

**LARGE-SCALE, DYNAMIC, MICROSCOPIC SIMULATION FOR
REGION-WIDE LINE SOURCE DISPERSION MODELING**

A Dissertation
Presented to
The Academic Faculty

by

Daejin Kim

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in the
School of Civil and Environmental Engineering

Georgia Institute of Technology
MAY 2020

COPYRIGHT © 2020 BY DAEJIN KIM

**LARGE-SCALE, DYNAMIC, MICROSCOPIC SIMULATION FOR
REGION-WIDE LINE SOURCE DISPERSION MODELING**

Approved by:

Dr. Randall L. Guensler, Advisor
School of Civil and Environmental
Engineering
Georgia Institute of Technology

Dr. Angshuman Guin
School of Civil and Environmental
Engineering
Georgia Institute of Technology

Dr. Haobing Liu, Co-advisor
School of Civil and Environmental
Engineering
Georgia Institute of Technology

Dr. Fang (Cherry) Liu
College of Computing
Georgia Institute of Technology

Dr. Michael O. Rodgers
School of Civil and Environmental
Engineering
Georgia Institute of Technology

Dr. Catherine Ross
School of City and Regional Planning
Georgia Institute of Technology

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

Date Approved: [March 25, 2020]

ACKNOWLEDGEMENTS

First and foremost, I would like to express my deepest gratitude to my advisors, Dr. Randall Guensler and Dr. Haobing Liu, for their continuous guidance, encouragement, support and understanding. I am also the most grateful for the advice and comments from the other members of my dissertation committee: Drs. Michael O. Rodgers, Angshuman Guin, James A. Mulholland, Fang (Cherry) Liu, and Catherine Ross. Their patience and flexibility have given me the opportunity to grow and succeed, and prepare me for life after graduation. I also wish to thank Dr. Joonho Ko who has always supported me and been an exceptional mentor in my life. I would also like to express appreciation to my professors at University of Seoul and Seoul National University: Drs. Young Ihn Lee, Sung Soo Kim, Justin S. Chang, Dong Joo Park, and Do Gyeong Kim.

I would also like to give my special thanks to my friends I have made while at Georgia Tech: Jongchan, Seongho, Koochul, Hyun Woong, Sung Hoo, Sungtaek, Ali, Geoffrey (Bo), Haobing, Xiaodan, Yingping, Ann, Hongyu, Abihilasha, to name a few, as well as all of the Dokdo United soccer team members. Their friendship and company have filled my doctoral experience with joy. I also wish to thank my best colleagues, Daisik, Bumsik, Dongjae, and Hyungjoo, whom I made friendship since I met them at Seoul National University. This work was sponsored by the USDOT's UTC program via the National Center for Sustainable Transportation.

I wish thank all of my family, my father Yonggyu Kim, my mother Insun Cha, and only sister Jiyeong Kim. Every accomplishment in my life would not be possible without their love and support. I am profoundly thankful to my fiancée, Yujin Park. She has always supported and cared for me since I first met her in April 2012. I also congratulate on her successful completion of her forthcoming doctorate degree, and wish her success.

TABLE OF CONTENTS

ACKNOWLEDGEMENTS	iii
LIST OF TABLES	vii
LIST OF FIGURES	viii
LIST OF SYMBOLS AND ABBREVIATIONS	x
SUMMARY	xii
CHAPTER 1. INTRODUCTION	1
1.1 Background	1
1.2 Research Objectives	3
1.3 Dissertation Outline	6
CHAPTER 2. LITERATURE REVIEW	8
2.1 Dispersion Models	8
2.1.1 AERMOD	14
2.1.2 CALINE	16
2.1.3 CAL3QHC/CAL3QHCR	17
2.1.4 R-LINE	18
2.1.5 ADMS	19
2.1.6 Summary	19
2.2 Regional-level Microscale Dispersion Modeling	21
2.3 Spatial and Temporal Resolution for Dispersion Modeling	24
2.4 Practical Modeling Implementation for Regional-level Microscale Dispersion Modelling	26
2.4.1 Integration of Traffic Data into Dispersion Modelling	26
2.4.2 Emission Source Type	28
2.4.3 Receptor Siting	29
2.4.4 Meteorological Inputs	31
2.4.5 Model Calibration	32
2.5 Summary of Regional-Level Microscale Dispersion Modeling Issues	34
CHAPTER 3. MODELING OVERVIEW	36
CHAPTER 4. DYNAMIC GRID RECEPTOR METHOD FOR REGIONAL-LEVEL NEAR-ROAD AIR QUALITY ANALYSIS	39
4.1 Introduction	39
4.2 Model Development Process	40
4.2.1 Hypothetical Network	40
4.2.2 Simulation Inputs	43
4.2.3 R-LINE Simulation	44
4.3 Simulation-informed Fundamental Relationship among Predicted Pollutant Concentration Profile and R-LINE Input Parameters	48

4.4	Step-wise Searching Method for Optimal Receptor Locations	53
4.5	Dynamic grid-receptor Model Development	58
4.5.1	Number of Optimal Receptor Model	59
4.5.2	Optimal Receptor Location Model	61
4.6	Model Verification	65
4.7	Case Studies	68
4.7.1	Practical Implementation of Dynamic grid-receptor Model to Transportation Network	75
4.7.2	Comparison of Dynamic grid-receptor Model and Static-grid-receptor Model	78
4.7.3	Dynamic Grid-Receptor Model vs. Low-density Receptor Model	86
4.8	Chapter Summary	92
CHAPTER 5. DEVELOPMENT OF ROADWAY LINK SCREENING MODEL		
	95	
5.1	Introduction	95
5.2	Model Development	101
5.2.1	Data	102
5.2.2	Variable Selection	110
5.2.3	Model Development	119
5.2.4	Model Results	121
5.3	Case Study	126
5.4	Summary of Findings	135
CHAPTER 6. STREAMLINED DATA PROCESSING FOR REGIONAL-LEVEL MICROSCALE DISPERSION MODELING		
	138	
6.1	Introduction	138
6.2	Modeling Framework	139
6.3	Data Preparation	142
6.3.1	Scenario Set-up	142
6.3.2	MOVES-Matrix	143
6.3.3	Geographic Input Data	145
6.3.4	Meteorology Data	153
6.3.5	Others	155
6.4	Emissions Inventories	157
6.4.1	Integration of MOVES-Matrix and Regional Travel Demand Model	158
6.4.2	Emissions Calculation for Dispersion Modelling	164
6.5	Link Screening	165
6.6	Dispersion Modeling	166
6.6.1	Screening Dispersion Modeling	166
6.6.2	Standard Dispersion Modeling	168
6.7	Distributed Computing Strategy	169
6.8	Case Study	171
6.9	Chapter Summary	178
CHAPTER 7. CONCLUSIONS AND FUTURE RESEARCH		
	181	
7.1	Regional-Level Microscale Dispersion Modeling Processes	181
7.2	Theoretical Contributions	184

7.3	Practical Contributions	186
7.4	Limitations and Future Research	188
7.5	Publication Outline	195
APPENDIX A. LOGISTIC REGRESSION MODEL RESULTS FOR DYNAMIC GRID-RECEPTOR MODEL		197
APPENDIX B. THE DEPLOYMENT STRATEGY OF MODELING SYSTEM ON AMAZON WEB SERVICES		207
REFERENCES		210

LIST OF TABLES

Table 1 – Pasquill Stability Categories	12
Table 2 – Ranges of Input Parameters for R-LINE Simulations	44
Table 3 – Linear Regression Model for Optimal Number of Receptors.....	61
Table 4 – Comparison Analysis Results between Dense-receptor vs. Dynamic-grid/Static-grid Receptor Models.....	68
Table 5 – Comparison Results of PM _{2.5} Concentration Profiles among High-density Receptor, Dynamic grid-receptor, and Static-grid-receptor Models	86
Table 6 – Comparison Results of PM _{2.5} Concentration Profiles among High-density Receptor, Dynamic grid-receptor, and Low-density grid-receptor Models.....	92
Table 7 – Descriptive Statistics of Meteorological Variables	107
Table 8 – Results of Variable Selection Models.....	115
Table 9 – Overall Prediction Accuracy of Link Screening Models.....	123
Table 10 – AERMOD Simulation Results.....	134
Table 11 – Typical Urban Boundary Layer Parameters	156
Table 12 – MOVES2014 On-road Source Types	163
Table 13 – Expected and Actual AERMOD Runtimes for Metro Atlanta (Standard Modeling).....	177
Table 14 - Logistic Regression Model Results for Each Candidate Receptor Site	198
Table 15 - Amazon EC2 On-Demand Pricing as of April, 2020: Compute Optimized & Linux/Unix Services	209

LIST OF FIGURES

Figure 1 – Concept of Gaussian Plume Dispersion	11
Figure 2 – Element Series Represented by Series of Finite Line Sources.....	13
Figure 3 - Stability Classes (A) σ_y and (B) σ_z	14
Figure 4 – Representation of Roads in AERMOD	16
Figure 5 – Overview of Modeling Processes.....	38
Figure 6 – Examples of Hypothetical Networks.....	41
Figure 7 – ARC-ABM Network’s Link Length Distribution	42
Figure 8 – Examples of PM _{2.5} Concentration Profiles ($\mu\text{g}/\text{m}^3$) Predicted by R-LINE under Varying Conditions.....	47
Figure 9 – Sensitivity Results of PM _{2.5} Concentration Profile ($\mu\text{g}/\text{m}^3$) to Wind Direction	50
Figure 10 – Sensitivity Results of PM _{2.5} Concentration Profile ($\mu\text{g}/\text{m}^3$) to Wind Speed. 51	51
Figure 11 – Sensitivity Results of PM _{2.5} Concentration Profile ($\mu\text{g}/\text{m}^3$) to Link Emission	52
Figure 12 – The Example of Step-wise Search for Selecting Optimal Receptor Locations	58
Figure 13 – Development of the Dynamic Grid-Receptor Model	59
Figure 14 – Dynamic Grid-receptor Model Development.....	64
Figure 15 – Examples of Model Verification Processes.....	67
Figure 16 – Case Study Areas for Dynamic Grid-Receptor Model Evaluation	70
Figure 17 – Receptor Settings for the City of Decatur	72
Figure 18 - Receptor Settings for Midtown Atlanta	73
Figure 19 - Receptor Settings for Downtown Atlanta	74
Figure 20 – Dynamic Grid-Receptor Model Implementation for the ARC-ABM15 Network.....	77
Figure 21 – Dynamic Grid-Receptor Model vs. Static Grid-Receptor Model (City of Decatur, GA).....	80
Figure 22 – Dynamic Grid-Receptor Model vs. Static Grid-Receptor Model (Midtown Atlanta, GA).....	81
Figure 23 – Dynamic Grid-Receptor Model vs. Static Grid-Receptor Model (Downtown Atlanta, GA).....	82
Figure 24 – PM _{2.5} Concentration Profiles for High-Density Receptor, Dynamic Grid- Receptor, and Static-Grid-Receptor Models.....	84
Figure 25 – Dynamic Grid Receptor Model vs. Low-density Grid-Receptor Model (City of Decatur, GA)	88
Figure 26 – Dynamic Grid-Receptor Model vs. Low-density Grid-Receptor Model (Midtown Atlanta, GA).....	89
Figure 27 – Dynamic Grid-Receptor Model vs. Low-density Grid-Receptor Model (Downtown Atlanta)	90
Figure 28 – Examples of AERMOD Processing Time by Number of Links	99
Figure 29 – Examples of AERMOD Processing Time by Number of Receptors	100

Figure 30 – Estimated Hourly Average PM _{2.5} Mass Flux (g/m ² /hour) for the Metropolitan Atlanta Area	105
Figure 31 – Location of the Significant Links for a Receptor as Identified by Random Forest.....	126
Figure 32 – Comparison of Annual Average PM _{2.5} Concentrations by Different Model Specifications	129
Figure 33 – Predicted Annual Average PM _{2.5} Concentration Profile of Study Area.....	131
Figure 34 – Proposed Modeling Framework for Regional-level Microscale Dispersion Analysis.....	141
Figure 35 – Example of MOVES CO ₂ Emission Rates by VSP Bin for Passenger Trucks (Model Year of 2016 in 2016)	145
Figure 36 – Example of Generating Link Geometry Data using GIS-Polyline.....	147
Figure 37 – Impact of Road Grade on Pollutant Concentrations (Downtown Atlanta) .	149
Figure 38 – Annual Average PM _{2.5} Concentrations from Geometry-based and Gridded Receptor Models	152
Figure 39 – Examples of Wind Rose Diagrams in Two Locations in Georgia, USA: Calendar Year of 2018	154
Figure 40 – Emissions Calculation Process	158
Figure 41 – ARC-ABM’s Daily Link-by-Link Total Traffic Volumes in 2015.....	160
Figure 42 – ARC-ABM’s Link-by-Link Congested Traffic Speeds in 2015	161
Figure 43 – MOVES-Matrix Run Module: Developing On-Road Fleet Emission Rates	165
Figure 44 – Distributed Modeling Process for Dispersion Modeling	170
Figure 45 – Temporal and Meteorological Factors Considered in Emissions Calculations	172
Figure 46 – Sensitivity Analysis of the Total PM _{2.5} Emissions in Metro Atlanta	174
Figure 47 – Estimated PM _{2.5} Concentration Results: (a) Screening model, (b) Standard Model	176

LIST OF SYMBOLS AND ABBREVIATIONS

ABM	Activity-based model
AdaBoost	Adaptive boosting
ADMS	Atmospheric dispersion modelling system
AERSCREEN	Program to run AERMOD in screening mode
AERMET	Meteorological data preprocessor for AERMOD
AERMINUTE	Pre-processor to AERMET to read 1-minute ASOS data to calculate hourly average winds for input into AERMET
AERMOD	American meteorological society/environmental protection agency regulatory model
AERSURFACE	Land cover data tool in AERMET
AMS	American meteorological society
ARC-ABM15	Atlanta regional commission activity-based model 2015
AREES	Atlanta roadside emissions exposure study
ASOS	Automated surface observing systems
ARC	Atlanta regional commission
AWS	Amazon web services
CAL3QHC	Screening version of the CALINE3 model
CAL3QHCR	Refined version of the CALINE 3 model
CALINE	California line source dispersion model
CERC	Cambridge environmental research consultants
CO	Carbon monoxide
DEM	Digital elevation model
DNR	Department of natural resources
DPM	Diesel particulate matter
EPD	Environmental Protection Division
GA	Georgia
GIS	Geographic information system

GUI	Graphical user interface
HEI	Health effects institute
HPC	High-performance computing
LR	Logistic regression
MAE	Mean absolute error
MAPE	Mean absolute percentage error.
MOVES	Motor vehicle emission simulator
MSE	Mean squared error
NAAQS	National ambient air quality standards
NEPA	National environmental policy act
NN	Neural network
NO	Nitric oxide
OA	Overall accuracy
PACE	Partnership for an advanced computing environment
PM	Particulate matter
RF	Random forest
R-LINE	A research line source model for near-surface releases
RT	Regression tree
SLS	Supervised link screening
SVM	Support vector machine
TDM	Travel demand model
USEPA	U.S. Environmental Protection Agency
USGS	United States geological survey
UTM	Universal Transverse Mercator
VMT	Vehicle miles travelled
VOCs	Volatile organic compounds

SUMMARY

Since the 1980's, researchers have developed a variety of air pollution microscale dispersion models to predict potential public exposure to harmful transport emissions immediately downwind from freeways and arterials. However, when these tools are applied at regional and sub-regional scales, critical concerns arise due to computational efficiency. Current dispersion models require extremely long runtimes for larger roadway networks when high-resolution pollutant concentrations are needed in space and time (e.g., hourly predictions for thousands of receptor sites). Motivated by the challenges encountered in the previous efforts, this study develops an advanced modeling framework for a regional-level line source dispersion modeling that employs several innovative modeling techniques.

