Constrained Positive Matrix Factorization: Elemental Ratios, Spatial Distinction, and Chemical Transport Model Source Contributions

Timothy M. Sturtz

A dissertation submitted in partial fulfillment of
the requirements for the degree of

Doctor of Philosophy

University of Washington

2014

Reading Committee:

Timothy V. Larson, Chair

Dean A. Hegg

Bret A. Schichtel

Program Authorized to Offer Degree:
Civil & Environmental Engineering Department

©Copyright 2014
Timothy M. Sturtz

University of Washington

Abstract

Constrained Positive Matrix Factorization: Elemental Ratios, Spatial Distinction, and Chemical Transport Model Source Contributions

Timothy M. Sturtz

Chair of the Supervisory Committee:

Atmospheric source apportionment models attempt to untangle the relationship between pollution sources and the impacts at downwind receptors. Two frameworks of source apportionment models exist: source-oriented and receptor-oriented. Source based apportionment models use presumed emissions and atmospheric processes to estimate the downwind source contributions. Conversely, receptor based models leverage speciated concentration data from downwind receptors and apply statistical methods to predict source contributions. Integration of both source-oriented and receptor-oriented models could lead to a better understanding of the implications pollution sources have on the environment and society. The research presented here investigated three different types of constraints applied to the Positive Matrix Factorization (PMF) receptor model within the framework of the Multilinear Engine (ME-2): element ratio constraints, spatial separation constraints, and chemical transport model (CTM) source attribution constraints.

$PM_{10-2.5}$ mass and trace element concentrations were measured in Winston-Salem, Chicago, and St. Paul at up to 60 sites per city during two different seasons in 2010. PMF was used to explore the underlying sources of variability. Information on previously reported $PM_{10-2.5}$ tire and brake wear profiles were used to constrain these features in PMF by prior specification of selected species ratios. We also modified PMF to allow for combining the measurements from all three cities into a single model

while preserving city-specific soil features. Relatively minor differences were observed between model predictions with and without the prior ratio constraints, increasing confidence in our ability to identify separate brake wear and tire wear features.

Source contributions to total fine particle carbon predicted by a CTM were incorporated into the PMF receptor model to form a receptor-oriented hybrid model. The level of influence of the CTM versus traditional PMF was varied using a weighting parameter applied to an object function as implemented in ME-2. The resulting hybrid model was used to quantify the contributions of total carbon from both wildfires and biogenic sources at two Interagency Monitoring of Protected Visual Environment monitoring sites, Monture and Sula Peak, Montana, from 2006 through 2008. At the weighting parameter associated with a minimum cross-validated RMSE for each site, the profiles and contributions were reasonably correlated with the CTM while maintaining a strong fit to the measurements. The cross-validated RMSE of total carbon for both sites was improved over the pure CTM or PMF predictions, indicating an improvement in the ability to fit total carbon.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

The author wishes to express sincere appreciation to his advisor Timothy Larson for providing thoughtful guidance and support throughout his tenure as a graduate student. Additionally, the author would like to thank his committee memebers for their time and efforts.

DEDICATION

This dissertation is dedicated to my father who impressed upon me the importance of education, my mother who has always helped me maintain perspective, my wife who has always supported and encouraged me, and to the rest of my family.

Chapter 1
BACKGROUND

1.1 Introduction

Source apportionment models attempt to untangle the relationship between pollution sources and the impacts at downwind receptors. Two frameworks of source apportionment models exist: source-oriented and receptor-oriented. Source based apportionment models use presumed emissions and atmospheric processes to estimate the downwind source contributions. Conversely, receptor based models leverage speciated concentration data from downwind receptors and apply statistical methods to predict source contributions. Frequently, one modeling approach acts in a supporting role to the other, helping to understand model error. Integration of both source-oriented and receptor-oriented models to leverage the positive aspects of each model type could lead to a better understanding of the implications each source has on the environment and society.

Understanding the contributions and chemical makeup of different source types in a region may benefit numerous entities. For example, knowledge of impacting sources allows regulatory agencies to develop mitigation strategies, industries to bolster a litigation defense, or researchers to improve health analyses. Further, the public can remain more informed about the sources impacting their region, feasibly influencing votes and public forums. However, realization of the benefits from source apportionment depends upon the accuracy of the modeling predictions.

Source-oriented apportionment methods apply chemical transport models (CTMs) to predict source contributions at monitor locations. CTMs, such as the Community

Multiscale Air Quality (CMAQ) model or the Comprehensive Air Quality Model with Extensions (CAMx), incorporate dispersion and chemical mechanisms that allow for the estimation of primary and transformed, or secondary, pollutant concentrations downwind. The current state of science provides a robust understanding of many atmospheric processes, however it is impossible to perfectly incorporate every transport and chemical reaction within a model. Source contributions derived from CTMs do not have associated error bounds which adds difficulty in establishing where the models err.

Receptor-based source apportionment analyzes speciated concentration data by applying statistical techniques to reveal underlying multivariate features. Receptor models, such as the Environmental Protection Agency (EPA) Positive Matrix Factorization (PMF) and EPA Chemical Mass Balance (CMB) model, decompose monitor data by solving a chemical mass balance equation for a predetermined number of sources. The use of measurement concentrations and associated uncertainties as model input allows receptor models to quantify modeling errors and better understand inaccuracies. However, receptor-oriented models cannot readily discern between primary and secondary sources due to the lack of a chemistry mechanism, accompanying meteorological information, and likely collinearity between primary and secondary features within the data.

Often, receptor modeling output are verified against source-oriented results or emission inventories linked with meteorological data. A one-to-one comparison can be made if CTM results are available. Lacking CTM results, verification can be determined by linking emission inventories and receptor modeling results through the use of back trajectories. These trajectories can be determined for individual dates and times or can be analyzed over a time period using geostatistical methods such as the potential source contribution function (PSCF) [50]. Results of the PSCF identifies the regions with the highest likelihood of contributing to the modeled receptor, thereby indicating the region of the emission inventory for comparison.

Schichtel et al. [98] proposed a hybrid PMF modeling technique that directly implements CTM results into a receptor modeling framework, integrating the advantages of both modeling approaches. Applied to a synthetic data set, the modeling technique estimated combined primary and secondary contribution of biomass burning to fine particulate matter with greater accuracy than a CTM or receptor model alone. Using this method in conjunction with real-world data could provide improved regional mitigation strategies and ability to decipher CTM modeling errors.

The proposed research will investigate a modified hybrid-receptor approach aimed at apportioning primary and secondary pollutants using two approaches: one aimed at providing verification through the use of a synthetic data set and the other aimed at testing the hybrid approach against real measurements. Additionally, the research will assess PMF modeling of spatially distributed sites and the separation of ubiquitous and unique source impacts. The results of this research could provide a platform for integration of source-oriented and receptor-oriented models and ultimately contribute to the scientific community's understanding of source-receptor relationships.

1.2 Source Apportionment

This section reviews various methods of both source-oriented and receptor-oriented source apportionment. The associated techniques, skills, and challenges of each approach are presented and discussed. Since the proposed research specifically focuses on the apportionment of primary and secondary pollutants, the section concludes with a discussion of previously pursued hybrid methods.

1.2.1 Receptor-Oriented Apportionment

Receptor based source apportionment is commonly conducted using principal component analysis (PCA), EPA Chemical Mass Balance Model (CMB), or EPA Positive Matrix Factorization (PMF) [30]. Each model is based primarily on speciated receptor data, x, and, in the case of CMB and PMF, the associated uncertainty, σ. The

methodologies, advantages, and disadvantages of each is described in the following sections.

Principal Component Analysis

The approach of PCA reduces speciated receptor data, x, into subset linear combinations of species, or features, which describe the natural variance and correlations that exist within the data [47]. The initial feature determined from PCA explains the greatest amount of variation within the data, while each subsequent feature explains less variation and remains mutually uncorrelated with the prior features[122]. The requirement of mutually uncorrelated features would not allow PCA to accurately apportion primary and secondary features since correlations often exist between the two. Secondly, uncertainty is not incorporated in PCA but varies between species and samples due to measurement methods. For these reasons PCA is not an ideal candidate for apportioning environmental data where collinearity and uncertainties are ever present.

Chemical Mass Balance Models

The generic chemical mass balance model [68] provides the basis for both EPA CMB and PMF. Conceptually, the model represents the idea that monitored concentration data, x_{ij}, is a function of source profiles, f_{kj}, and source contributions, g_{ik}, where the indices i,j,and k represent the number of samples, species, and contributing sources respectively. The generic chemical mass balance equation is presented in equation 1.1 below.

$$x_{ij} = \sum_{k=1}^{p} g_{ik} f_{kj} + e_{ij} \tag{1.1}$$

Where e_{ij} is defined as the residual variable, accounting for modeling error. While both CMB and PMF solve this model, the method deployed is significantly different.

EPA CMB

The approach of CMB requires speciated data, known source profiles, and the associated source profile uncertainties. The requirement of source profiles assumes the user has a thorough understanding of the sources impacting the monitor of study, often obtained through emission inventories or source profile databases. However, the development of emission inventories require numerous assumptions and likely does not capture the full extent of sources or emissions. Existing source profile databases (e.g., EPA SPECIATE) attempt to generalize the source profiles by source type which may not accurately represent the sources of interest. Further, transformed secondary pollutants impacting the monitor are not represented in emission inventories or profile databases. The potential errors in the *a priori* source profiles and lack of accurate secondary profiles eliminate CMB as a candidate for the hybrid approach.

EPA PMF

The EPA Positive Matrix Factorization (PMF) model attempts to solve equation 1.1 without prior knowledge of source profiles or contributions, decomposing the concentration data using a bilinear factor analytic approach (through a table-driven least squares methodology) [72], and outputting matrices of source profiles and contributions. This multivariate approach uses an iterative process to minimize an object function, providing results which best fit the data. The object function is defined in equation 1.2.

$$Q = \sum_{i=1}^{n} \sum_{j=1}^{m} \left[\frac{x_{ij} - \sum_{k=1}^{p} g_{ik} f_{kj}}{\sigma_{ij}} \right]^2 \tag{1.2}$$

An associated uncertainty, σ_{ij}, for the species, i, and samples, j, exists to allow for variation in confidence across the sampled species and is incorporated in the object

function. The uncertainties are often developed from the analytic uncertainty and method detection limit associated with the individual species. The mass balance model is fit using a conjugate gradient (CG) algorithm which minimizes the object function described in equation 1.2.

The model is initialized with random data and executed multiple times to ensure a consistent minimum in the CG results. Further, blocked bootstrapping is often applied to quantify the level of confidence for each profile and contribution pair.

The multivariate structure and flexibility of the Multilinear Engine 2 (ME-2), the backend to PMF, allows for a variety of new constraints to be implemented. The latest version of PMF allows for user input of some constraints but does not give the user access to the full potential of the ME-2 model capabilities.

1.2.2 Source-Oriented Apportionment

While not the focus of this proposal, source-oriented apportionment functions as a tool to aid in understanding the source-receptor relationship. As previously mentioned, source-oriented apportionment is frequently conducted using CTMs. In general, these models implement tracers to track species as they move downwind and document the species composition to provide apportionment by source type. In the case of primary pollutants, a single tracer is used, however, with secondary pollutants a suite of reactive tracers must be used to track the changing composition over time. In contrast to typical CTMs, the CAPITA Monte Carlo model tracks each discrete event [95], allowing for analysis of the chemistry and transport of each trajectory. With any source-oriented apportionment the result is the same, a set of pollutant contributions by source type. However, the contributions may not agree well with monitored concentrations because of incomplete transport and chemistry. Validation of source-oriented model results is often conducted using traditional receptor-oriented models such as those described in section 1.2.1.

1.2.3 Other Supporting Methods

Receptor modeling is often verified using emission inventories, CTMs, and/or spatial analysis. Using back trajectories or back dispersions from a modeled receptor, correlations can be made between large peaks in pollutant concentrations and known influential sources. Certain analyses may also find use in the Potential Source Contribution Function (PSCF) [12]. Using a large collection of back trajectories within a gridded domain, the frequency of the trajectories passing through grid cells is calculated, thereby indicating probable locations of sources. The formal definition of PSCF is provided in equation 1.3, where m_{ab} is the total number of trajectory endpoints, n_{ab} is the number of endpoints in a single grid cell, and the indices, a and b, represent the grid location.

$$PSCF_{ab} = \frac{m_{ab}}{n_{ab}} \tag{1.3}$$

Other spatial analyses, such as pollution roses, have also been used to support receptor modeling results.

1.2.4 Hybrid Approaches

Numerous approaches to CTM and spatial hybrid source apportionment modeling have been studied in recent years. Schichtel et al. [98] incorporated combined primary and secondary contributions from the CAPTIA Monte Carlo model in ME-2 using a synthetic data set. The results from this work indicated that constraining ME-2 with CTMs can improve the apportionment of total carbon. By using synthetic data the true source-specific contributions were known and were available for confirmation of the models skill.

The hybrid approach outlined here seeks to separate primary and secondary features by constraining ME-2 using CTM primary and secondary results. Initially, a synthetic data set was be leveraged to test and verify the ME-2 scripts. With a

successful separation of primary and secondary features in the synthetic data, the hybrid method was then applied to real-world ambient measurements to investigate the overall value of the model.

1.3 Dissertation Intent

Aim 1: Evaluate spatially dispersed measurements using PMF with elemental ratio constraints and spatial separation

One intent of this work was to use PMF to assess spatially dispersed data using constraints to separate specific source types and to delineate spatially unique features. Coarse mode particulate data sampled for two weeks across three cities during two different seasons was used as the basis of the analysis. Elemental ratios common to brake wear and tire wear were used as source constraints and the data were spatially subset by city to obtain unique wind-blown dust features. Investigation of the influence on the profiles from the ratio constraints and the influence on the wind-blown dust features from the spatial separation technique are explored.

Aim 2: Implement and evaluate the hybrid approach using IMPROVE monitor data at two sites

A methodology of combining source attributions from CTM and PMF was implemented to distinguish biomass combustion from secondary biogenic contributions. The method was tested on measurements from two different IMPROVE monitor sites located in Montana. Validation of the model was conducted using a novel 10-fold cross-validation approach. The relationship between a pure PMF solution, a pure CTM solution, and the hybrid methodology is investigated.

Chapter 2

CONSTRAINED SOURCE APPORTIONMENT OF COARSE PARTICULATE MATTER AND SELECTED TRACE ELEMENTS IN THREE CITIES[1]

2.1 Abstract

$PM_{10-2.5}$ mass and trace element concentrations were measured in Winston-Salem, Chicago, and St. Paul at up to 60 sites per city during two different seasons in 2010. Positive Matrix Factorization (PMF) was used to explore the underlying sources of variability. Information on previously reported $PM_{10-2.5}$ tire and brake wear profiles was used to constrain these features in PMF by prior specification of selected species ratios. We also modified PMF to allow for combining the measurements from all three cities into a single model while preserving city-specific soil features. Relatively minor differences were observed between model predictions with and without the prior ratio constraints, increasing confidence in our ability to identify separate brake wear and tire wear features. Brake wear, tire wear, fertilized soil, and re-suspended soil were found to be important sources of copper, zinc, phosphorus, and silicon respectively across all three urban areas.

---

[1]As of November 2013, this chapter has been accepted for publication in *Atmospheric Environment*, authors T.M. Sturtz (Department of Civil and Environmental Engineering, University of Washington, Box 352700, Seattle, WA 98195-2700), S.D. Adar (Department of Epidemiology, School of Public Health University of Michigan, 1415 Washington Heights, Ann Arbor, MI 48109-2029), T. Gould (Department of Civil and Environmental Engineering, University of Washington, Box 352700, Seattle, WA 98195-2700), and T.V. Larson (Department of Civil and Environmental Engineering, University of Washington, Box 352700, Seattle, WA 98195-2700 and Department of Environmental and Occupational Health Sciences, University of Washington, Box 357234, Seattle, WA 98195-7234).

2.2 Introduction

There is ample evidence that long-term exposure to fine airborne particles ($PM_{2.5}$) is detrimental to human health [86, 109]. In contrast, our understanding of the long-term effects of the coarse particle fraction ($PM_{10-2.5}$) is more limited [32, 85, 17, 121, 87, 60]. One major challenge for chronic epidemiological studies is in accurately describing the long-term spatial gradients in coarse mode mass and species concentrations within urban areas. Recent work has focused on characterizing $PM_{10-2.5}$ spatial concentration gradients [51, 43, 107, 69, 25, 34, 27, 103] and developing models to allow spatial interpolation [123, 83, 33]. Another challenge is to characterize the sources that influence these gradients as well as the species that are associated with these sources.

Prior source apportionment studies of $PM_{10-2.5}$ have relied on either fully constrained models such as chemical mass balance (CMB), principal component analysis (PCA) and mass closure [81, 62, 5, 102, 31, 74, 114, 78, 25], partially constrained models such as the constrained physical receptor model (COPREM) [115, 93], or relatively unconstrained models such as factor analysis or positive matrix factorization (PMF) [117, 41, 14, 13, 58, 45, 73, 106, 64, 22, 51, 43]. Several of these studies have employed multiple sites within a city to capture spatial as well as temporal variability in the source contributions [102, 64, 25, 22, 51, 43, 78].

While there has been a number of near-roadway studies examining the sources and components of non-exhaust $PM_{10-2.5}$ [108, 48, 10, 36, 93, 46, 57, 39, 52, 42, 44, 19, 113, 101, 2, 55, 7, 116, 18], only a few of the urban-scale source apportionment studies cited earlier have attempted to separate "road dust" into its separate components, including brake wear and tire wear [116, 7, 93, 18, 48]. The studies which did not separate road dust into its components commonly identified the dominant source of $PM_{10-2.5}$ as resuspended road dust for sites near roadways and as crustal material for non-roadway sites.

Here we use a partially constrained version of PMF [6, 88, 16] in order to examine

the sources of $PM_{10-2.5}$ collected simultaneously at multiple sites in three urban areas during two-week periods in two different seasons. We use PMF with constraints imposed by prior knowledge of several important, ubiquitous source profiles, namely brake and tire wear. We furthermore impose additional constraints on the source contributions in order to combine all measurements into a single model. To our knowledge, this is the first application of a combined-cities PMF modeling approach with profile-constraints to identify contributions of brake and tire wear in $PM_{10-2.5}$ across multiple urban areas. This work is part of a larger effort to examine the chronic health effects of $PM_{10-2.5}$ and selected species in these same cities under the auspices of the Multi-Ethnic Study of Atherosclerosis and Coarse Particulate Matter (MESA Coarse), an ancillary study of the Multi-Ethnic Study of Atherosclerosis and Air Pollution (MESA Air).

2.3 Materials & Methods

2.3.1 Filter sampling and analysis

The MESA Air study leveraged the National Heart, Lung, and Blood Institute's Multi-Ethinic Study of Atherosclerosis (MESA) cohort to provide data for assessing the relationship between long-term exposures to fine ambient particulates and related health effects. The MESA cohort [56] was comprised of 6,814 white, black, Hispanic, and Chinese participants located in six U.S. cities. As an ancillary study to MESA Air, MESA Coarse assesses the health implications associated with coarse mode particulate exposure in three of the MESA cohort cities, namely Chicago, Illinois, St. Paul, Minnesota, and Winston-Salem, North Carolina.

Paired, two-week average PM_{10} and $PM_{2.5}$ Teflon filter samples were simultaneously collected over two different two-week periods, in the winter and summer of 2009, in Chicago, IL, St. Paul, MN, and Winston-Salem, NC. The monitoring sites in each city (see Figure 2.1) were residential locations of the existing MESA cohort selected

to maximize variability in geographic features expected to influence coarse particles including land use, roadways, and vegetation as well as representative community monitoring sites. $PM_{10-2.5}$ mass concentrations were computed by the difference in collocated PM_{10} and $PM_{2.5}$ measurements. This "difference method" has been shown to be a reliable approach in estimating $PM_{10-2.5}$ in urban areas by the U.S. Environmental Protection Agency [24]. At affiliated field centers in each sampled city, the Teflon filters were loaded into Harvard personal environmental monitors (HPEMs, Harvard School of Public Health, Boston, MA). These monitors were connected to a Medo VP0125 (MEDO USA, Inc., Roselle, IL) vacuum pump drawing 1.8 L/min air sample and equipped with a timer valve system that obtained a 50% duty cycle sample, where the flow alternated between the PM_{10} and $PM_{2.5}$ filter every 5 minutes to avoid filter overload.

PM_{10} and $PM_{2.5}$ mass concentrations were gravimetrically determined from weighing of Teflon filters at the University of Washington in a temperature and humidity controlled environment (Allen et al, 2001), and from the total volumetric flow of air sampled through the HPEMs. A Mettler-Toledo UMT-2 balance was used to determine sample mass following standard filter weighing procedures. Overall, the precision of duplicate PM_{10}, $PM_{2.5}$ and $PM_{10-2.5}$ samples as measured by the average Relative Percent Difference was 2%, 10% and 18%, respectively. The filter samples were analyzed for a suite of 48 elements by X-Ray Fluorescence (XRF) at Cooper Environmental Services (Portland, OR). Method sensitivity was defined by a set of acceptable detection levels for a subset of 21 key elements from the Method IO-3.3 analyte list. The quality assurance and quality control data are provided in Tables 2.10 and 2.11.

2.3.2 PMF Model Inputs

Measurement uncertainty for coarse mode species j, σ_j, was calculated by combining the uncertainties of the PM_{10} and $PM_{2.5}$ measurements using standard error propa-

gation as follows.

$$\left(\sigma_j^2\right)_{PM_{10-2.5}} = \left(\sigma_j^2\right)_{PM_{10}} + \left(\sigma_j^2\right)_{PM_{2.5}} \tag{2.1}$$

The measured coarse mode species concentrations were pre-processed to remove frequently below detection species and species with a signal to noise [71], S/N, ¡10. In addition, pre-processing included removal of sulfur samples identified as outliers (exceeding 2 standard deviations from the mean). Four samples were removed based on this criterion. The S/N cutoff choice was motivated by the consistently high signal to noise ratios of a subset of species and relatively low and variable ratios for some species depending upon city. Enrichment of the coarse mode for certain elements is not unexpected and has been documented in other literature [7, 106, 21]. The S/N criteria eliminated the following species: Ag, As, Au, Cd, Ce, Co, Cs, Eu, Ga, Hf, Hg, In, Ir, La, Mo, Nb, Rb, S, Sc, Se, Sm, Sn, Ta, Tb, V, W, and Y (see Table 2.4 in Appendix A). Although Sb had S/N < 10, we chose to include it in the models because of its value as a brake wear constraint variable described in the next section. We retained $PM_{10-2.5}$ mass but increased its uncertainty by a factor of 30 to avoid redundancy with all other measured species. The retention of coarse mass allows for the production of feature profiles in a gram per gram $PM_{10-2.5}$ basis. There were no missing species measurements. We also ran the models including all species with S/N ¿ 2 without any significant difference in our final results.

2.4 Theory/calculation

We implemented the Positive Matrix Factorization (PMF) receptor model using the Multilinear Engine version 2 (ME-2) (Paatero, 1999). The PMF model solves the basic mass balance equation (Equation 2.2) for source contributions, g_{ik} , source profiles, f_{kj} , and model error, e_{ij} , for i=1,n samples, j=1,m species, and k=1,p sources. Species concentrations, x_{ij} , corresponding uncertainties, σ_{ij}, and the user-defined number of

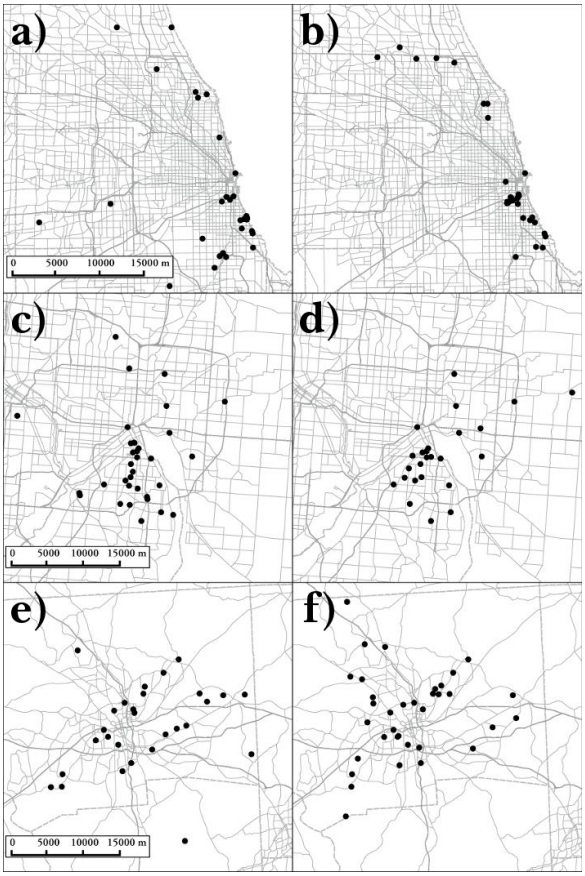

Figure 2.1: Map of Sampling Locations in Chicago (a, b), St. Paul (c, d) and Winston-Salem (e, f). Samples were taken in the Winter or Early Spring (left panels) and in the Summer (right panels)

sources, p, serve as the model input.

$$x_{ij} = \sum_{k=1}^{p} g_{ik} f_{kj} + \varepsilon_{ij} \qquad \text{where } g_{ik}, f_{kj} > 0 \tag{2.2}$$

and the g_{ik} are normalized by their average value across all samples such that

$$\bar{g}_{ij} = \sum_{i=1}^{n} \frac{g_{ik}}{n} = 1 \pm \delta \tag{2.3}$$

where $\delta = 0.01$ in this case.

Equations 2.2 and 2.3 comprise the basic PMF model. To add prior source profile constraints, we have added an additional set of equations that are solved simultaneously with equations 2.2 and 2.3. In this case we have added equations representing each of t=1,v species constraints using prior knowledge of two sources: brake wear (k=1) and tire wear (k=2). The t^{th} constraint is shown in equation 2.4.

$$f_{kq} - \lambda_t f_{kr} = 0 \tag{2.4}$$

where the kth source profile (k =1 or 2) is constrained using the rth and qth species, and λ_t represents the value of the species ratio for that source profile, f_{kq} / f_{kr}. The constraints were developed from a literature review of brake wear and tire wear source profiles (Table 2.1). The median values for each reported ratio were used.

We present, in equation 2.4, ratio constraints in the form of a difference with a target of zero. However, within the code of ME-2 we define sub-expressions to invert the denominator element for each ratio constraint and then define an auxiliary equation to represent the ratio. Thus, the ratio is constrained to the target value, λ_t, not zero. Due to the construct within ME-2 we applied error mode -12. The alternative, error mode -5, is limited to special cases where the target is zero (Paatero 2009). Through the use of error mode -12 and our ratio constraints we are able to control the order of the constrained model results. This approach differs from other similar studies which model the data unconstrained, identify source-like features, and then pull up or down (error mode -22) to a desired target.

In ME-2, equations 2.2 through 2.4 are solved by minimizing an object function, Q, through the use of a preconditioned conjugate gradient algorithm. The object function (equation 2.5) includes a penalty, q_t, associated with the applied constraints.

Table 2.1: Coarse PM source profile constraints.

Constraint type	Species	Source				Literature
		Brake Wear	Tire Wear	Road Salt	Soil(s)	Citations[e]
Species	Cu/Sb	3.9 (0.1)[b]				1-12
Ratio (λ_t)[a]	Cu/Ba	1.2 (0.1) b				1-9,12
	Cu/Fe	0.05 (0.1) b				1-8, 12
	Zn/Pb		1000 (0.1)[b]			13,14
	Zn/K		24 (0.1)[b]			13,14
	Zn/Ca		26 (0.1)[b]			13,14
Upper or lower	Ba	>0.001				1-9, 12
limit	Cu	>0.001				1-9,12
constraints	Pb		<0.001			16
on average	Zn		>0.001			13-15
species	Cl			>0.01		
concentration	Na			>0.01		
(ng/m3)	β_c				0 (1e-5)[c]	
Maximum	Si				400[d]	
constraint						
on Q						

[a] See Eq. (4) in text.

[b] Maximum allowable constraint error.

[c] Dimensionless value for c not equal to k - 2 see Eq. (6) in text.

[d] maximum allowable increase of Q for upward pulling of average species concentration using error mode -20 in ME2 see also Eq. (5) in text.

[e] 1. Kennedy and Gadd, 2003; 2.Garg et al., 2000; 3. Iijima et al., 2007; 4. Geitel et al., 2010; 5. Schauer et al., 2006a; 6. Grieshop et al., 2005; 7. Bukoweicki et al., 2010; 8. Von Uexkll et al., 2005; 9. Sternbeck et al., 2004; 10. Weckwerth, 2001; 11. Adachi and Tainosho, 2004, 12. Johansson et al., 2009; 13. Apeagyei et al., 2011; 14. EPA, 2003; 15. Han et al., 2011. 16. Wahlin et al., 2006.

$$Q = \sum_{i=1}^{n} \sum_{j=1}^{m} [\varepsilon_{ij}]^2 + \sum_{t=1}^{v} q_t \tag{2.5}$$

We implement this in ME-2 by specifying auxiliary equations for q_t using error mode -12 and penalty values of 0.1. For constraint t in equation 2.4, these penalty values define the maximum allowable error from the defined constraint. The penalty values associated with a given constraint are shown in Table 2.1.

In order to increase the number of measurements used in the model, we combined all samples into one larger combined-city model. We hypothesize that the brake and tire wear profiles are universally applicable across all three cities. However, we assume that the soil profiles differ by geographic region of the country and therefore differ by city. To address this issue, we modified the model to allow three separate soil profiles, one for each city, while keeping all other features in common across cities. For the three soil sources (k=3 to 5), we allowed only one unique source in each of the three cities (c=1 to 3) as follows:

$$\text{where } \beta_c = \left\{ \begin{array}{ll} 1 & \text{for } c = k-2 \\ 0 \pm 1 \times 10^{-5} & \text{for } c \neq k-2 \end{array} \right\} \tag{2.6}$$

We did this by enforcing hard constraints in the form of additional auxiliary equations with a target of zero and a tolerance for error of 1e-5 ug/m3 on the contributions from the soil sources of two of the three cities. We also included additional upward pulling of Si on the sources to ensure the features were soil related (see Table 2.1); the pulling was limited by a maximum change in Q of 400. In addition, to insure that it is the soil profile that we are restricting with equation 2.6, we add additional profile constraints for k = 3 to 5 (see Table 2.1) based on prior knowledge of the soil-derived $PM_{10-2.5}$ from the literature as well as from the individual city model predictions.

$$f_{kq} - \lambda_t f_{kr} > 0 \tag{2.7}$$

For an initial user-specified value of p, multiple model runs were conducted at 20 different starting points chosen randomly, and the chosen p-source baseline model was the one with the minimum value of Q (= Qmin). Separate model runs were then made assuming a range of values of p. The final value of p was chosen based on the following criteria: the ratio as a function of p between Qmin and Qtheoretical = n*ms +n*mw/3 - n*p, where ms is number of strong species and mw is the number of weak species; changes in Q as a function of p (Comero, Capitani, and Gawlik, 2009), the relationship between each f_{kj} and prior knowledge of source profiles; the known source types within the modeled region; and user judgment. The values of Q/Qtheoretical, degrees of freedom, and the selected number factors for each model are presented in Table 2.12 and plots of Q/Qtheoretical for various p are provided in Figure 2.6.

Blocked bootstrapping (Norris & Vedantham, 2008) was then applied to the baseline model results, providing an estimate of the confidence limits of the fkj and the average values of gik by city. Given that the 2-week samples were collected simultaneously over space in each city only twice and only a fraction were re-sampled in both seasons, a sample block size of 5 was defined to contain samples within a season for a given city. Profile matching was done on the predicted contributions of Ba, Br, Cl, Cu, Fe, Mg, Na, Ni, P, Pb, Si, Zn and Zr, with an acceptable match defined as an R^2 ¿ 0.6 across all contributions for a given model run.

