:10.1021/acs.est.5b03709

Cao, R., Ma, Y.Z., Gomez, E., 2014. Geostatistical applications in petroleum reservoir modelling. J. South. African Inst. Min. Metall. 114, 625–629.

Carlton, A.G., Bhave, P. V, Napelenok, S.L., Edney, E.O., Sarwar, G., Pinder, R.W., Pouliot, G.A., Houyoux, M., 2010. Model representation of secondary organic aerosol in CMAQv4.7. Environ. Sci. Technol. 44, 8553–60. doi:10.1021/es100636q

Chan, M.N., Surratt, J.D., Claeys, M., Edgerton, E.S., Tanner, R.L., Shaw, S.L., Zheng, M., Knipping, E.M., Eddingsaas, N.C., Wennberg, P.O., Seinfeld, J.H., 2010. Characterization and Quantification of Isoprene-Derived Epoxydiols in Ambient Aerosol in the Southeastern United States. Environ. Sci. Technol. 44, 4590–4596.

Chen, J.-C., Wang, X., Wellenius, G. a, Serre, M.L., Driscoll, I., Casanova, R., McArdle, J.J., Manson, J.E., Chui, H.C., Espeland, M. a, 2015. Ambient air pollution and neurotoxicity on brain structure: evidence from Women's Health Initiative Memory Study. Ann. Neurol. 78, 466–76. doi:10.1002/ana.24460

Christakos, G., 2000. Modern spatiotemporal geostatistics. Oxford University Press, New York.

Christakos, G., Serre, M.L., 2000. BME analysis of spatiotemporal particulate matter distributions in North Carolina. Atmos. Environ. 34, 3393–3406.

Christakos, G., Serre, M.L., Kovitz, J.L., 2001. BME representation of particulate matter distributions in the measurements. J. Geophys. Res. 106, 9717–9731.

Crooks, J., Isakov, V., 2013. A wavelet-based approach to blending observations with deterministic computer models to resolve the intraurban air pollution field. J. Air Waste Manage. Assoc. 63, 1369–1385. doi:10.1080/10962247.2012.758061

de Nazelle, A., Arunachalam, S., Serre, M.L., 2010. Bayesian Maximum Entropy integration of ozone observations and model predictions: An application for attainment demonstration in North Carolina. Environ. Sci. Technol. 44, 5707–5713.

Dennis, R., Fox, T., Fuentes, M., Gilliland, A., Hanna, S., Hogrefe, C., Irwin, J., Rao, S.T., Scheffe, R., Schere, K., Steyn, D., Venkatram, A., 2010. A framework for evaluating regional-scale numerical photochemical modeling systems. Environ. Fluid Mech. 10, 471–489. doi:10.1007/s10652-009-9163-2

Dennis, R.L., Byun, D.W., Novak, J.H., Galluppi, K.J., Coats, C.J., Vouk, M. a., 1996. The next generation of integrated air quality modeling: EPA's models-3. Atmos. Environ. 30, 1925–1938. doi:10.1016/1352-2310(95)00174-3

Dennison, P.E., Brewer, S.C., Arnold, J.D., Moritz, M.A., 2014. Large wildfire trends in the western United States, 1984–2011. Geophys. Res. Lett. 41, 2928–2933. doi:10.1002/2014GL061184.Received

Di-Toro, D.M., McGrath, J.A., Hansen, D.J., 2000. Technical basis for narcotic chemicals and polycyclic aromatic hydrocarbon criteria.I. Water and tissue. Environ. Toxicol. Chem. 19, 1951–1970. doi:10.1897/1551-5028(2000)019<1951:TBFNCA>2.3.CO;2

Fann, N., Lamson, A.D., Anenberg, S.C., Wesson, K., Risley, D., Hubbell, B.J., 2012. Estimating the national public health burden associated with exposure to ambient PM2.5 and Ozone. Risk Anal. 32, 81–95. doi:10.1111/j.1539-6924.2011.01630.x

Foley, K.M., Dolwick, P., Hogrefe, C., Simon, H., Timin, B., Possiel, N., 2015a. Dynamic evaluation of CMAQ part II: Evaluation of relative response factor metrics for ozone attainment demonstrations. Atmos. Environ. 103, 188–195. doi:10.1016/j.atmosenv.2014.12.039