The first part of this dissertation proposes a strategic receptor placement method that minimizes the number of receptors across a region, without undermining the pollutant high-resolution concentration profile generated with dense receptor placement. A dynamic-receptor-grid approach places receptor locations with respect to each link's geometry and emissions characteristics. The modeling results suggest that the optimal receptor placement using the dynamic receptor-grid models readily approximates the $PM_{2.5}$ concentration profiles predicted by dense-receptor-grid models.

As an innovative way to significantly improve modeling efficiency, the second part of this dissertation develops a methodology to screen roadway link sources from the analyses that do not significantly contribute to pollutant concentrations at a particular receptor, based upon mass flux from the roadway and distance to receptor (i.e., methods to exclude insignificant link-receptor pairs from the analyses). The link screening models are

developed by a supervised machine learning Random Forest (RF) classification model with a parsimonious variable selection process. The link screening model identifies roadway links that do not contribute significantly to receptor concentration with a high classification accuracy (greater than 95%), and significantly reduces the total dispersion modeling run-time (98.9% - 99.8% reduction in model run-time).

The third part of this dissertation proposes advanced techniques to efficiently prepare the extensive input datasets needed for line source modeling over entire metropolitan areas and addresses the technical issues associated with expanding the line source dispersion model from project-level to regional-level. For example, this study proposes the integration of MOVES-Matrix and regional travel demand models (TDMs) to efficiently estimate the link emissions at regional-scale, and demonstrates how to integrate road grade profiles generated by U.S. Geological Survey Digital Elevation Model. The research also demonstrates how to streamline data processing through a distributed computing cluster to boost modeling speed.

Lastly, a case study for the 20-county metropolitan Atlanta area, which accounts for an extremely large number of link-receptor pairs (161,188 links and 1,163,958 receptors), demonstrates that the developed modeling system generates reliable pollutant concentration estimates with high computational efficiency. The total processing time for running AERMOD was only 10 days which is a substantial improvement over traditional methods (which would take more than one year). The improvement in dispersion modeling efficiency is attributed to 1) the dynamic grid-receptor model that minimizes the number of receptors across a region, 2) the employment of link screening model that objectively excludes all irrelevant link sources that do not significantly affect the pollutant

concentration of each receptor, 3) the streamlined connection between regional TDM and MOVES-Matrix in the emissions calculation, and 4) the use of PACE clustering system, where multiple AERMOD simulation jobs are split and simultaneously processed to reduce total run-time.

CHAPTER 1. INTRODUCTION

1.1 Background

Clean Air Act Section 176(c) (42 U.S.C. 7506(c)), otherwise known as transportation and air quality conformity, requires all federal transportation projects to be properly incorporated into state/regional air quality planning. That is, transportation programs, plans, and projects must conform with statewide air quality management plans (known as state implementation plans (SIPs)). The Clean Air Act and U.S. Environmental Protection Agency (USEPA) regulations for transportation and air quality conformity require that federal transportation projects will not cause new air quality violations, worsen existing violations, or delay attainment of the National Ambient Air Quality Standards (NAAQS) (USEPA, 2006). In 2015, the USEPA published transportation conformity guidance for hot-spot analysis in particulate matter (PM) and carbon monoxide (CO) nonattainment and maintenance areas (USEPA, 2015a). The hot-spot analysis aims to compare future local pollutant concentrations with current pollutant concentrations and the NAAQS, and requires detailed modeling of the impacts of transportation project emission sources on the surrounding environment (i.e., downwind from transportation facilities) using microscale dispersion modeling.

The conformity guidance requires that the MOVES (Motor Vehicle Emission Simulator) model be used to estimate mobile source emissions for use in air quality analysis (USEPA, 2015a), and that the U.S. States should use MOVES to develop their SIPs (USEPA, 2014). The USEPA also recommends multiple models including AERMOD, CALINE3, CAL3QHCR, etc. for use in air quality dispersion modeling (USEPA, 2019a).

Because California operates a different on-road vehicle fleet, due to more stringent new vehicle certification standards in California, the USEPA guidance for California requires the use of the EMFAC emission rate model (Caltrans, 2019).

CALINE4 is currently used for CO hot-spot analysis in California (Caltrans, 2019), while USEPA's preferred and recommended model (i.e., AERMOD) should be used for analyzing other pollutants (USEPA, 2015a). R-LINE, which is specifically designed to simulate line-type source emissions and address previous issues with pollutant concentrations at low wind speeds and when wind runs parallel to roadways (Snyder, et al., 2013a), was developed by USEPA to provide "analytical solution" for quicker simulation under acceptable ranges of errors (Snyder and Heist 2013b). R-LINE is expected to supersede or be integrated into AERMOD (USEPA, 2019b).

Because emissions are a complex function of many local conditions (i.e., local fleet, operations, fuel supply, and I/M strategy), the interface is complex and requires numerous inputs to properly characterize any specific emission scenario modeled by a user. Thus, significant labor is required to prepare MOVES input files. In addition, running MOVES is time-consuming, because emission calculations always begin with base emission rates that are internally adjusted by various correction factors such as temperature, humidity, and fuel properties. Hence, MOVES is difficult to use for large-scale transportation networks that experience dynamic changes in on-road fleet composition and operating conditions that affect model outputs (e.g., impacts of speed and acceleration, temperature).

Similarly, complex setup procedures for dispersion modeling also have a high potential for introducing analytical error. Vallamsundar and Lin (2012) showcased the use

of MOVES and AERMOD for transportation conformity analysis in accordance with EPA's guidance as an example for others to better understand the proposed process of conformity analysis and model setup. The case study conducted by Vallamsundar and Lin (2012) showed that concentrations were found to be high 1) near the sources and gradually decreased as the distance from the source increases, 2) at locations where the traffic volumes are high, and 3) at locations where prevailing winds are blowing. Wu and Niemeier (2016) highlighted potential issues associated with excluding zones and improperly setting up receptors.¹ The dynamics and fluctuation of traffic flow and meteorology suggested a need for a better connection between traffic, emission rate, and dispersion models. Hence a systematic and automatic process for microscale line-source dispersion modeling for air quality analysis seems a reasonable goal. An automated method coupled with iterative modeling efforts to assess sensitivity analysis, should help to identify modeling uncertainties that arise from the dynamic nature of traffic and the near-roadway atmospheric environment.

1.2 Research Objectives

Motivated by the need to quickly provide fleet emission rates for the thousands of modeling scenarios required to assess model performance, this dissertation research uses MOVES-Matrix (Guensler, et al., 2016), a high-performance emission rate lookup system, with the USEPA recommended dispersion models for use in a large-scale air quality impact analysis. The primary goal of using MOVES-Matrix is to increase modeling efficiency and

¹ The term "receptor" refers to the geographic location in which pollutions emitted from emission source (e.g., roadway, plant) are concentrated.

speed while ensuring that the resulting pollutant concentration results are exactly the same as when MOVES is applied directly to the dispersion model for each link-receptor pair.

MOVES-Matrix is essentially a multidimensional array containing emission rate outputs from tens of thousands of MOVES model runs conducted on the Partnership for an Advanced Computing Environment (PACE) high-performance computing (HPC) cluster. MOVES is a USEPA's motor vehicle emission simulator, which is a state-of-the-science emission modeling system to estimate emissions for mobile sources at the national, county, and project levels for criteria air pollutants, greenhouse gases, and air toxics (USEPA, 2019c). MOVES-Matrix generates exactly the same emission rates as running MOVES for each application, but using MOVES-Matrix is more than 200 times faster than implementing the MOVES model, on a case-by-case basis (Guensler, et al., 2016). MOVES-Matrix has proven its capability in analyzing project-level air quality impact assessments with efficiency, by connecting MOVES-Matrix outputs with multiple air dispersion models such as AERMOD and CALINE4 (Liu et al., 2017; Xu, et al., 2016; Xu, et al., 2018a; Kim, et al., 2019a; Kim, et al., 2019b).

Because the transportation and air quality conformity requires a demonstration that future transportation plans (and all of the incorporated projects) will not violate the NAASQ, planners and analysts have sought methods to demonstrate conformity in a single process. That is, if planners can demonstrate that no NAAQS violations will occur anywhere in the region when all of the projects in the 2040 transportation plan are integrated into the transportation network, case-by-case project-level assessments will no longer need to be conducted (and staff time associated with undertaking such planning assessments would be significantly reduced). However, modeling the microscale

dispersion impacts associated with all of the roadways across an entire region on all potential receptor locations within a region (known as regional-level microscale dispersion modeling) has been computationally challenging, given the huge number of link source and receptor combinations. For example, a preliminary study conducted for major interstate highways in the Atlanta metropolitan area, Georgia, USA, showed that predicting $PM_{2.5}$ concentrations for a total of 1,163 miles of road networks and 57,017 receptors would take around 600 days on a regular desktop computer through AERMOD simulation. (Kim, et al. 2019c). Distributed computing can significantly reduce computational time. For example, the study by Kim, et al. (2019c) showed that the AERMOD simulation through a distributed computing cluster (Georgia Tech's PACE) could reduce the total simulation time to 7 days (around 85-times faster than using a single desktop computer). However, the receptor grid employed in this study was somewhat sparse, and future research still remains to enhance model performance and efficiency when conducting air quality dispersion modeling for large roadway networks (freeways and arterials).

This dissertation research aims to develop a regional-level microscale dispersion modeling system, which 1) improves the efficiency of the present modeling system regarding modeling speed, while preserving the spatial resolution of model outputs, 2) develops a streamlined program to automate the connection among travel demand model outputs, emissions modeling (through MOVES-Matrix), and dispersion modeling, 3) builds a user-friendly interface with minimal data input requirements. In developing the modeling system, the dissertation research will connect MOVES-Matrix directly with ARC-ABM15 (Atlanta Regional Commission Activity-Based Model 2015) travel demand

model outputs and then with dispersion models as a case study, to evaluate model performance.

1.3 Dissertation Outline

The chapters in the dissertation are organized as follows. Chapter 2 provides a literature review on various topics related to the development of a regional-level microscale dispersion modeling system, including the review of line source dispersion models, the previous modeling frameworks for regional-level traffic-related air quality analysis, and spatial and temporal resolution requirements for modeling input data.

Followed by Chapter 3 that addresses the modeling overview, Chapter 4 proposes a strategic receptor placement method that minimizes the number of receptors across a region, without undermining the pollutant high-resolution concentration profile generated with dense receptor placement. A dynamic grid-receptor approach identifies optimal receptor locations with respect to each link's geometry and emissions characteristics.

Chapter 5 introduces a state-of-the-art link screening model that significantly increases the modeling efficiency, which is developed as part of this dissertation research. The link screening model, developed based on a supervised machine learning classification algorithm (random forest), aims to identify and remove roadway links with negligible concentration contributions for each receptor with high precision. The new link screening method is also applied to the metropolitan Atlanta area to demonstrate the promising performance of supervised link screening for regional-scale applications of microscale dispersion modeling.

Chapter 6 addresses the system features of a streamlined processor that automates individual modeling steps ranging from data preparation, travel demand model (TDMs) connectivity, emissions calculations, and dispersion modeling. This chapter measures the model's performance by applying the developed model to the 20-county Metropolitan Atlanta Region in predicting $PM_{2.5}$ concentrations across the area, as a case study. The case study uses MOVES-Matrix emission rates coupled with outputs from the ARC-ABM15, and the pollutant concentrations were predicted by AERMOD.

The last chapter concludes the dissertation, by giving an overview of the contributions and a discussion for future research.

CHAPTER 2. LITERATURE REVIEW

2.1 Dispersion Models

Since the 1980's, a number of line-source microscale dispersion models have been developed to model the downwind impacts of transportation facilities, including AERMOD, CALINE3/4, CAL3QHC/CAL3QHCR, R-LINE, and ADMS, are available for air quality dispersion modeling for traffic-related emissions (USEPA, 2015b; USEPA, 2019a). Over time, each of these models evolved to incorporate as research provided a better understanding of theoretical background, impacts of formulation, results from model application, and assessment of comparative model performance. All of these dispersion models are capable of predicting concentrations at selected downwind receptor locations, based on existing data on emissions and meteorological and topologic conditions (Bellander, et al., 2001).

Line source microscale dispersion models are generally based upon a generalized Gaussian plume equation (Bellander, et al. 2001), where pollutants dispersing in the y and z axes via normal distributions as a function of distance along the x-axis, as described in Equation (1) and Figure 1 (Turner, 1994; Vallero, 2014; Hanna, et al., 1984). As such, air pollutant concentration at specific receptor site is basically a function of pollutant emission rate from source, and the stability of atmosphere which is depicted by wind speed and horizontal and vertical deviation of the emission distribution, in the Gaussian plume-based modeling framework.

$$C(x, y, z; H) = \frac{Q}{2\pi u \sigma_y \sigma_z} \cdot \exp\left[-\frac{y^2}{2\sigma_y^2}\right] \left\{ \exp\left[-\frac{(H-z)^2}{2\sigma_z^2}\right] + \exp\left[-\frac{(H+z)^2}{2\sigma_z^2}\right] \right\} \quad (1)$$

Here,

C = air pollutant concentration in mass per volume, in g/m^3 ,

Q = source pollutant emission rate, in g/s ,

u = wind speed at the point of release, in m/s ,

σ_y = horizontal standard deviation of the emission distribution (in m), that is, standard deviation of horizontal distribution of plume concentration, which is evaluated at the downwind distance x and for the appropriate stability

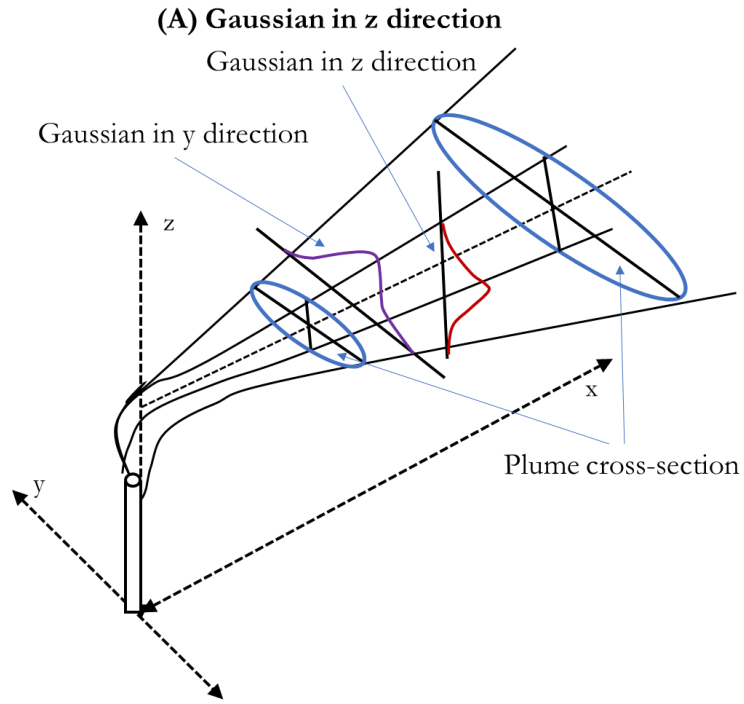
σ_z = vertical standard deviation of the emission distribution (in m), that is, standard deviation of vertical distribution of plume concentration, evaluated at the downwind distance x and for the appropriate stability, and

H = height of emission plume centreline above ground level, in m.