To assess the effect of the prior profile constraints, the model was run with and without these constraints. In the latter case, $q_t = 0$. We also ran the constrained and unconstrained models for each city separately (in this case equations 2.6 and 2.7 were not used).

2.5 Results

In addition to the brake wear, tire wear, and city-specific soil features, the combined cities model was able to identify three additional source-related features: fertilized

soil, road salt, and a feature enriched in Pb. For individual city models, the features in addition to brake and tire wear and soil were: fertilized soil in all cities, a metals-rich feature in Chicago, and a road salt feature in St. Paul. The source-related profiles are shown in Figures 2.2 to 2.5. The gik are normalized in such a way (equation 2.3) that the f_{kj} (vertical bars in each figure) represent the average species concentrations contributed by a given feature across all samples. Table 2.5 in Appendix A summarizes model performance statistics. Table 2.9 shows the estimated average $PM_{10-2.5}$ contributions percentages by feature, model and city. Table 2.6 provides the estimated average $PM_{10-2.5}$ contributions and associated bootstrapped errors. Table 2.7 shows the estimated contribution of selected species and Table 2.8 provides the correlations between the contributions of each feature. Table 2.2 summarizes the average ratio of tire wear to brake wear contributions to $PM_{10-2.5}$ by model and city. Finally, Table 2.3 shows the pairwise correlation coefficients between the g_{ik} and measured concentrations of selected species.

2.5.1 Brake Wear

The brake wear profiles for all models are shown in Figure 2.2. As expected, the Cu/Fe, Cu/Sb and Cu/Ba ratios are consistent with the prior constraints in all four constrained model-derived profiles (a-d). An additional "Metals-rich" factor (e) was identified in the Chicago individual model that contributes significantly to both Cu and Ba, but not to Sb. The contribution to Zn is low in all profiles except the Chicago individual profile (b). The predicted average contributions of brake wear to $PM_{10-2.5}$ are generally higher for the individual city models than for the combined city model (see Table 2.9).

The effect of the prior brake wear profile constraints on the final brake wear profiles is relatively minor for the combined cities model (a) with the exception of Sb which is pulled to a larger value in the constrained case. The effect of the prior constraints on this profile is also small for the individual city models, with the exception of

increasing the P concentration in Winston Salem (d). Comparing the constrained combined cities brake wear profile with its individual city counterparts, the differences are again small except for the absence on Al in the combined profile (a).

The brake wear contributions of Al, Ca, Fe, and Si are similar to the soil profiles across the individual city models. Within the combined model, the source contributions of these elements decreases and the bootstrapped variability of the contributions become significantly wider. The "Metals-rich" factor found in the Chicago individual model also contains these elements along with others found in the brake wear profile, but differentiates itself from brake wear due, not only to the differences in Sb mentioned above, but also due to a lack of Zn and the addition of P.

2.5.2 Tire Wear

The tire wear profiles are shown in Figure 2.3. The percent contribution to Zn is high in all tire wear profiles (a-d). The Pb to Zn ratio is consistent with prior constraints except for the Chicago individual model (b), where Pb/Zn is larger than specified by the soft constraint. In contrast, an additional "Pb-rich" factor is present in the combined model (e) that is distinctly separate from the accompanying tire wear profile (a). The effect of the prior profile constraints on other species is small for all models with the exception of Ca which is smaller in the constrained case for all profiles (a to d) and K which is smaller in St. Paul and Winston Salem (c and d). The predicted average contributions of tire wear to $PM_{10-2.5}$ are shown in Table 2.6. The confidence intervals for $PM_{10-2.5}$ from tire wear are generally smaller for the constrained, combined city model compared with the other models.

2.5.3 Soil

The soil profiles are shown in Figure 2.4. The effect of prior tire and brake wear constraints on the soil profile is small in all models. In addition, the combined versus individual city model profiles are similar with the exception of P in Chicago (a vs. b),

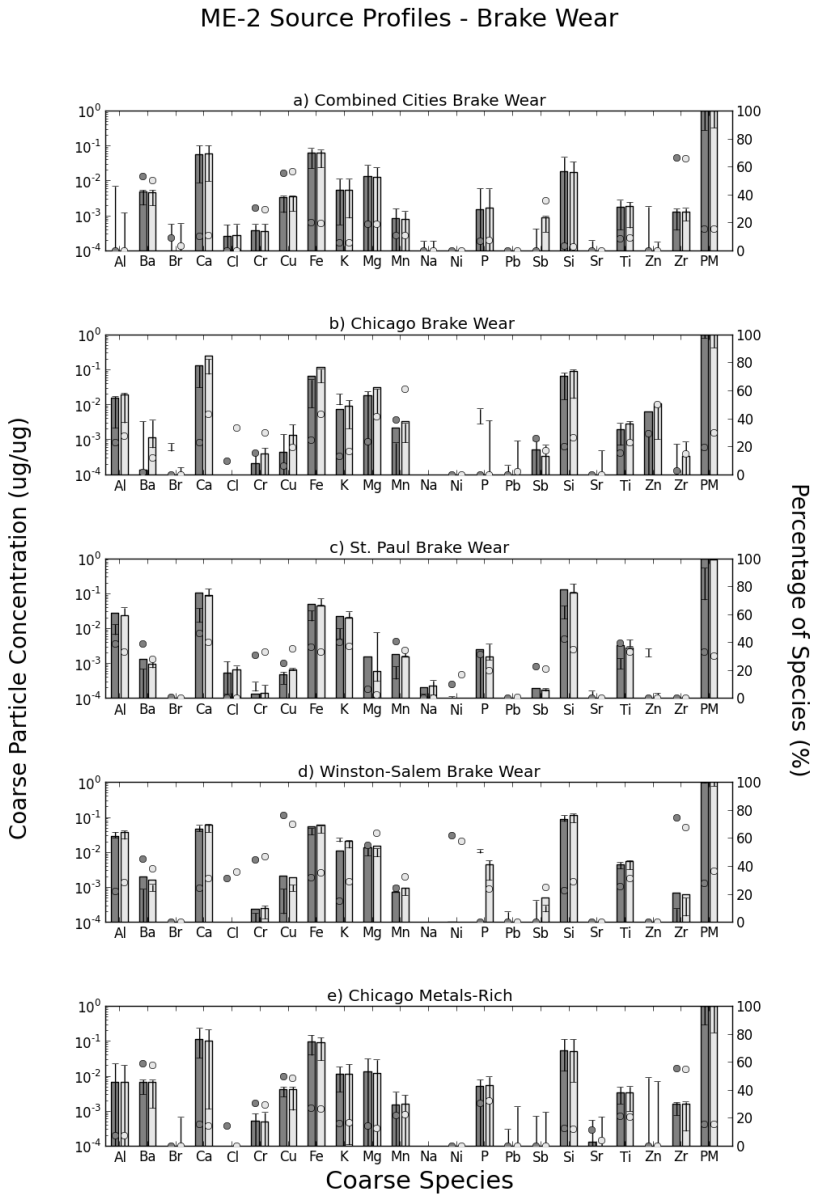

Figure 2.2: PMF-derived features identified as brake wear (aed). The black bars represent the average species contributions for the unconstrained models and the white bars represent the models with prior source profile constraints (see Table 2.1 and text). The bootstrapped 95% confidence limits are also shown. The circles refer to the percent of the total predicted concentration for a given species associated with that feature for the unconstrained (closed circles) and constrained (open circles) models. Also shown is an additional metals-rich source identified by the individual Chicago model (e) that is enriched in both Ba and Cu.

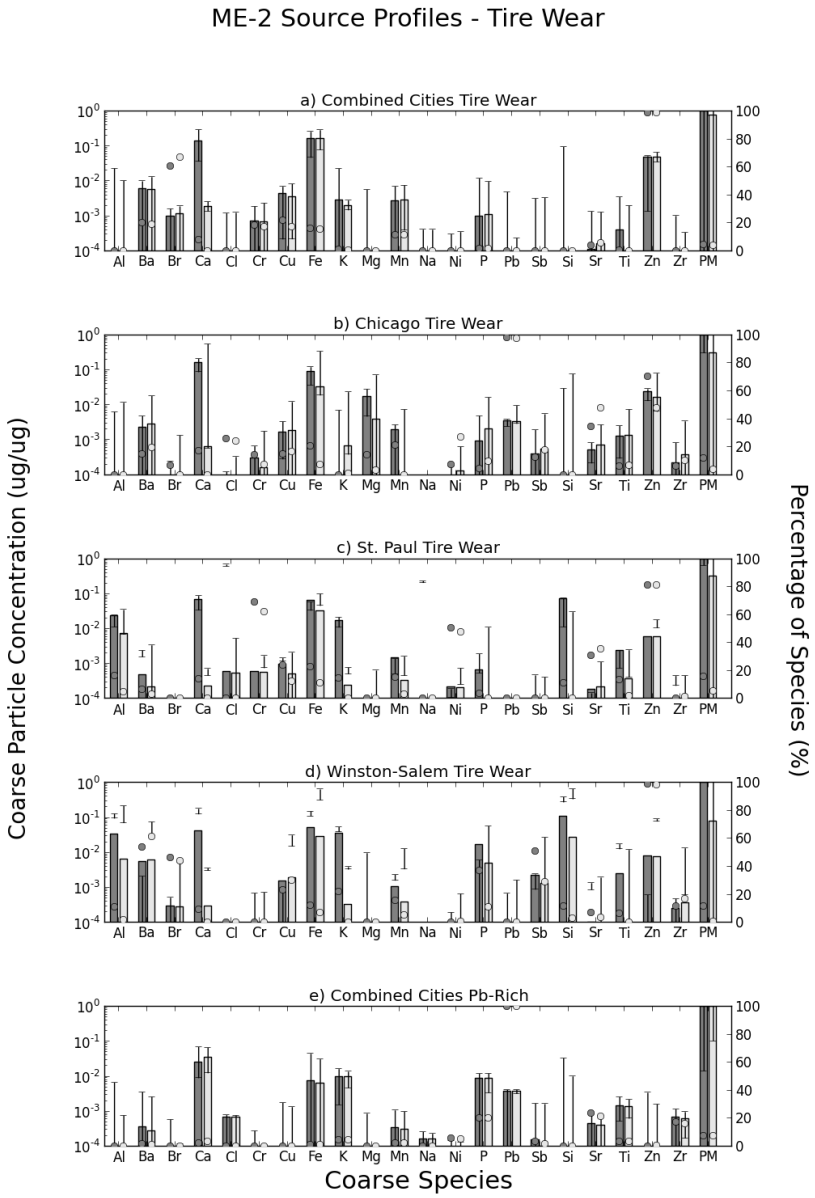

Figure 2.3: PMF-derived features identified as tire wear (aed). The black bars represent the average species contributions for the unconstrained models and the white bars represent the models with prior source profile constraints (see Table 2.1 and text). The bootstrapped 95% confidence limits are also shown. The circles refer to the percent of the total predicted concentration for a given species associated with that feature for the unconstrained (closed circles) and constrained (open circles) models. Also shown is an additional Pb-rich source identified by the combined-cities model (e).

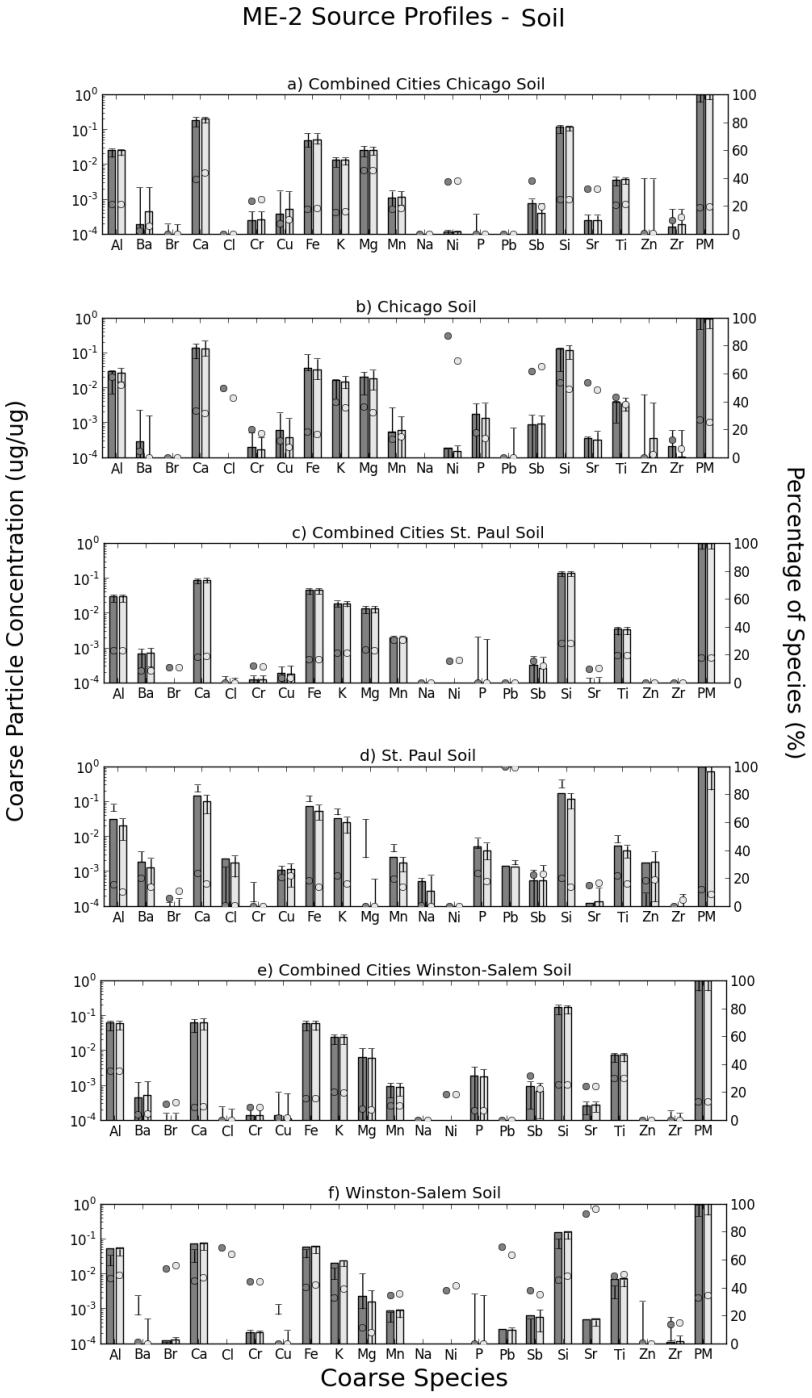

Figure 2.4: PMF-derived features identified as soil. The black bars represent the average species contributions for the unconstrained models and the white bars represent the models with prior source profile constraints (see Table 2.1 and text). The bootstrapped 95% confidence limits are also shown. The circles refer to the percent of the total predicted concentration for a given species associated with that feature for the unconstrained (closed circles) and constrained (open circles) models. The predictions from the combined-cities model (a, c, e) are shown by city along with the relevant predictions from the individual-city models (b, d, f).

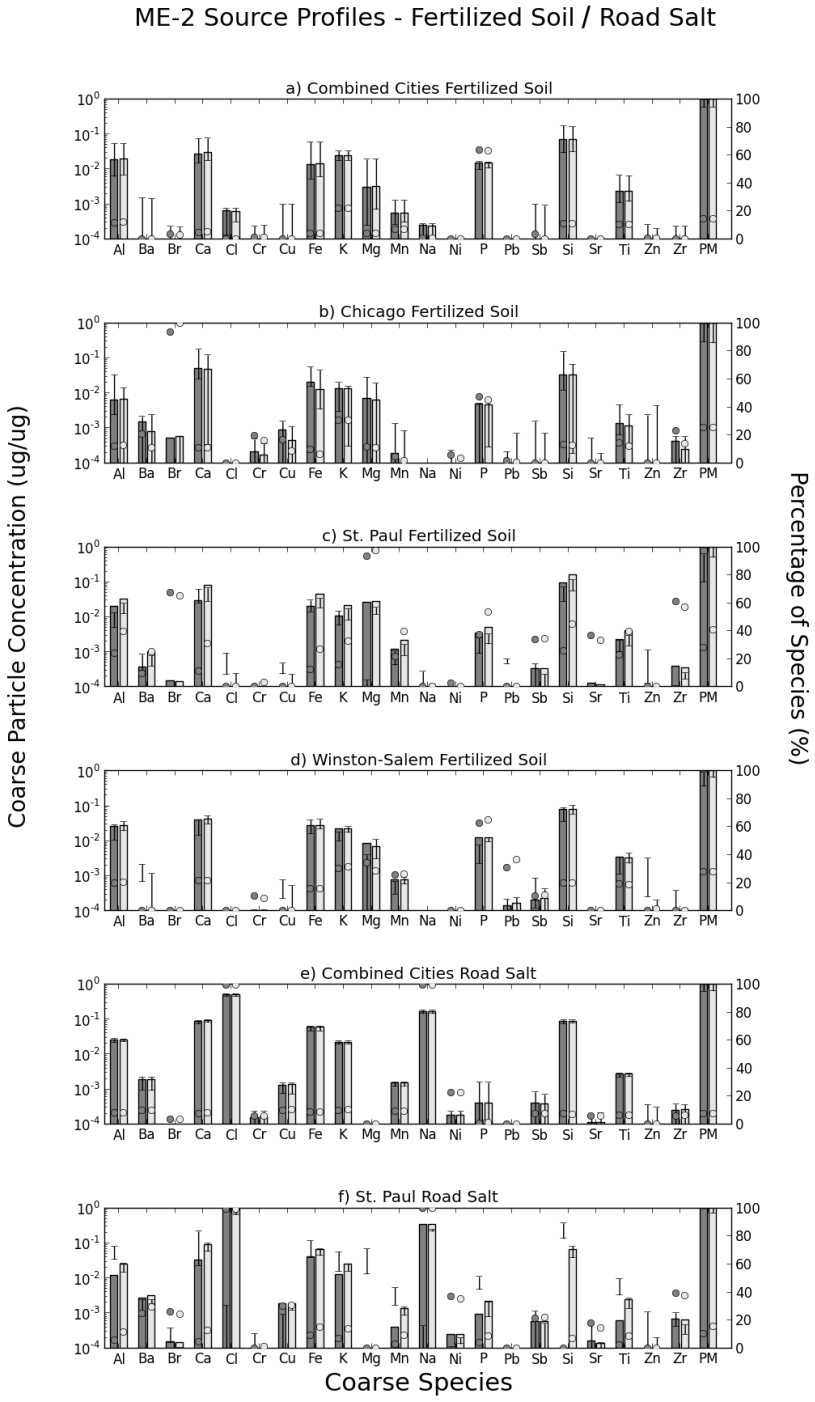

Figure 2.5: PMF-derived features identified as fertilized soil (aed) and road salt (e).. The black bars represent the average species contributions for the unconstrained models and the white bars represent the models with prior source profile constraints (see Table 2.1 and text). The bootstrapped 95% confidence limits are also shown. The circles refer to the percent of the total predicted concentration for a given species associated with that feature for the unconstrained (closed circles) and constrained (open circles) models.

Table 2.2: Ratio of tire wear to brake wear $PM_{10-2.5}$ contributions

Model		Tire wear to brake wear median ratio	
		With profile constraints	Unconstrained
Combined-cities model	Combined Model	0.24 (0.00,1.38)[a]	0.31 (0.00,1.85)
	Chicago	0.30 (0.00,1.70)	0.37 (0.00,2.27)
	St. Paul	0.35 (0.00,1.88)	0.48 (0.00,2.96)
	Winston-Salem	0.09 (0.00,0.65)	0.11 (0.00,0.86)
Individual-city models	Chicago	0.13 (0.00,0.64)	0.48 (0.13,1.17)
	St. Paul	0.17 (0.00,1.26)	0.65 (0.17,12.66)
	Winston-Salem	0.03 (0.00,0.98)	1.18 (0.29,2.05)

[a] () = 95% Confidence limits from bootstrapping.

and Mg and P and Pb in St Paul (c vs. d). Predicted average $PM_{10-2.5}$ contributions and their associated confidence intervals by city are generally lower for the individual city models than for the combined cities model (Tables 2.7 and 2.9).

2.5.4 Fertilized Soil

The fertilized soil profiles are shown in Figure 2.5. All four profiles (a-d) indicate that fertilized soil is a major contributor to P, Mg and K concentrations. The effect of prior tire and brake wear constraints on the fertilized soil feature is small in all models. Compared with the combined model (a), Ba is slightly enriched in the Chicago and St. Paul features (b and c) and Mg is enriched in St. Paul. The contribution of this feature to $PM_{10-2.5}$ is generally larger in the individual city models than the combined city model.

2.5.5 Other features

The road salt profiles are shown in Figure 2.5. The road salt feature is similar to that reported by Schauer and coworkers (2006) and contributed substantially to both Na and Cl levels. Its contributions were only observed in St. Paul during the winter sampling period but not in the summer nor in the other two cities. This city makes extensive use of NaCl as a road deicing agent during snowfall events [9]. Such events occurred during our winter sampling campaign in St. Paul, but not during winter sampling campaigns in either Chicago or Winston-Salem. The effect of the tire and brake wear constraints is small. For the road salt model, the difference between the combined and individual cities model is also small. This is consistent with the fact that the predicted road salt contributions are negligible in Chicago and Winston Salem.

The Pb-rich profile is shown in Figure 2.3. The effect of tire and brake wear constraints on this profile is small. The Pb-rich profile was not identified in any of the individual cities models.

2.6 Discussion

2.6.1 Effect of constraints

A comparison of the constrained versus unconstrained brake wear and tire wear profiles and their bootstrapped confidence intervals from either the individual or combined city models shows that the application of species ratio constraints had a relatively minor impact. The constraints impact was minimal not only on the specified species values, but also on those species in the profile that were not explicitly specified in the constraints. The ratio constraints are "soft" constraints in that they are limited by a maximum increase in Q rather than "hard" constraints that are forced specifically to their target values without consideration of the effect on Q. Yet both the brake wear and tire wear constrained profiles achieved their target ratios speci-

Table 2.3: Combined-cities model with profile constraints: pairwise correlations between predicted and observed species by feature, city and season.

	Feature	Pairwise correlations[a] (summer/winter)			
		Cu	Zn	P	Si
Chicago	Brake Wear	0.71 (0.68/0.76)	0.15 (0.09/0.22)	0.23 (0.20/0.24)	0.23 (0.17/0.45)
	Tire Wear	**0.67** (**0.70/0.67**)	**0.94** (**0.93/0.97**)	0.51 (0.44/**0.61**)	0.53 (**0.59/0.78**)
	Soil	**0.63** (**0.60/0.75**)	**0.62** (**0.61/0.80**)	0.29 (**0.56**/0.51)	**0.97** (**0.98/0.97**)
	Fertilized Soil	-0.48 (-0.58/-0.35)	-0.52 (-0.78/-0.24)	0.31 (-0.01/0.43)	-0.31 (-0.3/-0.12)
	Pb-Rich	**0.57** (**0.63**/0.52)	**0.89** (**0.93/0.84**)	**0.59** (**0.55/0.66**)	0.50 (**0.58/0.68**)
St. Paul	Brake Wear	0.33 (**0.90**/0.39)	0.43 (**0.55**/-0.08)	0.54 (0.26/-0.21)	**0.70** (0.44/-0.04)
	Tire Wear	**0.57** (**0.75**/0.41)	**1.00** (**1.00/0.99**)	0.43 (0.54/0.50)	0.41 (**0.64/0.57**)
	Soil	-0.10 (**0.71**/-0.30)	0.37 (**0.70**/-0.24)	**0.79** (**0.60/-0.56**)	**0.98** (**0.97**/-0.14)
	Fertilized Soil	-0.21 (0.15/-0.10)	0.29 (0.24/0.52)	**0.90** (**0.78/0.87**)	**0.76** (0.50/0.28)
	Road Salt	0.49 (**0.65/0.88**)	-0.04 (**0.85**/0.43)	**-0.65** (0.39/0.25)	**-0.78** (0.54/**0.82**)
Winston-	Brake Wear	**0.95** (**0.98/0.92**)	0.31 (0.02/**0.55**)	0.21 (0.17/**0.57**)	0.44 (**0.58**/0.37)
Salem	Tire Wear	0.39 (0.13/**0.62**)	**1.00** (**1.00/1.00**)	0.03 (0.38/0.35)	0.40 (0.28/0.43)
	Soil	0.30 (0.47/0.25)	0.30 (0.21/0.27)	-0.28 (0.25/0.42)	**0.94** (**0.95/0.96**)
	Fertilized Soil	0.07 (-0.05/0.37)	-0.01 (0.26/0.31)	**0.94** (**0.84/0.78**)	-0.21 (0.07/0.14)

[a] Pearson correlation coefficient between measured species concentration and predicted source contribution from baseline model; values with $p < 0.05$ are shown in bold

fied by the constraints. In other words, these profiles were robust to the application of prior information about their presumed sources, providing strong evidence that these two derived features were strongly influenced by tire and brake wear and thus appropriately named.

In contrast to the soft ratio constraints, we applied hard constraints on the soil contributions in the combined cities model in order to increase the number of observations while also deriving independent soil features for each city (equations 2.6 and 2.7). A comparison of the combined-city versus the individual-city models reveal that these hard constraints appear to have had a more significant effect on the derived tire and brake wear profiles than the soft constraints mentioned earlier, although these same hard constraints did not significantly alter the city- specific soil profiles compared to those derived from the individual city models. One major difference was the ability to separate Zn from Pb using the combined-city model, resulting in separate tire wear and Pb-rich profiles. This latter source's contributions to $PM_{10-2.5}$ are small but similar in magnitude to those from tire wear (see Table 2.6) and may be due in part to the abrasion of wheel weights [91]. Even though there is no evidence that tires themselves contain Pb, the individual-city tire wear profile in Chicago contained significant amounts of Pb. This may be driven by re-suspension of Pb from historically contaminated soil especially in southeast Chicago [105] and the resuspension of soils contaminated with leaded house paint [11]. The larger number and diversity of samples in the combined-cities model allowed these Pb-rich sources to be separated from the Zn rich tire profile.

2.6.2 Tire to Brake Wear Ratios

The ratio of the $PM_{10-2.5}$ contributions from tire wear relative to brake wear as predicted by the models can be compared with other independent estimates of this ratio. Based on a combination of particle size distribution and chemical species measurements in London at a heavily-trafficked curbside site and an urban background site,

Harrison and coworkers [48] found that brake wear and tire wear contributed 55.3(+-7) percent and 10.7(+-2.3) percent respectively to particle mass between 0.9 and 11.5 m, corresponding to an average tire-to-break wear ratio of 0.19. The traffic mix at their curbside site was dominated by light duty vehicles [23]. This estimate is consistent with our model predictions in Table 2.2 with the exception of those from the unconstrained, individual-city models. The relatively low absolute contributions of tire wear to $PM_{10-2.5}$ (see Table 2.7) is also consistent with those reported elsewhere by Kumata et al. (2011) using molecular markers.

The results in Table 2.2 can also be compared with emissions derived from the EPA MOVES model for 2009 across all vehicle and roadway categories in the relevant counties in the three cities. The MOVES tire wear to brake wear emission ratio estimate is 0.29, reasonably consistent with our model predictions with the exception of those from the unconstrained, individual-city models. The corresponding ratios used in the California EMFAC model are 0.61-0.63 depending upon vehicle category, somewhat higher than MOVES but still consistent with our combined-city model predictions. However, it should be noted that these MOVES and EMFAC emission estimates for tire wear and brake wear $PM_{10-2.5}$ are themselves uncertain [29].

2.6.3 Soil Profiles

The relative proportions of Si, Fe, K, Al, Ti and Mn in the St. Paul and Chicago soil profiles are similar to the Minneapolis resuspended coarse particle soil profile (MPNSoil) reported by Watson and co-workers [120], the Chicago urban dust profile (UDUST) reported by Vermette and co-workers [111], surface soil geochemistry of Chicago soils reported by Cannon and co-workers [20], and the surrounding surface layer geochemistry of upper Midwestern U.S. soils (mollisols and alfisols) [100, 110]. Winston-Salem soil particles have slightly higher mass fractions of Ti compared with the other two cities consistent with the composition of the soils surrounding Winston Salem (ultisols), predominant throughout Virginia, the Carolinas, Tennessee, Georgia

and Alabama [100].

The phosphorus-rich fertilized soil feature contributions are consistently higher during the summer versus winter sampling period in all three cities (not shown). Phosphorus is added as a fertilizer to the soils within the urban areas as well as to those in the agricultural areas surrounding all three cities [53]. Such windblown soil is a major source of airborne phosphorus in agricultural areas in the spring and summer [8].

2.6.4 Sources of Selected Species

An interesting question is how well a given measured species concentration represents the contribution from a given source. We addressed this by examining for each feature the relative contributions of and pairwise correlations with selected indicator species of Cu, Zn, P and Si.

Our model results clearly indicate that tire wear is highly correlated with and also an important source of Zn. In Chicago, Zn was strongly associated with tire wear but also contributed to the soil and the Pb-rich features. Since Pb was present in the tire wear feature for the Chicago individual model but not the individual models for St. Paul and Winston-Salem, it is logical that it would be separated from the combined model to ensure a ubiquitous tire wear source across cities. As a result, the Pb-rich contributions are highly correlated with the tire wear contributions in Chicago (Table 2.8).

We also found that the brake wear feature is a major contributor to Cu in all three cities and that crustal soil material is highly correlated with and also an important source of Si in all three cities (see Table 2.7). The Metals-rich feature identified in the Chicago individual models are very similar to the brake wear profile and differs in only a few elements, potentially the result of artificially splitting the brake wear feature based on the subtle differences of brake wear composition. However, the combined model does not identify a Metals-rich feature. By using a multi-city approach we

increase the number of samples and thereby improve the model's ability to separate these ubiquitous features.

A positive correlation between brake wear, tire wear, and soil contributions implies that re-suspension from roadways, or road dust, is involved. This idea is reinforced by the presence of Cu, Ba, and Sb in both the individual and combined model profiles (Figure 2.4). Interestingly, the road salt feature in St. Paul in the winter is also a major contributor to and also more highly correlated with the observed Cu than the brake wear feature in this season (Table 2.3). To the extent that road salt is acting as a tracer of re-suspended road particles, this would indicate that a measurable portion of Cu from brake wear is from re-suspended material, at least in St. Paul. This is consistent with the idea that some fraction of brake (and tire) wear dusts are deposited on or near the roadway and subsequently re-suspended, partially as coarse mode particles [80]. This is suggested by the accompanying enrichment of Ba in the same road salt feature in both the combined and individual model results (Figures 2.5e and 2.5f respectively). The soil and road salt contributions associated with Cu could also be due to the model's inability to separate these sources from a truly independent brake wear feature, but the fact that the soil and road salt features contain both Ba and Cu suggest that re-suspension of road dust is also playing an important role. This idea is reinforced by positive correlations of brake wear contributions with contributions from soil and road salt during the summer and winter seasons respectively (Table 2.8). Additionally tire wear and Pb-rich were slightly correlated with the Chicago soil feature indicating that re-suspension of road dust is playing a role in these features.

The fertilized soil feature is well correlated with P in St. Paul and Winston-Salem, however, Chicago exhibited elevated P contributions within the fertilized soil as well as the Pb-rich feature. A positive correlation of contributions between the soil and Pb-rich feature (Table 2.8) dilutes the ability of P to act as a strong indicator for fertilized soil in Chicago.

2.7 Conclusions

We were able to successfully use a modified version of PMF to identify contributions from brake wear, tire wear, crustal material, fertilized soil and a small Pb-rich feature. Our modified PMF model allowed not only for inclusion of prior source profile information for selected species, but also for locally specific results for soil (crustal) contributions in addition to generally applicable results for the other features. The effect of prior source profile constraints on model predictions was relatively small, increasing our confidence in correctly identifying the tire and brake wear features. The modified model also allowed us to include measurements from different cities with different soil compositions. The locally specific soil profiles in this model were consistent with those derived from city-specific models. Elements Cu, Zn, P, and Si were identified as general indicators of brake wear, tire wear, fertilized soil, and soil for the combined city analysis.