Foley, K.M., Hogrefe, C., Pouliot, G., Possiel, N., Roselle, S.J., Simon, H., Timin, B., 2015b. Dynamic evaluation of CMAQ part I: Separating the effects of changing emissions and changing meteorology on ozone levels between 2002 and 2005 in the eastern US. Atmos. Environ. 103, 247–255. doi:10.1016/j.atmosenv.2014.12.038

Foley, K.M., Roselle, S.J., Appel, K.W., Bhave, P. V., Pleim, J.E., Otte, T.L., Mathur, R., Sarwar, G., Young, J.O., Gilliam, R.C., Nolte, C.G., Kelly, J.T., Gilliland, a. B., Bash, J.O., 2010. Incremental testing of the Community Multiscale Air Quality (CMAQ) modeling system version 4.7. Geosci. Model Dev. 3, 205–226. doi:10.5194/gmd-3-205-2010

Fuentes, M., Raftery, A.E., 2005. Model evaluation and spatial interpolation by Bayesian combination of observations with outputs from numerical models. Biometrics 61, 36–45.

Guerreiro, C.B.B., Horálek, J., de Leeuw, F., Couvidat, F., 2016. Benzo(a)pyrene in Europe: Ambient air concentrations, population exposure and health effects. Environ. Pollut. 214, 657–667. doi:10.1016/j.envpol.2016.04.081

Jedynska, A., Hoek, G., Wang, M., Eeftens, M., Cyrys, J., Keuken, M., Ampe, C., Beelen, R., Cesaroni, G., Forastiere, F., Cirach, M., De Hoogh, K., De Nazelle, A., Nystad, W., Declercq, C., Eriksen, K.T., Dimakopoulou, K., Lanki, T., Meliefste, K., Nieuwenhuijsen, M.J., Yli-Tuomi, T., Raaschou-Nielsen, O., Brunekreef, B., Kooter, I.M., 2014. Development of land use regression models for elemental, organic carbon, PAH, and hopanes/steranes in 10 ESCAPE/TRANSPHORM European study areas. Environ. Sci. Technol. 48, 14435–14444. doi:10.1021/es502568z

Kang, D., Mathur, R., Schere, K., Yu, S., Eder, B., 2007. New categorical metrics for air quality model evaluation. J. Appl. Meteorol. Climatol. 46, 549–555. doi:10.1175/JAM2479.1

Kim, K.-H., Jahan, S.A., Kabir, E., Brown, R.J.C., 2013. A review of airborne polycyclic aromatic hydrocarbons (PAHs) and their human health effects. Environ. Int. 60, 71–80. doi:10.1016/j.envint.2013.07.019

Krewski, D., Jerrett, M., Burnett, R.T., Ma, R., Hughes, E., Shi, Y., Turner, M.C., Pope, C.A., Thurston, G., Calle, E.E., Thun, M.J., 2009. Extended follow-up and spatial analysis of the American Cancer Society study linking particulate air pollution and mortality. Respir. Rep Heal. Eff. Inst. 140, 5–114.

Lee, C.-L., Huang, H.-C., Wang, C.-C., Sheu, C.-C., Wu, C.-C., Leung, S.-Y., Lai, R.-S., Lin, C.-C., Wei, Y.-F., Lai, I.-C., Jiang, H., Chou, W.-L., Chung, W.-Y., Huang, M.-S., Huang, S.-K., 2016. A new grid-scale model simulating the spatiotemporal distribution of PM2.5-PAHs for exposure assessment. J. Hazard. Mater. 314, 286–294. doi:10.1016/j.jhazmat.2016.04.047

Lee, S., Serre, M.L., Donkelaar, A. Van, Martin, R. V, Burnett, R.T., Jerrett, M., 2012. Comparison of Geostatistical Interpolation and Remote Sensing Techniques for Estimating Long-Term Exposure to Ambient PM 2.5 Concentrations across the Continental United States. Environ. Health Perspect.

120, 1727–1732.