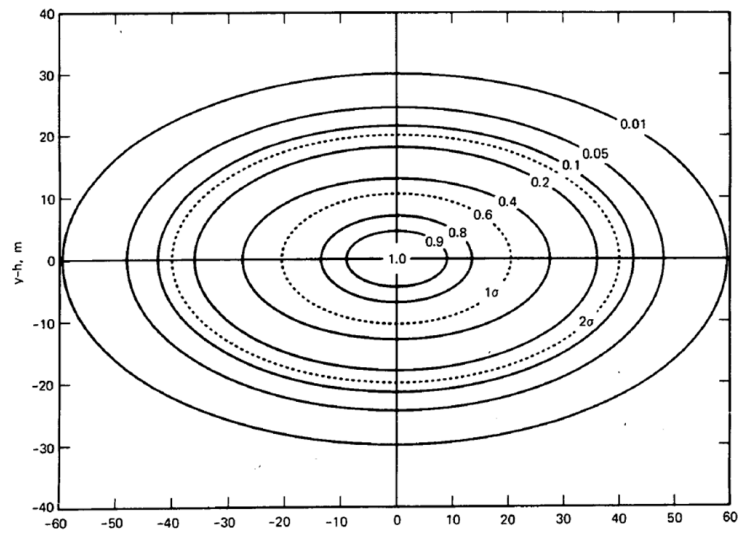
Some line source dispersion models including CALINE, AERMOD, and R-LINE assumed that a line source is structured as a finite number of point sources. For example, CALINE4 models each line source as an equivalent finite line element positioned normal to the wind direction and centered at the element midpoint (Figure 2), and the contributions from each element are summed to predict net concentration. Similar approaches were also applied to R-LINE and AERMOD where the concentration from a finite line source is estimated by approximating the line as a series of point sources (Snyder, et al., 2013a; Cimorelli, et al., 2005). As such, those models for dispersion of roadway emissions are analytical approximations to the integral associated with modeling a line source as a set of

point sources. Then, the contributions along the line source are computed with the Gaussian, steady-state plume formulation.

For those dispersion models, an estimate of σ_y or σ_z in the Gaussian plume formulation (equation (1)) is often explained by atmospheric fluctuation measurements (Pasquill and Michael 1977). In the early stage of air pollution research, Pasquill's dispersion parameters restated by Gifford (1961) were introduced, by allowing their use in the Gaussian plume equations, where the parameters σ_y and σ_z are found by estimation from the graphs (Figure 3) as a function of the distance between source and receptor, from the appropriate curve, one for each stability class. For this, the necessary parameters for the scheme consist of wind speed, insolation, and cloudiness, which are basically obtainable from routine observations (Table 1) (Pasquill, 1976).



(B) Cross-section of a Gaussian Plume with $\sigma_y=20\text{m}$, $\sigma_z=10\text{m}$, and centerline concentration of 1.0



Note: The figure on top is amended based on Vallero (2014), and the figure on the bottom is retrieved from Hanna, et al. (1984).

Figure 1 – Concept of Gaussian Plume Dispersion

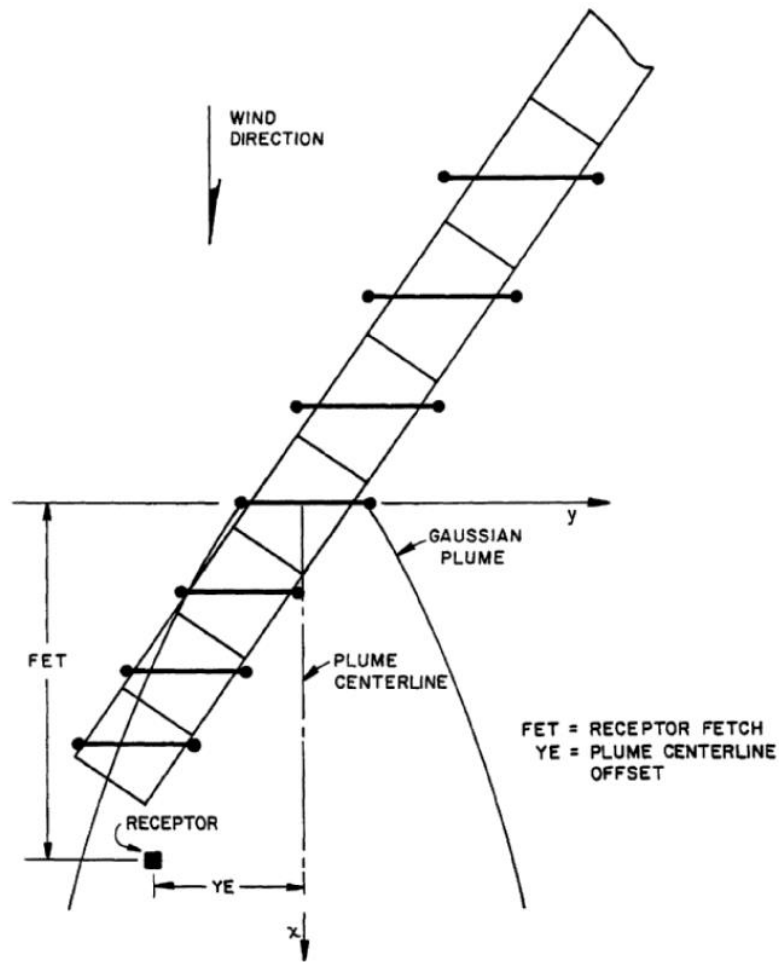
Table 1 – Pasquill Stability Categories

Surface wind speed (m/s)	Isolation			Night	
	Strong	Moderate	Slight	Thinly overcast or ≥ 4/8 low cloud	≤ 3/8 Cloud
<2	A	A-B	B	-	-
2–3	A-B	B	C	E	F
3–5	B	B-C	C	D	E
5–6	C	C-D	D	D	D
>6	C	D	D	D	D

Source: Pasquill (1976).

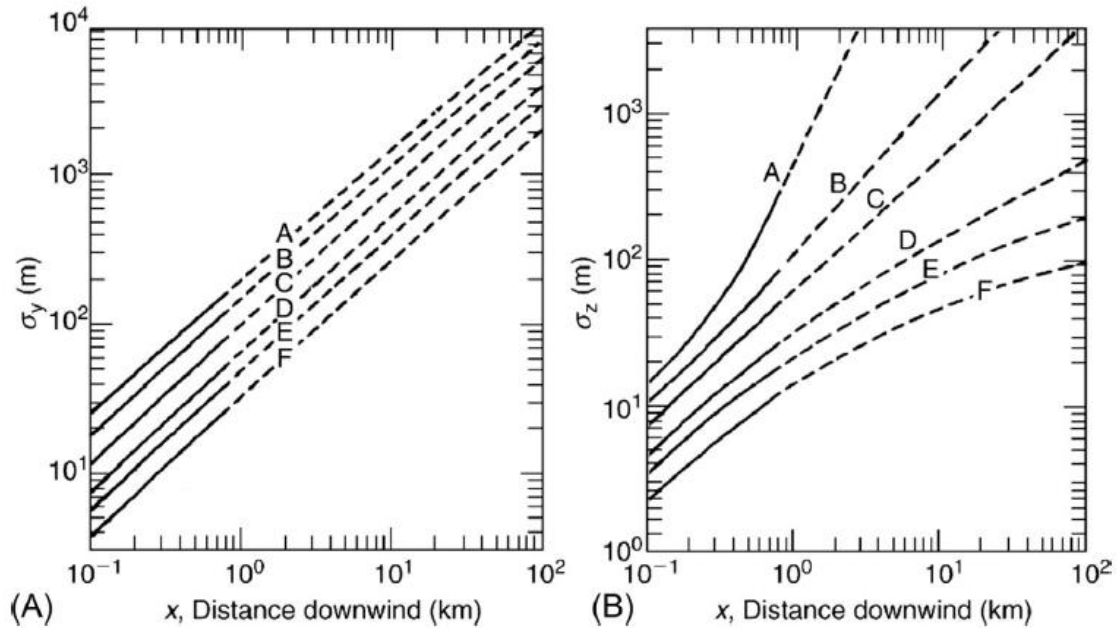
Over time, line source dispersion models have distinguished in terms of the estimation of horizontal and vertical spread parameters. For example, CALINE4 uses a modified version of the Pasquill-Smith vertical dispersion curves (Pasquill, 1974) to describe the Gaussian vertical dispersion parameter, σ_z , and uses a method developed by Draxler (Draxler, 1976) to compute values for the Gaussian horizontal dispersion parameter, σ_y . AERMOD has evolved with a similar theoretical background of the Gaussian formulation, while concerns regarding plume interaction with terrain, surface releases, building downwash, and urban dispersion were addressed in AERMOD to improve planetary boundary layer (PBL) parameterizations (Cimorelli, et al., 2005). In this regard, AERMOD characterizes the boundary layer with computation of the Monin–Obukhov length, surface friction velocity, surface roughness length, sensible heat flux, convective scaling velocity, and both the shear- and convection-driven mixing heights. These parameters are used in conjunction with meteorological measurements to

characterize the vertical structure of the wind, temperature, and turbulence. R-LINE has evolved based on the CALINE3 and AERMOD formulations, while also incorporating some new features that distinguish it from other models (Snyder, et al., 2013a). R-LINE includes new formulations for the vertical and horizontal plume spread of near-surface releases based on historical field data (Barad, 1958) as well as a tracer field study (Finn, et al., 2010) and recent wind tunnel studies (Heist, et al., 2009). Also, R-LINE contains a wind meander algorithm that accounts for dispersion in all directions during light and variable winds.



Source: Benson (1984).

Figure 2 – Element Series Represented by Series of Finite Line Sources



Source: Gifford (1961).

Figure 3 - Stability Classes (A) σ_y and (B) σ_z

2.1.1 AERMOD

AERMOD (American Meteorological Society (AMS)/EPA Regulatory Model) is the U.S. EPA's preferred and recommended steady-state plume model for use in demonstrating compliance with environmental regulatory programs, including transportation facility line sources (USEPA, 2019b). The USEPA recommends using either AERMOD or CAL3QHCR for highway and intersection projects, but using only AERMOD for transit, freight, and terminal projects and projects that involve highway intersections and terminals, nearby sources, or both (USEPA, 2015a). AERMOD was developed as a replacement for EPA's industrial source complex model (i.e., ISC3) by incorporating the planetary boundary layer, which is the turbulent air layer next to the earth's surface affected by surface heating, drag, turbulence, and friction because of its contact with the planetary

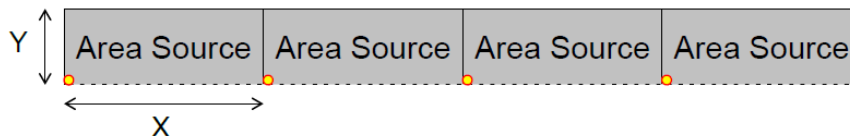
surface (USEPA, 2019b). There are two types of planetary boundary layer: convective boundary layer, driven by surface heating, and stable boundary layer, driven by surface cooling. AERMOD utilizes a Gaussian distribution in both horizontal and vertical directions in the stable boundary layer, but in the convective boundary layer, AERMOD uses a Gaussian distribution in the horizontal direction and a bi-Gaussian one in the vertical direction; the AERMOD concentration is calculated as a weighted average of two distributions in the convective boundary layer (USEPA, 2019d).

While AERMOD is a comprehensive model for various types of sources including point, area, line, and volume source (USEPA, 2019b), USEPA (2018) recommends modeling the roadway line source as a series of volume or area sources when using AERMOD (Figure 4). AERMOD is capable of modeling a number of sources and receptors, processing multiple years of meteorological data simultaneously, and providing the option of varying emission factors by different time scales, such as by season, month, and hour-of-day. (USEPA, 2019b).

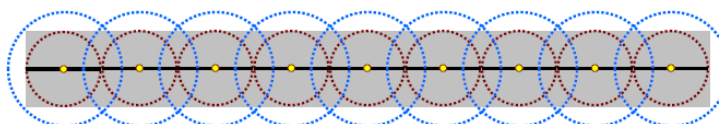
A road is a **line source**:



That can be modeled as a series of **area sources**:



Or as a series of **volume sources**:



Source: USEPA (2018).

Figure 4 – Representation of Roads in AERMOD

2.1.2 CALINE

CALINE4 is a line source dispersion model based on the Gaussian diffusion equation, developed by the California Department of Transportation for estimating air pollution levels within 500 meters of roadways (Benson, 1984; 1992). CALINE4 represents a line source as a series of finite length elements each oriented perpendicular to the wind, and it employs a mixing zone concept to characterize pollutant dispersion over the roadway. To improve computational efficiency, the length of each element is determined based on its distance from the receptor of interest (the further the element is from the perpendicular line from source to receptor, the larger the element, because any error associated with the larger element size will have very little impact on predicted receptor concentrations). CALINE uses Pasquill Gifford categories to characterize the stability of the atmosphere and uses a modified version of the Pasquill-Smith vertical dispersion curves (Benson, 1982) and horizontal dispersion estimates based on Draxler (1976). CALINE4 has algorithms to

model the effects of the mixing zone over the roadway and certain aspects of roadway geometry such as depressed and elevated sections, among others. CALINE4 can also be used in screening analysis, if permission is obtained from the USEPA Regional Office (40 CFR Appendix W).

CALINE3 is a steady-state Gaussian dispersion model designed to determine air pollution concentrations at receptor locations downwind of highways located in relatively uncomplicated terrain (Benson, 1979). CALINE3 is incorporated into the more refined CAL3QHC and CAL3QHCR models for regulatory use in specific roadway applications for CO and PM. The primary differences in the model formulations of CALINE3 and CALINE4 are related to the lateral dispersion curves and the introduction of vehicle-induced turbulence in CALINE4 (Heist, et al., 2013).

2.1.3 CAL3QHC/CAL3QHCR

CAL3QHC is a CALINE3 based CO model with queuing and hot spot calculations and with a traffic model to calculate delays and queuing using highway-capacity methods (the QHC portion of the acronym) that occur at signalized intersections. CAL3QHCR is a more refined model based on CAL3QHC that requires local meteorological data (USEPA, 2019a). Similar to other models, CAL3QHC/CAL3QHCR also uses Gaussian formulation. CAL3QHCR updated CAL3QHC by incorporating daily to seasonal runs, whereas CAL3QHC was designed to process one hour of traffic and meteorological data (Eckhoff and Braverman, 1995). While AERMOD was initially developed for industrial source applications and its use has been extended to highways, CAL3QHCR was specifically developed for highway applications based on research conducted near highways. As stated

earlier, USEPA recommends using either AERMOD or CAL3QHCR for highway and intersection projects, but CAL3QHCR is not recommended for use in transit, freight, and terminal projects and projects that involve highway intersections and terminals, nearby sources, or both (USEPA, 2015a). CAL3QHCR remains the model of choice by many State Departments of Transportation because of: 1) their familiarity with the CALINE series of models; 2) consistency with other dispersion modeling conducted in the highway air quality analysis; 3) the computational efficiency of CAL3QHCR over AERMOD (CAL3QHCR runs approximately six times faster); and 4) CAL3QHCR typically provides lower results – a factor of two lower than AERMOD for some applications (Vallamsundar and Lin, 2012).