2.8 Acknowledgements

This work was made possible by the U.S.EPA STAR Grant program R833741 and RD831697. Its contents are solely the responsibility of the grantee and do not necessarily represent the official views of the USEPA. Further, USEPA does not endorse the purchase of any commercial products or services mentioned in the publication. The MESA study is supported by contracts N01-HC-95159 through N01-HC-95169 from the National Heart, Lung, and Blood Institute (NHLBI). The authors acknowledge the other investigators, staff, and participants of MESA, MESA Coarse, and MESA Air for their valuable contributions to this work. A full list of MESA investigators and institutions is located at http://www.mesa-nhlbi.org. Additionally, this work was partially funded by the Joint Fire Science Program (09-1-03-1). The assumptions, findings, conclusions, judgments, and views presented herein are those of the authors and should not be interpreted as representing the National Park Service.

2.9 Supplemental Material

Table 2.4: Summary of Observed Coarse Particle Concentrations by City

Species	Mean Concentration (ng/m^3)[a]						Average
	Chicago		St. Paul		Winston-Salem		Signal-
	Summer (n = 31)	Winter (n = 33)	Summer (n = 33)	Winter (n = 25)	Summer (n = 30)	Winter (n = 35)	to-noise[b]
Al	61	92	154	78	116	138	135
Ba	11	10.1	6.2	6.2	4.1	5.3	16
Br	1.1	0.6	0.5	0.3	0.2	0.5	11
Ca	616	680	483	269	197	193	173
Cl	b.d.[c]	1.5	b.d.	1320	b.d.	2.6	25
Cr	1.85	1.54	0.87	0.65	0.47	0.62	22
Cu	7.45	7.92	2.85	4.1	2.69	2.66	76
Fe	306	314	280	191	154	183	152
K	63	62	114	68	70	68	85
Mg	87	90	74	3	28	22	58
Mn	6.31	6.58	10.4	4.74	2.95	2.9	77
Na	b.d.	b.d.	b.d.	460	b.d.	b.d.	19
Ni	0.32	0.41	0.29	0.51	0.15	0.22	11
P	17.9	13.6	18.7	8.2	25.9	12.8	60
Pb	2.7	2.06	1.34	0.76	0.87	0.59	14
Sb	1.8	2.8	2	1.1	1.8	2.3	1.6
Si	307	428	719	266	346	410	162
Sr	0.88	1.26	0.67	0.55	0.45	0.77	15
Ti	12.6	15.7	19.5	8.5	15.3	18.3	122
Zn	26	23.8	6.4	5.4	2.9	3.6	58
Zr	2.53	2.72	0.96	0.98	0.84	1.08	24
$PM_{10-2.5}$	5.9	5.5	6.8	3.5	3.8	3.5	0.7[d]

[a] $PM_{10-2.5}$ in μg/m^3;

[b] across all 3 cities;

[c] "b.d." = below detection limits;

[d] includes additional model down-weighting (see text)

Table 2.5: PMF Baseline Model Performance Statistics (R^2/RMSE[a]) by Species and City

Species	Individual City Models						Combined Cities Model			
	Chicago		St. Paul		Winston-Salem		Chicago		St Paul[f]	Winston- Salem[f]
	U[b]	C[c]	U	C	U	C	CU[d]	CC[e]	CC	CC
Al	0.97/7.4	0.97/7.5	0.97/12	0.97/12	0.97/8.2	0.97/8.2	0.97/8.1	0.96/8.1	0.97/14	0.97/8.5
Ba	0.88/2.4	0.88/2.4	0.54/2.3	0.50/2.4	0.60/2.6	0.59/2.7	0.88/2.4	0.87/2.4	0.48/2.5	0.50/2.8
Br	0.94/0.19	0.95/0.18	0.32/0.32	0.27/0.33	0.42/0.39	0.43/0.39	0.38/0.58	0.44/0.58	0.10/0.35	0.36/0.42
Ca	0.98/58	0.98/61	0.97/47	0.96/56	0.59/92	0.60/92	0.98/59	0.97/59	0.96/51	0.57/94
Cl	0.40/3.2	0.40/3.2	0.99/84	1.00/72	0.05/11	0.11/11	0.51/3.0	0.51/3.0	0.95/256	1.00/11
Cr	0.45/1.1	0.45/1.1	0.84/0.31	0.84/0.30	0.52/0.29	0.52/0.29	0.42/1.1	0.45/1.1	0.70/0.41	0.48/0.30
Cu	0.92/1.3	0.92/1.3	0.86/0.83	0.87/0.79	0.86/0.70	0.86/0.67	0.92/1.3	0.92/1.3	0.95/0.57	0.96/0.42
Fe	0.95/40	0.95/41	0.97/31	0.97/32	0.96/16	0.96/16	0.93/48	0.94/47	0.97/33	0.96/15
K	0.93/9.0	0.93/8.9	0.96/10.	0.96/9.6	0.88/8.2	0.86/8.8	0.90/10	0.90/10	0.96/10	0.87/8.5
Mg	0.91/18	0.92/17	0.85/23	0.93/16	0.76/7.8	0.75/8.1	0.89/19	0.90/19	0.94/14.7	0.65/9.7
Mn	0.89/1.3	0.88/1.4	0.98/0.68	0.98/0.69	0.84/0.59	0.84/0.59	0.77/2.0	0.79/2.0	0.98/0.77	0.83/0.59
Na	–	–	0.99/45	0.99/45	–	–	0.00/0.20	0.00/0.20	0.97/76.8	1.00/5.9
Ni	0.74/0.17	0.73/0.17	0.78/0.21	0.78/0.21	0.48/0.13	0.48/0.13	0.69/0.18	0.69/0.18	0.67/0.27	0.29/0.14
P	0.60/4.7	0.60/4.7	0.85/3.9	0.86/3.9	0.97/1.9	0.97/1.9	0.98/1.1	0.98/1.1	0.99/0.84	0.99/1.1
Pb	0.97/0.55	0.97/0.56	0.46/1.3	0.45/1.3	0.21/0.84	0.24/0.83	1.00/0.20	1.00/0.20	0.99/0.28	0.99/0.14
Sb	0.34/3.9	0.34/3.9	0.26/3.0	0.24/3.0	0.15/3.3	0.09/3.3	0.33/3.9	0.13/3.9	0.26/3.1	0.00/3.4
Si	0.99/23	0.99/22	0.99/32	0.99/37	0.99/17	0.99/17	0.99/25	0.98/25	0.99/41	0.98/20
Sr	0.89/0.35	0.89/0.35	0.55/0.47	0.61/0.45	0.74/0.31	0.74/0.32	0.87/0.40	0.87/0.40	0.52/0.49	0.66/0.36
Ti	0.93/1.9	0.93/1.9	0.99/1.3	0.99/1.3	0.95/1.7	0.95/1.7	0.92/2.1	0.92/2.01	0.98/1.3	0.96/1.6
Zn	0.96/6.7	0.97/6.2	0.97/1.3	0.97/1.2	0.98/0.36	0.98/0.40	0.97/5.4	0.94/5.4	1.00/0.34	1.00/0.11
Zr	0.90/0.51	0.90/0.51	0.71/0.50	0.64/0.55	0.70/0.47	0.69/0.48	0.91/0.47	0.91/0.47	0.36/0.68	0.70/0.47
$PM_{10-2.5}$	0.89/0.95	0.88/0.97	0.79/1.9	0.78/1.9	0.49/1.3	0.49/1.3	0.85/1.1	0.85/1.1	0.78/1.9	0.47/1.3

[a] RMSE in ng/m3 except PM10-2.5 is in g/m3;

[b] individual city unconstrained model (equations 2.2,2.3,2.5); $q_t = 0$)

[c] individual city constrained model (equations 2.2-2.5, qt > 0);

[d] combined-cities, unconstrained model (equations 2.2, 2.3, 2.5, 2.6, 2.7, qt $= 0$);

[e] combined-cities constrained model (equations 2.2-2.7, qt > 0);

[f] constrained and unconstrained results essentially identical

Table 2.6: Average $PM_{10-2.5}$ Source Contributions ($\mu g/m^3$)[a]

Model	Constraints		Feature					
			Brake Wear	Tire Wear	Soil	Fertilized Soil	Road Salt	Pb-rich or metals rich
Individual City Models	Unconstrained	Chicago	1.44 (0.88,1.84)	0.69 (0.22,1.22)	1.11 (0.77,2.08)	1.53 (0.65,2.01)	0.00 (0.00,0.00)	0.88 (0.23,1.58)
		St.Paul	0.83 (0.00,1.53)	0.54 (0.16,1.37)	1.73 (0.82,2.84)	0.62 (0.00,1.22)	1.46 (0.48,2.74)	–
		Winston-Salem	0.97 (0.52,1.73)	1.14 (0.37,1.49)	0.97 (0.10,1.62)	0.42 (0.00,1.29)	0.00 (0.00,0.00)	–
	With Profile Constraints	Chicago	1.69 (0.72,2.51)	0.21 (0.00,0.79)	1.42 (0.74,2.23)	1.43 (0.93,1.89)	0.00 (0.00,0.00)	0.89 (0.14,1.59)
		St.Paul	1.56 (0.44,2.88)	0.27 (0.00,1.14)	0.45 (0.00,1.41)	2.11 (0.54,2.95)	0.79 (0.11,1.36)	–
		Winston-Salem	1.28 (0.51,2.01)	0.03 (0.00,0.79)	1.21 (0.41,1.62)	0.98 (0.35,1.64)	0.00 (0.00,0.00)	–
Combined Cities Model	Unconstrained	Chicago	1.33 (0.33,1.85)	0.50 (0.00,1.51)	2.62 (1.58,3.72)	0.49 (0.25,0.88)	0.00 (0.00,0.00)	0.63 (0.01,1.19)
		St.Paul	0.26 (0.10,0.60)	0.12 (0.00,0.54)	2.69 (1.82,3.30)	0.70 (0.31,1.13)	1.10 (0.68,1.42)	0.28 (0.00,0.55)
		Winston-Salem	0.60 (0.12,0.96)	0.07 (0.00,0.33)	1.82 (0.87,2.40)	0.84 (0.45,1.84)	0.00 (0.00,0.00)	0.19 (0.00,0.37)
	With Profile Constraints	Chicago	1.29 (0.35,1.74)	0.39 (0.00,1.22)	2.78 (1.92,3.85)	0.48 (0.25,0.91)	0.00 (0.00,0.00)	0.63 (0.02,1.20)
		St.Paul	0.27 (0.11,0.62)	0.09 (0.00,0.40)	2.70 (1.83,3.41)	0.71 (0.30,1.16)	1.11 (0.71,1.42)	0.28 (0.00,0.55)
		Winston-Salem	0.61 (0.13,0.97)	0.05 (0.00,0.22)	1.82 (0.86,2.42)	0.85 (0.45,1.84)	0.00 (0.00,0.00)	0.19 (0.00,0.35)

[a] ( ) = bootstrapped 95% confidence limits

Table 2.7: Combined-Cities Model with Profile Constraints: Predicted Contributions for Selected Species

	Feature	Average Contributions (ng/m^3)			
		Cu	Zn	P	Si
Chicago	Brake Wear	4.55 (1.60,5.00)[a]	0.00 (0.00,0.21)	2.19 (0.00,6.42)	23.32 (0.00,44.12)
	Tire Wear	1.72 (0.07,3.16)	23.7 (10.9,27.3)	0.56 (0.00,3.52)	0.00 (0.00,0.00)
	Soil	1.38 (0.25,4.59)	0.16 (0.00,10.67)	0.00 (0.00,0.00)	310 (252,351)
	Fertilized Soil	0.00 (0.00,0.48)	0.03 (0.00,0.09)	7.24 (4.12,9.40)	34.0 (18.1,71.1)
	Pb-rich	0.00 (0.00,0.80)	0.06 (0.00,1.06)	5.58 (2.01,7.62)	0.00 (0.00,6.45)
St. Paul	Brake Wear	0.95 (0.56,1.58)	0.00 (0.00,0.08)	0.46 (0.00,2.97)	4.88 (0.00,12.25)
	Tire Wear	0.42 (0.02,0.91)	5.81 (4.67,6.74)	0.14 (0.00,1.02)	0.00 (0.00,0.00)
	Soil	0.49 (0.00,0.83)	0.00 (0.00,0.08)	0.00 (0.00,4.78)	379 (308,421)
	Fertilized Soil	0.00 (0.00,0.58)	0.05 (0.00,0.13)	10.60 (5.40,11.29)	49.8 (19.8,112.7)
	Road Salt	1.47 (0.80,1.64)	0.03 (0.00,0.33)	0.44 (0.00,1.72)	96.3 (88.4,104.2)
Winston-Salem	Brake Wear	2.16 (0.79,2.35)	0.00 (0.00,0.09)	1.04 (0.00,4.14)	11.07 (0.00,22.32)
	Tire Wear	0.23 (0.01,0.47)	3.23 (2.70,3.54)	0.08 (0.00,0.57)	0.00 (0.00,0.00)
	Soil	0.20 (0.00,1.07)	0.00 (0.00,0.07)	3.28 (0.00,5.27)	312 (206,356)
	Fertilized Soil	0.00 (0.00,1.00)	0.05 (0.00,0.17)	12.69 (10.12,14.30)	59.7 (23.9,168.4)

[a] ( ) = bootstrapped 95% confidence limits

Table 2.8: Combined-Cities Model with Profile Constraints: Source Contribution Correlations by City and Season.

City	Feature		1	2	3	4	5
Chicago	1	Brake Wear	1[a] (1/1)[b]	.15 (.07/.23)	.18 (.10/.39)	-.11 (-.14/-.11)	.08 (.02/.13)
	2	Tire Wear		1 (1/1)	.63 (.70/.79)	-.53 (-.81/-.23)	.88 (.93/.80)
	3	Soil			1 (1/1)	-.44 (-.45/-.27)	.59 (.69/.70)
	4	Fertilized Soil				1 (1/1)	-.54 (-.79/-.35)
	5	Pb-Rich					1 (1/1)
St. Paul	1	Brake Wear	1 (1/1)	.43 (.56/-.03)	.71 (.46/-.35)	.45 (.09/-.29)	-.58 (.52/.31)
	2	Tire Wear		1 (1/1)	.36 (.70/-.25)	.29 (.24/.50)	-.03 (.83/.48)
	3	Soil			1 (1/1)	.71 (.35/-.51)	-.86 (.65/-.20)
	4	Fertilized Soil				1 (1/1)	-.67 (.05/-.01)
	5	Road Salt					1 (1/1)
Winston-	1	Brake Wear	1 (1/1)	.31 (.02/.55)	.27 (.47/.19)	.09 (-.06/.42)	
Salem	2	Tire Wear		1 (1/1)	.31 (.23/.26)	-.04 (.24/.30)	
	3	Soil			1 (1/1)	-.50 (-.16/-.04)	
	4	Fertilized Soil				1 (1/1)	

[a] Correlations greater than 0.5 represented by bold underlined format

[b] () = (Summer / Winter)

Table 2.9: Comparison of the average source contribution percentages provided by each modeling scenario and described by city.

City	Model	Brake Wear	Tire Wear	Fertilized Soil	Soil	Pb/Metals-Rich	Road Salt
Chicago	Combined Constrained	23.20%	7.00%	8.70%	49.90%	11.30%	–
Chicago	Combined Unconstrained	23.90%	8.90%	8.80%	47.10%	11.30%	–
Chicago	Individual Constrained	29.90%	3.80%	25.30%	25.10%	15.80%	–
Chicago	Individual Unconstrained	25.50%	12.20%	27.10%	19.60%	15.60%	–
St. Paul	Combined Constrained	5.60%	1.90%	14.50%	55.30%	–	22.70%
St. Paul	Combined Unconstrained	5.30%	2.50%	14.40%	55.20%	–	22.60%
St. Paul	Individual Constrained	30.10%	5.20%	40.70%	8.70%	–	15.30%
St. Paul	Individual Unconstrained	16.00%	10.40%	11.90%	33.40%	–	28.20%
Winston-Salem	Combined Constrained	18.40%	1.60%	25.40%	54.60%	–	–
Winston-Salem	Combined Unconstrained	18.10%	2.00%	25.20%	54.70%	–	–
Winston-Salem	Individual Constrained	36.50%	1.00%	27.90%	34.50%	–	–
Winston-Salem	Individual Unconstrained	27.70%	32.60%	12.00%	27.70%	–	–

Table 2.10: Range, Average, Average Percent Recovery from Measurement Quality Assurance Data

Analyte	SRM	n	Certified Value (μg/cm^2)	Measured Value (μg/cm^2)			Avg. % Rec.
				High	Low	Average	
Al	1228	8	12.6 ± 1.3	14.22	12.55	13.41	106.09
Si	1228	8	34 ± 1.1	37.17	33.9	35.17	103.49
Ca	1228	9	19 ± 1.3	22	18.66	19.57	103.14
V	1228	9	3.95 ± 0.5	4.24	3.83	3.96	100.19
Mn	1228	9	4.11 ± 0.47	4.53	4.09	4.35	105.76
Cu	1228	9	2.53 ± 0.16	2.32	2.13	2.22	87.79
Si	987	8	35.8 ± 2.3	40.56	36.73	38.95	108.9
K	987	9	19.12 ± 1.8	22.43	17.59	19.79	103.5
Ti	987	9	11.86 ± 1.6	12.39	11.19	11.8	99.49
Fe	987	9	14.5 ± 0.53	15.11	13.99	14.46	99.58
Zn	987	9	5.49 ± 0.53	5.84	5.3	5.48	99.88
Pb	987	9	22.83 ± 1.24	23.2	21.84	22.71	99.49

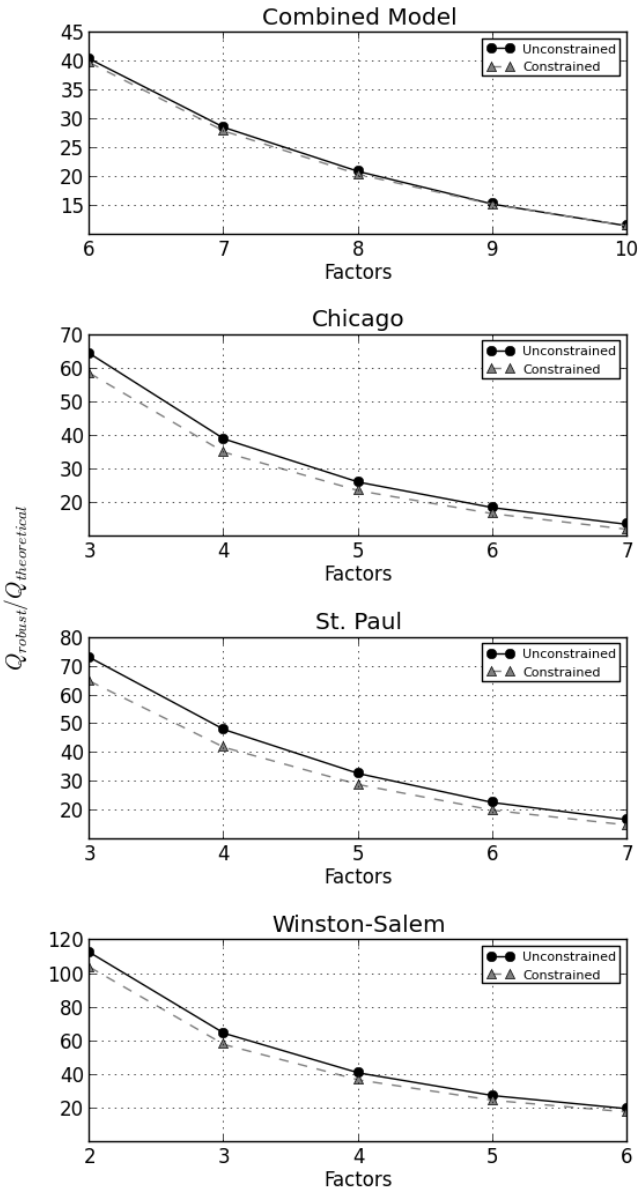

Figure 2.6: The value of p, degrees of freedom, and values of Q are: 8, 2864, 26966 for the unconstrained combined site model; 8, 1969, 27205 for the constrained combined site model; 5, 986, 7532 for the unconstrained Chicago model; 5, 696, 7618 for the constrained Chicago model; 5, 918, 8468 for the unconstrained St. Paul model; 5, 648, 8412 for the constrained St. Paul model; 4, 1062, 9219 for the unconstrained Winston-Salem model; and 4, 767, 9287 for the constrained Winston-Salem model.

Table 2.11: Average Measurement and Percent Error from Measurement Quality Control Data

Analyte	n	Conc. (μg/cm^2)		Avg.
		Calib.	Avg. Meas.	%E
Si	167	9.86	9.22	0.08
V	176	10.07	10.11	-0.31
Ni	176	9.98	10.13	-0.93
Pb	176	21.02	21.84	2.03
Cd	176	6.04	6.19	0.77
Se	176	3.88	4.11	2.01

Table 2.12: Analysis of modeled Q versus $Q_{theoretical}$ across differing quantities of features for combined, individual, constrained and unconstrained scenarios.

Model	Constraints	Number of features	Degrees of Freedom	Q_{robust} /$Q_{theoretical}$
Combined	Unconstrained	6	3258	40.42
		7	3060	28.56
		8	2864	20.91
		9	2670	15.25
		10	2478	11.47
Combined	Constrained	6	2353	39.8
		7	2160	27.93
		8	1969	20.42
		9	1780	15.2
		10	1593	11.48
Chicago	Unconstrained	3	1140	64.56
		4	1062	39.17
		5	986	26.33
		6	912	18.73
		7	840	13.75
Chicago	Constrained	3	840	58.7
		4	767	35.23
		5	696	23.79
		6	627	16.93
		7	560	12.3
St. Paul	Unconstrained	3	1064	73.2
		4	990	48.2
		5	918	32.94
		6	848	22.86
		7	780	16.91
St. Paul	Constrained	3	784	64.98
		4	715	42.02
		5	648	29.11
		6	583	20.2
		7	520	15.06
Winston-Salem	Unconstrained	2	1220	112.9
		3	1140	64.82
		4	1062	41.27
		5	986	27.76
		6	912	19.95
Winston-Salem	Constrained	2	915	104.15
		3	840	58.12
		4	767	37.11
		5	696	24.97
		6	627	18

Chapter 3

COUPLING CHEMICAL TRANSPORT MODEL SOURCE ATTRIBUTIONS WITH POSITIVE MATRIX FACTORIZATION[2]

3.1 Abstract

Source contributions to total fine particle carbon predicted by a chemical transport model (CTM) were incorporated into the positive matrix factorization (PMF) receptor model to form a receptor-oriented hybrid model. The level of influence of the CTM versus traditional PMF was varied using a weighting parameter applied to an object function as implemented in the Multilinear Engine (ME-2). The resulting hybrid model was used to quantify the contributions of total carbon from both wildfires and biogenic sources at two Interagency Monitoring of Protected Visual Environment monitoring sites, Monture and Sula Peak, Montana, from 2006 through 2008. CTM source impacts were used to aid in the separation of biogenic sources from biomass combustion due to wildfires. Two additional features were identified at each site, a soil derived feature with elevated contributions in the summer and feature enriched in both sulfate and nitrate with significant, but sporadic contributions across the sampling period.

---

[2]This chapter has been drafted for submission to *Environmental Science & Technology*, authors T.M. Sturtz (Department of Civil and Environmental Engineering, University of Washington, Box 352700, Seattle, WA 98195-2700), B.A. Schichtel (Cooperative Institute for Research in the Atmosphere/NPS, Colorado State University, Fort Collins, Colorado), and T.V. Larson (Department of Civil and Environmental Engineering, University of Washington, Box 352700, Seattle, WA 98195-2700 and Department of Environmental and Occupational Health Sciences, University of Washington, Box 357234, Seattle, WA 98195-7234).

3.2 Introduction

Models that accurately describe the source contributions to ambient fine particle mass and composition are an important air quality management tool. These models span a spectrum from purely deterministic models based on a priori knowledge of emissions, meteorology and chemistry, to multivariate receptor models based on ambient pollutant measurements at a given receptor site.[15, 35, 79, 89, 92, 99, 112, 118, 125, 126] Frequently, one of these two modeling approaches acts in an independent, supporting role to the other. [63, 119]

Some investigators have focused on combining deterministic models with receptor-based particle measurement approaches to form a single "hybrid" model. Different approaches to combining these models include the use of genetic algorithms, [3, 4, 49, 59] ensemble methods, [26, 40, 61, 66, 67] multiplicative bias correction [97, 104] and non-linear optimization. [124] A subset of these hybrid modeling approaches include a deterministic chemical transport model (CTM) that includes secondary formation of particle mass. Of these, even fewer also include receptor information on particle composition in addition to particle mass. [61, 97, 98, 104]

Here we present a hybrid model that explicitly combines predictions from a CTM with those from the Positive Matrix Factorization receptor model within the framework of the Multilinear Engine.[75] Our model is an extension of one initially proposed by Schichtel et.al. [98] that was developed using a synthetic data set. In this case, we apply an extension of this latter hybrid model to actual data at two rural IMPROVE monitoring sites in Montana with the goal of distinguishing contributions to total fine particle carbon from biogenic sources versus those from biomass combustion due to wildfires. Based on correlations between soluble potassium and particulate organic carbon, the impact of wildfires at Western U.S. IMPROVE sites is known to be significant. [54, 70, 82] Recent CTM modeling supports this conclusion, although correlations between predicted and observed particulate carbon values at these same

Western U.S. sites are lower than in other regions, [94] providing additional motivation for the development of our hybrid model.

3.3 Methods

3.3.1 Monitoring Data

We used three years (2006 - 2008) of speciated PM2.5 data from two IMPROVE sites located in western Montana - Monture and Sula Peak (vista.cira.colostate.edu/improve/) whose locations are shown in the inset of Figure 3.1. The methodology described in Polissar et. al. [84] was implemented to determine measurement uncertainties. We assessed the average concentration to measurement uncertainty ratio (signal-to-noise, S/N) using the methodology of Norris and Vedantham [37] and removed any species with average S/N ¡ 0.2, and down weighted by a factor of 3 species with $0.2 < S/N < 2.0$. [77] In addition, species with reported concentrations below their detection limit or missing in over half the samples were removed from the analysis. The remaining species we considered are summarized in Table 3.1. Finally, an additional 8% and 25% of the respective samples from Monture and Sula Peak were removed if the mass reconstruction was outside IMPROVE limits. [38, 119]

3.3.2 Chemical Transport Model

At the two monitoring sites, we used predictions of fine particle carbon based on the CAPITA Monte Carlo Lagrangian chemical transport model (CTM). [96] The model considered 5 day upwind trajectories with accompanying emissions and atmospheric reactions along each trajectory. It was recently implemented at each IMPROVE site in the United States and shown to have similar performance metrics to CMAQ in predicting fine particle carbon across the U.S.. [94] The model provided source contribution estimates of both primary and secondary carbonaceous fine particles from the following source categories: biomass combustion, biogenic, mobile, area, oil,

point and other. Uncertainty estimates for each source were taken from the work previously conducted by Schichtel et. al.. [98]

3.3.3 Combined CTM/PMF Model

Main Equations

We implemented a modified version of the Positive Matrix Factorization (PMF) receptor model using the Multilinear Engine version 2. [75] The standard PMF model solves the basic mass balance equation (Equation 3.1) for source contributions, g_{ik} , source profiles, f_{kj} , and model error, e_{ij} , for i=1,n samples, j=1,m species, and k=1,p sources. Species concentrations, x_{ij} , corresponding uncertainties, σ_{ij}, and the user-defined number of sources, p, serve as model inputs.

$$x_{ij} = \sum_{k=1}^{p} g_{ik} f_{kj} + \varepsilon_{ij} \qquad \text{where } g_{ik}, f_{kj} > 0 \tag{3.1}$$

To add prior source contribution constraints from the CTM, an additional set of equations were solved simultaneously with equation 3.1. Specifically, we have added equations representing each of the t=1,v contributions, g'_{it}, to total fine particulate carbon predicted by the CTM model, in this case, for v=2 sources: total biomass combustion (k=1) and biogenic emissions (k=2). In general, the tth CTM constraint is given by equation 3.2.

$$g'_{it} = g_{it} I_t + \varepsilon'_{it} \qquad \text{where } t \subseteq k \tag{3.2}$$

where It represents a diagonal matrix which is solved by the model to account for potential multiplicative bias in the CTM predictions.

Profile Constraints

The thermal fractions of carbon for each source are normalized as follows,

$$\sum_{s=1}^{w} f_{ks} = 1 \pm u_s \qquad \text{where } s \subseteq j \tag{3.3}$$

where s=1,w carbon fractions. In this case, we set $u_s = 0.001$. Equation 3.3 rescales f such that g_{ik} in equations 3.1 and 3.2 represent total fine particle carbon and f_{kj} represents mass fraction of species j in source k relative to total carbon.

Based on prior knowledge of fire emissions in this region, we constrained the biomass combustion profile such that the value of f for potassium is > 0.01 [70] and that the corresponding values for NO3 and SO4 are < 0.05.[90]

Additionally, we constrain each of the l=1,b secondary feature profiles, in this case the biogenic source, so that each of the r=1,c non-carbonaceous species are near zero.

$$f_{lr} = 0 \pm u_{lr} \qquad \text{where } l \subseteq k \text{ and } r \subseteq j \tag{3.4}$$

In this case, we set $u_{lr} = 1 \times 10^{-5}$.

For the biomass combustion source we have limited the species to contain the carbon thermal fractions, potassium, nitrate, sulfate, and hydrogen. The decision to limit these species was based on the biomass combustion-like source profile resolved for these sites by the pure PMF model and the EPA SPECIATE database. To impose this constraint we applied Equation 3.4 with the above set of r=11 species. Going forward, PMF will refer to the $\gamma = 0$ scenario.

Penalized Object Function

In ME-2, equations 3.1 through 3.4 are solved by minimizing an object function, Q, through the use of a preconditioned conjugate gradient algorithm. The object function (equation 3.5) includes a weighting parameter associated with each of the applied constraints.

$$Q = (1-\gamma)\sum_{i=1}^{n}\sum_{j=1}^{m}\left[\frac{\varepsilon_{ij}}{\sigma_{ij}}\right]^2 + (\gamma)\sum_{i=1}^{n}\sum_{t=1}^{v}\left[\frac{\varepsilon'_{ij}}{\omega_{ij}}\right]^2 + \sum_{s=1}^{w}u_s + \sum_{l=1}^{b}\sum_{r=1}^{c}u_{lr} \qquad (3.5)$$

The first two terms in equation 3.5 represent the residuals from equations 3.1-3.3 where σ_{ij} is the species measurement uncertainty, ω_{it} is the CTM uncertainty, and γ is a user-defined weighting parameter that allows one to weigh the CTM predictions (equation 3.2) relative to those from the bivariate mass balance model (equations 3.1, 3.3, and 3.4). We implement this in ME-2 by specifying auxiliary formulas using error mode -14 for the first two terms. The second two terms in equation 3.5 represent the prior source profile constraints (equations 3.3 and 3.4) and are implemented with auxiliary formulas using error mode -12. Definition of the ME-2 error modes [76] are provided within the supplemental material (Table 3.3).