Liu, L.Y., Kukucka, P., Venier, M., Salamova, A., Klanova, J., Hites, R.A., 2013. Differences in spatiotemporal variations of atmospheric PAH levels between North America and Europe: Data from two air monitoring projects. Environ. Int. 64, 48–55. doi:10.1016/j.envint.2013.11.008

Menzie, C.A., Potocki, B.B., Santodonato, J., 1992. Exposure to Carinogenic PAHs in the Environment. Environ. Sci. Technol. 26, 1278–1284. doi:10.1021/es00031a002

Messier, K.P., Campbell, T., Bradley, P.J., Serre, M.L., 2015. Estimation of Groundwater Radon in North Carolina Using Land Use Regression and Bayesian Maximum Entropy. Environ. Sci. Technol. 49, 9817–9825. doi:10.1021/acs.est.5b01503

Morris, R.E., Koo, B., Guenther, A., Yarwood, G., McNally, D., Tesche, T.W., Tonnesen, G., Boylan, J., Brewer, P., 2006. Model sensitivity evaluation for organic carbon using two multi-pollutant air quality models that simulate regional haze in the southeastern United States. Atmos. Environ. 40, 4960–4972. doi:10.1016/j.atmosenv.2005.09.088

Motallebi, N., Taylor Jr, C.A., Croes, B.E., 2003. Particulate matter in California: Part 2-Spatial, temporal, and compositional patterns of PM2.5, PM10-2.5, and PM10. J. Air Waste Manage. Assoc. 53, 1517–1530. doi:10.1080/10473289.2003.10466322

Noth, E.M., Hammond, S.K., Biging, G.S., Tager, I.B., 2011. A spatial-temporal regression model to predict daily outdoor residential PAH concentrations in an epidemiologic study in Fresno, CA. Atmos. Environ. 45, 2394–2403. doi:10.1016/j.atmosenv.2011.02.014

Padula, A.M., Balmes, J.R., Eisen, E.A., Mann, J., Noth, E.M., Lurmann, F.W., Pratt, B., Tager, I.B., Nadeau, K., Katharine, S., 2015. Ambient polycyclic aromatic hydrocarbons and pulmonary function in children. J. Expo. Sci. Environ. Epidemiol. 25, 295–302. doi:10.1038/jes.2014.42.Ambient

Pleil, J.D., Vette, A.F., Rappaport, S.M., 2004. Assaying particle-bound polycyclic aromatic hydrocarbons from archived PM2.5 filters. J. Chromatogr. A 1033, 9–17. doi:10.1016/j.chroma.2003.12.074

Pope, C.A., Burnett, R.T., Thurston, G.D., Thun, M.J., Calle, E.E., Krewski, D., Godleski, J.J., 2004. Cardiovascular Mortality and long-term exposure to particulate air pollution: Epidemiological evidence of general pathophysiological pathways of disease. Circulation 109, 71–7. doi:10.1161/01.CIR.0000108927.80044.7F

Pope, C.A., Ezzati, M., Dockery, D.W., 2009. Fine-particulate air pollution and life expectancy in the United States. N. Engl. J. Med. 360, 376–86. doi:10.1056/NEJMsa0805646

Ravindra, K., Sokhi, R., van Grieken, R., 2008. Atmospheric polycyclic aromatic hydrocarbons: Source attribution, emission factors and regulation. Atmos. Environ. 42, 2895–2921. doi:10.1016/j.atmosenv.2007.12.010

Reyes, J., Xu, Y., Vizuete, W., Serre, M.L., 2016. Regionalized PM2.5 Community Multiscale Air Quality model performance evaluation across a continuous spatiotemporal domain. Atmos. Environ. submitted.