2.1.4 R-LINE

R-LINE is a research dispersion modeling tool under development by the US EPA's Office of Research and Development, which is specifically designed to model roadways as line segments (Snyder et al, 2013a; Venkatram et al., 2013). The model uses Gaussian, steady-state plume-dispersion formulation that incorporates new algorithms for predicting concentrations from road sources at receptors near roads by numerically integrating point source emissions. The model uses the surface meteorology provided by the AERMET model (the meteorological pre-processor for AERMOD) and simplified road-link specifications. The model computes concentrations by integrating point sources along a source line and has been formulated for appropriate simulation for receptors very near the source line. The current beta version of the model is designed for flat roadways (no surrounding complexities), though the model framework is designed to accommodate

future algorithms for simulating the near-source effects of complex roadway configurations (noise barriers, depressed roadways, etc.) (Snyder and Heist, 2013b).

2.1.5 ADMS

ADMS (atmospheric dispersion modeling system) is a Gaussian plume air dispersion model for modeling point, line, area and volume source types in a variety of atmospheric conditions, which is developed by Cambridge Environmental Research Consultants (CERC) and the United Kingdom (UK) Meteorological Office (Carruthers et al., 1994; McHugh et al., 1997). It uses the Monin-Obukhov similarity to define the structure of the boundary layer and computes steady-state Gaussian solutions to describe the diffusion of pollutants. For line sources, ADMS decomposes the source into a series of elements whose spacing depends on the source-receptor distance. ADMS-Roads is a version of ADMS, which is specifically developed for simulating traffic sources. In terms of air quality modeling for transportation emissions sources, ADMS-Roads has incorporated new features, including algorithms that account for traffic-produced turbulence, and the presence of roadside noise barriers, and has an integrated street canyon model (CERC, 2019). It also includes modules that account for the spatial variation of terrain height and surface roughness, NO_x and sulfate chemistry, and dry and wet deposition of pollutants.

2.1.6 Summary

The literature review identified that several models are capable of modeling microscale air pollutant dispersion for line source emissions (e.g., mobile sources along roadways) at regional and sub-regional scales. Based on the literature review, this study adopted three models (AERMOD, CALINE4, and R-LINE) into the regional-level microscale dispersion

modeling framework. AERMOD was chosen for the representative model in the model development as it is the USEPA's preferred and recommend model for use in transportation project evaluations under the guidance of project-level conformity and hot-spot analyses. In addition, AERMOD is also one of the well-maintained dispersion models with periodical model updates by the efforts of USEPA (e.g., R-LINE is also integrated into AERMOD beginning with AERMOD version 19191 in 2019).

CALINE4 was also chosen for its model performance in terms of model runtime. Because CALINE4 requires minimal data, such as simplified roadway configuration and meteorological data (e.g., fewer parameters than other models, and a categorized atmospheric stability condition), CALINE4 significantly reduces total model runtime. CALINE4 also tends to over-predict observed concentrations in field studies (Heist, et al., 2013). Thus, CALINE4 may be a viable option for identifying potential hot-spots with the minimum requirement for input data preparation, while minimizing the total model runtime (i.e., can be put into service as a rapid screening tool in hot-spot assessment work).

R-LINE is also a good option for simulating line source emissions. While the theoretical background and model application of R-LINE are quite similar to AERMOD, R-LINE has unique modules for formulating near-surface releases and vertical plume buoyancy. The R-LINE algorithms account for particular roadway configurations: roadside barriers (e.g., noise barriers) and depressed roadways. In addition, the analytical option in R-LINE runs more quickly than any other line source dispersion models, while producing comparable results.

Through this dissertation research, the application of each of these models to the developed modeling system will be addressed. In this regard, this study develops a streamlined data processing model to generate a shared database containing various data sources required for dispersion modeling, and to process the input data to be used in each of these dispersion models.

2.2 Regional-level Microscale Dispersion Modeling

Motor vehicles are one of the major sources of urban air pollution and are increasingly important contributors to carbon dioxide and other greenhouse gases (HEI, 2010). Traffic-related air pollutants are emitted at or near ground level and mostly in urban areas where they can cause locally-elevated concentrations (USEPA, 2008). The short- and long-term human exposure to traffic-related primary air pollutants, such as carbon monoxide (CO) or diesel particulate matter (DPM), can cause adverse health outcomes, including respiratory and heart disease, impaired lung development, premature mortality, and reduced life expectancy (Bernstein, et al., 2004; Tainio, et al., 2016; Brook, et al., 2010, Kampa and Castanas, 2007).

Historically, a number of modeling tools have been developed to predict potential public exposure to harmful transport emissions at the regional and sub-regional scales, using the outputs from line source pollutant dispersion modeling to represent the spatial structure of the dynamics of ambient pollutant concentrations (Guensler, et al., 2000; Guensler, et al., 2008; Kall, et al., 2008; Vallamsundar and Lin, 2012; D'Onofrio, et al., 2016; Wu, 2018; Zhai, et al., 2016; Zhai, et al., 2019; Liu, et al., 2017). A recent body of research on pollutant dispersion modeling has further incorporated USEPA's

recommended emissions (i.e., MOVES) and dispersion models (e.g., AERMOD) for air quality assessment at regional scales in accordance with the U.S. Environmental Protection Agency (USEPA)'s transportation conformity regulations as outlined in the paragraphs that follow.

Vallamsundar and Lin (2012) showcased the use of MOVES and AERMOD for transportation conformity analysis with details about the setup and running of the models with respective data inputs in accordance with the transportation conformity guidance. The study addressed the complexity of emissions modeling using MOVES and dispersion modeling using AERMOD for given traffic data, pointing out the challenges involved with the input data preparation and the importance of the interagency consultation process.

Wu (2018) focused on the health impact assessment of regional transportation system at a fine spatial scale, examining the spatial and demographic variations in human exposure to near-road pollutions. The study developed a modeling system that predicts traffic-related air pollutant concentration levels based on MOVES and AERMOD, by integrating an activity-based model (ABM) developed for the City of Sacramento, California. The analysis applied medium-spatial-resolution data of around 7,000 links representing 1,600 miles of major roadways in the City of Sacramento. To estimate the PM_{2.5} emission rates for each link, the model performed a number of MOVES runs (16 model runs for four seasons and four time periods) to estimate the annual average emission rates. For the AERMOD modeling, the study applied uniform-grid receptor locations at medium-level spatial resolution (100 m to 200 m resolution, depending on the proximity to roadways, resulting in a total of 18,175 receptors). The modeling approaches applied in this study considered annual average traffic conditions and medium-spatial-resolution of

receptor locations appeared to help understand the overall human exposure levels to traffic-related air pollution, while a more high-resolution data (regarding time-dependent traffic and meteorology conditions) is required to estimate varied human exposure levels, which may be more concerned by governmental officials and urban planners.

Atlanta Roadside Emissions Exposure Study (AREES) (D'Onofrio, et al., 2016) is another example that modeled traffic-related air pollutant dispersion at a regional-scale. The study estimated PM_{2.5} concentrations emitted from road traffic sources in the 20-county metropolitan Atlanta area. AREES estimated annual average PM_{2.5} emissions for each link (approximately 25,000 roadway segments in total) through iterative MOVES runs by linking data from Atlanta Regional Commission (ARC)'s 4-step travel demand model to MOVES model. The dispersion modeling was conducted with the R-LINE model. To reduce model run-time, AREES reduced the domain of each individual R-LINE run by dividing the entire 20-county Atlanta area into a number of grid areas, where only a subset of roadway links belonging to each of the grid areas was considered in the R-LINE modeling (which can undermine pollutant concentration predictions in certain areas).

A major obstacle to using spatially-realistic and temporally-realistic data for regional-scale dispersion modeling has been the difficulty of integrating the outputs from more sophisticated modeling tools with the USEPA's regulatory-required and recommended emission rate and dispersion models (MOVES and AERMOD respectively) for transportation conformity and hot-spot analysis in the United States (USEPA, 2015a). The MOVES modeling runs for a regional transportation network can take days to process and are prone to input errors, due to the complex nature of model input parameters (Liu, et al., 2017). Similarly, the complex setup procedures required for AERMOD have a high

potential for introducing analytical errors (Liu, et al., 2017). Even if the model setups are successful, running dispersion models for a large geographic region still poses significant computational challenges, requiring an impractically large processing time. For instance, a case study conducted for major interstate highways in the metropolitan Atlanta area showed that the total run-time for predicting PM_{2.5} concentrations for the 20-county area would take more than one year on a typical desktop computer (Kim, et al., 2019c).

This situation recalls the use of a high-performance emission model and distributed computing for line source dispersion modeling (Liu, et al., 2017). Liu, et al. (2017) introduced a distributed computing method for line source dispersion modeling that integrates MOVES-Matrix, a high-performance emission modeling tool, with the microscale dispersion models CALINE4 and AERMOD. The modeling system built on MOVES-Matrix that generates exactly the same emission rates as using MOVES, but the approach is more than 200 times faster than using the MOVES graphical user interface (Guensler, et al., 2016). The study also pointed out the significant reduction in dispersion modeling run-time by running multiple dispersion models simultaneously on a distributed computing cluster.

2.3 Spatial and Temporal Resolution for Dispersion Modeling

The integration of models is complex, and computational efficiency is of critical concern in designing modeling tools for use in conducting federal/state air quality impact assessment. To date, the temporal/spatial resolution of receptor location and transportation sources (links) used in impact assessment modeling has been low (Zhai, et al., 2016; Xu, et al., 2019; D'Onofrio, et al., 2016; Wu, 2018) to limit the need for access to large

computational resources. In predicting pollutant concentration profiles over an entire region or sub-region, receptor grid spacing is typically greater than 200 meters between each receptor on all sides (e.g., Zhai, et al., 2016; D'Onofrio, et al., 2016; Wu, 2018). While uniform grids of receptors at low-spatial-resolution may be useful in exploring overall spatial distributions of pollutant concentrations, the resulting concentration predictions are necessarily biased, because each receptor's concentration is sensitive (in a non-linear fashion) to the distance between receptor and pollutant source links (Wu and Niemeier, 2016). Because road network density and facility configuration vary across a region, the locations of receptors must reflect this complex spatial distribution of mass flux from the nearby roadway links for receptor predictions to accurately reflect local pollutant concentrations.

In terms of temporal resolution, most previous models also assumed static traffic operation conditions (e.g., annual average daily traffic, see the examples of Zhai, et al., 2016; D'Onofrio, et al., 2016; Wu, 2018). However, traffic volumes and on-road operating conditions, which generate emissions, vary significantly by hour of day, day of week, and season of year (traffic operations are quite dynamic). In this regard, Samaranayake, et al. (2014) and Aziz and Ukkusuri (2012) assumed dynamic traffic conditions in estimating on-road vehicle emissions at small time intervals, but the computational efficiency of these models in predicting near-road pollutant concentration was not validated for a large-scale road network. Furthermore, vehicle fleet compositions also vary in time and space. For example, weekend fleets operating on local roadways look very different than morning commute fleets on freeways, which tend to be composed of newer vehicles (Granell, et al.,

2002; Bachman, et al., 1998). Finally, meteorological conditions also vary in time and space throughout the year.

2.4 Practical Modeling Implementation for Regional-level Microscale Dispersion Modelling

Because temporal and spatial resolutions of dispersion model inputs have consistently been a concern regarding model runtime and predicted concentration accuracy, the previous efforts have tried to conceive some methods that address these challenges. This chapter particularly focuses on addressing the modeling techniques addressed in previous region-wide line source dispersion modeling. From this review, the study will also address some of the potential solutions that improve model predictions and computational efficiencies.

2.4.1 Integration of Traffic Data into Dispersion Modelling

Traditional dispersion modeling for transportation emission sources (e.g., roadway traffic-related pollutants) has used varying forms of transportation data for calculating emission inputs for dispersion models. Depending on the scope of the research and data availability, the level of aggregation for traffic data has varied from study to study. For project-level analyses, microscopic traffic simulation data can be used to estimate vehicle emissions based on individual vehicle's trajectories, providing a better understanding of vehicle characteristics and operating conditions and resultant pollutant concentrations (e.g., Kim, et al., 2019b; Xu, et al., 2016). However, this approach has rarely been used in a regional or sub-regional analysis because it requires tremendous computational time to integrate link-by-link traffic operations at such large scales and for calculating emissions for a large number of individual vehicle source types.

Thus, a body of traditional modeling methods have linked travel demand model data to mobile source emission calculator such as MOVES to estimate link-level emissions (Batterman, et al., 2014; Zhai, et al., 2016; Hartley, et al., 2009; Wu, 2018). Travel demand models can provide travel activity data for a large-scaled area with a certain degree of aggregation (e.g., annual average link traffic volumes and speeds), making it possible to efficiently estimate regional-scale link-by-link emissions. However, it is important to note that all emission calculations depend on the travel data predicted by the travel demand model. For example, Zhai, et al. (2016) estimated annual average link emissions for the 20-county metropolitan Atlanta network in 2010 based on 2010 ARC-ABM outputs and MOVES2010b model, and simply scaled to annual average levels for other years. Aggregation of such activity data tends to ignore operational variability.

Such approaches provide modeling efficiency in calculating link emissions for a large-scale area. However, the use of aggregated traffic data may underestimate the impacts of seasonal peak emissions, introduced by the dynamic traffic characteristics across the hour, day, and year. In addition, rather simplified vehicle fleet composition defined in general travel demand models has undermined the predicted emissions calculation results because the vehicle emission rates vary significantly depending on the vehicle types (Bachman, et al., 2000; Guensler, et al., 2017; Liu, et al., 2015; Liu and Kim, 2019a). In this regard, some studies including Liu, et al. (2017) introduced the use of the local fleet and operation data (e.g., defining local fleets based upon license-plate data from video data collection matched to vehicle registration database, highway performance monitoring system data, USEPA certification data for light-duty vehicles, national vehicle sub-fleet composition data for heavy-duty vehicles). The study suggested that such various

supplementary data can be linked to travel demand model outputs, to improve the emissions calculation predictions.

In addition, the previous models have rarely tested the impacts of road grades in vehicle emissions and resultant pollutant concentrations, potentially due to the lack of road grade profile covering the whole study area. For example, Liu, et al. (2018) introduced the development of road grade using the United States Geological Survey (USGS) digital elevation model that can be used for regional-scale line source dispersion modeling. The model results suggested that near-road pollutant concentrations are affected by road grades to some extent, depending on the degree of road grade and fleet composition.

As such, there has been a concern about estimating accurate vehicle emissions profiles for a regional-scale analysis. However, the modeling complexity and the limited input datasets have constrained modelers from estimating more refined traffic-related emissions profiles. This situation suggests that streamlined data processing that combines multiple data sources, thereby efficiently estimating much more refined emissions profiles.

2.4.2 Emission Source Type

Dispersion models provide several types of sources to represent a road emission source. For example, AERMOD can consider roadway as various types, including AREA (rectangle-shaped line source), VOLUME (three-dimensional space), AREAPOLY (area source as an irregularly-shaped polygon of 3 to 20 sides), and any of these options are allowed for transportation conformity analysis (USEPA, 2018). However, the results and computation time depend to some extent on the source type. Wu and Niemeier (2016) addressed that VOLUME source in AERMOD has a faster runtime than AREA source, but

it is likely to introduce analytical errors largely due to the effects of exclusion zones. In the meantime, generating source geometry based on VOLUME source is likely to significantly increase the total model runtime because it decomposes a single line source into many small VOLUME sources that need to be modeled separately. On the other hand, the AREA source in AERMOD has a limitation in that it does not have a horizontal meander algorithm that decreases the likelihood of introducing a coherent plume after long travel times (USEPA, 1995). All these aspects require a balance in choosing a proper source type when using AERMOD for line source dispersion modeling.