The uncertainties associated with the CTM results (equation 3.6) were defined as a function of estimated fractional error, α_t, and a fixed minimum CTM error, β_t, for the CTM model.

$$\omega_{it} = \sqrt{\left(\lambda_t \times g'_{it}\right)^2 + \beta^2} \qquad (3.6)$$

Model Implementation

For an initial user-specified number of sources, p, and a given value of the weighting parameter, γ, multiple model runs were conducted at 40 different starting points chosen randomly, and the chosen p-source baseline model was selected based on the minimum value of Q (= Qmin). The model was executed using a standard and a 10-fold cross-validation approach (as described in the next paragraph). These profile constrained model runs were conducted assuming a range of values for p with γ set equal to zero, representing the basic PMF solution (equations 3.1, 3.3 and 3.4) without additional information from the CTM (equation 3.2). The final value of p was chosen

based on the following criteria: the smallest value of p where a change in the ratio of n cross-validated Qmin to Qtheoretical approaches zero, where Qtheoretical = n*ms +n*mw/3 - n*p, ms is number of strong species and mw is the number of weak species; [28] and user judgment based upon qualitative agreement between each f_{kj} and prior knowledge of source profiles from known source types within the modeled region. Plots of the ratio between cross-validated Qmin and Qtheortical as a function of p are provided in the supplemental materials (Figure 3.6).

A final value of γ was chosen based on a minimum in the estimated RMSE of predicted versus measured total carbon. For a given value of γ, the RMSE was computed via a 10-fold cross validation procedure to insure a robust result. The measurements and CTM results were initially assigned to one of 10 separate groups. The model was then run 10 times, leaving one group out each time. The contributions of total carbon were predicted for the 10% missing values using ME-2 with the derived profiles from a given run. The resulting 10 groups of predictions were then combined into a single data set and the RMSE of total carbon was determined for each value of γ.

After selecting the γ associated with the minimum RMSE, blocked bootstrapping [37] was then used to estimate the uncertainties of the f_{kj} and the average values of g_{ik}. A sample block size of 4 was defined and profile matching was conducted on the predicted contributions of each species with an acceptable match defined as an $R^2 > 0.6$ across all contributions for a given model run.

Here, the hybrid approach was applied using 17 different values of γ and solved for 4 features. The Monture and Sula Peak IMPROVE monitor data were modeled using CTM predicted contributions from biomass combustion and biogenic sources as constraints (see equation 3.2). Total biomass combustion was used instead of specifying *a priori* separate secondary and primary biomass combustion features because of the high Pearson correlations between the CTM predictons of these two source contributions to total carbon, 0.97 and 0.98 at Monture and Sula Peak respectively.

This high correlation would likely cause difficulty in effectively separating the two sources, thus the sum of the CTM predicted contributions from these two sources was used to constrain the single biomass combustion PMF feature as described earlier.

To determine the CTM uncertainty, the fractional error was set to $\lambda = 1.18$ and 0.66 for the biomass combustion and biogenic features, respectively, and a minimum CTM error of $\beta = 0.1$ was used. A sensitivity analysis of these assumed values was also conducted.

3.4 Results

At both sites, we chose a four feature model based on the criteria described earlier (see Figure 3.4 in the supplemental material for plots of $Q/Q_{theoretical}$ versus p). At each site, the hybrid model was defined by a minimum value of the cross-validated total carbon RMSE as a function of γ (Figure 3.1). The $Q/Q_{theoretical}$ for Monture and Sula Peak at the minimum γ was 2.19 and 1.94 respectively. A sensitivity analysis of the cross-validated total carbon RMSE curve to modifications of the uncertainty error fraction and minimum model error at Monture was conducted and the resulting plots are provided in the supplementary material (figures 3.10 and 3.11). In both the PMF and hybrid models, two features were identified in addition to the constrained biomass combustion and biogenic source features: a sulfate/ nitrate-rich feature and a soil feature. The profiles for all four of these features at each site for the final hybrid model are shown in Figure 3.2. Overall, the derived profiles at the two sites are similar and, where differences occur, the associated bootstrapped uncertainties span these differences with the exception of Al in the sulfate and nitrate rich feature. Profiles of the four features derived from the PMF model are provided in the supplemental material (Figure 3.5).

Performance statistics of both the hybrid and PMF model across all modeled species are provided in Table 3.1. In general, the hybrid model performs slightly better than the PMF model across all species. In addition, the predicted average

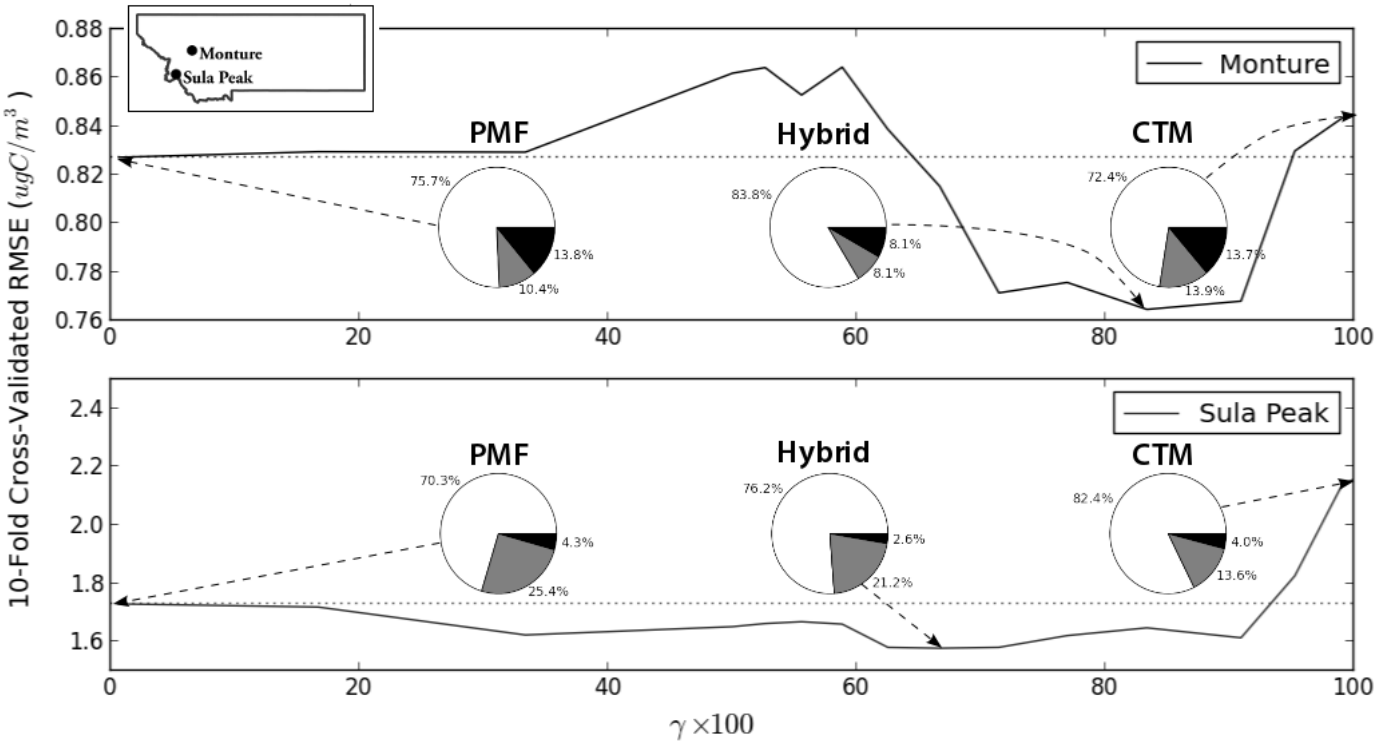

Figure 3.1: Cross-Validated Root Mean Sqaure Error (RMSE) of Total Carbon versus weighting parameter (see equation 3.5) for Monture and Sula Peak. The black line represents the RMSE and the dotted line represents the PMF RMSE value. In addition, the pie charts for the PMF, CTM, and hybrid (minimum γ) models are provided for each site. A general map of Montana with the site locations is provided as an inset.

contributions to total carbon from biomass combustion, biogenic sources, and "other" sources (sulfate/nitrate-rich and soil) are shown in Figure 3.1 as pie charts for γ values associated with PMF ($\gamma = 0$), CTM ($\gamma = 1$) and the hybrid model. Figure 3.3 shows predicted versus observed total carbon for each sample at each site for all three models. The hybrid model displays the best agreement, given that it by definition has the lowest RMSE. For the PMF and hybrid models, a time series of the predicted contributions at each site can found in the supplemental material (Figures 3.5 and 3.6). Seasonal performance statistics of modeled total carbon are provided in Table 3.2 for the hybrid and CTM model; associated scatter plots are provided in Figure 3.9.

3.5 Discussion

Our models did not attempt to distinguish between primary and secondary biomass because the CTM model predictions of these two features were highly correlated. These CTM results were combined due to their high correlation and the potential of the constrained ME-2 features to swap between the two. However, the separation of these two features using the hybrid model would in principle be possible using other CTM models, data from other sites further from the fire origins, or additional marker species specific to primary and/or secondary particles.

Recall that the biomass combustion feature represents both primary and secondary particles and the contributions are dominated by summer fire events in all three seasons. The hybrid model biomass combustion profiles, shown in Figure 3.2, are largely comprised of organic carbon thermal fractions. The wildfire CTM constraint on this feature is consistent with the high percentage of pyrolytic carbon relative to the other thermal fractions. Additionally, the biomass combustion contributions predicted by the hybrid model are reasonably well correlated with the CTM predictions of total wildfire (Pearson correlation coefficients of 0.72 and 0.91 for Monture and Sula Peak, respectively) even though the CTM predicts consistently larger contributions than

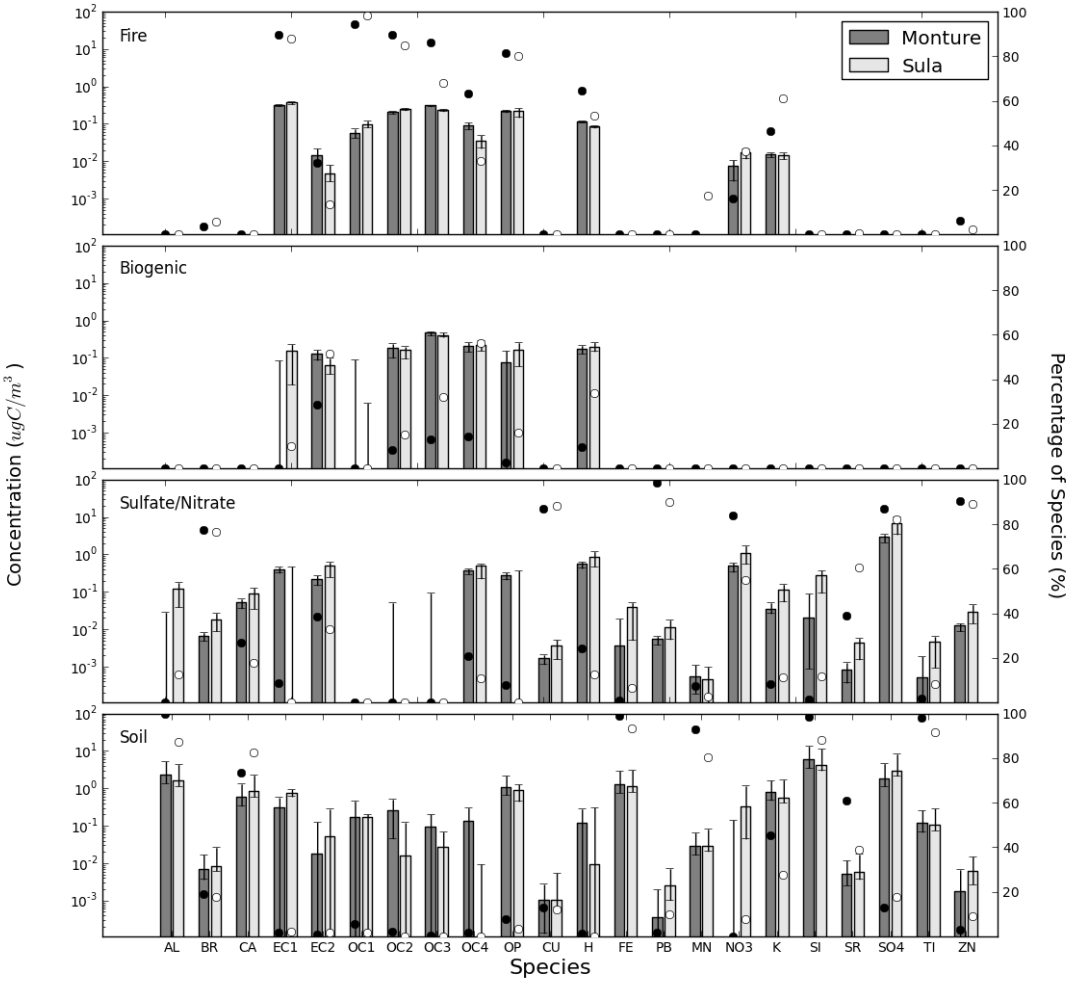

Figure 3.2: Hybrid model derived source profiles for Monture and Sula Peak. The dark bars represent the average species contributions to total carbon for Monture and the light bars represent Sula Peak. The bootstrapped 95% confidence intervals are also shown. The circles represent the percent of overall predicted concentrations for a given species associated with a feature at Monture (black circles) and Sula Peak (white circles).

the hybrid mode (see Figure 3.3). The hybrid model fit the measured total carbon better than the CTM (see Figure 3.3) and, based on a seasonal comparison, fit the data better during times of high total carbon contributions (see Figure 3.9 and Table 3.2).

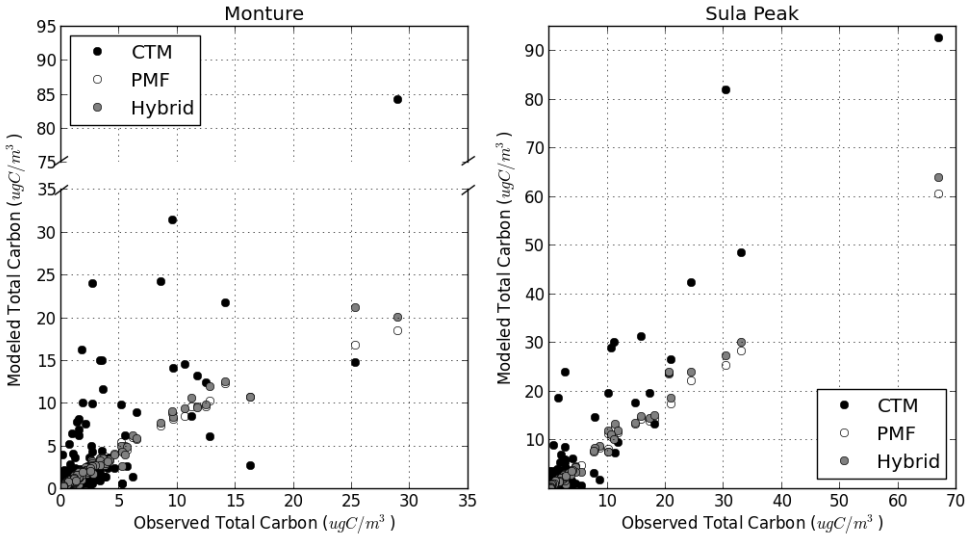

Figure 3.3: Predicted Versus Observed Total Carbon for the CTM, PMF, and hybrid models at Monture and Sula Peak. The black dots represent the CTM, the grey dots represent the hybrid model, and the white dots represent PMF. The white dots are somewhat obscured due to similarities in the fit of PMF and the hybrid solution.

The average ratio of K to OC from the biomass combustion profiles at Monture and Sula were 0.036 (s.d. = 0.017) and 0.026 (s.d. = 0.016), respectively. Assuming that the K in this feature is water soluble, these values are consistent with contributions from primary fire emissions, specifically with: 1) the range of values of primary emissions of woody debris sampled near these sites (Northern Rockies region) as reported by Munchack et.al. [70]; 2) our model predictions of major contributions from the primary fire feature to total carbon and organic carbon that were observed in the summer and originated from fires in this region ; 3) the relatively small amount of secondary biomass particles predicted by the CTM model that could have otherwise

decreased the primary K/OC ratio; and, 4) the fact that the largest impacts from this feature to total carbon did not occur at the same time (see Figure 3.8) at both sites, implying the lack of a broad regional haze from long range transport.

The biogenic feature was constrained to have a purely carbon and hydrogen based profile, shown in Figure 3.2, and contributions with strong seasonality (see Figures 3.4 and 3.5) which is expected of biogenic emissions. The ability to distinguish the biogenic feature from the biomass combustion is a result of the CTM constraints and the profile constraints. The CTM provides the guidance required to interpret this feature as biogenic, without it the feature could potentially be interpreted as an arbitrary secondary source.

The soil profiles are provided in Figure 3.2. The feature identified as soil is the majority contributor to Al, Ca, Fe, Mn, Si, and Ti. In addition to elements commonly found in soils, we identified the presence of K and Sr, a result in agreement with the surface geochemistry identified by Shacklette and Boerngen [100]. A comparison of the relative composition of K, Fe, and Ca reported by Shacklette is consistent with our profile. The contributions from the soil feature have a high level of seasonality with the peaks in the summer due to dry conditions (see Figures 3.4 and 3.5). It is plausible that some of these particles were generated via updrafts during fire events given that the modest correlations between the biomass combustion feature and the soil feature (r = 0.52 and 0.56 at Monture and Sula Peak, respectively).

The nitrate and sulfate dominated profiles, shown in Figure 3.2, are enriched with Cu, Pb, and Zn and present as misaligned, sharp peaks, potentially implicating nearby industrial sources. Generally, the composition of the two sites are in agreement, however the confidence levels of Sr, Zn, and Al indicate a differing level of contributions for these element. This feature was found to have a low correlation with biomass combustion source.

To examine the possible source(s) associated with the sulfate/nitrate-rich feature, we used the NOAA hysplit model to assess 48 hour back trajectories leading to the

Table 3.1: RMSE and R^2 at Monture and Sula Peak for the γ_{min} hybrid model and constrained PMF against observed concentrations for each element and total carbon.

Element	Monture			Sula Peak		
	S/N	RMSE/R^2		S/N	RMSE/R^2	
		Hybrid	Const. PMF		Hybrid	Const. PMF
AL	2.85	0.037/0.93 (0.062/0.77)[a]	0.039/0.92 (0.068/0.73)	2.42	0.012/0.95 (0.015/0.93)	0.012/0.94 (0.014/0.94)
BR	9.8	0.000/0.73 (0.000/0.56)	0.000/0.73 (0.001/0.54)	11.45	0.000/0.77 (0.001/0.52)	0.000/0.77 (0.001/0.53)
CA	18.42	0.028/0.60 (0.031/0.57)	0.028/0.60 (0.032/0.51)	18.25	0.020/0.61 (0.023/0.51)	0.020/0.61 (0.023/0.51)
EC1	5.48	0.198/0.96 (0.368/0.88)	0.281/0.96 (0.434/0.84)	5.19	0.353/0.99 (0.657/0.86)	0.250/0.99 (0.751/0.85)
EC2	1.55	0.042/0.47 (0.030/0.74)	0.041/0.55 (0.045/0.57)	1.43	0.027/0.80 (0.037/0.69)	0.027/0.80 (0.040/0.69)
OC1	1.12	0.297/0.81 (0.333/0.67)	0.312/0.78 (0.341/0.65)	1.25	0.464/0.89 (0.702/0.69)	0.470/0.87 (0.739/0.63)
OC2	2.57	0.410/0.93 (0.546/0.79)	0.454/0.91 (0.564/0.77)	2.56	0.610/0.96 (1.174/0.81)	0.694/0.96 (1.279/0.70)
OC3	2.39	0.141/0.96 (0.173/0.94)	0.104/0.98 (0.218/0.91)	2.21	0.254/0.98 (0.408/0.88)	0.222/0.99 (0.471/0.85)
OC4	2.37	0.110/0.77 (0.097/0.82)	0.102/0.81 (0.108/0.78)	1.95	0.106/0.82 (0.138/0.75)	0.101/0.84 (0.146/0.72)
OP	2.48	0.211/0.91 (0.281/0.87)	0.218/0.93 (0.309/0.85)	2.43	0.537/0.82 (0.229/0.96)	0.457/0.83 (0.261/0.95)
CU	1.9	0.000/0.14 (0.000/0.19)	0.000/0.14 (0.000/0.18)	1.86	0.000/0.06 (0.000/0.10)	0.000/0.06 (0.000/0.10)
H	13.62	0.039/0.98 (0.097/0.91)	0.046/0.98 (0.121/0.86)	15.52	0.066/0.99 (0.257/0.86)	0.078/0.99 (0.299/0.80)
FE	18.75	0.005/0.99 (0.016/0.90)	0.005/0.99 (0.020/0.86)	18.71	0.005/0.98 (0.012/0.91)	0.005/0.98 (0.011/0.91)
PB	3.47	0.000/0.55 (0.000/0.45)	0.000/0.55 (0.000/0.46)	3.11	0.000/0.36 (0.000/0.35)	0.000/0.34 (0.000/0.35)
MN	11.32	0.002/0.42 (0.003/0.37)	0.002/0.43 (0.003/0.32)	7.44	0.001/0.69 (0.001/0.39)	0.001/0.68 (0.001/0.31)
NO3	4.21	0.082/0.35 (0.098/0.15)	0.079/0.40 (0.097/0.16)	6.4	0.106/0.79 (0.166/0.39)	0.100/0.80 (0.159/0.50)
K	15.92	0.018/0.93 (0.031/0.82)	0.010/0.98 (0.047/0.50)	18.65	0.016/0.96 (0.038/0.80)	0.018/0.97 (0.042/0.82)
SI	11	0.070/0.95 (0.136/0.80)	0.075/0.95 (0.151/0.75)	8.52	0.021/0.97 (0.040/0.91)	0.021/0.97 (0.040/0.91)
SR	1.81	0.000/0.62 (0.000/0.63)	0.000/0.62 (0.000/0.57)	1.85	0.000/0.65 (0.000/0.49)	0.000/0.65 (0.000/0.48)
SO4	10.67	0.177/0.70 (0.264/0.51)	0.177/0.70 (0.234/0.57)	8.66	0.118/0.77 (0.233/0.45)	0.120/0.76 (0.226/0.48)
TI	5.25	0.001/0.98 (0.001/0.90)	0.001/0.98 (0.002/0.86)	6.15	0.000/0.98 (0.001/0.91)	0.000/0.98 (0.001/0.92)
ZN	16.06	0.002/0.34 (0.002/0.20)	0.002/0.31 (0.002/0.20)	9.28	0.002/0.54 (0.002/0.30)	0.002/0.52 (0.002/0.25)
TC	–	0.574/0.98 (0.765/0.91)	0.636/0.97 (0.827/0.88)	–	0.888/0.99 (1.577/0.86)	0.908/0.99 (1.729/0.82)

[a] () = Cross-validated RMSE / R^2

three large total carbon spikes observed in Monture and 2 large spikes at Sula Peak (see Figures 3.4 and 3.5). At Monture two of the spikes were associated with winds coming from the West and most recently passing over the Missoula region. During the sampling period a large pulp and paper mill was active in Missoula that may have contributed to these spikes. The third spike at Monture was associated with winds from the Northwest near Libby where current mining operations [65] may have also contributed. At Sula Peak, the winds arrived from the West during these spikes, passing over active gold, cobalt, and molybdenum mines and processing facilities.[1] A comparison of this feature (Figure 3.2) to the EPA SPECIATE database found that the "Regional smelter background" SPECIATE profile is in agreement with the enriched values of Cu, Pb, Sr, Zn, NO3 and SO4 in this feature.

The sensitivity of the hybrid model to variations in the CTM uncertainty equation were explored by modifying the relative error fraction between biomass combustion and biogenic and by modifying the minimum model error values (figures 3.10 and 3.11). Changes in the relative error fractions caused only small variations in the RMSE curve and identified similar minimum values. The feature profiles and compositions at the minimum in each case was identical or near identical, indicating that the model is robust to the error fraction estimates for dominant features. Minimum model error values of $\beta = 0.01$, 0.1, and 1.0 demonstrated consistent minimum identification. With β values of 10 or 100 the RMSE minimum was no longer identifiable. The lack of a minimum in the cross-validated RMSE indicates no improvement over PMF, likely due to the large error associated with the CTM and thus the lack of control by the constraining CTM features.

Here we provided a framework for direct coupling of source apportionment from a CTM model with PMF via ME-2. The CTM modeling results used here provided estimates of total carbon, however our modeling framework provides the flexibility to assess particle mass or other species of interest as predicted by a CTM. The hybrid model's use of CTM constraints provides guidance and increased confidence in

identifying derived features.

Previous work by Schichtel et. al. [98] introduced this hybrid modeling approach using a synthetic data set. The synthetic data sets performance was compared against a synthetic truth using direct RMSE calculations. The model described here differs from that of Schichtel et. al. by including hard profile constraints and by incorporating a cross-validation procedure to locate a minimum RMSE as a function of the weighting parameter. Since the true source contributions are unknown, we are limited to a comparison of total carbon from all hypothesized sources. However, like the work of Schichtel et. al., our resulting cross-validated RMSE curve has a modest minimum indicating a better fit of the data than either CTM or PMF alone.

Other investigators have coupled CTM results with receptor-based models. The approach most similar to ours is that of Maier and colleagues [61]. Their ensemble model uses CTMs and PMF to derive average contributions from sources, uses these contributions to derive source profiles using an inverse chemical mass balance (CMB) approach, and finally uses the new source profiles in a common CMB approach to estimate contributions. Our model differs by directly applying the CTM predictions with uncertainty to the PMF model. This allows us to resolve sources that are not necessarily identified by PMF alone.

3.6 Acknowledgements

This work was supported by the Joint Fire Science Program (09-1-03-1). The assumptions, findings, conclusions, judgments, and views presented herein are those of the authors and should not be interpreted as representing the National Park Service polices.

3.7 Supplemental Material

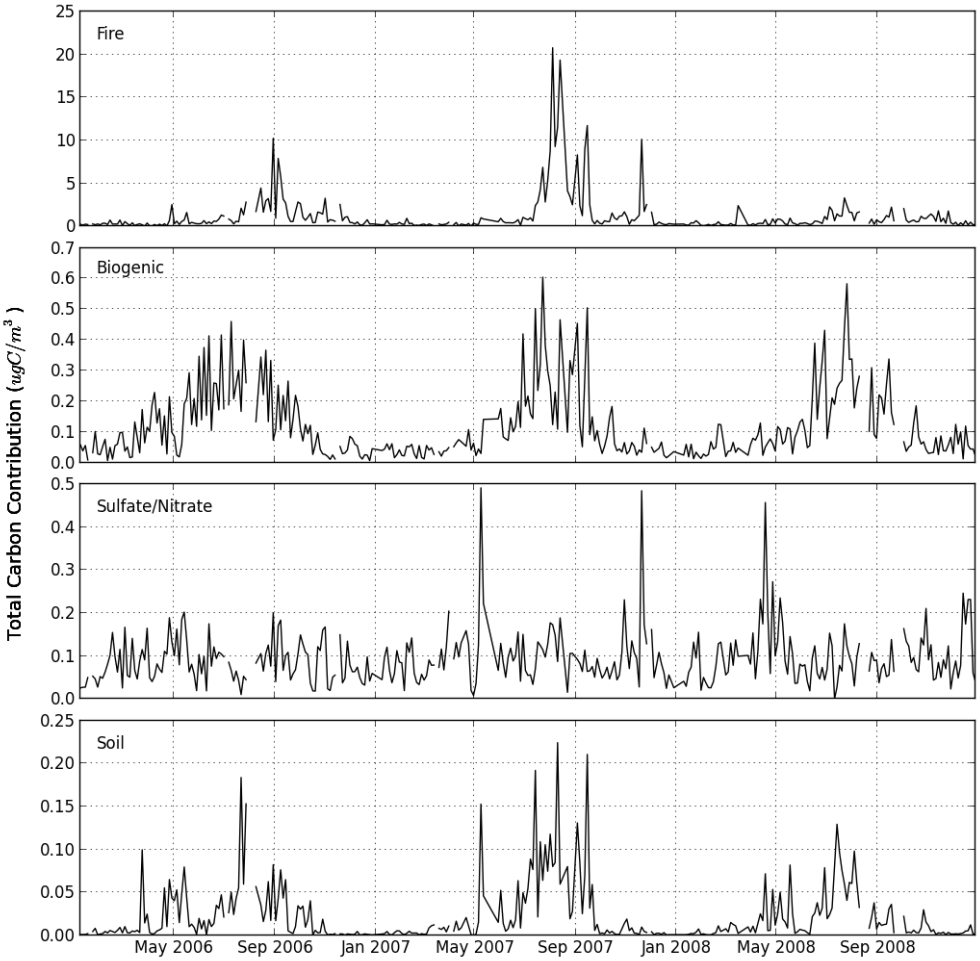

Figure 3.4: Hybrid model derived feature contributions for the Monture site. From top to bottom, the features represent the biomass combustion, biogenic, nitrate/sulfate dominant, and soil features. Broken sections within the time series are due to missing samples.

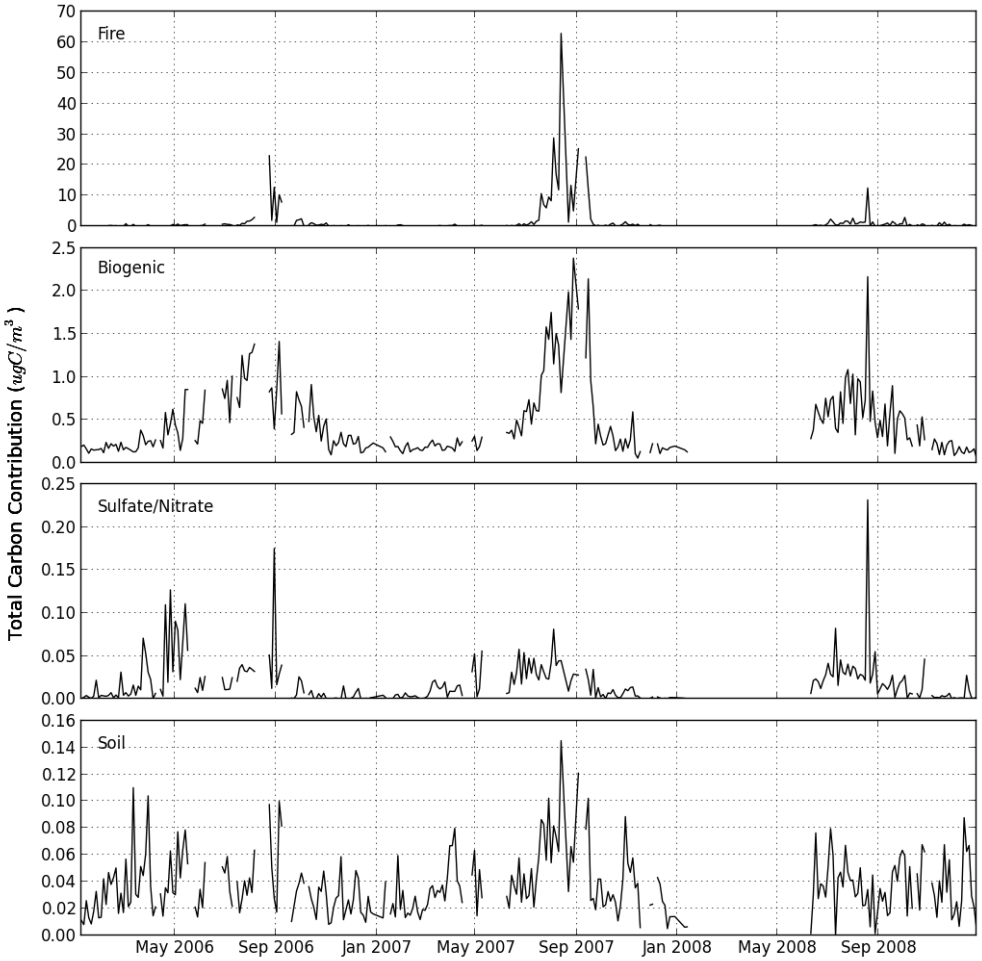

Figure 3.5: Hybrid model derived feature contributions for the Sula Peak site. From top to bottom, the features represent the biomass combustion, biogenic, nitrate/sulfate dominant, and soil features. Broken sections within the time series are due to missing samples.