Reyes, J.M., Serre, M.L., 2014. An LUR/BME framework to estimate PM2.5 explained by on road mobile and stationary sources. Environ. Sci. Technol. 48, 1736–44. doi:10.1021/es4040528

Ribeiro, M.C., Pinho, P., Llop, E., Branquinho, C., Pereira, M.J., 2015. Geostatistical uncertainty of assessing air quality using high-spatial-resolution lichen data: A health study in the urban area of Sines, Portugal. Sci. Total Environ. 562, 740–750. doi:10.1016/j.scitotenv.2016.04.081

Simon, H., Baker, K.R., Phillips, S., 2012. Compilation and interpretation of photochemical model performance statistics published between 2006 and 2012. Atmos. Environ. 61, 124–139.

doi:10.1016/j.atmosenv.2012.07.012

Solazzo, E., Galmarini, S., 2016. Error apportionment for atmospheric chemistry-transport models: a new approach to model evaluation. Atmos. Chem. Phys. Discuss. 1–39. doi:10.5194/acp-2016-15

Steyn, D.G., de Wekker, S.F.J., Kossmann, M., Martilli, A., 2013. Boundary Layers and Air Quality in Mountainous Terrain, in: Chow, F.K., de Wekker, S.F.J., Snyder, B.J. (Eds.), Mountain Weather Research and Forecasting. Springer Netherlands, pp. 219–260. doi:10.1007/978-94-007-4098-3

Tang, Q., Bo, Y., Zhu, Y., 2016. Spatio-temporal fusion of multiple satellite aerosol optical depth (AOD) products using Bayesian Maximum Entropy method. J. Geophys. Res. Atmos. n/a–n/a. doi:10.1002/2015JD024571

Thunis, P., Pederzoli, A., Pernigotti, D., 2012. Performance criteria to evaluate air quality modeling applications. Atmos. Environ. 59, 476–482. doi:10.1016/j.atmosenv.2012.05.043

United States Geological Survey, 2016. Federal Wildland Fire Occurrence Data [WWW Document]. URL http://wildfire.cr.usgs.gov/firehistory/data.html

US EPA, n.d. Air Quality System (AQS) [WWW Document]. URL http://www.epa.gov/ttn/airs/airsaqs/ (accessed 9.11.10).

USEPA, 2005. CMAQ Model Performance Evaluation for 2001: Updated March 2005, annual report.

USEPA, 2001. National Air Quality and Emission Trends Report, 1999. Research Triangle Park, NC.

van Donkelaar, A., Martin, R. V., Brauer, M., Boys, B.L., 2015. Use of satellite observations for long-term exposure assessment of global concentrations of fine particulate matter. Environ. Health Perspect. 123, 135–143. doi:10.1289/ehp.1408646

Venkatram, A., 2008. Computing and displaying model performance statistics. Atmos. Environ. 42, 6862–6868. doi:10.1016/j.atmosenv.2008.04.043

Wolff, M.S., Teitelbaum, S.L., Lioy, P.J., Santella, R.M., Wang, R.Y., Jones, R.L., Caldwell, K.L., Sjödin, A., Turner, W.E., Li, W., Georgopoulos, P., Berkowitz, G.S., 2005. Exposures among pregnant women near the World Trade Center Site on 11 September 2001. Environ. Health Perspect. 113, 739–748. doi:10.1289/ehp.7694

Xu, Y., Serre, M.L., Reyes, J.M., Vizuete, W., 2016. Bayesian Maximum Entropy integration of ozone observations and model predictions: A national application. Environ. Sci. Technol. 50, 4393–4400. doi:10.1021/acs.est.6b00096

Yu, S., Mathur, R., Pleim, J., Pouliot, G., Wong, D., Eder, B., Schere, K., Gilliam, R., Rao, S.T., 2012. Comparative evaluation of the impact of WRF/NMM and WRF/ARW meteorology on CMAQ simulations for PM 2.5 and its related precursors during the 2006 TexAQS/GoMACCS study. Atmos. Chem. Phys. 12, 4091–4106. doi:10.5194/acp-12-4091-2012

Yu, S., Mathur, R., Schere, K., Kang, D., Pleim, J., Young, J., Tong, D., Pouliot, G., Mckeen, S.A., Rao, S.T., 2008. Evaluation of real-time PM 2.5 forecasts and process analysis for PM 2.5 formation over the eastern United States using the Eta-CMAQ forecast model during the 2004 ICARTT study. J. Geophys. Res. 113. doi:10.1029/2007JD009226

Zhang, Y., Tao, S., 2009. Global atmospheric emission inventory of polycyclic aromatic hydrocarbons (PAHs) for 2004. Atmos. Environ. 43, 812–819. doi:10.1016/j.atmosenv.2008.10.050