2.4.3 Receptor Siting

The receptor placement for evaluating traffic-related pollutant concentrations has been one of the major concerns because concentrations of traffic-related air pollutants show dramatic temporal and spatial variation in on-road and near-field environments. For example, some traffic-related pollutions of particulate matters, volatile organic compounds (VOCs), and nitric oxide (NO) demonstrate steep gradients in concentrations, with elevated levels near and on roads, and remain as background levels at distances of roughly 150 meters to 200 meters (Barzyk et al., 2009; Hagler et al., 2009; Hu et al., 2009). The great variation leads to significant uncertainty in evaluating near-road concentrations and human exposures (HEI, 2010).

To address this issue, traditional region-wide dispersion analyses have contrived some strategic methods based on a grid receptor approach with varying resolution in relation to roadway characteristics. Wu (2018) suggested two-layered receptor placements with 100 m resolution receptors within 1000 m of the major corridors and downtown area

for the city of Sacramento, and with 200 m spacing for the rest of the study area. Hartley, et al. (2009) spaced receptors at Census block centroids and adjacent to the midpoints of major link sources, accounting for around 750 m receptor resolution for modeling metropolitan Atlanta area. Batterman, et al. (2014) undertook a much more intensive analysis of the impacts of receptor grid density, evaluating pollutant concentration profiles using receptor grids with decreasing spatial resolution (10, 20, 40, 80 and 160 meters spacing), suggesting that near-road pollutant concentration profiles depend on the receptor grid resolution. In particular, the study suggested that interpolations between receptors and points of interest should not exceed about 40 m near major roads, and 100 m at larger distances.

Receptor density has always been a concern with respect to computational resources and time for conducting dispersion modeling, because model runtime tends to increase linearly with the number of link-receptor pairs (Kim, et al, 2019c; D'Onofrio, et al., 2016). In addition, low-density gridded receptor grid approaches introduce biased concentration profiles, because predicted pollutant concentrations are highly affected by the proximity to link sources (the distance between the source and grid receptor are necessarily random). Previous modeling efforts have suggested that dense receptor placement should be used near major roads, but that these densities can decrease further from the roadway sources given the non-linear basis of Gaussian dispersion. Thus, this situation calls for the development of an advanced receptor placement method to predict an optimal number and placement of receptors, relative to road geometry, roadway emissions, and meteorological conditions.

2.4.4 *Meteorological Inputs*

The performance of the dispersion models depends on the type of meteorological data (Cimorelli, et al., 1998). USEPA's preferred and recommended models such as AERMOD processes hourly meteorological data which is normally generated using AERMOD meteorological processors, AERMET (USEPA, 2004) and AERMINUTE (USEPA, 2015c). USEPA (2015a) emphasized that one of the key factors in producing credible results in a hot-spot analysis is the use of meteorological data that is as representative as possible of the project area. The USEPA PM hot-spot guidance (USEPA, 2015a) recommends using at least one year of site-specific data or five consecutive years of the most recent representative meteorological data (e.g., from National Weather Service), preferably in consultant with respective state and local air quality agencies for choosing representative meteorological data.

Traditional regional-level dispersion modeling efforts have balanced the temporal resolution of meteorological data regarding the desired modeling computation time and resources. For sub-regional transportation network or relatively low-resolution receptor settings, a body of studies applied hourly meteorological data through the modeling year (e.g., Batterman, et al., 2014; Hartley, et al., 2009). Meanwhile, some research has tried to reduce the temporal resolution by categorizing the hourly meteorological data into several categories, in a way to minimize model runtimes for analyses with a large-scaled network (Zhai, et al., 2016). This approach appears to be effective to predict pollutant concentrations in a worst-case scenario, by introducing the worst-case meteorological condition in the analysis.

Worst-case meteorological conditions have also been used for the purpose of screening analysis. Some dispersion models (e.g., CALINE3/4, AERMOD) embedded screening modes to find the worst-case wind direction for each receptor (Benson, 1984; USEPA, 2016). CALINE3/4 searches the worst-case wind angle for each source, by cycling through wind directions (Benson, 1984). In the case of AERMOD, it embeds an internal screening mode that forces the model calculations to represent values for the plume centerline, regardless of the source-receptor-wind direction orientation (USEPA, 2016). This option is included in AERMOD to facilitate the use of the model in a screening mode to estimate worst-case impacts. Although the screening options are not recommended by the USEPA for use conformity analysis (USEPA, 2015a), the screening options are an effective way to identify potential hot spots while minimizing the model runtime because these options do not necessarily perform the calculations considering dynamic meteorology conditions (Liu, et al., 2017). As such, the screening options may help reducing total model runtimes, while suggesting the areas for further investigation.

2.4.5 Model Calibration

Dispersion models have been widely used to model concentrations of primary pollutants for regional-scale. The widely-used dispersion models such as AERMOD and R-LINE formulate pollutant dispersion by solving a simplified form of the pollutant transport equation (e.g., Gaussian models) and have limited descriptions of chemical transformation (Zhai, et al., 2016). (Meanwhile, chemical transformation is occasionally considered for some reactive plume chemistries, e.g., NO₂ in CALINE (Benson, 1984).) However, previous research addressed that dispersion models tend to overestimate for high

concentrations and underestimate for the lower range of concentrations (Venkatram et al., 2004).

To address model over-prediction, a few studies such as Zhai, et al. (2016) and Venkatram, et al. (2004) have focused on model calibration by comparing dispersion model predictions with observed concentration profiles obtained from on-sites. The study by Zhai, et al. (2016) revealed that the trends of R-LINE estimates and pollutant concentrations at monitor locations are similar (Pearson R^2 are 0.64 to 0.83), while the R-LINE estimates are overestimated as compared to the observational data. Thus, the study proposed a calibrated R-LINE model developed based on the trend lines accounting for the relationship between the R-LINE estimates and observational data. Venkatram, et al. (2004) particularly suggested that onsite turbulence information in a simple model for meandering can lead to adequate estimates of observed concentrations.

These calibration methods showed availability to calibrate modeled concentrations with efficiency (simply applies the scaling factors), but this approach is limited due to its dependence on available monitor data in developing the calibration models. Depending on regions, monitor data may not be sufficient to calibrate regional-level microscale dispersion modeling outputs. In addition, it should be noted that the calibrated dispersion models may also depend on the site-specific monitor data, suggesting that the scaling factor can overestimate or underestimate the calibrated model concentrations. In particular, Hodan and Barnard (2004) suggested that understanding the characteristics of pollutant concentrations is essential for proper model calibration. For example, $PM_{2.5}$ concentrations measured from monitoring stations are attributed to: 1) primary $PM_{2.5}$ emissions emitted from vehicle tailpipes, and 2) secondary $PM_{2.5}$ formation from precursor emissions such

as sulfur dioxide, nitrogen oxides, volatile organic compounds, and ammonia. In practice, the primary $PM_{2.5}$ emissions can be modeled by dispersion models; however, most dispersion models do not account for secondary $PM_{2.5}$ formation. Thus, understanding of potential gaps between modeled and monitored pollutant concentrations is essential for proper processes of calibrating the predicted concentrations modeled by dispersion models. Meanwhile, all these aspects suggest that attention must be paid to interpreting the pollutant concentrations predicted by the line source dispersion models.

2.5 Summary of Regional-Level Microscale Dispersion Modeling Issues

Research has relied on microscale dispersion models such as AERMOD to evaluate the short- and long-term human exposure to traffic-related primary air pollutants. While microscale dispersion models have shown the capability of predicting concentrations supported by its theoretical background, refined formulation, and model application, a number of challenges in modeling traffic-related pollutant dispersion have been addressed in the literature, especially for large-scale analysis. The complex setup procedures for required emissions calculation (e.g., using MOVES) and microscale dispersion modeling have a high potential for introducing analytical errors. In addition, running dispersion models for a large geographic region has demanded an extremely large computation, requiring an impractically large processing time.

In particular, the computational issues in dispersion modeling have restrained practitioners from conducting detailed analyses with high temporal and spatial resolution data. For example, previous studies have often ignored the operational variability in on-road traffics to boost modeling speed for link emissions calculation; however, it is likely

to underestimate the impacts of seasonal variations in predicted emissions. In addition, some studies have reduced the meteorological variability, in a way to minimize model runtimes for dispersion modeling with a large-scale network. Although such approaches have been considered as an effective way for screening analysis, they are limited in transportation conformity analysis (the USEPA guidance requires to process at least one year of hourly meteorological data). In this regard, some recent studies have introduced some advanced methods that could improve the efficiency of regional-level dispersion modeling, e.g., MOVES-Matrix, dispersion modeling through a distributed computing, link screening, etc., which will be addressed through this dissertation.

In particular, receptor placement has also been one of the major concerns in the dispersion analysis because it is highly related to an increase in computational time. For example, research indicated that the dispersion model runtime tends to increase linearly with the number of link-receptor pairs. Meanwhile, literature also addressed that careful selection of receptor locations is important to predict precise near-road concentration profile (predicted pollutant concentrations are highly affected by the proximity to link sources). Thus, this situation calls for the development of an advanced receptor placement method to predict an optimal number and placement of receptors, relative to road geometry, roadway emissions, and meteorological conditions.

CHAPTER 3. MODELING OVERVIEW

The literature review addressed that previous modeling efforts for predicting potential public exposure to harmful transport emissions at regional and sub-regional scales have posed a concern regarding model prediction and computational efficiency in the design of modeling tools. Motivated by the challenges encountered in the previous modeling efforts, this work focused on developing an advanced modeling framework for region-wide applications of line source dispersion models that employs a number of innovative modeling techniques.

As a part of this dissertation, Chapter 4 proposes a strategic receptor placement method, called dynamic grid-receptor model, that determines the optimal placements of receptors with respect to each link's geometry and emissions characteristics. The dynamic grid-receptor model is developed based on training data sets from numerous R-LINE runs simulated under diverse scenarios. The dynamic grid-receptor model is incorporated into the region-wide line source dispersion modeling system in a way to reduce the total modeling run-time (by minimizing the number of receptors across a region), without undermining the pollutant concentration profile.

As an innovative way to significantly improve the modeling efficiency, the second part of this dissertation addressed in Chapter 5 develops a methodology to screen roadway link sources that do not significantly contribute to pollutant concentrations at a particular receptor. The link screening method is developed by a supervised machine learning Random Forest (RF) classification model from the analysis conducted by AERMOD simulations. The link screening model is also incorporated into the modeling system.

Incorporating the dynamic grid-receptor and supervised link screening models as well as some advanced techniques into the modeling framework, Chapter 6 streamlines the data processing that efficiently prepares the extensive input datasets needed for line source modeling over entire metropolitan areas (Figure 5). In particular, this work proposes the integration of MOVES-Matrix and regional travel demand models (TDMs) to efficiently estimate the link emissions at regional-scale, and demonstrates how to integrate road grade profiles generated by U.S. Geological Survey Digital Elevation Model. The work also demonstrates how to streamline data processing through a distributed computing cluster to boost the modeling speed. Lastly, a case study for the 20-county metropolitan Atlanta area, which accounts for an extremely large number of link-receptor pairs (161,188 links and 1,163,958 receptors), is conducted to validate the developed modeling system regarding its model prediction and efficiency.

Streamlined Data Processing for Region-wide Line Source Dispersion Modeling (Chapter 6)

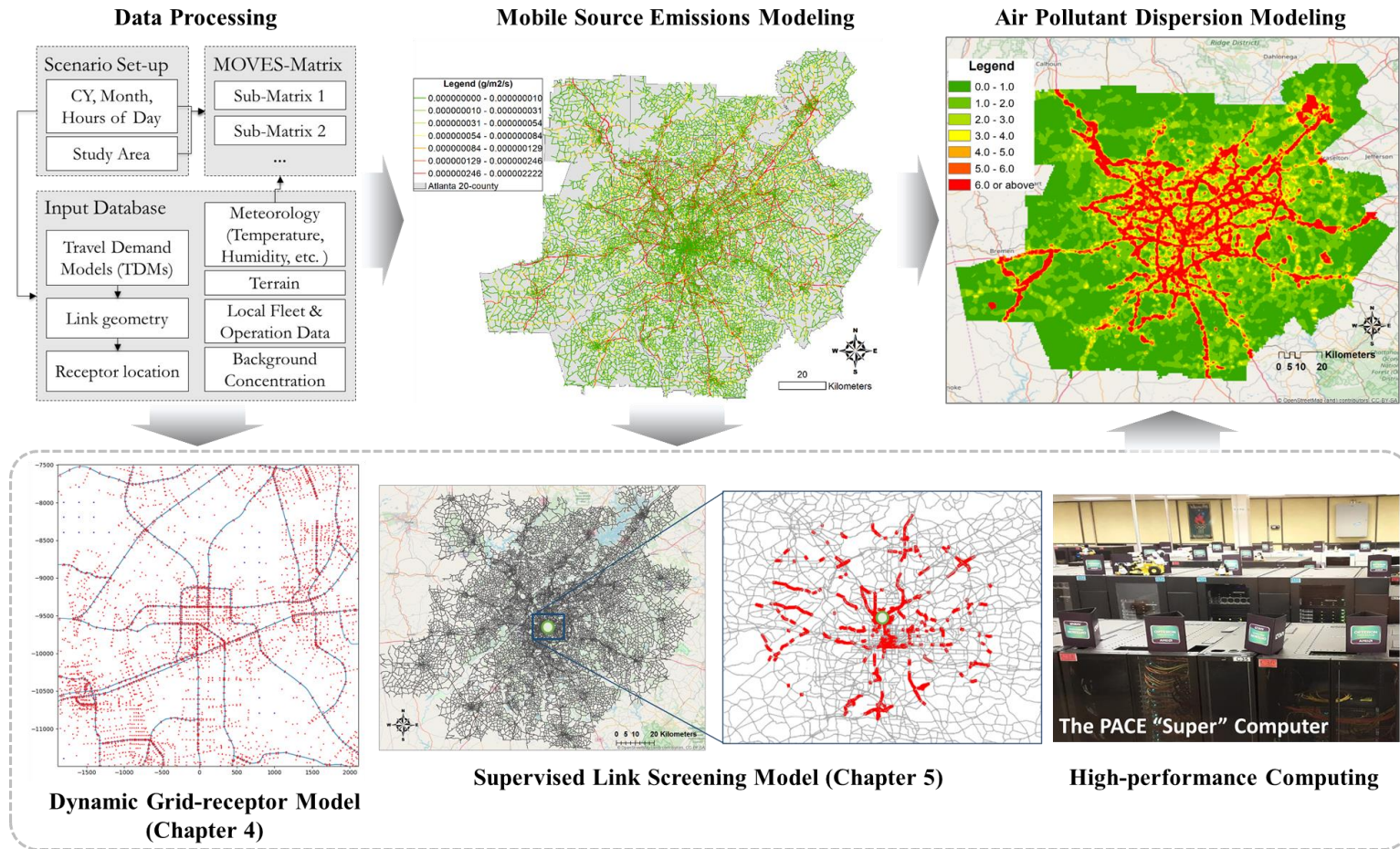


Figure 5 – Overview of Modeling Processes

CHAPTER 4. DYNAMIC GRID RECEPTOR METHOD FOR REGIONAL-LEVEL NEAR-ROAD AIR QUALITY ANALYSIS

4.1 Introduction

Computational efficiency has been a major concern in previous studies of near-road air quality impact assessment conducted over large geographic scales. Most metropolitan area or regional applications of microscale models have employed simplified sets of model assumptions, with fairly low-resolution receptor grids (e.g., 200 m by 200 m resolution). While gridded and sparse receptor placement helps reduce computational complexity and model run time, it can lead to biased concentration profiles (over-prediction or under-prediction), because predicted concentration estimates are very sensitive to the distance from receptor to the nearest roadway links and each link's mass flux. In this regard, some recent studies have used the state-of-the-practice methodology recommended by USEPA to assess near-road pollutant concentration profiles (USEPA, 2015b), by spatially adjusting the density and location of receptors in proximity to link emission sources in all directions, based upon the spatial layout of the roadway network. However, this approach necessarily requires a great number of near-road receptors to be processed, increasing the total processing time as a function of the number of link-receptor pairs. Motivated by the challenges identified in previous studies, this study aims to develop a strategic receptor placement method that minimizes the number of receptors without undermining the pollutant high-resolution concentration profile that is generated with dense receptor placement.