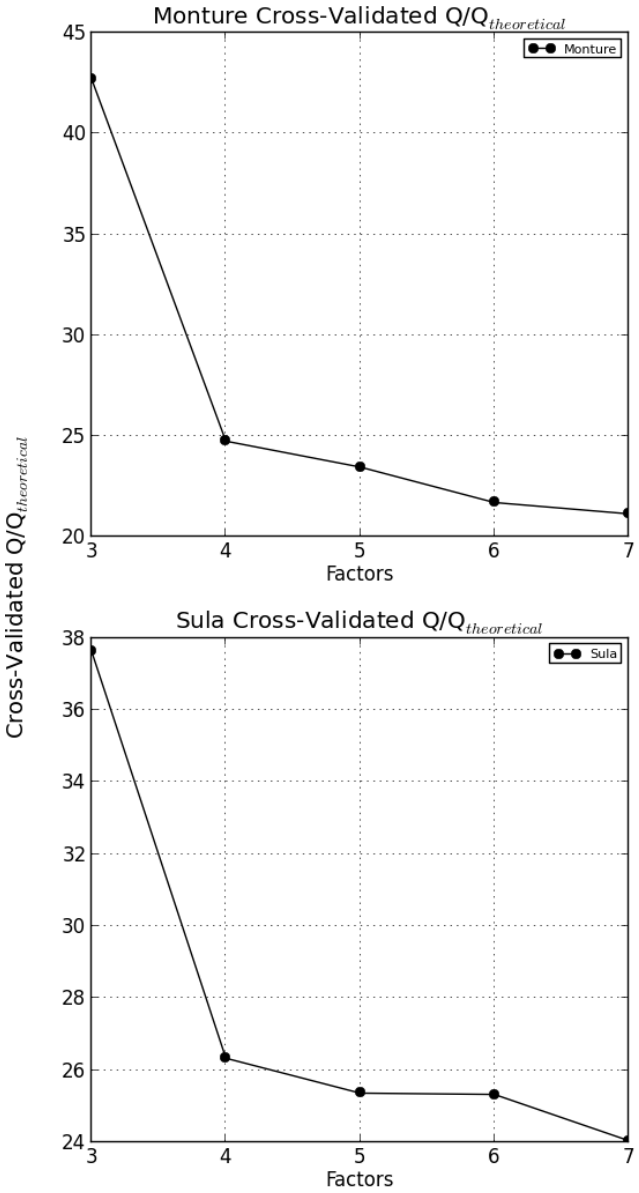

Figure 3.6: Cross-Validated Q/Qtheoretical versus number of factors. The analysis of different values of Q provides an indication of how the model performs under different numbers of sources. An inflection in these plots, if followed by a consistent trend, provides evidence of deteriorating gains beyond the number of source associated with the inflection point.

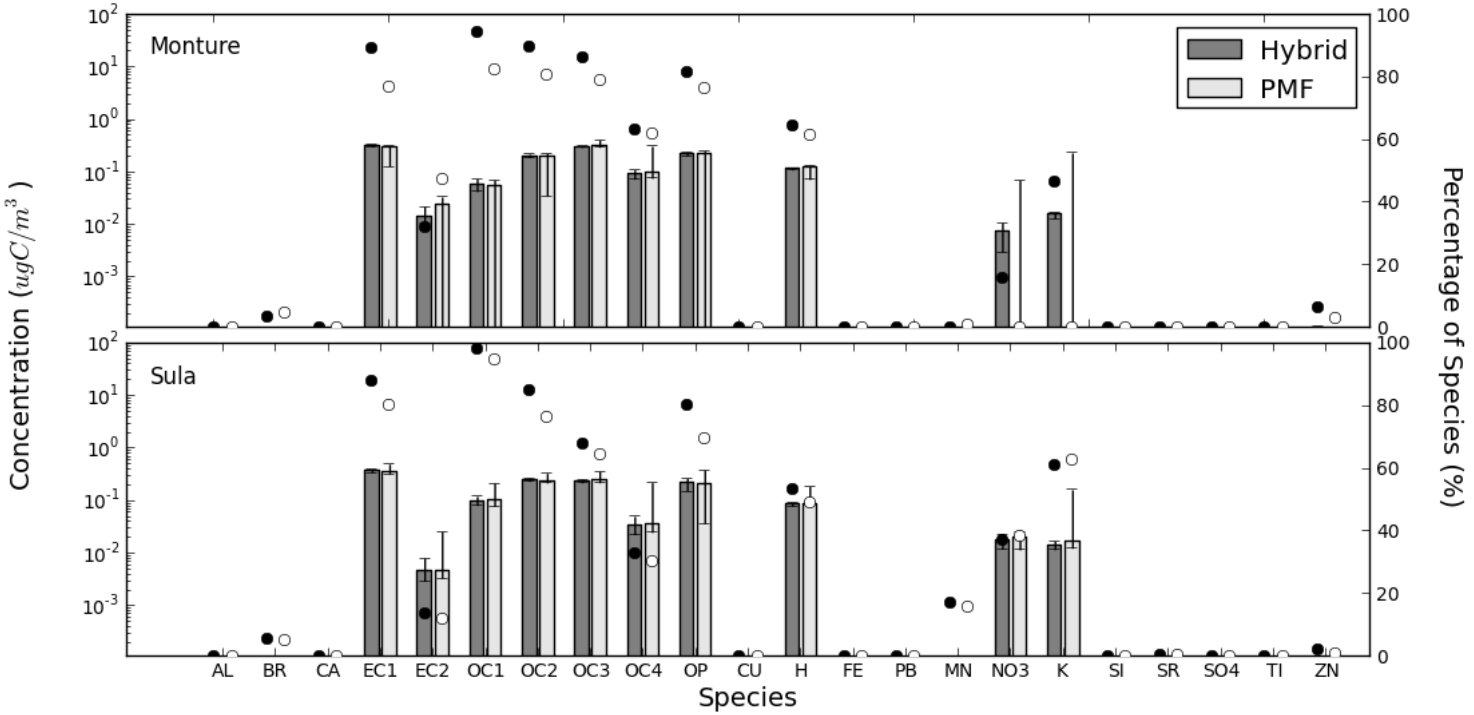

Figure 3.7: Hybrid model derived source profiles of the hybrid and PMF solutions at Monture and Sula Peak. The dark bars represent the average species contributions to total carbon from the hybrid model and the light bars represent PMF. The bootstrapped 95% confidence intervals are also shown. The circles represent the percent of overall predicted concentrations for a given species associated with a feature from the hybrid model (black circles) and PMF (white circles).

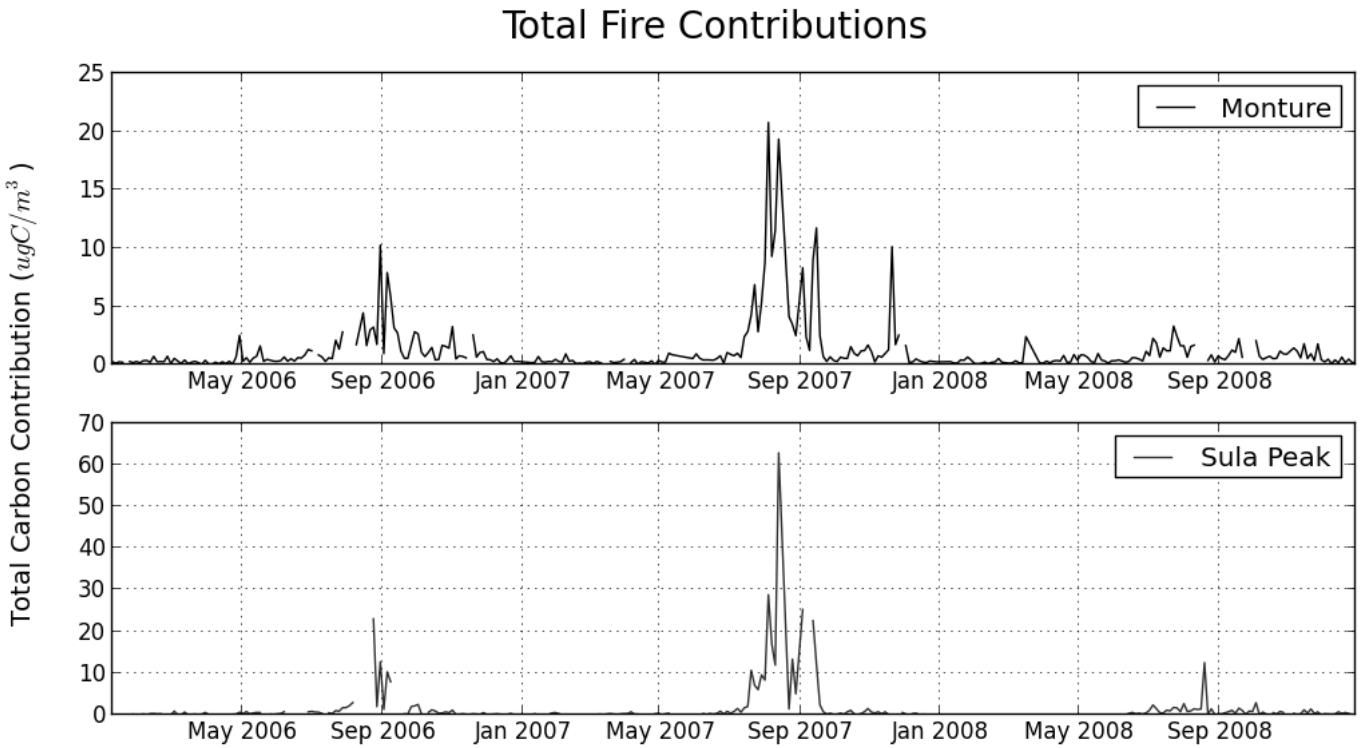

Figure 3.8: Side-by-side comparison of the biomass combustion feature contributions from Monture and Sula Peak. Of note, the peaks of the contributions are frequently misaligned, indicating that the features are likely dominated by primary wildfire combustion.

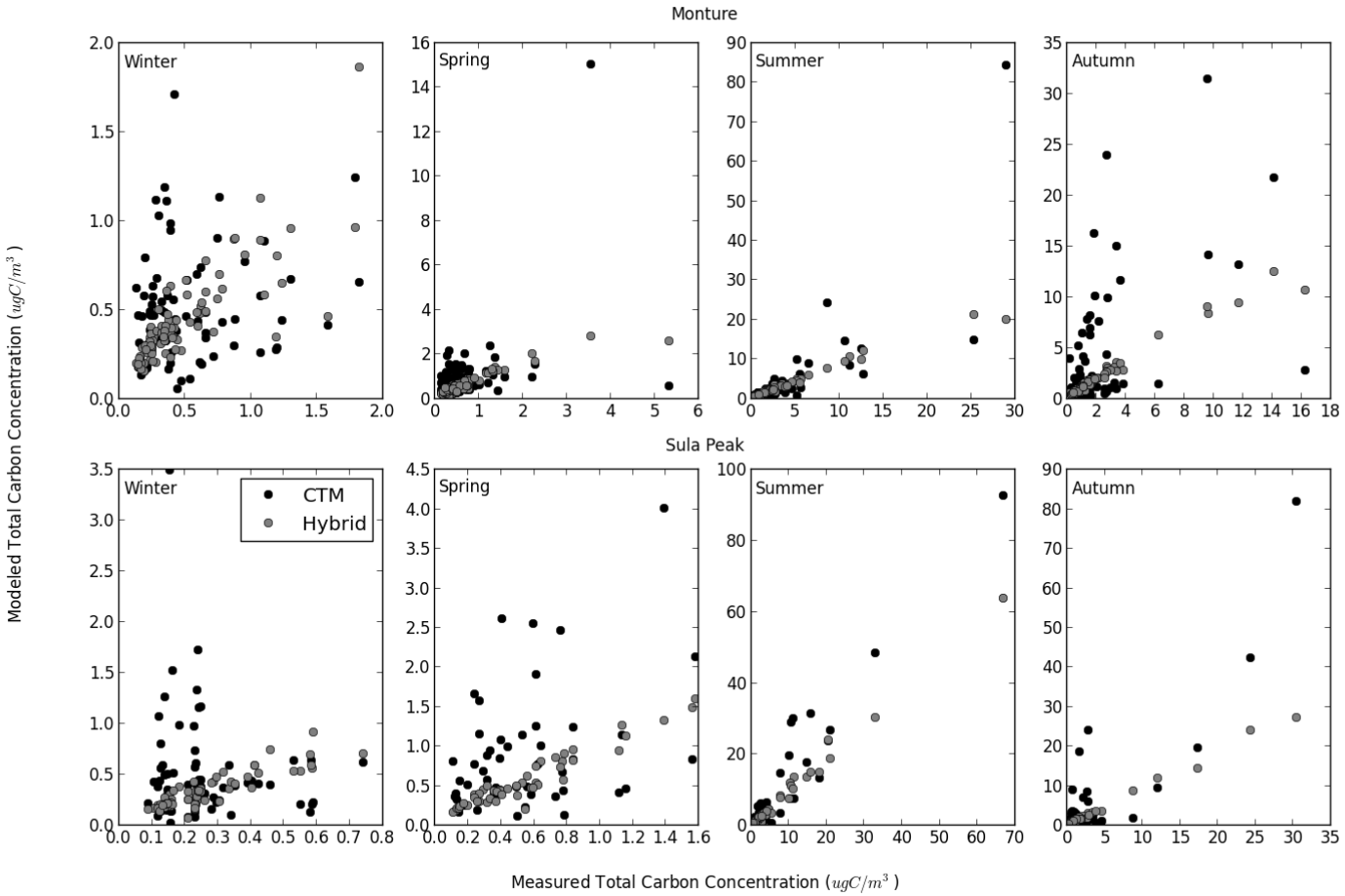

Figure 3.9: Comparison of CTM and hybrid modeled total carbon versus observed total carbon for Monture and Sula Peak across seasons. The winter season represents December, January, and February; the spring season represents March, April, and May; the summer season represents June, July, and August; and the autumn season represents September, October, and November.

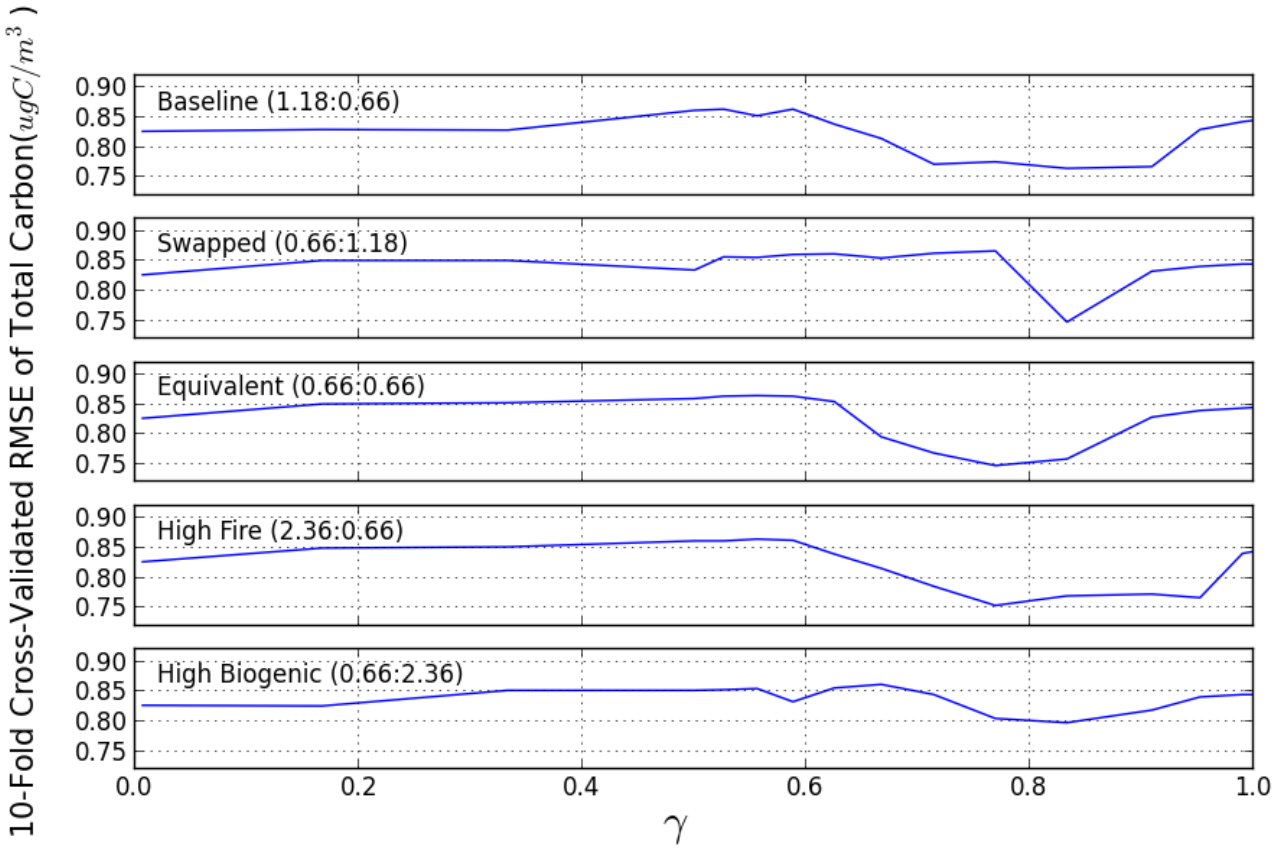

Figure 3.10: Impact on the cross-validated total carbon RMSE curve due to variations in the relative error fraction (biomass combustion : biogenic) at the Monture site.

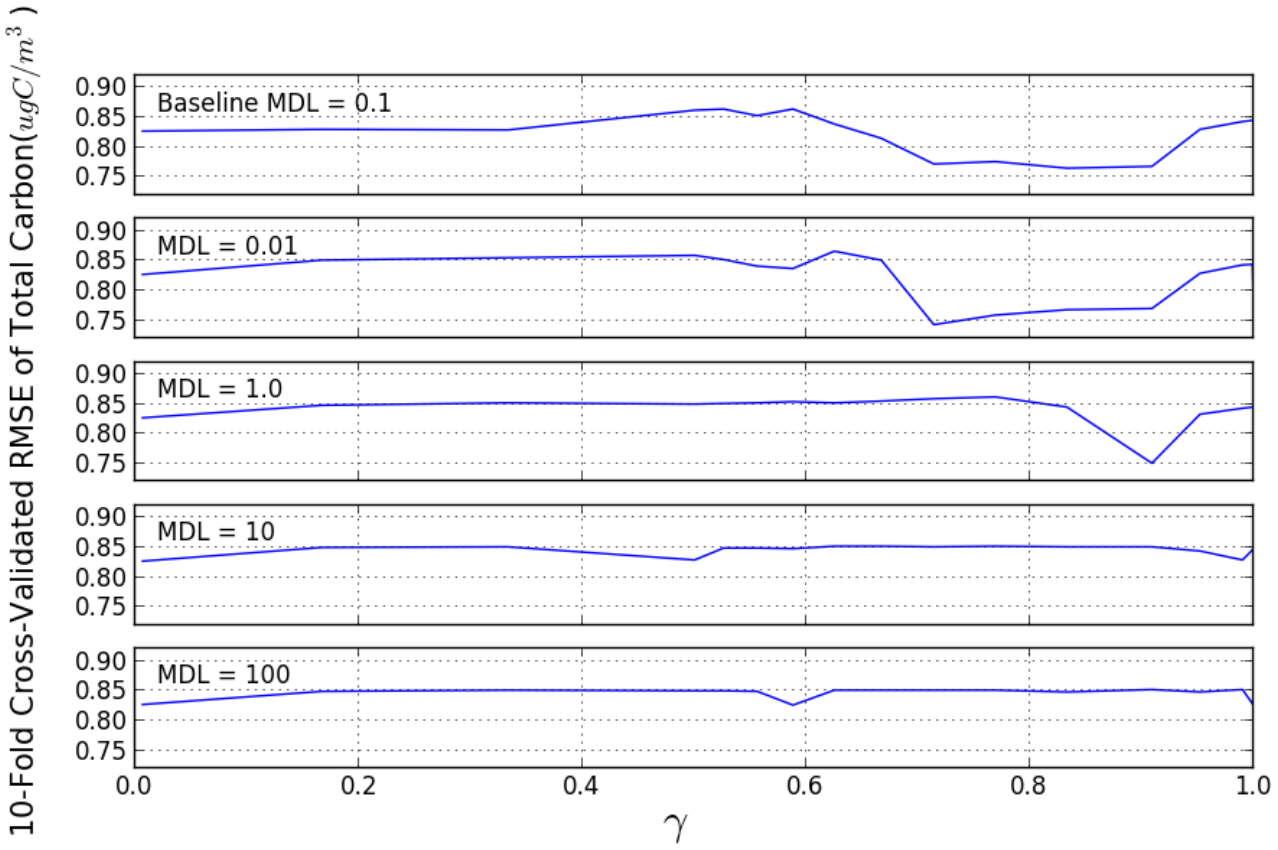

Figure 3.11: Impact on the cross-validated total carbon RMSE curve due to variations in the minimum model error (ugC/m3) at the Monture site.

Table 3.2: Modeled total carbon seasonal performance statistics from the CTM and the hybrid model (γ_{min}) at Monture and Sula Peak.

		Hybrid		CTM	
Site	Season	R^2	RMSE	R^2	RMSE
Monture	Winter	0.64	0.13	0.03	0.81
	Spring	0.87	0.18	0.21	1.73
	Summer	0.98	1.00	0.65	2.84
	Autumn	0.97	0.50	0.31	21.66
Sula Peak	Winter	0.77	0.07	0.02	1.09
	Spring	0.92	0.09	0.14	1.35
	Summer	0.99	1.54	0.92	5.10
	Autumn	0.99	0.65	0.78	25.64

Table 3.3: ME-2 Error Mode Equations[a]

Error Mode	Equation
-12	$Std = C1^{b} + C2^{c} \times \sqrt{\lvert Data \rvert} + C3^{d} \times \lvert Data^{e} \rvert$
-14	$Std = C1^{f} + C2^{c} \times \sqrt{max\left[\lvert Data \rvert, \lvert Fitted \rvert\right]} + C3^{g} \times max\left[\lvert Data^{h} \rvert, \lvert Fitted \rvert\right]$

[a] Paatero, 2009 [76]

[b] Normalizing equation C1 = 0.001; Profile constraint C1 = 1×10^{-5}

[c] C2 = 0

[d] Normalizing and profile constraints C3 = 0

[e] Normalizing equation Data = 1; Profile constraint Data = 0

[f] PMF C1 = measurement uncertainty; CTM constraint C1 = estimated CTM uncertainty

[g] PMF and CTM constraint C3= 0.05

[h] PMF Data = measurements; CTM constraint Data = CTM predicted contributions

Chapter 4

SUMMARY AND FINDINGS

4.1 Summary

This dissertation sought to improve upon standard approaches to receptor-oriented source apportionment by incorporating novel sets of constraints within the ME-2 framework. Specifically, elemental ratio constraints, spatial separation constraints, and CTM source attribution constraints were used to influence the standard PMF model. Tracer-like elemental ratios were developed to aid in the separation of a brake wear feature and a tire wear feature while modeling two-week samples of coarse PM collected for two seasons across three cities. Inclusion of the element ratio constraints improved the models ability to identify the constrained features and provided increased confidence in the solution. However, initial modeling of these data indicated that all of the sources were ubiquitous with the exception of windblown dust. To address this, ME-2 was configured to model a different soil-like feature for each city while allowing the other features to span all three cities. This separation of city-specific soil features allowed distinct soil profiles to be realized while providing a larger set of samples than modeling the individual cities, thus improving model robustness. Overall, these profile constraints and contribution constraints improved the ability to interpret the model output.

Using similar ME-2 scripting methods, independent predictions of total fine particle carbon from a CTM model were used to help separate biomass combustion aerosol from secondary biogenic aerosol. The CTM results were incorporated into ME-2 using an additional equation which set the ME-2 contributions proportional to the CTM contributions with a specified level of uncertainty. Additional profile constraints were

used to isolate the species associated with either biomass combustion or biogenic formation. Using a weighting parameter, the CTM model uncertainty was varied, resulting in a set of solutions spanning a pure PMF solution and a pure CTM solution. This approach was applied to two rural sites in Montana which are known to have a heavy impact from biomass combustion during the summer months. In order to test the validity of this hybrid approach, and to identify the ideal weighting parameter, a 10-fold cross-validation was used. At the weighting parameter associated with a minimum cross-validated RMSE for each site, the profiles and contributions were reasonably correlated with the CTM while maintaining a strong fit to the measurements. The cross-validated RMSE of total carbon for both sites was improved over the pure CTM or PMF predictions. While other researches have coupled CTM results with receptor-oriented models applied to real data, this model provides a direct coupling of the CTM results within the ME-2 framework.

4.2 Strengths & Weaknesses

Evaluation of this hybrid ME-2 model provided promising results and included important features. The profile constraints present in both the coarse PM analysis and the hybrid modeling offered an approach for constraining PMF with prior profile information which, in turn, provides an inherent ordering to the constrained factors and removes any ambiguity in their interpretation. The ability to impart these profile constraints is not new and is becoming common place in the latest versions of EPA PMF. The spatial separation methodology used for the coarse PM soil features and the CTM constraints used in the hybrid modeling also provide some improvement in the ability to interpret these constrained features. The application of these two constraints is new and requires the use of the ME-2 scripting language. Additionally, by constraining with the CTM contributions, sources can be modeled which would otherwise not be present using traditional receptor modeling techniques (e.g., biogenic SOA). The ability to separate these features with real data is a key finding of

this work. In addition to these strengths, the hybrid model also supplies a method of varying CTM uncertainty to achieve a best fit result, accounts for multiplicative bias in CTM results, and is flexible enough to model any species resulting from any CTM.

Importantly, the use of this hybrid approach also has limitations. Applied elemental ratio constraints or source profile constraints implies that the constrained species are well understood, which is not always true. Care must be taken when developing profile constraints. With regard to the hybrid model, the uncertainties associated with the CTM constraints require an initial estimate . The γ values tested by the hybrid approach multiply these uncertainties by a scalar, thus the relative uncertainties between different sources always remains the same. This becomes problematic if the initial uncertainty of a single source is poorly estimated. Similarly, if the CTM constraint is significantly different from reality the hybrid model may not fit the CTM constraint well. However, if the CTM constraint differs only in magnitude (i.e., not in temporal shape), then the hybrid model will adjust for the multiplicative bias. Finally, the hybrid model, much like PMF, has difficulty separating highly correlated features. The model becomes prone to factor swapping in the presence of highly correlated CTM constraints and the reliability of proper separation, lacking strong profile constraints, becomes difficult.

4.3 Suggested Future Research

4.3.1 Combined Site Hybrid Model

Combining multiple sites using the hybrid approach would be a natural next step from this work. To conduct the modeling, sites which are in reasonable proximity would be identified, each separately modeled with the hybrid model, and differences in the profiles would be identified. Presumably nearby sites share some ubiquitous features but may also be influenced by local sources. Combining the measurement data and running the hybrid method with spatial separation on local sources could

enhance the signal of the ubiquitous features due to increase sample size, provide an increased number of constraint samples, and potentially identify features which would have otherwise been overlooked. The use of multiple sites may provide enough information to more readily separate primary and secondary biomass combustion. However care must be taken when modeling multiple sites as the profiles associated with the combustion of different materials may vary. Proper identification of these different materials would provide a basis for using the spatial separation method to model each type of combustion separately.

4.3.2 Source Specific γ

As indicated previously, under the current hybrid model as γ changes the source-specific CTM uncertainties remain constant relative to one another. By modeling a separate γ for each source, γ_t, the model would provide separate estimates of the ideal uncertainty for each constrained feature. The results of this work my provide more insight into the uncertainties of the CTM model used on a source-by-source basis and offer guidance in diagnosing potential issues with the CTM. However, this approach does not provide the same information of an overall CTM prediction as the hybrid model does here, rather it inspects the individual sources.

4.3.3 Other CTMs

Numerous CTMs exists, each having their own strengths and weaknesses. Here we've tested the hybrid model using the results from the CAPITA Monte Carlo model. Exploration of other modes such as CAMx, CMAQ, and WRF/Chem could provide additional refinements for the hybrid modeling approach. Additionally, an ensemble of results from multiple CTMs could be used to (where available) to refine the constraints further.

4.4 Conclusions

The interpretation of traditional PMF receptor modeling results is often fraught with ambiguity due to convergence issues, lack of data, poorly defined uncertainties, or unexpected results. By incorporating information which helps tie the model to reality, in this case a separate, deterministic CTM or set of emission profiles, an improvement in these interpretations can be realized. The ratio, spatial, and CTM constraint methods described here provided results which helped the model interpretation and overall understanding of sources at two IMPROVE sites.

BIBLIOGRAPHY

[1] Idaho Geological Survey Map of 2006 Mining-Exploration Activity, 2006.

[2] Kouji Adachi and Yoshiaki Tainosho. Characterization of heavy metal particles embedded in tire dust. *Environment international*, 30(8):1009–17, October 2004.

[3] C Allen, G Young, and S Haupt. Improving pollutant source characterization by better estimating wind direction with a genetic algorithm. *Atmospheric Environment*, 41(11):2283–2289, April 2007.

[4] Christopher T. Allen, Sue Ellen Haupt, and George S. Young. Source Characterization with a Genetic AlgorithmCoupled DispersionBackward Model Incorporating SCIPUFF. *Journal of Applied Meteorology and Climatology*, 46(3):273–287, March 2007.

[5] S M Almeida, C a Pio, M C Freitas, M a Reis, and M a Trancoso. Approaching PM(2.5) and PM(2.5-10) source apportionment by mass balance analysis, principal component analysis and particle size distribution. *The Science of the total environment*, 368(2-3):663–674, 2006.

[6] F. Amato, M. Pandolfi, a. Escrig, X. Querol, a. Alastuey, J. Pey, N. Perez, and P.K. Hopke. Quantifying road dust resuspension in urban environment by Multilinear Engine: A comparison with PMF2. *Atmospheric Environment*, 43(17):2770–2780, June 2009.

[7] F Amato, M Pandolfi, T Moreno, M Furger, J Pey, A Alastuey, N Bukowiecki, A S H Prevot, U Baltensperger, and X Querol. Sources and variability of inhalable road dust particles in three European cities. *Atmospheric Environment*, 45(37):6777–6787, 2011.

[8] Kelsy a. Anderson and John a. Downing. Dry and wet atmospheric deposition of nitrogen, phosphorus and silicon in an agricultural region. *Water, Air, and Soil Pollution*, 176(1-4):351–374, July 2006.

[9] S T Anthony, Andrew Sander, Eric Novotny, Omid Mohseni, and Heinz Stefan. Inventory of Road Salt Use in the Minneapolis / St . Paul Metropolitan Area by. (503), 2007.

[10] Eric Apeagyei, Michael S Bank, and John D Spengler. Distribution of heavy metals in road dust along an urban-rural gradient in Massachusetts. *Atmospheric Environment*, 45(13):2310–2323, 2011.

[11] Mark a.S. Laidlaw, Sammy Zahran, Howard W. Mielke, Mark P. Taylor, and Gabriel M. Filippelli. Re-suspension of lead contaminated urban soil as a dominant source of atmospheric lead in Birmingham, Chicago, Detroit and Pittsburgh, USA. *Atmospheric Environment*, 49:302–310, March 2012.

[12] Lowell L Ashbaugh, William C Malm, and Willy Z Sadeh. A residence time probability analysis of sulfur concentrations at grand Canyon National Park. *Atmospheric Environment (1967)*, 19(8):1263–1270, 1985.

[13] Bilkis a. Begum. Identification of Sources of Fine and Coarse Particulate Matter in Dhaka, Bangladesh. *Aerosol and Air Quality Research*, (2004):345–353, August 2010.

[14] Bilkis a. Begum, Swapan K. Biswas, and Philip K. Hopke. Key issues in controlling air pollutants in Dhaka, Bangladesh. *Atmospheric Environment*, 45(40):7705–7713, October 2010.

[15] C.a. Belis, F. Karagulian, B.R. Larsen, and P.K. Hopke. Critical review and meta-analysis of ambient particulate matter source apportionment using receptor models in Europe. *Atmospheric Environment*, 69:94–108, April 2013.

[16] S. G. Brown, T. Lee, G. a. Norris, P. T. Roberts, P. Paatero, and D. R. Worsnop. Receptor modeling of near-roadway aerosol mass spectrometer data in Las Vegas, Nevada, with EPA PMF. *Atmospheric Chemistry and Physics*, 12(1):309–325, January 2012.

[17] B Brunekreef and B Forsberg. Epidemiological evidence of effects of coarse airborne particles on health. *The European respiratory journal : official journal of the European Society for Clinical Respiratory Physiology*, 26(2):309–18, August 2005.

[18] N. Bukowiecki, P. Lienemann, M. Hill, M. Furger, a. Richard, F. Amato, a.S.H. Prévôt, U. Baltensperger, B. Buchmann, and R. Gehrig. PM10 emission factors for non-exhaust particles generated by road traffic in an urban street canyon and along a freeway in Switzerland. *Atmospheric Environment*, 44(19):2330–2340, June 2010.

[19] Nicolas Bukowiecki, Peter Lienemann, Matthias Hill, Renato Figi, Agnes Richard, Markus Furger, Karen Rickers, Gerald Falkenberg, Yongjing Zhao, Steven S Cliff, Andre S H Prevot, Urs Baltensperger, Brigitte Buchmann, and Robert Gehrig. Real-world emission factors for antimony and other brake wear related trace elements: size-segregated values for light and heavy duty vehicles. *Environmental science & technology*, 43(21):8072–8, November 2009.

[20] W.F. Cannon and John D. Horton. Soil geochemical signature of urbanization and industrialization Chicago, Illinois, USA. *Applied Geochemistry*, 24(8):1590–1601, August 2009.

[21] Abel O.M. Carvalho and Maria Do Carmo Freitas. Sources of trace elements in fine and coarse particulate matter in a sub-urban and industrial area of the Western European Coast. *Procedia Environmental Sciences*, 4(3):184–191, January 2011.

[22] Yiu-Chung Chan, David D Cohen, Olga Hawas, Eduard Stelcer, Rod Simpson, Lyn Denison, Neil Wong, Mary Hodge, Eva Comino, and Stewart Carswell. Apportionment of sources of fine and coarse particles in four major Australian cities by positive matrix factorisation. *Atmospheric Environment*, 42(2):374–389, 2008.

[23] Aurelie Charron and Roy M. Harrison. Fine (PM2.5) and Coarse (PM2.5-10) Particulate Matter on A Heavily Trafficked London Highway : Sources and Processes. 39(20):7768–7776, 2005.

[24] Fu-Lin Chen, Robert Vanderpool, Ronald Williams, Fred Dimmick, Brett D Grover, Russell Long, and Robert Murdoch. Field evaluation of portable and central site PM samplers emphasizing additive and differential mass concentration estimates. *Atmospheric Environment*, 45(26):4522–4527, 2011.

[25] Kalam Cheung, Nancy Daher, Winnie Kam, Martin M Shafer, Zhi Ning, James J Schauer, and Constantinos Sioutas. Spatial and temporal variation of chemical composition and mass closure of ambient coarse particulate matter (PM102.5) in the Los Angeles area. *Atmospheric Environment*, 45(16):2651–2662, May 2011.

[26] Ming-Tung Chuang, Yang Zhang, and Daiwen Kang. Application of WRF/Chem-MADRID for real-time air quality forecasting over the Southeastern United States. *Atmospheric Environment*, 45(34):6241–6250, November 2011.

[27] Nicholas Clements, Ricardo Piedrahita, John Ortega, Jennifer L. Peel, Michael Hannigan, Shelly L. Miller, and Jana B. Milford. Characterization and Non-parametric Regression of Rural and Urban Coarse Particulate Matter Mass Concentrations in Northeastern Colorado. *Aerosol Science and Technology*, 46(1):108–123, January 2012.

[28] Sara Comero, Luisa Capitani, and Bernd Manfred Gawlik. Positive Matrix Factorisation (PMF) environmental monitoring data using PMF. *JRC Scientific and Technical Reports*, 2009.

[29] Coordinating Research Council. CRC Report No . E-68a REVIEW OF THE 2009 DRAFT MOTOR VEHICLE EMISSIONS SIMULATOR (MOVES) MODEL November 2010. Technical Report November, 2010.

[30] Thomas C. Coulter. CMB8.2 Users Manual. *EPA*, EPA-452/R-, 2004.

[31] Nancy Daher, Ario Ruprecht, Giovanni Invernizzi, Cinzia De Marco, Justin Miller-Schulze, Jong Bae Heo, Martin M. Shafer, Brandon R. Shelton, James J. Schauer, and Constantinos Sioutas. Characterization, sources and redox activity of fine and coarse particulate matter in Milan, Italy. *Atmospheric Environment*, 49:130–141, March 2012.

[32] Douglas W Dockery, C Arden Pope, Xiping Xu, John D Spengler, James H Ware, Martha E Fay, Benjamin G Ferris, and Frank E Speizer. An Association Between Air Pollution and Mortality in Six U.S. Cities. *The New England journal of medicine*, 329(24):1753–1759, 1993.

[33] Marloes Eeftens, Rob Beelen, Kees de Hoogh, Tom Bellander, Giulia Cesaroni, Marta Cirach, Christophe Declercq, Audrius Ddel, Evi Dons, Audrey de Nazelle, Konstantina Dimakopoulou, Kirsten Eriksen, Grégoire Falq, Paul Fischer, Claudia Galassi, Regina Gražulevičien, Joachim Heinrich, Barbara Hoffmann, Michael Jerrett, Dirk Keidel, Michal Korek, Timo Lanki, Sarah Lindley, Christian Madsen, Anna Mölter, Gizella Nádor, Mark Nieuwenhuijsen, Michael Nonnemacher, Xanthi Pedeli, Ole Raaschou-Nielsen, Evridiki Patelarou, Ulrich Quass, Andrea Ranzi, Christian Schindler, Morgane Stempfelet, Euripides Stephanou, Dorothea Sugiri, Ming-Yi Tsai, Tarja Yli-Tuomi, Mihály J Varró, Danielle Vienneau, Stephanie Von Klot, Kathrin Wolf, Bert Brunekreef, and Gerard Hoek. Development of Land Use Regression models for PM(2.5), PM(2.5) absorbance, PM(10) and PM(coarse) in 20 European study areas; results of the ESCAPE project. *Environmental science & technology*, 46(20):11195–205, October 2012.

[34] Marloes Eeftens, Ming-Yi Tsai, Christophe Ampe, Bernhard Anwander, Rob Beelen, Tom Bellander, Giulia Cesaroni, Marta Cirach, Josef Cyrys, Kees de Hoogh, Audrey De Nazelle, Frank de Vocht, Christophe Declercq, Audrius Ddel, Kirsten Eriksen, Claudia Galassi, Regina Gražulevičien, Georgios Grivas, Joachim Heinrich, Barbara Hoffmann, Minas Iakovides, Alex Ineichen, Klea Katsouyanni, Michal Korek, Ursula Krämer, Thomas Kuhlbusch, Timo Lanki, Christian Madsen, Kees Meliefste, Anna Mölter, Gioia Mosler, Mark Nieuwenhuijsen, Marieke Oldenwening, Arto Pennanen, Nicole Probst-Hensch, Ulrich Quass, Ole Raaschou-Nielsen, Andrea Ranzi, Euripides Stephanou, Dorothee Sugiri, Orsolya Udvardy, Éva Vaskövi, Gudrun Weinmayr, Bert Brunekreef, and Gerard Hoek. Spatial variation of PM2.5, PM10, PM2.5 absorbance and PMcoarse concentrations between and within 20 European study areas and the relationship with NO2 Results of the ESCAPE project. *Atmospheric Environment*, 62:303–317, December 2012.

[35] Mohanad El-harbawi. Air quality modelling , simulation , and computational methods :. 179(November 2012):149–179, 2013.

[36] EPA. SPECIATE Data Browser, 1989.

[37] U S Epa, Office Transportation, Air Quality, and Standards Division. Regulatory Impact Analysis: Control of Emissions of Air Pollution from Category 3 Marine Diesel Engines (EPA-420-R-09-019).

[38] Neil H Frank. Retained nitrate, hydrated sulfates, and carbonaceous mass in federal reference method fine particulate matter for six eastern U.S. cities. *Journal of the Air & Waste Management Association (1995)*, 56(4):500–11, April 2006.

[39] Bhagwan D. Garg, Steven H. Cadle, Patricia a. Mulawa, Peter J. Groblicki, Chris Laroo, and Graham a. Parr. Brake Wear Particulate Matter Emissions. *Environmental Science & Technology*, 34(21):4463–4469, November 2000.

[40] Kristi a. Gebhart, Bret a. Schichtel, William C. Malm, Michael G. Barna, Marco a. Rodriguez, and Jeffrey L. Collett. Back-trajectory-based source apportionment of airborne sulfur and nitrogen concentrations at Rocky Mountain National Park, Colorado, USA. *Atmospheric Environment*, 45(3):621–633, January 2011.

[41] Johanna K Gietl and Otto Klemm. Source Identification of Size-Segregated Aerosol in Münster, Germany, by Factor Analysis. *Aerosol Science and Technology*, 43(8):828–837, 2009.

[42] Johanna K. Gietl, Roy Lawrence, Alistair J. Thorpe, and Roy M. Harrison. Identification of brake wear particles and derivation of a quantitative tracer for brake dust at a major road. *Atmospheric Environment*, 44(2):141–146, January 2010.

[43] Maria Luiza D P Godoy, José Marcus Godoy, Luiz Alfredo Roldão, Daniela S Soluri, and Raquel a. Donagemma. Coarse and fine aerosol source apportionment in Rio de Janeiro, Brazil. *Atmospheric Environment*, 43(14):2366–2374, 2009.

[44] Andrew P. Grieshop, Eric M. Lipsky, Natalie J. Pekney, Satoshi Takahama, and Allen L. Robinson. Fine particle emission factors from vehicles in a highway tunnel: Effects of fleet composition and season. *Atmospheric Environment*, 40:287–298, 2006.

[45] J. S. Han, K. J. Moon, S. J. Lee, Y. J. Kim, S. Y. Ryu, S. S. Cliff, and S. M. Yi. Size-resolved source apportionment of ambient particles by positive matrix factorization at Gosan background site in East Asia. *Atmospheric Chemistry and Physics*, 6(1):211–223, January 2006.

[46] Sehyun Han, Jong-Sang Youn, and Yong-Won Jung. Characterization of PM10 and PM2.5 source profiles for resuspended road dust collected using mobile sampling methodology. *Atmospheric Environment*, 45(20):3343–3351, 2011.

[47] Harry Horace Harman. *Modern factor analysis*. University of Chicago Press, Chicago, 1967.

[48] Roy M Harrison, Alan M Jones, Johanna Gietl, Jianxin Yin, and David C Green. Estimation of the contributions of brake dust, tire wear, and resuspension to nonexhaust traffic particles derived from atmospheric measurements. *Environmental science & technology*, 46(12):6523–9, June 2012.

[49] Sue Ellen Haupt, George S. Young, and Christopher T. Allen. Validation of a ReceptorDispersion Model Coupled with a Genetic Algorithm Using Synthetic Data. *Journal of Applied Meteorology and Climatology*, 45(3):476–490, March 2006.

[50] P K Hopke, M Cheng, C Li, and Y Xie. Possible sources and preferred pathways for biogenic and non-sea-salt sulfur for the high Arctic. *Journal of Geophysical Research Geophysical Research*, 100(95), 1995.

[51] Injo Hwang, Philip K Hopke, and Joseph P Pinto. Source apportionment and spatial distributions of coarse particles during the Regional Air Pollution Study. *Environmental science & technology*, 42(10):3524–30, May 2008.

[52] Akihiro Iijima, Keiichi Sato, Kiyoko Yano, Masahiko Kato, Kunihisa Kozawa, and Naoki Furuta. Emission factor for antimony in brake abrasion dusts as one of the major atmospheric antimony sources. *Environmental science & technology*, 42(8):2937–42, April 2008.

[53] IPNI. A Nutrient Use Geographic Information System (NuGIS) for the U.S. (Final Version), 2012.

[54] Dan Jaffe, William Hafner, Duli Chand, Anthony Westerling, and Dominick Spracklen. Interannual variations in PM2.5 due to wildfires in the Western United States. *Environmental science & technology*, 42(8):2812–8, April 2008.

[55] Christer Johansson, Michael Norman, and Lars Burman. Road traffic emission factors for heavy metals. *Atmospheric Environment*, 43(31):4681–4688, October 2009.

[56] Joel D Kaufman, Sara D Adar, Ryan W Allen, R Graham Barr, Matthew J Budoff, Gregory L Burke, Adrian M Casillas, Martin a Cohen, Cynthia L Curl, Martha L Daviglus, Ana V Diez Roux, David R Jacobs, Richard a Kronmal, Timothy V Larson, Sally Lee-Jane Liu, Thomas Lumley, Ana Navas-Acien, Daniel H O'Leary, Jerome I Rotter, Paul D Sampson, Lianne Sheppard, David S Siscovick, James H Stein, Adam a Szpiro, and Russell P Tracy. Prospective study of particulate air pollution exposures, subclinical atherosclerosis, and clinical cardiovascular disease: The Multi-Ethnic Study of Atherosclerosis and Air Pollution (MESA Air). *American journal of epidemiology*, 176(9):825–37, November 2012.

[57] Paul Kennedy and Jennifer Gadd. Preliminary Examination of Trace Elements in Tyres, Brake Pads and Road Bitumen in New Zealand. Technical report, New Zealand Ministry of Transport, 2000.

[58] Zs. Kertész, Z. Szoboszlai, a. Angyal, E. Dobos, and I. Borbély-Kiss. Identification and characterization of fine and coarse particulate matter sources in a middle-European urban environment. *Nuclear Instruments and Methods in Physics Research Section B: Beam Interactions with Materials and Atoms*, 268(11-12):1924–1928, June 2010.

[59] Anis Khlaifi, Anda Ionescu, and Yves Candau. Pollution source identification using a coupled diffusion model with a genetic algorithm. *Mathematics and Computers in Simulation*, 79(12):3500–3510, August 2009.

[60] Morton Lippmann and Lung-Chi Chen. *Health effects of concentrated ambient air particulate matter (CAPs) and its components.*, volume 39. January 2009.

[61] Marissa L Maier, Sivaraman Balachandran, Stefanie E Sarnat, Jay R Turner, James a Mulholland, and Armistead G Russell. Application of an ensemble-trained source apportionment approach at a site impacted by multiple point sources. *Environmental science & technology*, 47(8):3743–51, April 2013.

[62] E Manoli, D Voutsa, and C Samara. Chemical characterization and source identification/apportionment of fine and coarse air particles in Thessaloniki, Greece. *Atmospheric Environment*, 36(6):949–961, 2002.

[63] Amit Marmur, Sun-Kyoung Park, James a. Mulholland, Paige E. Tolbert, and Armistead G. Russell. Source apportionment of PM2.5 in the southeastern United States using receptor and emissions-based models: Conceptual differences and implications for time-series health studies. *Atmospheric Environment*, 40(14):2533–2551, May 2006.

[64] Federico Mazzei, Franco Lucarelli, Silvia Nava, Paolo Prati, Gianluigi Valli, and Roberta Vecchi. A new methodological approach: The combined use of two-stage streaker samplers and optical particle counters for the characterization of airborne particulate matter. *Atmospheric Environment*, 41(26):5525–5535, 2007.

[65] Robin McCulloch. Montana Mines and Exploration - 2012, 2012.

[66] S. McKeen, S. H. Chung, J. Wilczak, G. Grell, I. Djalalova, S. Peckham, W. Gong, V. Bouchet, R. Moffet, Y. Tang, G. R. Carmichael, R. Mathur, and S. Yu. Evaluation of several PM 2.5 forecast models using data collected during the ICARTT/NEAQS 2004 field study. *Journal of Geophysical Research*, 112(D10):D10S20, March 2007.

[67] S. McKeen, G. Grell, S. Peckham, J. Wilczak, I. Djalalova, E.-Y. Hsie, G. Frost, J. Peischl, J. Schwarz, R. Spackman, J. Holloway, J. de Gouw, C. Warneke, W. Gong, V. Bouchet, S. Gaudreault, J. Racine, J. McHenry, J. McQueen, P. Lee, Y. Tang, G. R. Carmichael, and R. Mathur. An evaluation of real-time air quality forecasts and their urban emissions over eastern Texas during

the summer of 2006 Second Texas Air Quality Study field study. *Journal of Geophysical Research*, 114:D00F11, June 2009.

[68] M S Miller, S K Friedlander, and G M Hidy. A Chemical Element Balance for the Pasadena Aerosol. *Journal of Colloid and Interface Science*, 39(1):165–176, 1972.

[69] Katharine F Moore, Vishal Verma, María Cruz Minguillón, and Constantinos Sioutas. Inter- and Intra-Community Variability in Continuous Coarse Particulate Matter (PM 10-2.5) Concentrations in the Los Angeles Area. *Aerosol Science and Technology*, 44(7):526–540, 2010.

[70] Leigh a. Munchak, Bret a. Schichtel, Amy P. Sullivan, Amanda S. Holden, Sonia M. Kreidenweis, William C. Malm, and Jeffrey L. Collett. Development of wildland fire particulate smoke marker to organic carbon emission ratios for the conterminous United States. *Atmospheric Environment*, 45(2):395–403, January 2011.

[71] Gary Norris and Ram Vedantham. EPA Positive Matrix Factorization (PMF) 3.0 Fundamentals & User Guide. 2008.

[72] Gary Norris, Ram Vedantham, Katie Wade, Patrick Zahn, Steve Brown, Shelly Eberly, and Chuck Foley. Guidance Document for PMF Applications with the Multilinear Engine. *EPA*, 2009.

[73] Mi-Seok Oh. Quantitative Source Apportionment of Size-segregated Particulate Matter at Urbanized Local Site in Korea. *Aerosol and Air Quality Research*, pages 247–264, 2011.

[74] César Oliveira, Casimiro Pio, Alexandre Caseiro, Patrícia Santos, Teresa Nunes, Hongjun Mao, Lakhumal Luahana, and Ranjeet Sokhi. Road traffic impact on urban atmospheric aerosol loading at Oporto, Portugal. *Atmospheric Environment*, 44(26):3147–3158, August 2010.

[75] Pentti Paatero. The Multilinear Engine: A Table-Driven, Least Squares Program for Solving Multilinear Problems, including the n-Way Parallel Factor Analysis Model. *Journal of Computational and Graphical Statistics*, 8(4):pp. 854–888, 1999.

[76] Pentti Paatero. The Multilinear Engine (ME-2) script language (v. 1.313). 2009.

[77] Pentti Paatero and Philip K. Hopke. Discarding or downweighting high-noise variables in factor analytic models. *Analytica Chimica Acta*, 490(1-2):277–289, August 2003.

[78] Payam Pakbin, Zhi Ning, Martin M Shafer, James J Schauer, and Constantinos Sioutas. Seasonal and Spatial Coarse Particle Elemental Concentrations in the Los Angeles Area. *Aerosol Science and Technology*, 45(8):949–963, August 2011.

[79] Pallavi Pant and Roy M. Harrison. Critical review of receptor modelling for particulate matter: A case study of India. *Atmospheric Environment*, 49:1–12, March 2012.

[80] Pallavi Pant and Roy M. Harrison. Estimation of the contribution of road traffic emissions to particulate matter concentrations from field measurements: A review. *Atmospheric Environment*, 77:78–97, October 2013.

[81] Rajendra D Paode, Usama M Shahin, Jakkris Sivadechathep, Thomas M Holsen, and William J Franek. Source Apportionment of Dry Deposited and Airborne Coarse Particles Collected in the Chicago Area. *Aerosol Science and Technology*, 31(6):473–486, 1999.

[82] Rokjin J. Park, Daniel J. Jacob, and Jennifer a. Logan. Fire and biofuel contributions to annual mean aerosol mass concentrations in the United States. *Atmospheric Environment*, 41(35):7389–7400, November 2007.

[83] Richard E Peltier, Kevin R Cromar, Yingjun Ma, Zhi-Hua Tina Fan, and Morton Lippmann. Spatial and seasonal distribution of aerosol chemical components in New York City: (2) road dust and other tracers of traffic-generated air pollution. *Journal of exposure science & environmental epidemiology*, 21(5):484–94, September 2011.

[84] Alexandr V Polissar, Philip K Hopke, Pentti Paatero, William C Malm, and James F Sisler. Atmospheric aerosol over Alaska 2 . Elemental composition and sources. *Journal of Geophysical Research*, 103(98), 1998.

[85] C A Pope, R T Burnett, M J Thun, E Calle, D Krewski, and G D Thurston. Lung Cancer, Cardiolpulmonary Mortality, and Long-term Exposure to Fine Particulate Air Pollution. *Journal of the American Medical Association*, 287(9):1132–1141, 2002.

[86] C Arden Pope and Douglas W Dockery. Health Effects of Fine Particulate Air Pollution : Lines that Connect. 56(6):709–742, 2006.

[87] Robin C Puett, Jaime E Hart, Jeff D Yanosky, Christopher Paciorek, Joel Schwartz, Helen Suh, Frank E Speizer, and Francine Laden. Chronic fine and coarse particulate exposure, mortality, and coronary heart disease in the Nurses' Health Study. *Environmental health perspectives*, 117(11):1697–701, November 2009.

[88] Cristina Reche, Teresa Moreno, Fulvio Amato, Mar Viana, Barend L van Drooge, Hsiao-Chi Chuang, Kelly Bérubé, Tim Jones, Andrés Alastuey, and Xavier Querol. A multidisciplinary approach to characterise exposure risk and toxicological effects of PM and PM. samples in urban environments. *Ecotoxicology and environmental safety*, 78:327–35, April 2012.

[89] Adam Reff, Shelly I Eberly, and Prakash V Bhave. Receptor modeling of ambient particulate matter data using positive matrix factorization: review of existing methods. *Journal of the Air & Waste Management Association (1995)*, 57(2):146–54, February 2007.

[90] J. S. Reid, R. Koppmann, T. F. Eck, and D. P. Eleuterio. A review of biomass burning emissions part II: intensive physical properties of biomass burning particles. *Atmospheric Chemistry and Physics*, 5(3):799–825, March 2005.

[91] R a Root. Lead loading of urban streets by motor vehicle wheel weights. *Environmental health perspectives*, 108(10):937–40, October 2000.

[92] J J Schauer, G C Lough, and M M Shafer. Characterization of metals emitted from motor vehicles. *Health Effects Institute*, 133(133), 2006.

[93] James J Schauer, N Park St, and Madison Wi. Characterization of Metals Emitted from Motor Vehicles Appendix D . Source Profiles : Tire Wear , Brake Housing Dust , and Resuspended Road Dust. *Environmental Chemistry*, 2006.

[94] B.a. Schichtel, M.a. Rodriguez, M.G. Barna, K.a. Gebhart, M.L. Pitchford, and W.C. Malm. A semi-empirical, receptor-oriented Lagrangian model for simulating fine particulate carbon at rural sites. *Atmospheric Environment*, 61:361–370, December 2012.

[95] Bret A Schichtel and Rudolf B Husar. Regional Simulation of Atmospheric Pollutants with the CAPITA Monte Carlo Model. *Journal of the Air and Waste Management Association*, 47(3):301–333, 1997.

[96] Bret a. Schichtel and Rudolf B. Husar. Regional Simulation of Atmospheric Pollutants with the CAPITA Monte Carlo Model. *Journal of the Air & Waste Management Association*, 47(3):301–333, March 1997.

[97] Bret a. Schichtel, William C. Malm, Kristi a. Gebhart, Michael G. Barna, and Eladio M. Knipping. A hybrid source apportionment model integrating measured data and air quality model results. *Journal of Geophysical Research*, 111(D7):1–19, 2006.

[98] Bret A. Schichtel, William G. Malm, Jeffery L. Collett, Amy P. Sullivan, Amanada S. Holden, Leigh A. Patterson, Marco A. Rodriguez, and Michael G. Barna. Estimating the Contribution of Smoke to Fine Particulate Matter using a Hybrid-Receptor Model. *Air and Waste Management Aerosol & Atmospheric Optics: Visual Air Quality and Radiation*, 2008.

[99] Christian Seigneur, Prasad Pai, Philip K. Hopke, and Daniel Grosjean. Modeling Atmospheric Particulate Matter. *Environmental Science and Technology*, 33:80A–86A, 1999.

[100] Hansford T Shacklette and Josehpine G Boerngen. Element Concentrations in Soils and Other Surfici 1 Materials of the Conterminous United Stat Element Concentrations in Soils and Other Surficial Materials of the Conterminous United States. Technical report, U.S. Geological Survey Professional Paper 1270, Washington, D.C., 1984.

[101] John Sternbeck. Metal emissions from road traffic and the influence of resuspension results from two tunnel studies. *Atmospheric Environment*, 36:4735–4744, 2002.

[102] Elizabeth Stone, James Schauer, Tauseef a. Quraishi, and Abid Mahmood. Chemical characterization and source apportionment of fine and coarse particulate matter in Lahore, Pakistan. *Atmospheric Environment*, 44(8):1062–1070, March 2010.

[103] Maciej Strak, Maaike Steenhof, Krystal J. Godri, Ilse Gosens, Ian S. Mudway, Flemming R. Cassee, Erik Lebret, Bert Brunekreef, Frank J. Kelly, Roy M. Harrison, Gerard Hoek, and Nicole a.H. Janssen. Variation in characteristics of ambient particulate matter at eight locations in the Netherlands The RAPTES project. *Atmospheric Environment*, 45(26):4442–4453, August 2011.

[104] C. a. Stroud, M. D. Moran, P. a. Makar, S. Gong, W. Gong, J. Zhang, J. G. Slowik, J. P. D. Abbatt, G. Lu, J. R. Brook, C. Mihele, Q. Li, D. Sills, K. B. Strawbridge, M. L. McGuire, and G. J. Evans. Evaluation of chemical transport model predictions of primary organic aerosol for air masses classified by particle component-based factor analysis. *Atmospheric Chemistry and Physics*, 12(18):8297–8321, September 2012.

[105] Clyde W Sweet, Stephen J Vermette, and Sheldon Landsberger. Sources of Toxic Trace Elements in Urban Air in Illinois. 27(14):2502–2510, 1993.

[106] Lokman Hakan Tecer, Gürdal Tuncel, Ferhat Karaca, Omar Alagha, Pnar Süren, Abdullah Zararsz, and Rdvan Krmaz. Metallic composition and source apportionment of fine and coarse particles using positive matrix factorization in the southern Black Sea atmosphere. *Atmospheric Research*, 118:153–169, November 2012.

[107] Jonathan Thornburg, Charles E. Rodes, Phillip a. Lawless, and Ron Williams. Spatial and temporal variability of outdoor coarse particulate matter mass concentrations measured with a new coarse particle sampler during the Detroit Exposure and Aerosol Research Study. *Atmospheric Environment*, 43(28):4251–4258, September 2009.

[108] Alistair Thorpe and Roy M Harrison. Sources and properties of non-exhaust particulate matter from road traffic: a review. *The Science of the total environment*, 400(1-3):270–282, 2008.

[109] U.S. EPA. Integrated Science Assessment for Particulate Matter (Final Report) EPA/600/R-08/139F. Technical report, Washington, D.C., 2009.

[110] USDA. Distribution Maps of Dominant Soil Orders NRCS Soils.

[111] Stephen J Vermette, Clyde W Sweet, and Sheldon Landsberger. Airborne fine particulate matter (PM-10) in southeast Chicago : preliminary report II. Champaign, IL : Illinois State Water Survey , viii, 56 p. : ill. ; 28 cm. Illinois State Water Survey. Contract Report ; 481. *Energy*, (September), 1988.

[112] M. Viana, T.a.J. Kuhlbusch, X. Querol, a. Alastuey, R.M. Harrison, P.K. Hopke, W. Winiwarter, M. Vallius, S. Szidat, a.S.H. Prévôt, C. Hueglin, H. Bloemen, P. Wå hlin, R. Vecchi, a.I. Miranda, a. Kasper-Giebl, W. Maenhaut, and R. Hitzenberger. Source apportionment of particulate matter in Europe: A review of methods and results. *Journal of Aerosol Science*, 39(10):827–849, October 2008.

[113] Ole von Uexküll, Staffan Skerfving, Reed Doyle, and Michael Braungart. Antimony in brake pads-a carcinogenic component? *Journal of Cleaner Production*, 13(1):19–31, January 2005.

[114] S Waheed, MZ Jaafar, N Siddique, A Markwitz, and RG Brereton. PIXE analysis fo PM2.5 and PM2.5-10 for air quality assessment of Islamabad, Pakistan:

Application of chemometrics for source identification. *Journal of Environmental Science and Health Part A*, 47(13):2016–2027, 2012.

[115] P Wahlin, R Berkowicz, and F Palmgren. Characterisation of traffic-generated particulate matter in Copenhagen. *Atmospheric Environment*, 40(12):2151–2159, April 2006.

[116] Jens Wahlström, Lars Olander, and Ulf Olofsson. Size, Shape, and Elemental Composition of Airborne Wear Particles from Disc Brake Materials. *Tribology Letters*, 38(1):15–24, December 2009.

[117] Haobo Wang and David Shooter. Source apportionment of fine and coarse atmospheric particles in Auckland, New Zealand. *The Science of the total environment*, 340(1-3):189–98, March 2005.

[118] John G. Watson. Overview of Receptor Model Principles. *Journal of the Air Pollution Control Association*, 34(6):619–623, June 1984.

[119] John G. Watson. Visibility: Science and Regulation. *Journal of the Air & Waste Management Association*, 52(6):628–713, June 2002.

[120] John G Watson, Judith C Chow, L Antony Chen, and David W Dubois. Receptor Modeling for the Minnesota Particulate Matter 2.5 (PM2.5) Source Apportionment Study. Technical report, Desert Research Institute, 2008.

[121] Jennifer Weuve, Robin C Puett, Joel Schwartz, Jeff D Yanosky, Francine Laden, and Francine Grodstein. Exposure to particulate air pollution and cognitive decline in older women. *Archives of internal medicine*, 172(3):219–27, February 2012.

[122] Daniel S. Wilks. *Statistical Method in the Atmospheric Sciences*. Elsevier, 2006.

[123] Jeff D Yanosky, Christopher J Paciorek, and Helen H Suh. Predicting chronic fine and coarse particulate exposures using spatiotemporal models for the Northeastern and Midwestern United States. *Environmental health perspectives*, 117(4):522–9, April 2009.

[124] Shlomo Bekhor Yuval and David M. Broday. Data-driven nonlinear optimisation of a simple air pollution dispersion model generating high resolution spatiotemporal exposure. *Atmospheric Environment*, 79:261–270, November 2013.

[125] Yang Zhang, Marc Bocquet, Vivien Mallet, Christian Seigneur, and Alexander Baklanov. Real-time air quality forecasting, part I: History, techniques, and current status. *Atmospheric Environment*, 60:632–655, December 2012.

[126] Yang Zhang, Marc Bocquet, Vivien Mallet, Christian Seigneur, and Alexander Baklanov. Real-time air quality forecasting, part II: State of the science, current research needs, and future prospects. *Atmospheric Environment*, 60:656–676, December 2012.