4.2 Model Development Process

This study proposes a dynamic grid-receptor modeling approach in the placement of receptor locations with respect to link geometry (e.g., proximity to link emission source) and metrological conditions. In developing the dynamic grid-receptor model, the study runs numerous R-LINE models to generate training and testing datasets where each model predicts PM_{2.5} concentration profile originated from a hypothetical freeway corridor with a dense grid receptor setting. For each model run, this work applied varying PM_{2.5} link emission rates, link lengths and meteorological conditions to identify overall trends in PM_{2.5} concentration profiles under varying conditions.

4.2.1 Hypothetical Network

The study estimated the PM_{2.5} concentrations at numerous receptor sites along a hypothetical freeway network (Figure 6). The hypothetical network consists of three consecutive roadway segments with varying PM_{2.5} emissions rates for each R-LINE run. The link segment of focus (Link 2 in Figure 6) is located between its adjacent link segments (Link 1 and Link3). Above those link segments, numerous receptors are placed at high-spatial-resolution (every 5m) on the X- and Y-plane (dots in Figure 6), and the receptor height is set to 1.5m for human nose height. For model development, the receptor area is set between both ends of the focused link segment.

As shown in Figure 6, the design of the hypothetical network is rather simple than practical transportation network (e.g., curved roadways, complex intersections, etc.). However, the simplified network appears to be more convenient to figure out the direct relationship between the receptor settings and its predicted concentration profiles. As such,

the work first aims to identify the optimal receptor settings under simplified network conditions, and apply them to practical transportation network with some treatments on a case-by-case basis, which will be addressed in the following sections.

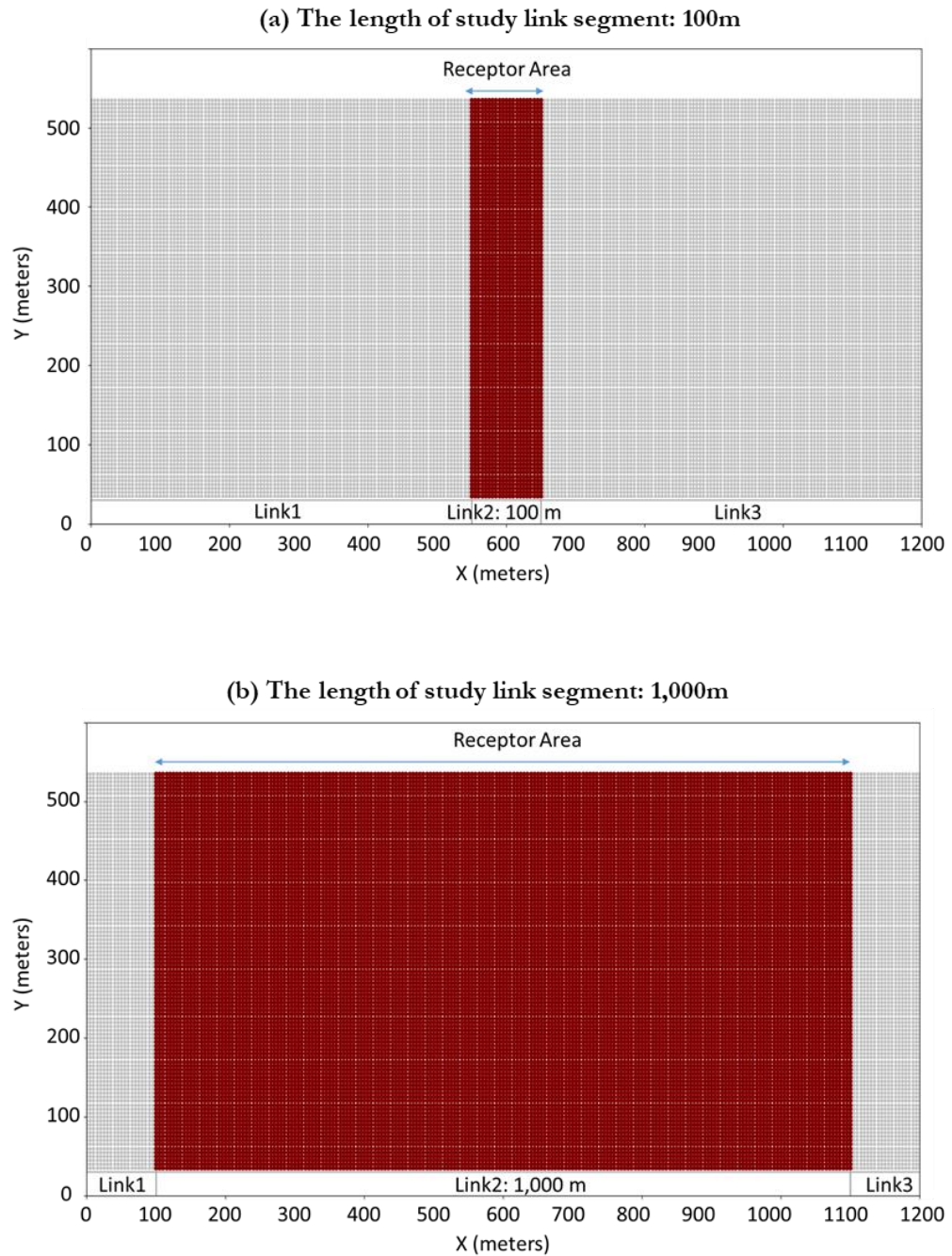


Figure 6 – Examples of Hypothetical Networks

To link the developed dynamic grid-receptor model to general transportation network composed of links of varying length, the length of the focus link segment in each run is randomized from 100m and 1,000m (uniformly distributed at one-meter resolution), accounting for the link length distribution of ARC-ABM15 network (99% of the link lengths are less than or equal to 1,000 m) (Figure 7).

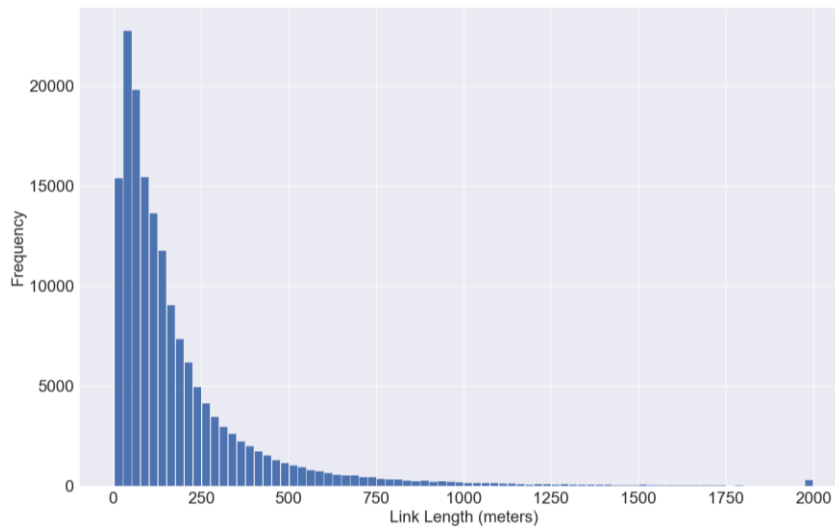


Figure 7 – ARC-ABM Network’s Link Length Distribution

For each link segment, the $PM_{2.5}$ link emission rates are assigned based on the $PM_{2.5}$ link emissions rates estimated by ARC-ABM15 traffic data. The work streamlined the processes of connecting the detailed link traffic data (e.g., volumes, speed, fleet composition) provided by ARC-ABM15 to a high-performance emissions calculator (i.e., MOVES-Matrix) to estimate the link-by-link $PM_{2.5}$ emission rates (see the section 3.4 for more details). Because ARC-ABM15 covers the whole metropolitan Atlanta network, and the work addressed in section 3.4 conducted $PM_{2.5}$ emissions calculations for every link in ARC-ABM15 under dynamic traffic and meteorological conditions across hours of the

year, the estimated PM_{2.5} emission rates can provide a sufficient variability in link emissions reflecting the real-world traffic conditions. For R-LINE simulation, the work assigns the randomized PM_{2.5} emission rates, ranging between the upper and lower limits of the link PM_{2.5} emission rates identified by the emissions calculation processes, to each link segment in the hypothetical network.

4.2.2 *Simulation Inputs*

Each model run generates the PM_{2.5} downwind concentration profiles for the given link lengths, emission rates, and meteorological conditions. Table 2 shows the limits of the input parameters for the R-LINE simulations. In generating R-LINE simulation input files, input values for each R-LINE run are randomly chosen based on the uniformly-distributed random values ranging between the pre-set limits in Table 2.

- The link lengths for the focused link segment (the link located between the two adjacent links in the hypothetical network) is set ranging from 100 m and 1,000 m, uniformly distributed in 1-meter unit increments.
- The upper and lower limits for PM_{2.5} emission rates were established from the PM_{2.5} link emissions rates estimated for use with ARC-ABM15 data. The upper limit of PM_{2.5} emission rates is set to 0.000122 g/m/s which was identified as the highest link emission rates among all links in the metropolitan Atlanta network. The lower limit of PM_{2.5} emission rates is set to 0.0 g/m/s reflecting some of the links that traffic volumes are not assigned from the ARC-ABM15. The PM_{2.5} emission rate for each link segment is randomly assigned between the limits (uniformly-distributed continuous values). Note that the R-LINE emission rate input is

specified in grams/meters/second and used by R-Line with the road width input parameter to automatically generate mass flux.

- In terms of wind direction, the range is set considering the effect of downwind in relation to the hypothetical network, which ranges from 90° (west-bound), 180° (north-bound), and 270° (east-bound), at a one-degree interval.
- The ranges of some meteorological conditions, wind speed, mixing height, and temperature, are taken from meteorological processor data, which is processed for the metropolitan Atlanta area by Georgia EPD (2019).

Table 2 – Ranges of Input Parameters for R-LINE Simulations

Description	Min	Max	Interval
Link length (m)	100.0	1,000.0	1.0
PM _{2.5} emission rates (g/m/s)	0.000000	0.000122	Continuous
Wind direction (°)	90.0 (west-bound)	270.0 (east-bound)	1.0
Wind speed (m/s)	0.5	20.0	Continuous
Mixing height (m)	100.0	3,000.0	1.0
Temperature (°C)	-10.0	40.0	1.0

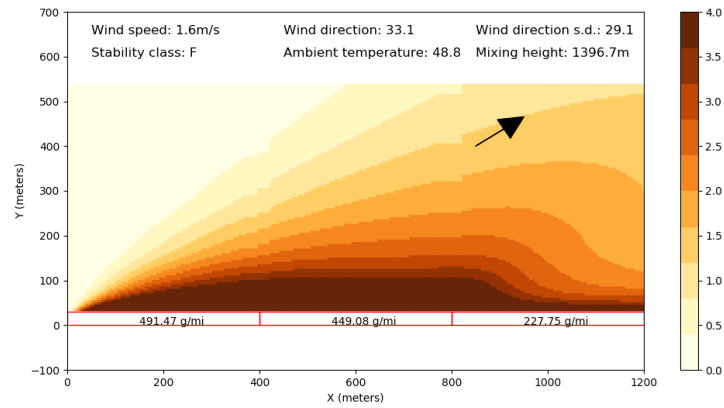
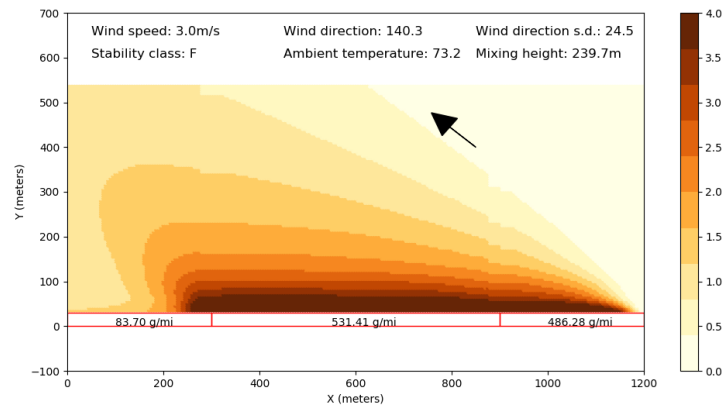
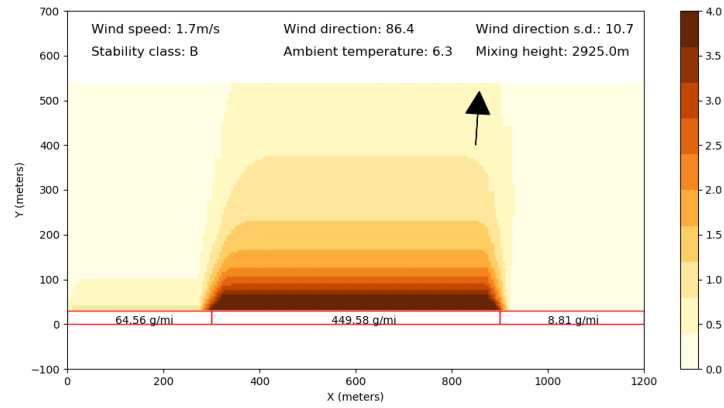
4.2.3 R-LINE Simulation

The PM_{2.5} concentrations were estimated by R-LINE. R-LINE was chosen for this analysis among other USEPA’s recommended dispersion models because it is easy to control the meteorology file, and maintain the direct relationship between the meteorological

parameters and predicted pollutant concentration. R-Line employs similar theoretical Gaussian dispersion equations (Snyder, et al., 2013a; Snyder and Heist 2013b) to USEPA's current preferred/recommended model (AERMOD), but R-LINE is running faster than AERMOD and was more practical for this effort. Using R-LINE in this manner can help researchers to better understand the fundamental relationship between dispersion model input parameters and model-predicted pollutant concentration profiles.

Because this study conducts dispersion modeling for a large number of input scenarios, it requires a huge computation resource and time. The study performs a total of 1,300 R-LINE runs (1,000 runs for training samples and 300 runs for testing samples), and each simulation predicts $PM_{2.5}$ concentration at a total of 24,340 receptor sites. In this circumstance, the study utilized the Georgia Tech's PACE computing cluster by submitting multiple jobs simultaneously, thereby reducing the total processing time.

Figure 8 shows some examples of the $PM_{2.5}$ concentration profiles predicted by the R-LINE modeling runs. (Each colored point represents the predicted $PM_{2.5}$ concentration at each receptor site.) The figures show the predicted $PM_{2.5}$ concentrations for the high-resolution grid of receptor sites, where darker colors represent higher $PM_{2.5}$ concentrations. The results suggest that line source pollutant concentration profiles vary depending on source emission rates and meteorology conditions. Further analysis regarding the relationship between various input parameters and predicted $PM_{2.5}$ concentration will be addressed in the following sections.



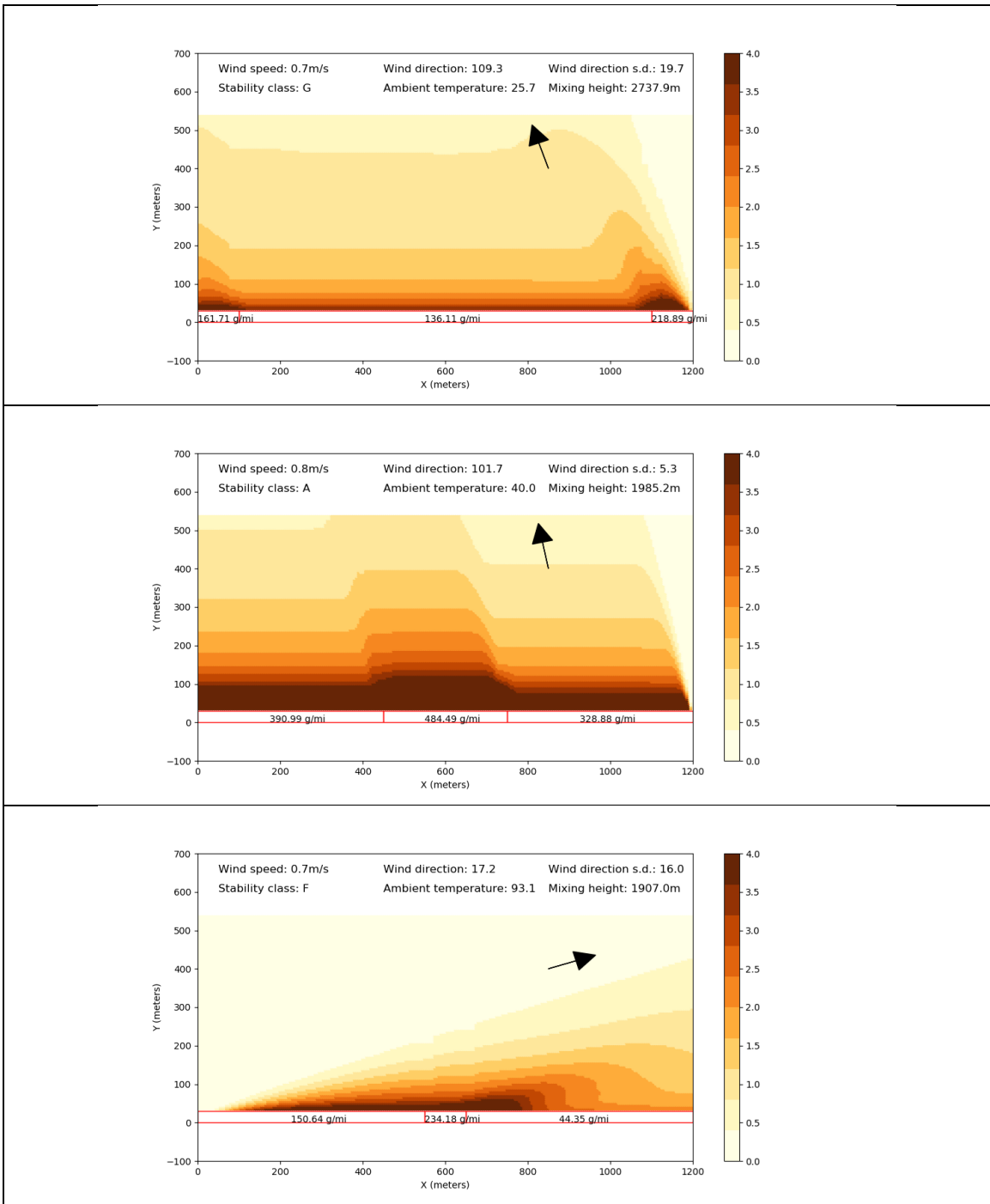


Figure 8 – Examples of PM_{2.5} Concentration Profiles ($\mu\text{g}/\text{m}^3$) Predicted by R-LINE under Varying Conditions

4.3 Simulation-informed Fundamental Relationship among Predicted Pollutant Concentration Profile and R-LINE Input Parameters

Understanding the relationships between the spatial distribution of predicted pollutant concentrations and related input parameters is essential to identify the optimal locations of receptors that could capture the well-distributed pollutant concentration profile. Equations governing the atmospheric flow are represented in many earth and environmental science references and help readers to understand the fundamental relationships among meteorological conditions and surface pollution levels. However, given the complexity of the equations (and interactions across such equations), it is somewhat challenging to identify the optimal locations of receptors for line source dispersion modeling using equations alone. In this regard, the simulation-informed fundamental relationship among the predicted pollutant concentration profile and its dispersion model input parameters can help to identify the optimal receptor locations.

This study specifically conducted the sensitivity analysis for the $PM_{2.5}$ concentration profile to R-LINE input parameters (see Figure 9 through Figure 11). The sensitivity analysis results indicate that the spatial distribution of $PM_{2.5}$ concentrations depends on some of the R-LINE input parameters including wind direction, wind speed, and link emissions. For example, the wind direction was found to be one of the most influential parameters that affect the spatial distribution of $PM_{2.5}$ concentrations (Figure 9). As expected, the predicted $PM_{2.5}$ concentrations are higher in the areas that are downwind of the line source. The near-road $PM_{2.5}$ concentrations are also significantly affected by wind speed (Figure 10). Lower wind speed is associated with higher $PM_{2.5}$ concentration

levels near road areas. As such, the changes in $PM_{2.5}$ concentrations with the distance from the link segment is greater when the wind speed is lower.

These findings suggest that the optimal receptor placements are required to better-capture the dynamic characteristics of pollutant concentration profiles depending on link emissions and meteorological conditions. For example, relatively dense receptors need to be set near road areas to better-capture the steep gradients in pollutant concentrations under low wind speed conditions. In addition, more receptor placements may be required at the sections where two link meet (high variability in concentrations is found, potential due to the change in the link emission rates).

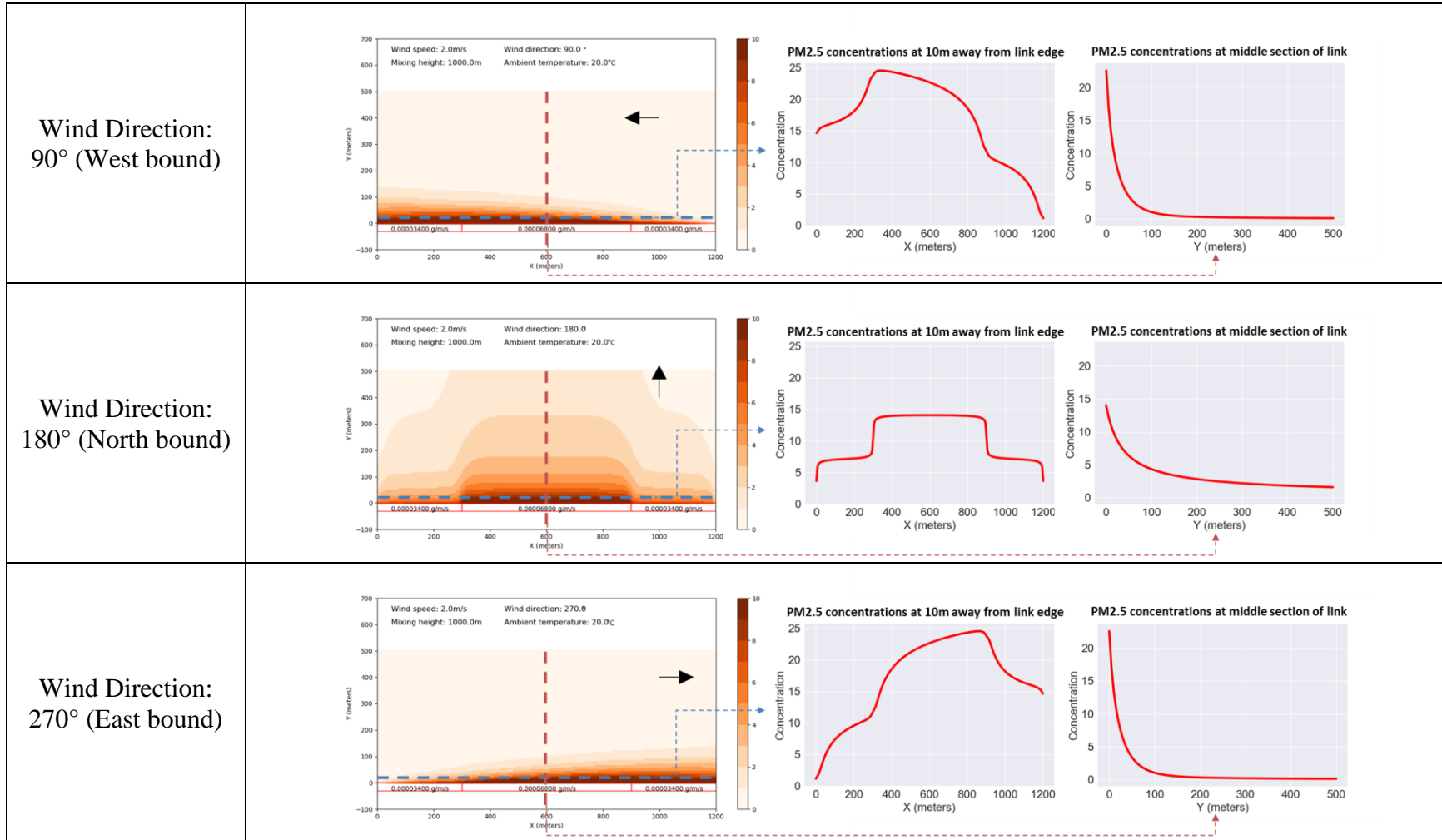


Figure 9 – Sensitivity Results of PM_{2.5} Concentration Profile ($\mu\text{g}/\text{m}^3$) to Wind Direction

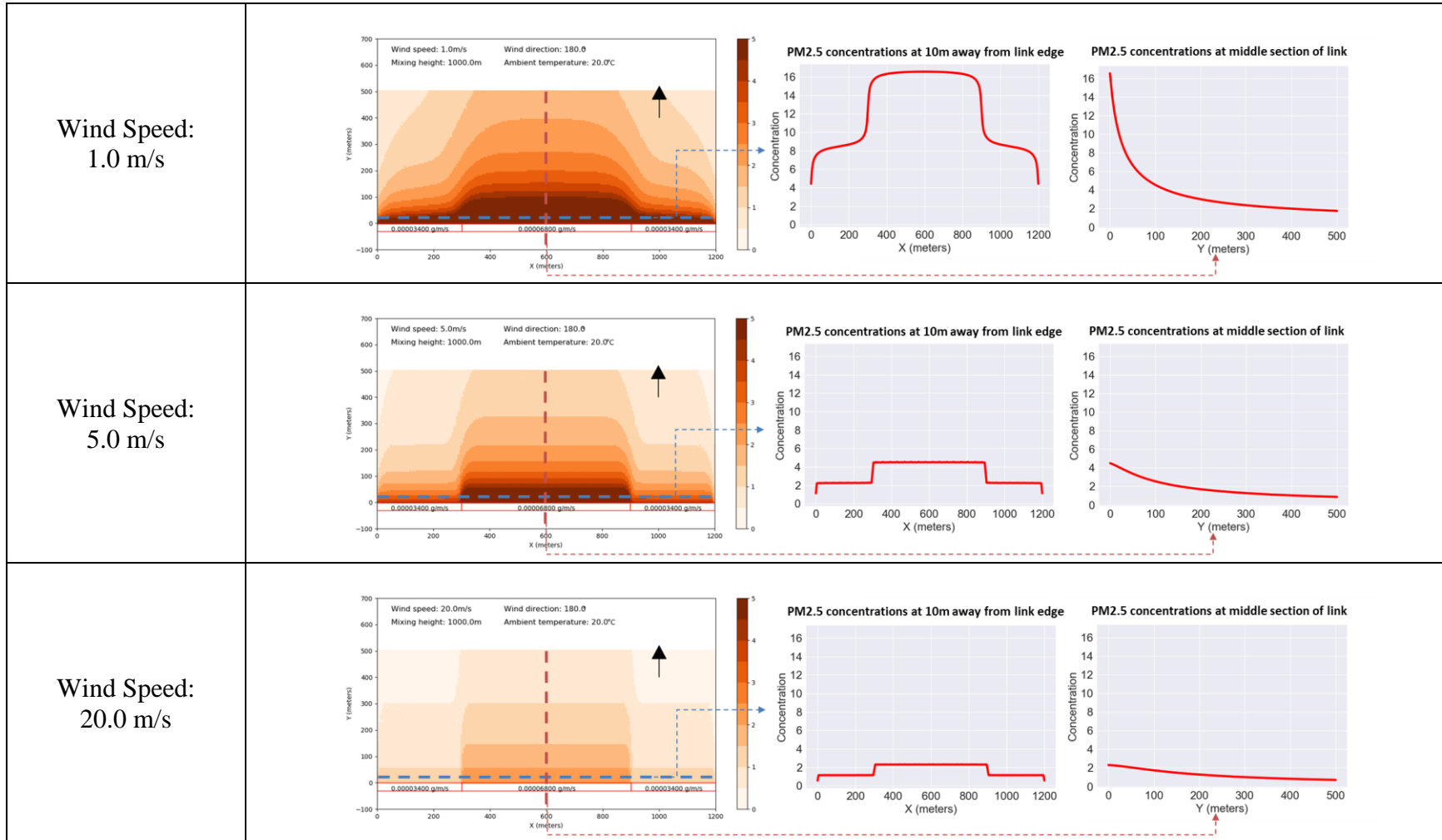


Figure 10 – Sensitivity Results of PM_{2.5} Concentration Profile ($\mu\text{g}/\text{m}^3$) to Wind Speed