Appendix A

HYBRID CODE: BUILDING MODEL INPUTS

```
# -*- coding: utf-8 -*-
"""
Created on Thu Jun 20 10:40:49 2013

@author: tmsturtz
"""
#########################################################
### Global Values used for preprocessing raw data ###
#########################################################
me2Lib    = "./ME2-Files/ME2libr.txt"
me2Key    = "./ME2-Files/me2key.key"
me2exe    = "../1-DataPreprocess/me2wG17.exe"
views     = "../1-DataPreprocess/MONT-RawData_2006-08.txt" # Comma separated VIEWS download file.
                                                    # Must have Values, MDL, and Uncertainties.
model     = "../1-DataPreprocess/SrcCont_Primary_Sec_IMP_2006-08_Truth.csv" # CTM Modeled Results
modelunc  = "../1-DataPreprocess/SrcCont_Uncert.csv"                # CTM Uncertainties
Constraint = ['FireTot','Bio_SOC']                          # Sources to constrain
Ccode     = ['FIT*','BIO*']                                 # Associated Source Codes
SecondSrcs = ['FireTot','Bio_SOC']                          # Secondary/Profile constrained Sources
secondary = {'FireTot':["OC1","OC2","OC3","OC4","EC1","EC2","EC3","OP","H","K","SO4","NO3"], #
     Allowed species in constrained profiles
           'Bio_SOC' :["OC1","OC2","OC3","OC4","EC1","EC2","EC3","OP","H"]}

#########################################################
### Global value used for developing hybrid model ###
#########################################################
np    = 4  # Number of sources to Solve
nt    = 40 # Number of Tasks
XV10  = 1 # 10-Fold Cross Validation? 0 or 1
knob  = ['PMF','CPMF', 0.00001, 0.0001, 0.001, 0.01, 0.05, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8,
     0.9, 1, 2, 5] # Dials to test
norm  = ["EC1","EC2","OC1","OC2","OC3","OC4"] # Species to normalize over
Clow  = 0    # Low value for multiplicative scaling
```

```
Chigh = 200 # High value for multiplicative scaling

Concentrations = './Data/ME2-Concentrations.csv'
Uncertainties = './Data/ME2-Uncertainties.csv'
PriorContribs = './Data/ME2-CTM-Contributions.csv'
PriorUncerts  = './Data/ME2-CTM-Uncertainties.csv'
SecondSpecies = './Data/ME2-CTM-Species.csv'

############ END USER INPUT ########################


def LoadMeas(n):
    ########################################
    ### Read-in Data Exported from VIEWS ###
    ########################################
    print ' ---'
    print ' PROCESSING VIEWS DATA'

    # Determine where the data starts and the number of columns
    lookup = 'Dataset,SiteCode'
    with open(n,'r') as myFile:
        for num, line in enumerate(myFile, 1):
            if lookup in line:
                FileStart = num
                ncols = len(line.rstrip().split(','))

    with open(n,'r') as myFile:
        for num, line in enumerate(myFile, 1):
            if num == FileStart+1:
                print" ---"
                print " VIEWS Site Name:",line.rstrip().split(',')[1]
                site = line.rstrip().split(',')[1].replace("1","")


    # Load data into memory
    views = genfromtxt(n,delimiter=',',
                       skiprows=FileStart-1,usecols=range(2,ncols),
                       converters = {2: datestr2num},names=True)


    # Break out Values, Uncertainties, MDLs, Species, and Dates
    Val_key = [line for line in views.dtype.names if "Value" in line]
```

```
Val_dat = views[Val_key].view(float).reshape(views[Val_key].shape + (-1,))

Unc_key = [line for line in views.dtype.names if "Unc" in line]
Unc_dat = views[Unc_key].view(float).reshape(views[Unc_key].shape + (-1,))

MDL_key = [line for line in views.dtype.names if "MDL" in line]
MDL_dat = views[MDL_key].view(float).reshape(views[MDL_key].shape + (-1,))

Species_dat = array([z.replace('Value','') for z in [w.replace('fValue','') for w in Val_key]])
Dates  = array([i[0] for i in views])

print " --- "
print " All Species Available in Raw Data:"
print Species_dat



###############################################
### Process VIEWS Data for use with ME-2 ###
###############################################

# Get indices where all data is missing or species which contain >50% BDL
MissList      = where((Val_dat == -999).sum(axis=0) > 0.5*Val_dat.shape[0])[0]
MDLList       = where((Val_dat <= MDL_dat).sum(axis=0) >= 0.5*Val_dat.shape[0])[0]
NoUncMDL      = where(average(MDL_dat,axis=0) == -999)[0]
NotRCFM       = where(Val_dat[:,where(Species_dat == 'RCFM')[0]] == -999)[0]
AdditionalDrop = where((Species_dat == 'CM_calculated',
                     Species_dat == 'RCFM',
                     Species_dat == 'SOIL',
                     Species_dat == 'fabs',
                     Species_dat == 'MT',
                     Species_dat == 'MF',
                     Species_dat == 'SeaSalt',
                     Species_dat == 'S',
                     Species_dat == 'RB',
                     Species_dat == 'V'))[1]
Exceptions    = where((Species_dat == 'OC2', Species_dat == 'OC1'))[1]
print " ---"
print " Exceptions to Processing Criteria as defined by user:", Species_dat[Exceptions[0]]

Removals = unique(concatenate((MissList,MDLList,NoUncMDL,AdditionalDrop)))
for i in range(len(Exceptions)):
```

```
    Removals = delete(Removals,where(Removals == Exceptions[i]))
Keep    = delete(array(range(len(Species_dat))),Removals)
Samples = delete(range(len(Val_dat)),NotRCFM)


# Exclude problematic species from Values, Uncertainties, and MDLs
Val     = Val_dat[:,Keep]
Unc     = Unc_dat[:,Keep]
MDL     = MDL_dat[:,Keep]
Val     = Val[Samples,:]
Unc     = Unc[Samples,:]
MDL     = MDL[Samples,:]
Species = Species_dat[Keep]
Dates   = Dates[Samples]

print " --- "
print " Species Kept:"
print Species
print " --- "
print " Samples:",Val.shape[0]
print " Species:",Val.shape[1]

### Calculate the geometric and arithmetic means of columns where Val > MDL != -999
from scipy.stats.mstats import gmean
arimeans = array([average(Val[(Val[:,i] > MDL[:,i]) & (Val[:,i] != -999),i]) for i in
     range(len(Val[0,:]))])
geomeans = array([gmean(Val[(Val[:,i] > MDL[:,i]) & (Val[:,i] != -999),i]) for i in
     range(len(Val[0,:]))])
MDL_med = array([median(MDL[(MDL[:,i] != -999),i]) for i in range(len(MDL[0,:]))])
Unc_med = array([median(Unc[(Unc[:,i] != -999),i]) for i in range(len(Unc[0,:]))])

### Val Processing ###
# For values greater than detection limit we keep as is
# For BDL values we use MDL/2
# For missing values we use the geometric mean of the measured concentrations that are above MDL
     initially
Val[((Val < MDL) & (Val != -999))] = (MDL/2)[where(((Val < MDL) & (Val != -999)))]
Val[(Val == -999)]              = geomeans[where((Val == -999))[1]]

### Unc Processing ###
# Check MDL and Unc for -999 values and fill with median of acceptable values for each
     respectively
```

```
# For determined concentrations use the uncertainty plus 1/3 MDL
# For BDL concentrations use 1/2 arithmetic mean of MDL plus 1/3 MDL of sample
# For Missing concentrations use 4 times the geometric mean of concentration
Unc[Unc == -999]                = Unc_med[where(Unc == -999)[1]]
MDL[MDL == -999]                = MDL_med[where(MDL == -999)[1]]
Unc[((Val > MDL) & (Val != -999))] = (Unc+MDL/3)[((Val > MDL) & (Val != -999))]
Unc[((Val < MDL) & (Val != -999))] = arimeans[where(((Val < MDL) & (Val !=
     -999)))[1]]+(MDL/3)[((Val < MDL) & (Val != -999))]
Unc[(Val == -999)]              = geomeans[where((Val == -999))[1]]*4


### Signal-to-Noise Assessment (Reference?)
#       S/N >= 2 :  Keep as is
# 0.2 <= S/N < 2 :  Downweight
#       S/N < 0.2 : Remove

SN  = ((Val-Unc)**2).sum(axis=0)/(Unc**2).sum(axis=0)
SNw = Species[where((SN < 2) & (SN > 0.2))]
SNb = Species[where(SN < 0.2)]
print " ---"
print " Signal-to-Noise"
print " Weak Species:", SNw
print " Bad Species:", SNb

### Scale the weak species uncertainties by 3 and remove the bad species
Unc[where((SN < 2) & (SN >0.2))] = Unc[where((SN < 2) & (SN >0.2))]*3
delete(Unc,where(SN < 0.2),1)
delete(Val,where(SN < 0.2),1)


### Write out files and return values for processing later
import csv
if not os.path.exists('./Data'):
       os.makedirs('./Data')
with open("./Data/VIEWS-Concentration-Data.csv",'w') as fout:
    writer = csv.writer(fout,delimiter=',')
    writer.writerow(Species)
    savetxt(fout,Val,delimiter=',')

with open("./Data/VIEWS-Uncertainty-Data.csv",'w') as fout:
    writer = csv.writer(fout,delimiter=',')
    writer.writerow(Species)
```

```
        savetxt(fout,Unc,delimiter=',')

    ViewOut = dstack((Val,Unc))
    return ViewOut, Dates, Species, site



def LoadCTMModel(n,m):
    ###########################################################################
    ### Read in Model Results, match dates, determine initial uncertainty ###
    ###########################################################################
    with open(n,'r') as myFile:
        mod_dat_raw = [line.rstrip().split(',') for line in myFile if Site in line]

    with open(n,'r') as myFile:
        mod_head_raw = array([line.rstrip().split(',') for line in myFile if "Date" in line][0])

    dataLoc = where(mod_head_raw == 'Date')[0]
    mod_head = array(mod_head_raw[dataLoc+1:len(mod_head_raw)-4])
    mDates  = datestr2num(array([i[dataLoc] for i in mod_dat_raw]))
    mod_dat = array([i[dataLoc+1:len(i)-4] for i in mod_dat_raw])

    ### Match Dates with VIEWS
    date_match = intersect1d(vDates,mDates)
    mod_dates  = array([i for i in mDates if i in date_match])
    views_dates = array([i for i in vDates if i in date_match])

    model_out = mod_dat[in1d(mDates,mod_dates),:].astype(float)
    views_out = VIEWS[in1d(vDates,views_dates),:,:]
    savetxt("./Data/CTM-AllSources-Contributions-4RMSE_Calcs.csv",model_out,delimiter=',')

    print ' '
    print ' ---'
    print ' PROCESSING CTM RESULTS'
    print ' ---'
    print ' The CTM Model contains the following sources:'
    print mod_head


    ### Get model uncertainty (PMF equation method)
    # PMF Users Guide 3.0: Unc = sqrt((ErrFrac * Conc)^2 + MDL^2)
    # For JFSP Work: Uncertainty taken from synthetic work by Bret Schichtel (PMF_Unc_Eq-1.xls)
```

```
with open(m,'r') as myFile:
    mod_unc_eq = array([line.rstrip().split(',')[1:len(line)] for num, line in
        enumerate(myFile,1) if num > 1]).astype(float)

with open(m,'r') as myFile:
    mod_unc_head = array([line.rstrip().split(',')[1:len(line)] for line in myFile if "Data" in
        line][0])

constloc = array([where(i == mod_head)[0] for i in Constraint]).T[0]
uconstloc = array([where(i == mod_unc_head)[0] for i in Constraint]).T[0]
mod_cut      = model_out[:,constloc]
mod_cut_unceq = array([map(float,i) for i in mod_unc_eq[:,uconstloc]])
mod_cut_unc = ones(mod_cut.shape,dtype=float)

for a in range(mod_cut.shape[0]):
    for b in range(mod_cut.shape[1]):
        if mod_cut[a,b] <= mod_cut_unceq[0,b]:
            mod_cut_unc[a,b] = 0.83333333*mod_cut_unceq[0,b]
        else:
            mod_cut_unc[a,b] = sqrt(((mod_cut_unceq[1,b]/100)*(mod_cut[a,b]))**2 +
                mod_cut_unceq[0,b]**2)

mod_head     = mod_unc_head[uconstloc]

### Format Dates and account for species/sources to be considered secondary
Dates = array([i.strftime("%Y-%m-%d") for i in num2date(date_match)]).reshape(len(date_match),1)
global Species_mod
Species_mod = [i+"*" if i in secondary else i for i in Species]
Species_mod.insert(0,'Date')


print " ---"
print " Final Sources:", mod_head
print " Final Species (* indicates Secondary)", Species_mod

### Write tables out for hybrid modeling
if not os.path.exists('./Data'):
       os.makedirs('./Data')
import csv
with open("./Data/ME2-Concentrations.csv",'w') as fout:
    writer = csv.writer(fout,delimiter=',')
    writer.writerow(Species_mod)
```

```
    savetxt(fout,column_stack((Dates,views_out[:,:,0])),delimiter=',',fmt="%s")

with open("./Data/ME2-Uncertainties.csv",'w') as fout:
    writer = csv.writer(fout,delimiter=',')
    writer.writerow(Species_mod)
    savetxt(fout,column_stack((Dates,views_out[:,:,1])),delimiter=',',fmt="%s")

with open("./Data/ME2-CTM-Contributions.csv",'w') as fout:
    writer = csv.writer(fout,delimiter=',')
    writer.writerow(insert(mod_head,0,"Date"))
    [writer.writerow(i) for i in [append('-999',Ccode)]]
    savetxt(fout,column_stack((Dates,mod_cut.astype(float))),delimiter=',',fmt="%s")

with open("./Data/ME2-CTM-Uncertainties.csv",'w') as fout:
    writer = csv.writer(fout,delimiter=',')
    writer.writerow(insert(mod_head,0,"Date"))
    [writer.writerow(i) for i in [append('-999',Ccode)]]
    savetxt(fout,column_stack((Dates,mod_cut_unc.astype(float))),delimiter=',',fmt="%s")

with open("./Data/ME2-CTM-Species.csv",'w') as fout:
    writer = csv.writer(fout,delimiter=',')
    [writer.writerow(append(i,secondary[i])) for i in SecondSrcs]

if XV10 == 1:
    if not os.path.exists('./XV-Data'):
        os.makedirs('./XV-Data')
    # Develop Data for 10-Fold Cross Validation
    from sklearn.cross_validation import KFold
    xvid = KFold(len(Dates), n_folds=10, indices=True)

    trainer = []
    tester = []
    for train, test in xvid:
        trainer.append(train)
        tester.append(test)

    train = array(trainer)
    test  = array(tester)

    for z in range(10):
        with open("./XV-Data/XV-Train-Conc-"+str(z)+".csv",'w') as fout:
            writer = csv.writer(fout,delimiter=',')
```

```
            writer.writerow(Species_mod)
            savetxt(fout,column_stack((Dates[train[z]],views_out[train[z],:,0])),delimiter=',',fmt="%s")
        with open("./XV-Data/XV-Test-Conc-"+str(z)+".csv",'w') as fout:
            writer = csv.writer(fout,delimiter=',')
            writer.writerow(Species_mod)
            savetxt(fout,column_stack((Dates[test[z]],views_out[test[z],:,0])),delimiter=',',fmt="%s")
        with open("./XV-Data/XV-Train-Unc-"+str(z)+".csv",'w') as fout:
            writer = csv.writer(fout,delimiter=',')
            writer.writerow(Species_mod)
            savetxt(fout,column_stack((Dates[train[z]],views_out[train[z],:,1])),delimiter=',',fmt="%s")
        with open("./XV-Data/XV-Test-Unc-"+str(z)+".csv",'w') as fout:
            writer = csv.writer(fout,delimiter=',')
            writer.writerow(Species_mod)
            savetxt(fout,column_stack((Dates[test[z]],views_out[test[z],:,1])),delimiter=',',fmt="%s")
        with open("./XV-Data/XV-Train-CTMCont-"+str(z)+".csv",'w') as fout:
            writer = csv.writer(fout,delimiter=',')
            writer.writerow(insert(mod_head,0,"Date"))
            [writer.writerow(i) for i in [append('-999',Ccode)]]
            savetxt(fout,column_stack((Dates[train[z]],mod_cut[train[z],:].astype(float))),delimiter=',',fmt="%s")
        with open("./XV-Data/XV-Train-CTMUnc-"+str(z)+".csv",'w') as fout:
            writer = csv.writer(fout,delimiter=',')
            writer.writerow(insert(mod_head,0,"Date"))
            [writer.writerow(i) for i in [append('-999',Ccode)]]
            savetxt(fout,column_stack((Dates[train[z]],mod_cut_unc[train[z],:].astype(float))),delimiter=',',fmt="%s")
```

```
def CreateFolders(XV=0):
    """ Create folder structure for Hybrid modeling.
        Optional: XV Number for subfolder creation """
    folders = ["./Data", "./ME2-Inputs", "ME2-Outputs", "./ME2-Files"]
    fcheck = [os.path.exists(i) for i in folders]

    if sum(fcheck) != 4 and os.path.exists("./Data"):
        [os.makedirs(i) for i in folders[1:]]
    elif sum(fcheck) != 4:
        [os.makedirs(i) for i in folders]
        sys.exit("New folders added. Please add preprocessing data to the 'Data' folder.")
    else:
        print "Folders in place."
```

```
    if XV10 == 1:
        if not os.path.exists('./ME2-Outputs/XV'+str(XV)):
            os.makedirs('./ME2-Outputs/XV'+str(XV))

        if not os.path.exists('./ME2-Inputs/XV'+str(XV)):
            os.makedirs('./ME2-Inputs/XV'+str(XV))



def LoadHybridData(icon=Concentrations, iunc=Uncertainties, ictm=PriorContribs, ictmu=PriorUncerts,
     isec=SecondSpecies,XV=0):
    """ Read in required data for hybrid modeling.
        Default inputs defined as Concentrations, Uncertainties, Prior Contribs, PriorUncerts,
             SecondSpecies.
        Other variables or direct filename input can be used"""

    global n1, n2, n3, ctmcodes, ctmsec, species, nonsec_sp, secsrc, normid

    Conc     = icon
    Conc_unc = iunc
    CTM      = ictm
    CTM_unc  = ictmu

    # Get species and dates
    species = genfromtxt(icon,delimiter=',',dtype='S')[0,1:]
    dates   = loadtxt(icon,delimiter=',',skiprows=1, converters={0:strpdate2num('%Y-%m-%d')})[:,0]
    dates   = [i.strftime('%Y-%m-%d') for i in num2date(dates)]

    # Load concentrations and uncertainties
    conc    = loadtxt(icon,delimiter=',',skiprows=1,usecols=(range(1,len(species)+1)))
    concu   = loadtxt(iunc,delimiter=',',skiprows=1,usecols=(range(1,len(species)+1)))

    # Load CTM names, codes, and secondary species
    ctmnames = genfromtxt(ictm,delimiter=',',dtype='S')[0,1:]
    ctmcodes = genfromtxt(ictm,delimiter=',',skiprows=1,dtype='S')[0,1:]
    with open(isec,'r') as fin:
        ctmsec = array([i.rstrip().split(',') for i in fin])

    # Load CTM contributions and contribution uncertainties
    ctm     = loadtxt(ictm,delimiter=',',skiprows=2,usecols=(range(1,len(ctmcodes)+1)))
    ctmu    = loadtxt(ictmu,delimiter=',',skiprows=2,usecols=(range(1,len(ctmcodes)+1)))
```

```
# Define model characteristics
characteristics = (conc.shape[0],conc.shape[1],len(ctmcodes),np)

# Get non-secondary species
secsrc   = array([where(ctmcodes == i)[0] for i in ctmcodes if "*" in i]).T[0]+1
noaster  = array([i.replace('*','') for i in species])
nonsec_sp = [[where(noaster == i)[0] for i in noaster if i not in ctmsec[j][1:]] for j in
     range(len(ctmsec))]
nonsec_sp = [[i.tolist()+1 for i in nonsec_sp[j] for i in i] for j in range(len(ctmsec))]

# Get normalization species
normid   = array([where(array([i.replace('*','') for i in species]) == j)[0] for j in
     norm]).T[0]+1

# Save data to files in ./ME2-Files/
if XV10 == 0:
    savetxt('./ME2-Files/Conc.me2',conc,delimiter=',')
    savetxt('./ME2-Files/Unc.me2',concu,delimiter=',')
    savetxt('./ME2-Files/Priors.me2',ctm,delimiter=',')
    savetxt('./ME2-Files/PriorsUnc.me2',ctmu,delimiter=',')
    savetxt('./ME2-Files/Species.me2',species,delimiter=',',fmt='%s')
    savetxt('./ME2-Files/Sources.me2',ctmnames,delimiter=',',fmt='%s')
    savetxt('./ME2-Files/Dates.me2', dates, delimiter=',',fmt='%s')
    savetxt('./ME2-Files/Codes.me2', ctmcodes, delimiter=',',fmt='%s')
    savetxt('./ME2-Files/Characteristics.me2',characteristics,delimiter=',')
else:
    savetxt('./ME2-Files/Conc-'+str(XV)+'.me2',conc,delimiter=',')
    savetxt('./ME2-Files/Unc-'+str(XV)+'.me2',concu,delimiter=',')
    savetxt('./ME2-Files/Priors-'+str(XV)+'.me2',ctm,delimiter=',')
    savetxt('./ME2-Files/PriorsUnc-'+str(XV)+'.me2',ctmu,delimiter=',')
    savetxt('./ME2-Files/Species-'+str(XV)+'.me2',species,delimiter=',',fmt='%s')
    savetxt('./ME2-Files/Sources-'+str(XV)+'.me2',ctmnames,delimiter=',',fmt='%s')
    savetxt('./ME2-Files/Dates-'+str(XV)+'.me2', dates, delimiter=',',fmt='%s')
    savetxt('./ME2-Files/Codes-'+str(XV)+'.me2', ctmcodes, delimiter=',',fmt='%s')
    savetxt('./ME2-Files/Characteristics-'+str(XV)+'.me2',characteristics,delimiter=',')

n1 = conc.shape[0]
n2 = conc.shape[1]
n3 = len(ctmnames)

### Print Model Setup to Screen
print "    ================================="
```

```
    print "    Data files populated successfully."
    print "    ================================="
    print "    Number of Samples:", n1
    print "    Number of Species:", n2
    print "    Number of Constraints:", n3
    print "    Number of Sources:", np
    print "    ----------------------------"
    print "    Constraining with:", ctmnames



################################
# CREATE THE INPUT ME-2 SCRIPT #
################################

def CreateInp(it,XV=0):
    script = []
    script.append("##ME-2 script for 2-way PMF. Licence: "+me2Key+" \n")
    script.append('''
section> defines;
version=1.100;
monitor=5;
robust=1;
posoutdist=4; negoutdist=4;
missdatlim=-990;
bdlneg=0;
convtests
  0.100,         40,       5000,        0,     0,    0.0001,
  0.010,         50,       8000,        0,     0,    0.0001,
  0.005,        80,       20000,       0,     0,    0.00002;
cgresets 10,  80,  1,  1,  1,  1; \n''')

    # Number of Tasks
    script.append("precmode=15; numtasks="+str(nt)+"; \n")

    script.append('''
  variables
  'numoldsol'=1,
  'alowlim'=0.0,
  'blowlim'=0.0,
  'seed1'=412,
  'normc1'=0.01,
```

```
  'contrun'=0;
if> (contrun>0); numtasks=1; if!; \n''')

    # Basic properties
    script.append("n1="+str(n1)+"; n2="+str(n2)+"; n3="+str(n3)+"; np="+str(np)+"; \n")

    script.append('''
 c1=0.0; c2=0.0; c3=0.05; em=-14;
 defarr maindata, XX[n1, n2];
 defarr auxdata, YY[n1, n3];
 defarr auxdata, NORM[1,np];
 defarr auxdata, ZERO[1,1];
 defarr freefact, AA[n1, np];
 defarr freefact, BB[n2, np];
 defarr freefact, CC[n3];
 defarr scripttext, AAHEAD[n1];
 defarr scripttext, BBHEAD[n2];
 defarr scriptdata, COLUMNAVG[n2];
 defarr scripttext, FACTHEAD[np];
 defarr scripttext, FORMATS[6]; \n ''')

    script.append("$include $Xwritex '"+ me2Lib+"' \n")
    script.append('''
$include $Fwrite2
 subroutine> Fhead{fi}{};
 local jp;
   write fi, '(16X)';
   for> jp=1:1:np;
     write fi, '(1X,A8)', FACTHEAD[jp];
   for!;
 subroutine!;
 dummyarr freefact, FACTNORM[];
 dummyarr auxdata, AUXNORM[];
section!;
section> equations; \n''')

    # Files to open for reading/writing
    if XV10 == 0:
        script.append("openfile 30, './ME2-Files/Conc.me2',  R, 'old', 6000; \n")
        script.append("openfile 31, './ME2-Files/Unc.me2',   R, 'old', 6000; \n")
        script.append("openfile 32, './ME2-Files/Priors.me2', R, 'old', 6000; \n")
        script.append("openfile 33, './ME2-Files/PriorsUnc.me2', R, 'old', 6000; \n")
```

```
        script.append("openfile 40, './ME2-Outputs/hyME2-u"+it+"/contributions.raw.dat', W,
            'replace', 2000; \n")
        script.append("openfile 41, './ME2-Outputs/hyME2-u"+it+"/profiles.raw.dat', W, 'replace',
            2000; \n")
        script.append("openfile 42, './ME2-Outputs/hyME2-u"+it+"/performance.raw.dat', W, 'replace',
            2000; \n")
        script.append("openfile 43, './ME2-Outputs/hyME2-u"+it+"/scaling.raw.dat', W, 'replace',
            2000; \n")
        if not os.path.exists('./ME2-Outputs/hyME2-u'+it):
            os.makedirs('./ME2-Outputs/hyME2-u'+it)
    else:
        script.append("openfile 30, './ME2-Files/Conc-"+str(XV)+".me2', R, 'old', 6000; \n")
        script.append("openfile 31, './ME2-Files/Unc-"+str(XV)+".me2', R, 'old', 6000; \n")
        script.append("openfile 32, './ME2-Files/Priors-"+str(XV)+".me2', R, 'old', 6000; \n")
        script.append("openfile 33, './ME2-Files/PriorsUnc-"+str(XV)+".me2', R, 'old', 6000; \n")
        script.append("openfile 40, './ME2-Outputs/XV"+XV+"/hyME2-u"+it+"/contributions.raw.dat', W,
            'replace', 2000; \n")
        script.append("openfile 41, './ME2-Outputs/XV"+XV+"/hyME2-u"+it+"/profiles.raw.dat', W,
            'replace', 2000; \n")
        script.append("openfile 42, './ME2-Outputs/XV"+XV+"/hyME2-u"+it+"/performance.raw.dat', W,
            'replace', 2000; \n")
        script.append("openfile 43, './ME2-Outputs/XV"+XV+"/hyME2-u"+it+"/scaling.raw.dat', W,
            'replace', 2000; \n")
        if not os.path.exists('./ME2-Outputs/XV'+XV+'/hyME2-u'+it):
            os.makedirs('./ME2-Outputs/XV'+XV+'/hyME2-u'+it)


    script.append('''
if> (contrun>0);
openfile 39,   ##p,    R,  'old',      2000;
if!;
FORMATS[1]='(/(', np,'(I2)', 'E14.5))';
XX[0,0]=0;
XX.C1[0,0]=0;
for> j1=1:1:n1;
    for> j2=1:1:n2;
        read 30, ' ', XX[j1,j2];
        read 31, ' ', XX.C1[j1,j2];
    for!;
for!;
YY[0,0]=0;
YY.C1[0,0]=0;
```

```
for> j1=1:1:n1;
     for> j2=1:1:n3;
          read 32, ' ', YY[j1,j2];
          read 33, ' ', YY.C1[j1,j2];
     for!;
for!;
for> j2=1:1:n2;
    for> j1=1:1:n1;
      equ> XX[j1,j2], C1=XX.C1[j1,j2], C3=c3, errmod=em;
       for> jp=1:1:np;
         term> pos; @AA[j1,jp]; @BB[j2,jp]; term!;
       for!;
      equ!;
    for!;
for!; \n''')


    # Write Hybrid Equations
    for i in range(n3):
        if it != "PMF" and it != "CPMF":
            script.append("for> j1=1:1:n1; equ> "+
                          "YY[j1,"+str(i+1)+"], C1=YY.C1[j1,"+str(i+1)+"]*"+it+
                          ", C3=c3, errmod=-14; term>"+
                          " pos; @AA[j1,"+str(i+1)+"]; @CC["+str(i+1)+"]; term!; equ!; for!; \n \n")

    # Write Zeroing Equations
    if it != "PMF":
        script.append("equ> ZERO[1,1], Data=0, C1=1e-05, errmod=-12; \n")
        for i in range(len(ctmsec)):
            script.append("for> jj={"+str(secsrc[i])+
                          "}; \n for> ii={"+', '.join(str(s) for s in nonsec_sp[i])+
                          "}; \n term> pos; @BB[ii,jj]; term!; for!; for!; \n")
        script.append("equ!; \n")

    # Write Normalization Equation
    script.append("for> jj=1:1:np; \n equ> NORM[1,jj], Data=1, C1=0.001, C3=0, errmod=-12; \n "
                  "for> ii={"+', '.join(str(s) for s in normid)+"}; \n "
                  "term> pos; @BB[ii,jj]; term!; for!; equ!; for!; \n")


    script.append('''
AA.fkey[0,0]=lolimit;  AA.flow[0,0]=alowlim; AA.fhigh[0,0]=5.0;
```

```
BB.fkey[0,0]=lolimit;  BB.flow[0,0]=blowlim;
BB.fkey[17,1]=lolimit; BB.flow[17,1]=0.01;
BB.fkey[16,1]=lohilimits; BB.flow[16,1]=0.0; BB.fhigh[16,1]=0.05;
BB.fkey[21,1]=lohilimits; BB.flow[21,1]=0.0; BB.fhigh[21,1]=0.05; \n''')

    script.append('CC.fkey[0]=lohilimits; CC.flow[0]='+str(Clow)+'; CC.fhigh[0]='+str(Chigh)+';')

    script.append('''
AA.fprecc[0,1]=np;
BB.fprecc[0,1]=np;
section!;
section> preproc;
if> (taskcount==1);
  for> j2=1:1:n2;
    cn=0; sm=0;
    for> j1=1:1:n1;
      if> (XX[j1,j2]>=0);
        sm=sm+XX[j1,j2]; cn=cn+1;
      if!;
    for!;
    COLUMNAVG[j2]=Maxval{sm,0.0001}/Maxval{cn,1};
    BB.fhigh[j2,0]=COLUMNAVG[j2];
  for!;
if!;
if> (contrun==0);
  NORM.aux1[1,np]=0.0;
  setrand 1, uniform, seed1;
  AA[0,0] = Urandom{0.01,2.0,1};
  for> j2=1:1:n2;
    BB[j2,0] = Urandom{0.01*COLUMNAVG[j2],0.5*COLUMNAVG[j2],1};
  for!;
  seed1=seed1+100;
elseif (contrun==1 | contrun==2);
  if> (taskcount==1);
    for> ii=1:1:numoldsol;
      read 39, ' ', AA[0,0], BB[0,0];
    for!;
  if!;
else;
  stop 'contrun should be =0 or =1 or =2';
if!;
if> (contrun==2);
```

```
  openfile 38,   ##c,    R,  'old',      2000;
  for> jp=1:1:np;
    for> j2=1:1:n2;
      read 38, ' ', tt;
      BB.fkey[j2,jp]=tt;
      if> (tt==locked);
        NORM.C1[1,np]=10*n1;
      if!;
    for!;
  for!;
$skiplines
  28*0        /fact # 1: all elements are lolimit
  28*-5       /fact # 2: all elements are locked (not variable)
  28*0        /fact # 3: all elements are lolimit
  28*0        /fact # 4: all elements are lolimit
  7*0, -6, -6, 19*0
      /fact # 5: all elements are lolimit, except #8 and #9
      /are fixed to zero
  28*0        /fact # 6: all elements are lolimit
  28*0        /fact # 7: all elements are lolimit
$endskip
if!;
section!;
section> postproc;
write 40, FORMATS[1], AA[0,0];
write 40, '(/A)', ' ' ;
write 41, FORMATS[1], BB[0,0];
write 41, '(/A)', ' ' ;
write 43, FORMATS[1], CC[0];
write 43, '(/A)', ' ' ;
write 42, '(//A/)',
' task#, seed,   Qrobust,  Q,     Qmain,     Qaux,    Iterations,  Self-Cancel';
write 42, '(I5)', taskcount, seed1;
write 42, '(F12.3)', qvalue,trueqvalue, mainqvalue, auxqvalue, itercount, selfcancel;
section!;
section> callback;
section!;
/ Slash comments can be used among data values, such as this line.''')

    return script
```

```
####################################
### Use Functions and Run Script ###
####################################

# COPY ME2 Library, Key, and Executable into model folders
import shutil
shutil.copy2(me2exe, me2exe.split('/')[len(me2exe.split('/'))-1])

# LOAD MEASUREMENTS
ViewsOut = LoadMeas(views)
VIEWS   = ViewsOut[0] # Conc and Unc as 3D Array
vDates  = ViewsOut[1] # VIEWS Dates
Species = ViewsOut[2] # Species
Site    = ViewsOut[3]

# LOAD CTM MODEL RESULTS
# & PROCESS WITH SITE MEASUREMENTS
LoadCTMModel(model,modelunc)

if XV10 == 0:
    # CREATE MODELING FOLDERS
    CreateFolders()

    # LOAD HYBRID MODEL DATA
    LoadHybridData()

    iniList = []
    for i in knob:
        me2In = CreateInp(str(i))
        File = './ME2-Inputs/hyME2-u'+str(i)+'.ini'
        iniList.append(File)
        with open(File,'w') as fin:
            for j in me2In:
                fin.write(j)

    # Write executables for running model
    for lines in iniList:
        with open(lines+".csh",'w') as qsub:
            print>>qsub, "#!/bin/tcsh -f"
            print>>qsub, "wine me2wG17.exe "+lines
    with open("RunHybridModel.csh", 'w') as Farout:
```

```
            print>>Farout, "#!/bin/tcsh -f"
            for lines in iniList:
                print>>Farout, 'qsub -cwd -l h="compute-2-*" -e ./ME2-Inputs/ -o ./ME2-Inputs/ -S
                    /bin/tcsh '+lines+'.csh'
    os.chmod("./RunHybridModel.csh",0744)


else:
    for z in range(10):
        # CREATE MODELING FOLDERS
        CreateFolders(z)

        # LOAD HYBRID MODEL DATA
        ConcentrationsXV = './XV-Data/XV-Train-Conc-'+str(z)+'.csv'
        UncertaintiesXV = './XV-Data/XV-Train-Unc-'+str(z)+'.csv'
        PriorContribsXV = './XV-Data/XV-Train-CTMCont-'+str(z)+'.csv'
        PriorUncertsXV = './XV-Data/XV-Train-CTMUnc-'+str(z)+'.csv'
        SecondSpeciesXV = './Data/ME2-CTM-Species.csv'
        LoadHybridData(ConcentrationsXV,UncertaintiesXV,PriorContribsXV,PriorUncertsXV,SecondSpeciesXV,XV=z)

        iniList = []
        for i in knob:
            me2In = CreateInp(str(i),str(z))
            File = './ME2-Inputs/XV'+str(z)+'/hyME2-u'+str(i)+'.ini'
            iniList.append(File)
            with open(File,'w') as fin:
                for j in me2In:
                    fin.write(j)

        # Write executables for running model
        for lines in iniList:
            with open(lines+".csh",'w') as qsub:
                print>>qsub, "#!/bin/tcsh -f"
                print>>qsub, "wine me2wG17.exe "+lines
        with open("./XV"+str(z)+"-RunHybridModel.csh", 'w') as Farout:
                print>>Farout, "#!/bin/tcsh -f"
                for lines in iniList:
                    print>>Farout, 'qsub -cwd -l "h=compute-2-*" -e ./ME2-Inputs/XV'+str(z)+' -o
                        ./ME2-Inputs/XV'+str(z)+' -S /bin/tcsh '+lines+'.csh'
        os.chmod("./XV"+str(z)+"-RunHybridModel.csh",0744)
```

Appendix B

HYBRID CODE: POST-PROCESSING

```
# -*- coding: utf-8 -*-
"""
Created on Fri Jul 5 13:08:24 2013

@author: tmsturtz
"""

import os, sys
import h5py
from sklearn import linear_model
import csv
from pylab import *
import matplotlib.pyplot as plt
from subprocess import call
import shutil
import pandas as pd


################################################################################
### Define RMSE & NRMSE (as %) for comparison with concentration values ###
################################################################################


def nrmse(sim,obs):
    sim2 = sim[:,array([where(species == i)[0] for i in TC]).T[0]] #.sum(axis=1)
    obs2 = obs[:,array([where(species == i)[0] for i in TC]).T[0]] #.sum(axis=1)
    return ((sqrt(sum((sim2-obs2)**2)/len(obs.sum(axis=1))))) #/(obs2.max()-obs2.min()))

def nrmseALL(sim,obs):
    #sim2 = sim[:,array([where(species == i)[0] for i in TC]).T[0]] #.sum(axis=1)
    #obs2 = obs[:,array([where(species == i)[0] for i in TC]).T[0]] #.sum(axis=1)
    return ((sqrt(sum((sim-obs)**2)/len(obs.sum(axis=1))))) #/(obs.max()-obs.min()))
```

```
def CTMnrmse(sim,obs):
    sim2 = sim.sum(axis=1)
    obs2 = obs[:,array([where(species == i)[0] for i in TC]).T[0]].sum(axis=1)
    return 100*((sqrt(sum((sim2-obs2)**2)/len(obs.sum(axis=1))))) #/(obs2.max()-obs2.min()))

def COE(sim,obs):
    'Coefficient of Efficiency (McCabe & Legates 2012): E1 method '
    sim2 = sim[:,array([where(species == i)[0] for i in TC]).T[0]].sum(axis=1)
    obs2 = obs[:,array([where(species == i)[0] for i in TC]).T[0]].sum(axis=1)
    return 1-(sum(abs(obs2-sim2))/sum(abs(obs2-mean(obs2))))

def GetObsTC():
    info = loadtxt('./ME2-Files/Characteristics.me2')
    n2   = int(info[1])
    TC   = ["EC1", "EC2", "OC1", "OC2", "OC3", "OC4"]
    conc = loadtxt('./Data/ME2-Concentrations.csv',delimiter=',',skiprows=1,usecols=range(1,n2+1))
    OTC  = conc[:,array([where(species == i)[0] for i in TC]).T[0]].sum(axis=1)
    return OTC

def GetObs():
    info = loadtxt('./ME2-Files/Characteristics.me2')
    n2   = int(info[1])
    conc = loadtxt('./Data/ME2-Concentrations.csv',delimiter=',',skiprows=1,usecols=range(1,n2+1))
    return conc

def GetObsUnc():
    info = loadtxt('./ME2-Files/Characteristics.me2')
    n2   = int(info[1])
    unc = loadtxt('./Data/ME2-Uncertainties.csv',delimiter=',',skiprows=1,usecols=range(1,n2+1))
    return unc

def GetQ(x):
    q = sum(((GetObs()-Simulations[x,:,:])/GetObsUnc())**2)
    return q


#!!! BEGIN PROCESSING DATA !!!#

##########################################################################
### Loop Over Directories, determine minQ, and save associated data ###
##########################################################################
def ProcessResults():
```

```
global TC, species, contributions, profiles, qdat, scaling, knobs, ModelRMSE
dirList = [i for i in os.listdir('./ME2-Outputs/') if "hyME2" in i]
TC      = ["EC1", "EC2", "OC1", "OC2", "OC3", "OC4"]

Qdat         = []
knobs        = []
contributions = []
oddcontribs  = []
profiles     = []
scaling      = []
ModelRMSE    = []

dirList = [i[1] for i in sorted(enumerate(dirList), key=lambda x:x[1])]
dirList = ['ME2-Outputs/'+i for i in dirList]
for z in dirList:
    z1 = z.split('-u')[1]
    knob = z1
    knobs.append(knob)

    # Get Modeling Setup
    # Samples = n1, Species = n2, constraints = n3, sources = np
    info  = loadtxt('./ME2-Files/Characteristics.me2')
    n1    = int(info[0])
    n2    = int(info[1])
    n3    = int(info[2])
    np    = int(info[3])

    # Get Dates, Species List, Concentrations, Constraint Names, Constraint Values, CTMs
    dates    = loadtxt('./ME2-Files/Dates.me2',converters={0:strpdate2num('%Y-%m-%d')})
    species  = loadtxt('./ME2-Files/Species.me2',dtype='string')
    conc     =
         loadtxt('./Data/ME2-Concentrations.csv',delimiter=',',skiprows=1,usecols=range(1,n2+1))
    priors_id = loadtxt('./ME2-Files/Sources.me2',dtype='string')
    priors   =
         loadtxt('./Data/ME2-CTM-Contributions.csv',delimiter=',',skiprows=2,usecols=range(1,n3+1))
    CTM_All  = loadtxt("./Data/CTM-AllSources-Contributions-4RMSE_Calcs.csv",delimiter=',')


    # Performance: task#, seed, Qrobust, Q, Qmain, Qaux, Iterations, Se
    perform = loadtxt('./'+z+'/performance.raw.dat',comments='task#')
    if len(perform[where(perform[:,4] == perform[:,4].min()),0][0]) == 1:
        minTask = int(perform[where(perform[:,4] == perform[:,4].min()),0][0])-1;
```

```
        else:
            minTask = int(perform[where(perform[:,4] == perform[:,4].min()),0][0][0])-1
        Qdat.append(perform[minTask-1,:].astype(list)) ####!!!!! MOVE TO END FOR COLLECTION?

        contribs = loadtxt('./'+z+'/contributions.raw.dat')
        contribs = reshape(contribs,(contribs.shape[0]/n1,n1,np))
        contributions.append(contribs[minTask-1,:,:])

        profs = loadtxt('./'+z+'/profiles.raw.dat')
        profs = reshape(profs,(profs.shape[0]/n2,n2,np))
        profiles.append(profs[minTask-1,:,:])

        scale = loadtxt('./'+z+'/scaling.raw.dat')
        scale = scale[minTask-1,:]
        scaling.append(scale)

        RMSE = CTMnrmse(contribs[minTask-1,:,:],conc)
        ModelRMSE.append(RMSE)

    contributions = array(contributions)
    profiles      = array(profiles)
    Qdat          = array(Qdat).astype('float')
    scaling       = array(scaling)
    knobs         = array(knobs)
    ModelRMSE     = array(ModelRMSE)

    # Write Data to HDF5 File
    f = h5py.File('./ME2-Outputs/ModelData.hdf5','w')

    h5contrib       = f.create_dataset("Contributions",contributions.shape,'f')
    h5contrib[:,:,:] = contributions

    h5prof          = f.create_dataset("Profiles",profiles.shape,'f')
    h5prof[:,:,:]   = profiles

    h5qdat          = f.create_dataset("QData",Qdat.shape,'f')
    h5qdat[:,:]     = Qdat

    h5scale         = f.create_dataset("Scaling",scaling.shape,'f')
    h5scale[:,:]    = scaling

    str_type        = h5py.new_vlen(str)
```

```
    h5knobs          = f.create_dataset("Knobs",knobs.shape,dtype=str_type)
    h5knobs[:]       = knobs

    h5rmse           = f.create_dataset("NRMSE",ModelRMSE.shape,'f')
    h5rmse[:]        = ModelRMSE

    f.close()



def ProcessResultsXV():
    global TC, species, contributions, oddcontrib, profiles, qdat, scaling, knobs, XVid, Simulations,
         ModelRMSE, XVQ, TrueSim
    import glob
    fullDir = glob.glob("./ME2-Outputs/XV*/*")
    fullDir = array(fullDir)
    fullDir = fullDir[fullDir.argsort()]
    XV      = [i.split('/')[2] for i in fullDir]
    knobls = [i.split('/')[3] for i in fullDir]
    knobs  = unique(knobls)
    knobs = array([i.split('-u')[1] for i in knobs])
    TC      = ["EC1", "EC2", "OC1", "OC2", "OC3", "OC4"]

    for y in range(10):
        dat =
             loadtxt('./XV-Data/XV-Test-Conc-'+str(y)+'.csv',delimiter=',',skiprows=1,converters={0:strpdate2num('%Y
        dat2 =
             loadtxt('./XV-Data/XV-Test-Unc-'+str(y)+'.csv',delimiter=',',skiprows=1,converters={0:strpdate2num('%Y-
        savetxt('./XV-Data/XV-Test-Conc-'+str(y)+'.me2',dat[:,1:],delimiter=',')
        savetxt('./XV-Data/XV-Test-Unc-'+str(y)+'.me2',dat[:,1:],delimiter=',')

    Qdat         = []
    contributions = []
    oddcontrib   = []
    profiles     = []
    scaling      = []
    XVid         = []
    XVQ          = []
    Simulations  = []
    RMSEXV       = []
```

```
for z in fullDir:

    z1 = z.split('/')[3].split('hyME2-')[1]
    knob = z1
    print z1
    XV = str(z.split('/')[2].split('XV')[1])
    print XV
    XVid.append(XV)

    # Get Modeling Setup
    # Samples = n1, Species = n2, constraints = n3, sources = np
    info  = loadtxt('./ME2-Files/Characteristics-'+XV+'.me2')
    n1    = int(info[0])
    n2    = int(info[1])
    n3    = int(info[2])
    np    = int(info[3])

    # Get Dates, Species List, Concentrations, Constraint Names, Constraint Values, CTMs
    dates    = loadtxt('./ME2-Files/Dates-'+XV+'.me2',converters={0:strpdate2num('%Y-%m-%d')})
    species  = loadtxt('./ME2-Files/Species-'+XV+'.me2',dtype='string')
    conc     =
        loadtxt('./Data/ME2-Concentrations.csv',delimiter=',',skiprows=1,usecols=range(1,n2+1))
    priors_id = loadtxt('./ME2-Files/Sources-'+XV+'.me2',dtype='string')
    priors   =
        loadtxt('./Data/ME2-CTM-Contributions.csv',delimiter=',',skiprows=2,usecols=range(1,n3+1))
    CTM_All  = loadtxt("./Data/CTM-AllSources-Contributions-4RMSE_Calcs.csv",delimiter=',')
    XVConc   = loadtxt('./XV-Data/XV-Test-Conc-'+str(XV)+'.me2',delimiter=',')


    # Performance: task#, seed, Qrobust, Q, Qmain, Qaux, Iterations, Se
    perform = loadtxt('./'+z+'/performance.raw.dat',comments='task#')
    if len(perform[where(perform[:,4] == perform[:,4].min()),0][0]) == 1:
        minTask = int(perform[where(perform[:,4] == perform[:,4].min()),0][0])-1;
    else:
        minTask = int(perform[where(perform[:,4] == perform[:,4].min()),0][0][0])-1
    Qdat.append(perform[minTask-1,:].astype(list)) ####!!!!! MOVE TO END FOR COLLECTION?

    contribs = loadtxt('./'+z+'/contributions.raw.dat')
    contribs = reshape(contribs,(contribs.shape[0]/n1,n1,np))
    if XV != '9':
        contributions.append(contribs[minTask-1,:,:])
    else:
```

```
    oddcontrib.append(contribs[minTask-1,:,:])

profs = loadtxt('./'+z+'/profiles.raw.dat')
profs = reshape(profs,(profs.shape[0]/n2,n2,np))
profiles.append(profs[minTask-1,:,:])

scale = loadtxt('./'+z+'/scaling.raw.dat')
scale = scale[minTask-1,:]
scaling.append(scale)

savetxt('./'+z+'/performance.minQ.dat',perform[minTask-1,:],delimiter=',')
savetxt('./'+z+'/contributions.minQ.dat',contribs[minTask-1,:,:],delimiter=',')
savetxt('./'+z+'/profiles.minQ.dat',profs[minTask-1,:,:],delimiter=',')
savetxt('./'+z+'/scaling.minQ.dat',scale,delimiter=',')

## Write ME-2 Inputs for cross-validation
with open('../1-DataPreprocess/Regress2.ini','r') as iniF:
    iniFile = [i.rstrip() for i in iniF]
iniNew = []
for i in iniFile:
    if "n1=328;" in i:
        tmp = i.replace("=328","="+str(XVConc.shape[0]))
    elif 'n2=22;' in i:
        tmp = i.replace("=22","="+str(n2))
    elif 'np=5;' in i:
        tmp = i.replace("=5","="+str(np))
    elif 'openfile 30,' in i:
        tmp = i.replace("'./ME2-Files/Conc.me2'","'./XV-Data/XV-Test-Conc-"+str(XV)+".me2'")
    elif 'openfile 31,' in i:
        tmp = i.replace("'./ME2-Files/Unc.me2'","'./XV-Data/XV-Test-Unc-"+str(XV)+".me2'")
    elif 'openfile 32,' in i:
        tmp = i.replace("'./ME2-Files/Conc.me2'","'"+z+"/profiles.minQ.dat'")
    elif 'openfile 40,' in i:
        tmp = i.replace("/REGRESS.FOLDER/",z+"/")
    elif 'openfile 41,' in i:
        tmp = i.replace("/REGRESS.FOLDER/",z+"/")
    elif 'openfile 42,' in i:
        tmp = i.replace("/REGRESS.FOLDER/",z+"/")
    else:
        tmp = i

    iniNew.append(tmp)
```

```
    with open('./ME2-Inputs/XV'+str(XV)+'/XVSim-'+str(knob)+'.ini','w') as iniN:
        [iniN.write(i+' \n') for i in iniNew]

    ## Run cross-validation
    with open(z+'/ME2-StdOut.txt','w') as stanout:
        call(["./me2wG17.exe",'./ME2-Inputs/XV'+str(XV)+'/XVSim-'+str(knob)+'.ini'],stdout=stanout)

    XVQdat = loadtxt(z+"/performance.XVSim.dat",skiprows=3)[2]
    XVmodcontribs = loadtxt(z+"/contributions.XVSim.dat",skiprows=1)
    Simul = XVmodcontribs.dot(profs[minTask-1,:,:].T)

    XVQ.append(XVQdat)
    Simulations.append(Simul)


Simulations  = array(Simulations)
TrueSim = []
for c in range(len(knobs)):
    newdata = Simulations[arange(c,len(knobs)*10,len(knobs)),:,:]
    fulldat = array([item for sublist in newdata.tolist() for item in sublist])
    TrueSim.append(fulldat)
TrueSim = array(TrueSim)

RMSEXV = array([nrmse(TrueSim[i,:,:],GetObs()) for i in range(len(unique(knobs)))])
CVQ = array([sum(((GetObs()-TrueSim[x,:,:])/GetObsUnc())**2) for x in range(len(knobs))])

contributions = array(contributions)
oddcontrib   = array(oddcontrib)
profiles     = array(profiles)
Qdat         = array(Qdat).astype('float')
scaling      = array(scaling)
XVid         = array(map(int,XVid))
XVQ          = array(XVQ)
RMSEXV       = array(RMSEXV)



# Write Data to HDF5 File
f = h5py.File('./ME2-Outputs/ModelDataXV.hdf5','w')
```

```
    h5contrib       = f.create_dataset("Contributions",contributions.shape,'f')
    h5contrib[:,:,:] = contributions

    h5contrib2      = f.create_dataset("OddContributions",oddcontrib.shape,'f')
    h5contrib2[:,:,:] = oddcontrib

    h5prof          = f.create_dataset("Profiles",profiles.shape,'f')
    h5prof[:,:,:]   = profiles

    h5qdat          = f.create_dataset("QData",Qdat.shape,'f')
    h5qdat[:,:]     = Qdat

    h5scale         = f.create_dataset("Scaling",scaling.shape,'f')
    h5scale[:,:]    = scaling

    str_type        = h5py.new_vlen(str)
    h5knobs         = f.create_dataset("Knobs",knobs.shape,dtype=str_type)
    h5knobs[:]      = knobs

    h5XV            = f.create_dataset("XV",XVid.shape,'i')
    h5XV[:]         = XVid

    h5simul         = f.create_dataset("SimulationData",TrueSim.shape,'f')
    h5simul[:,:,:]  = TrueSim

    h5cvq           = f.create_dataset("CVQ",CVQ.shape,'f')
    h5cvq[:]        = CVQ

    h5rmse          = f.create_dataset("NRMSE",RMSEXV.shape,'f')
    h5rmse[:]       = RMSEXV


    f.close()


def Bootstrap(dial='PMF'):
    from scipy.stats import pearsonr
    # Create Boot directory
    global AllProfMod, AllProf, BaseProf

    if not os.path.exists('./Boots'):
           os.makedirs('./Boots')
```

```
dial = dial

# Properties #
n1    = contributions.shape[1] # Number of Samples
n2    = profiles.shape[1] # Number of Species
np    = profiles.shape[2]  # Number of Factors
boots = 100  # Number of Bootstraps
blocksize = 4

# Get block indicies
grpbse= range(0,n1,blocksize)
grpind= range(blocksize,n1,blocksize)
grpind.append(n1)# Index ending for each city/season

# Input Files #
Conc = './ME2-Files/Conc.me2'
Unc  = './ME2-Files/Unc.me2'
CTM  = './ME2-Files/Priors.me2'
CTMu = './ME2-Files/PriorsUnc.me2'
Fb   = profiles[where(knobs == dial)[0],:,:][0] # Base Profiles
Gb   = contributions[where(knobs == dial)[0],:,:][0] # Base Contributions
with open('./Boots/ContRunData.csv','w') as fout:
    RMWrite = csv.writer(fout,delimiter=',')
    [RMWrite.writerow(lines) for lines in Gb]
    fout.write("\n")
    [RMWrite.writerow(map(str,lines)) for lines in Fb]

# Output Files #
Fn  = './Boots/Fnew.txt'     # New Profiles
Gn  = './Boots/Gnew.txt'     # New Contributions

# Copy original INI file to Boot directory and modify for bootstrapping
shutil.copy2('./ME2-Inputs/hyME2-u'+dial+'.ini','./Boots/hyME2-Boot-u'+dial+'.ini')
with open('./Boots/hyME2-Boot-u'+dial+'.ini','r') as iniF:
    iniFile = [i.rstrip() for i in iniF]
iniNew = []
for i in iniFile:
    if "'contrun'=0;" in i:
        tmp = i.replace("=0","=1")
    elif 'openfile 30' in i:
        tmp = i.replace("./ME2-Files/Conc.me2","./Boots/Conc.BS.me2")
```

```
    elif 'openfile 31' in i:
        tmp = i.replace("./ME2-Files/Unc.me2","./Boots/Unc.BS.me2")
    elif 'openfile 32' in i:
        tmp = i.replace("./ME2-Files/Priors.me2","./Boots/Priors.BS.me2")
    elif 'openfile 33' in i:
        tmp = i.replace("./ME2-Files/PriorsUnc.me2","./Boots/PriorsUnc.BS.me2")
    elif 'openfile 40' in i:
        tmp = i.replace("./ME2-Outputs/hyME2-u"+dial+"/contributions.raw.dat","./Boots/Gnew.txt")
    elif 'openfile 41' in i:
        tmp = i.replace("./ME2-Outputs/hyME2-u"+dial+"/profiles.raw.dat","./Boots/Fnew.txt")
    elif 'openfile 42' in i:
        tmp =
            i.replace("./ME2-Outputs/hyME2-u"+dial+"/performance.raw.dat","./Boots/performance.txt")
    elif 'openfile 43' in i:
        tmp = i.replace("./ME2-Outputs/hyME2-u"+dial+"/scaling.raw.dat","./Boots/scaling.txt")
    elif 'openfile 39' in i:
        tmp = i.replace("##p","'./Boots/ContRunData.csv'")
    else:
        tmp = i

    iniNew.append(tmp)

with open('./Boots/hyME2-Boot-u'+dial+'.ini.tmp','w') as iniN:
        [iniN.write(i+' \n') for i in iniNew]

### Read in Concentrations and Uncertainty used for base run
with open(Conc,'r') as fin:
    conc = [i.rstrip().split(',') for i in fin]
    conc = [map(float,i) for i in conc]

with open(Unc,'r') as fin:
    unc = [i.rstrip().split(',') for i in fin]
    unc = [map(float,i) for i in unc]

with open(CTM,'r') as fin:
    ctm = [i.rstrip().split(',') for i in fin]
    ctm = [map(float,i) for i in ctm]

with open(CTMu,'r') as fin:
    ctmu = [i.rstrip().split(',') for i in fin]
    ctmu = [map(float,i) for i in ctmu]
```

```
    ### Get bootstrap indicies and counts
    Bootind = []
    indcnt = []
    for k in range(boots):
        run = []

        bootcase = []
        for i in range(len(grpbse)):
            blkcut  = randint(grpbse[i],grpind[i],grpind[i]-grpbse[i])
            blkcut = blkcut.tolist()
            bootcase = bootcase+blkcut
        run = run+bootcase
        bsc = [len([i for i in run if i==j]) for j in range(n1)]
        Bootind.append(run)
        indcnt.append(bsc)

        with open('./Boots/ConstrainedIndicies.csv','w') as bsi:
            savetxt(bsi,Bootind)

    Bootind = array(Bootind)
    # END
    #####

### Loop over all bootstrap scenarios, run ME-2, and read in results
    bsprofiles = []
    bscontributions = []

    for k in range(len(Bootind)):
        bsconc = [conc[i] for i in Bootind[k]]
        bsunc = [unc[i] for i in Bootind[k]]
        bsctm = [ctm[i] for i in Bootind[k]]
        bsctmu = [ctmu[i] for i in Bootind[k]]
        bscnt = [indcnt[k][i] for i in Bootind[k]]
        bsuncm = [[j/sqrt(bscnt[i]) for j in bsunc[i]] for i in range(len(Bootind[k]))] #Uncertainty
             modified by counts
        bsctmum= [[j/sqrt(bscnt[i]) for j in bsctmu[i]] for i in range(len(Bootind[k]))]

        with open('./Boots/Conc.BS.me2','w') as fout:
            RMWrite = csv.writer(fout,delimiter=',')
            [RMWrite.writerow(map(str,lines)) for lines in bsconc]
```

```
    with open('./Boots/Unc.BS.me2','w') as fout:
        RMWrite = csv.writer(fout,delimiter=',')
        [RMWrite.writerow(map(str,lines)) for lines in bsuncm]

    with open('./Boots/Priors.BS.me2','w') as fout:
        RMWrite = csv.writer(fout,delimiter=',')
        [RMWrite.writerow(map(str,lines)) for lines in bsctm]

    with open('./Boots/PriorsUnc.BS.me2','w') as fout:
        RMWrite = csv.writer(fout,delimiter=',')
        [RMWrite.writerow(map(str,lines)) for lines in bsctmum]

    with open('./Boots/StdOut.txt','w') as stanout:
        call(["./me2wG17.exe",'./Boots/hyME2-Boot-u'+dial+'.ini.tmp'],stdout=stanout)

    with open(Fn,'r') as fin:
        fnew = loadtxt(fin)

    with open(Gn,'r') as fin:
        gnew = loadtxt(fin)

    bsprofiles.append(fnew)
    bscontributions.append(gnew)

bsprofiles = dstack(bsprofiles)
bscontributions = dstack(bscontributions)

AllProf = dstack((Fb,bsprofiles))
AllCont = dstack((Gb,bscontributions))
Bootind = append(array(range(Bootind.shape[1])).reshape(-1,1),Bootind.T,1)


### PROCESS BOOTSTRAP RESULTS ###
species = loadtxt('./ME2-Files/Species.me2',dtype='string')

np = AllProf.shape[1]
n1 = AllCont.shape[0]
n2 = AllProf.shape[0]

# Determine base profiles and contributions
BaseProf = AllProf[:,:,0]
BaseCont = AllCont[:,:,0]
```

```
# Calculate the species contributions for the base case and all replicates over each source
BaseSpCont = array([BaseCont[:,i].reshape(-1,1).dot(BaseProf[:,i].reshape(-1,1).T) for i in
    range(np)])
ReplicateSpCont = array([[AllCont[:,j,i].reshape(-1,1).dot(AllProf[:,j,i].reshape(-1,1).T) for j
    in range(np)] for i in range(1,boots+1)])

# Loop over bootstraps and sources to determine the correlation between sources for each bootstrap
MatchedIndicies = []
for i in range(boots):
    TestSamples = sort(unique(Bootind[:,i+1]))
    RepSampleInd = array([where(Bootind[:,i+1] == j)[0][0] for j in TestSamples])

    # Align base and boot samples
    MatchSamplesBase = BaseSpCont[:,TestSamples,:]
    MatchSamplesRepl = ReplicateSpCont[i,:,RepSampleInd,:].transpose((1,0,2))

    # Calculate correlations between base and repl
    Corrs = array([[pearsonr(MatchSamplesBase[j,:,:].reshape(MatchSamplesBase[j,:,:].size,1),
                    MatchSamplesRepl[k,:,:].reshape(MatchSamplesRepl[k,:,:].size,1))[0][0]
                    for k in range(np)] for j in range(np)])

    # Cycle over correlation matrix select max index from first row, then delete index and
        continue
    MaxCorr = []
    cnt = range(np)
    for j in range(np):
        chk = where(Corrs[j,cnt] == Corrs[j,cnt].max())[0][0]
        MaxCorr.append(cnt[chk])
        cnt = delete(cnt,chk)
    MatchedIndicies.append(MaxCorr)

MatchedIndicies = array(MatchedIndicies)
AllProfMod = array([AllProf[:,MatchedIndicies[i],i+1] for i in range(boots)])

# Calculate percentiles of profiles for plotting
from scipy.stats import scoreatpercentile
AllProf5 = array([scoreatpercentile(AllProfMod[:,:,i],5) for i in range(np)])
AllProf50 = array([scoreatpercentile(AllProfMod[:,:,i],50) for i in range(np)])
AllProf95 = array([scoreatpercentile(AllProfMod[:,:,i],95) for i in range(np)])
ProfLow = AllProf50-AllProf5
ProfHigh = AllProf95-AllProf50
```

```
#### How to save 3D Array in numpy
    if dial == 'PMF' or dial == 'CPMF':
        if os.path.exists('./ME2-Outputs/ModelDataBS-PMF.hdf5'):
            os.remove('./ME2-Outputs/ModelDataBS-PMF.hdf5')
        f = h5py.File('./ME2-Outputs/ModelDataBS-PMF.hdf5','w')
    else:
        if os.path.exists('./ME2-Outputs/ModelDataBS.hdf5'):
            os.remove('./ME2-Outputs/ModelDataBS.hdf5')
        f = h5py.File('./ME2-Outputs/ModelDataBS.hdf5','w')

    h5bsprof        = f.create_dataset("BootProfs",AllProf.shape,'f')
    h5bsprof[:,:,:] = AllProf

    h5bscont        = f.create_dataset("BootConts",AllCont.shape,'f')
    h5bscont[:,:,:] = AllCont

    h5bsind         = f.create_dataset("BootIndicies",array(Bootind).shape,'i')
    h5bsind[:,:]    = array(Bootind)

    h5boots         = f.create_dataset("BootCount",array([boots]).shape,'i')
    h5boots[:]      = array([boots])

    h5low           = f.create_dataset("BootLow",array(ProfLow).shape,'f')
    h5low[:,:]      = array(ProfLow)

    h5med           = f.create_dataset("BootMedian",array(AllProf50).shape,'f')
    h5med[:,:]      = array(AllProf50)

    h5high          = f.create_dataset("BootHigh",array(ProfHigh).shape,'f')
    h5high[:,:]     = array(ProfHigh)

    h5base          = f.create_dataset("BaseProfiles",array(BaseProf).shape,'f')
    h5base[:,:]     = array(BaseProf)

    f.close()
# END
#####
```

VITA

Timothy Sturtz is a PhD Candidate in the Civil & Environmental Engineering Department at the University of Washington and an Associate consultant at ENVIRON International Corporation. He obtained a MS in Civil & Environmental Engineering from the University of Washington in 2008 and a BS in Civil & Environmental Engineering with a minor in Mathematical Sciences from Michigan Technological University in 2006.