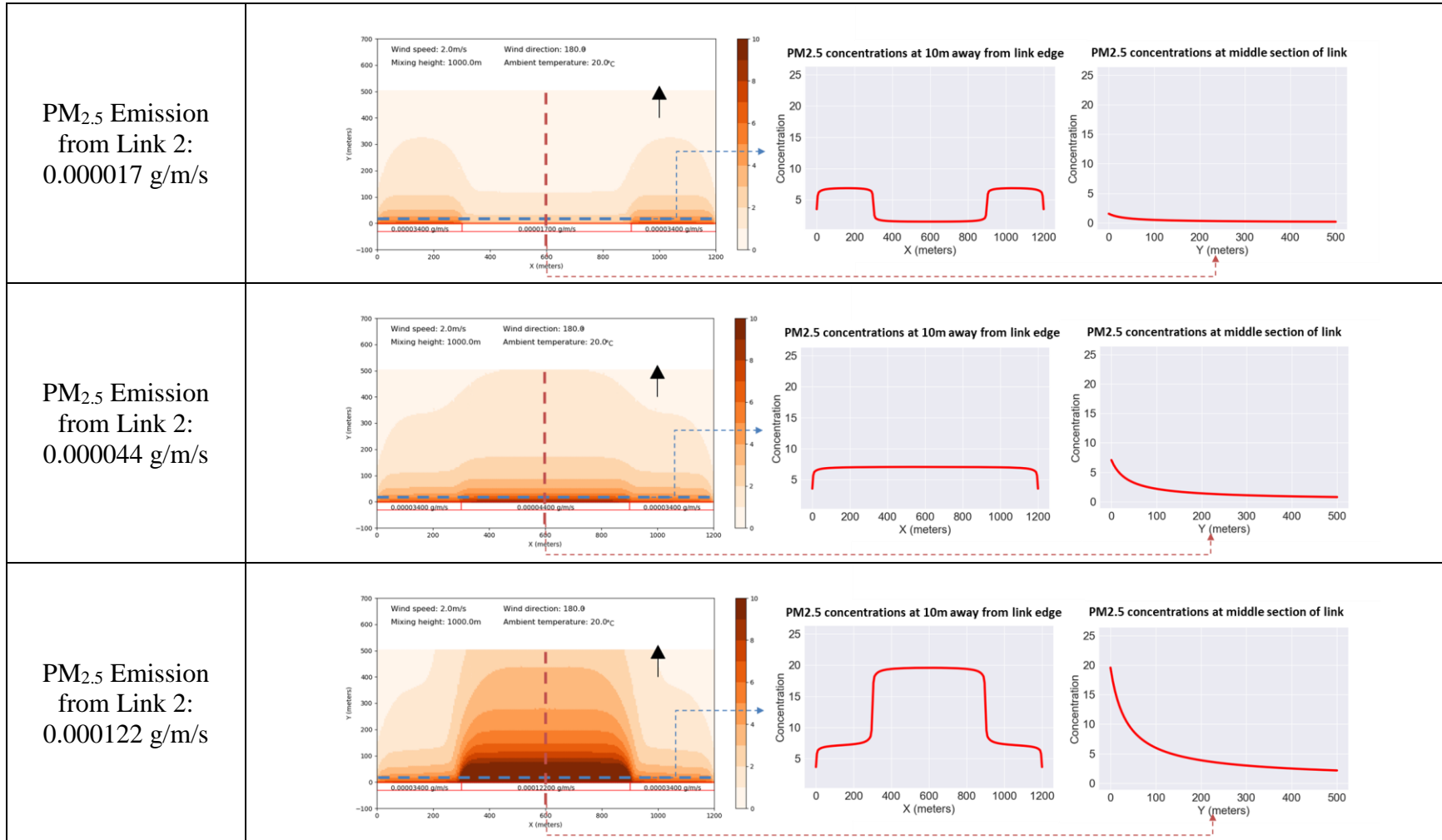


Figure 11 – Sensitivity Results of PM_{2.5} Concentration Profile ($\mu\text{g}/\text{m}^3$) to Link Emission

4.4 Step-wise Searching Method for Optimal Receptor Locations

For the training samples from 1,000 R-LINE runs, optimal receptor locations are identified through a systematic receptor search algorithm. The algorithm applied in this study utilizes a step-wise search method where each step finds the receptor that best-fits the $PM_{2.5}$ concentration profile that is predicted by the dense receptor model (5m resolution). To implement this algorithm, the method sets a total of 121 candidate receptor sites (11 by 11) for each training sample, where a combination of optimal receptor locations is identified within the candidate receptor sites. For each step, the $PM_{2.5}$ concentration profile predicted by the dense receptor model is compared with the $PM_{2.5}$ concentration profile predicted by the model with the sub-set of receptors. In this case, the $PM_{2.5}$ concentration profile predicted by the dense receptor model is generated by a heatmap with 100 by 100 grid spaces. Then, the method tests each of the candidate receptors, and selects the best receptor that minimizes the error in the $PM_{2.5}$ concentration profiles as compared to the $PM_{2.5}$ concentration profile predicted by the dense receptor model. The comparison is performed based on the $PM_{2.5}$ concentration heatmaps, which are generated based on a linear interpolation method, mapped onto a 100 by 100 grid space. In particular, the $PM_{2.5}$ concentration heatmaps were generated using the python module “LinearTriInterpolator” that performs linear interpolation on a triangular grid where each triangle is represented by a plane so that an interpolated value at point (x, y) lies on the plane of the triangle containing (x, y). As such, the interpolated values are therefore continuous across the triangulation, but their first derivatives are discontinuous at edges between triangles.² The

² The program used in this study can be found at:
<https://docs.scipy.org/doc/scipy/reference/generated/scipy.interpolate.griddata.html>

linear interpolation method was used to generate PM_{2.5} concentration profiles throughout this chapter because the method requires significantly less computational resources and time due to its simpler technical formations compared to other more advanced interpolation methods such as Kriging. As such, the method efficiently conducts numerous concentration profile predictions that can be used for developing an optimal receptor placement model. Throughout this chapter, hundreds of thousands of linear interpolations were required to generate PM_{2.5} concentration profiles and support the development of optimal receptor placement strategy. Because linear interpolation is relatively weaker at generating smooth profiles than more advanced non-linear interpolation methods, future research should be conducted to assess whether model predictions could be improved by applying more computationally intense non-linear interpolation methods.

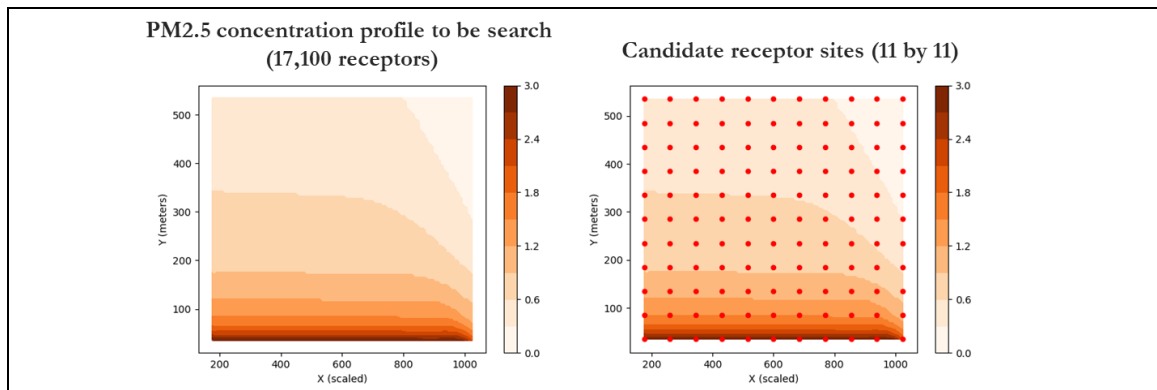
Mean square error (MSE) is then used as the measurement of the error in the PM_{2.5} concentration profiles (equation (2)).

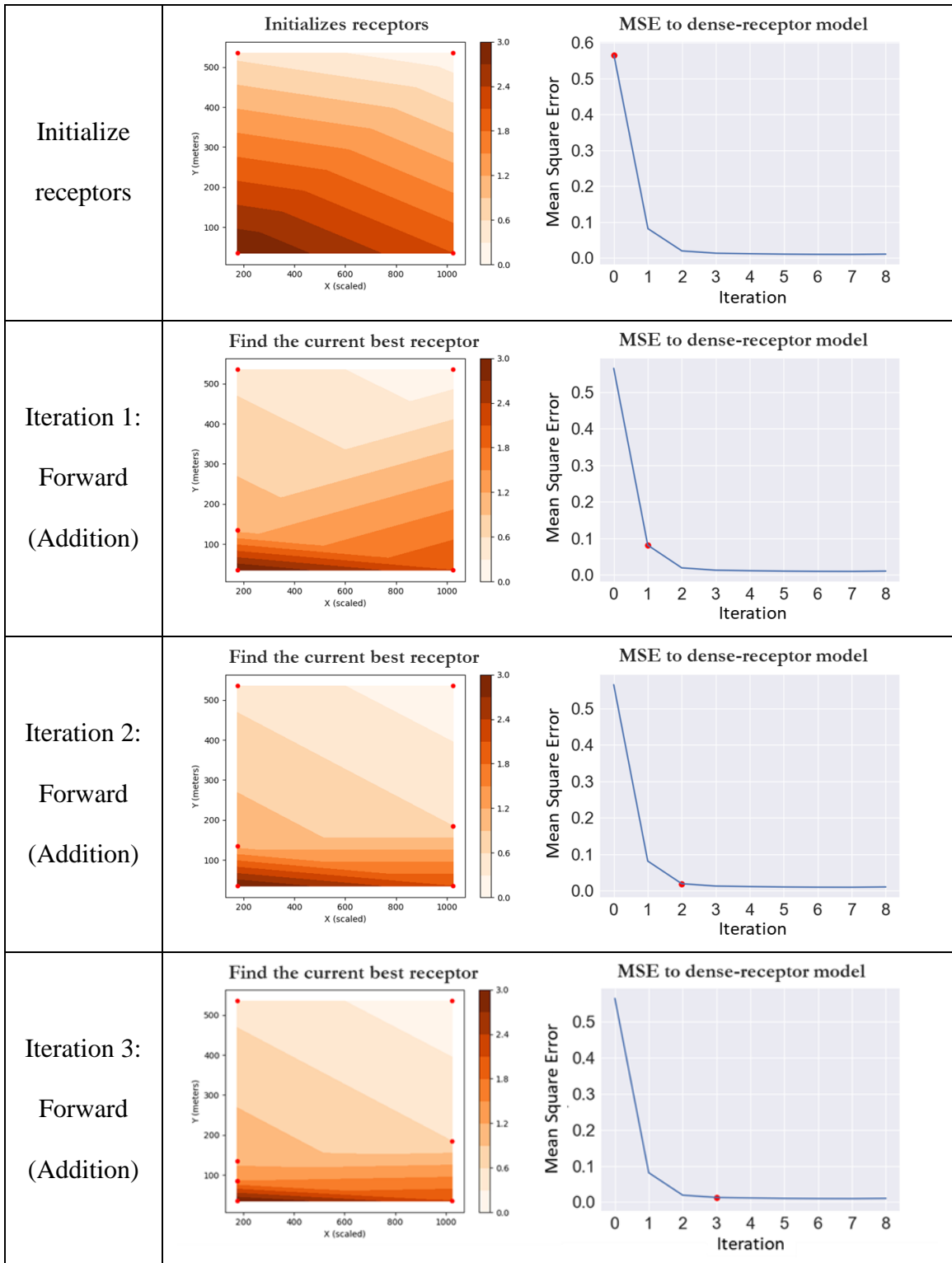
$$MSE = \frac{1}{n} \sum_{x,y}^n (PM2.5 (dense\ receptor\ model)_{x,y} - PM2.5 (subset\ receptor\ model)_{x,y})^2 \quad (2)$$

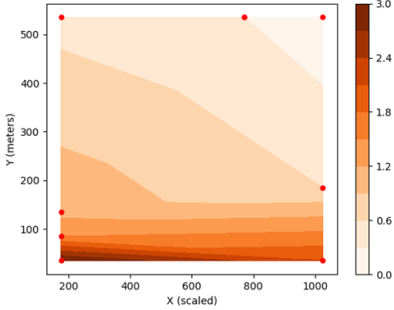
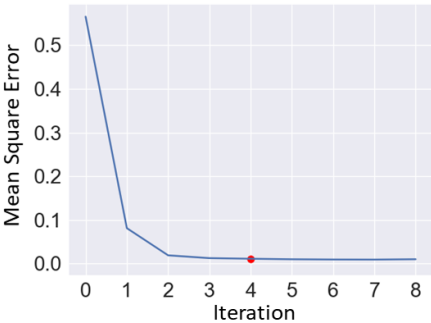
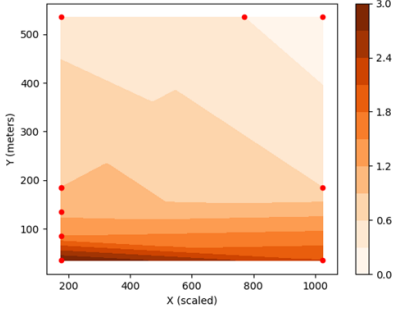
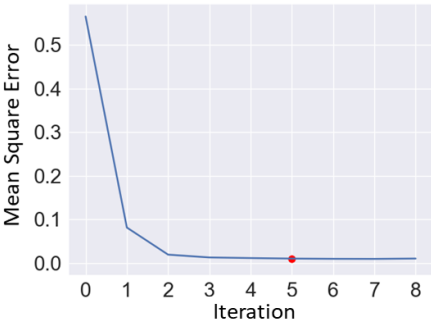
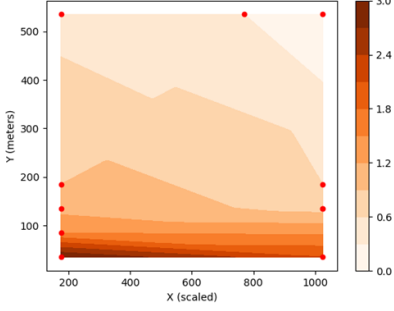
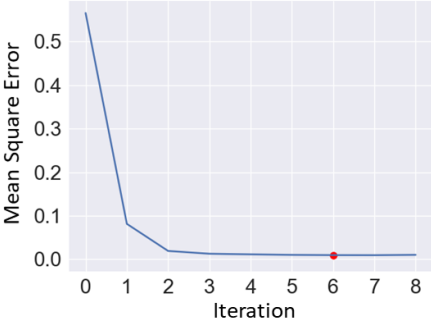
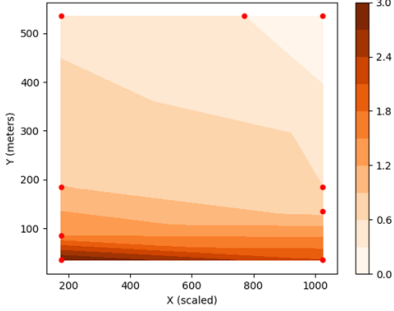
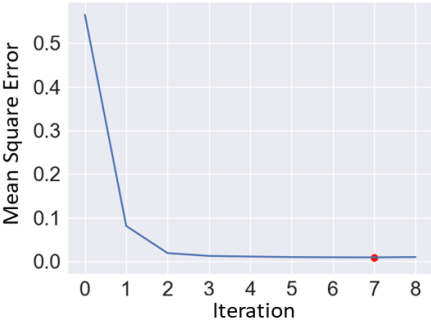
Here, x, y indicate each grid on the 100 by 100 grid space, and n is the number of grids which is 10,000 (100×100).

The step-wise search algorithm is composed of three main steps: 1) set reference receptors, 2) a forward step to add the current best receptor, and 3) a backward step to remove the current worst receptor (Figure 12). The process begins with initializing

receptors that set four reference receptors located at the corners of the receptor area. Based on those four reference points, the next searches essentially find infill points to add to these reference points. Then, the forward search is conducted, by adding current-best receptors until the marginal change in the MSE becomes less than the pre-set threshold (0.001 in this study). Afterward, the backward step removes the current-worst receptors until the marginal change in the MSE does not exceed the threshold. Figure 12 shows the example of the step-wise receptor searching processes, suggesting that the process can select an optimal combination of receptors that approximates the concentration profiles predicted by the high dense receptor models, with only a small set of receptors. This study applied the algorithm to 1,000 training samples, and identified the optimal receptor locations for each case. The results indicate that the algorithm was able to find the optimal receptor locations, with only 13.2 receptors on average (vs. 11,241 receptors on average for the high-density $PM_{2.5}$ concentration profiles).





<p>Iteration 4: Forward (Addition)</p>	<p>Find the current best receptor</p> 	<p>MSE to dense-receptor model</p> 
<p>Iteration 5: Forward (Addition)</p>	<p>Find the current best receptor</p> 	<p>MSE to dense-receptor model</p> 
<p>Iteration 6: Forward (Addition)</p>	<p>Find the current best receptor</p> 	<p>MSE to dense-receptor model</p> 
<p>Iteration 7: Backward (Deletion)</p>	<p>Find the current best receptor</p> 	<p>MSE to dense-receptor model</p> 

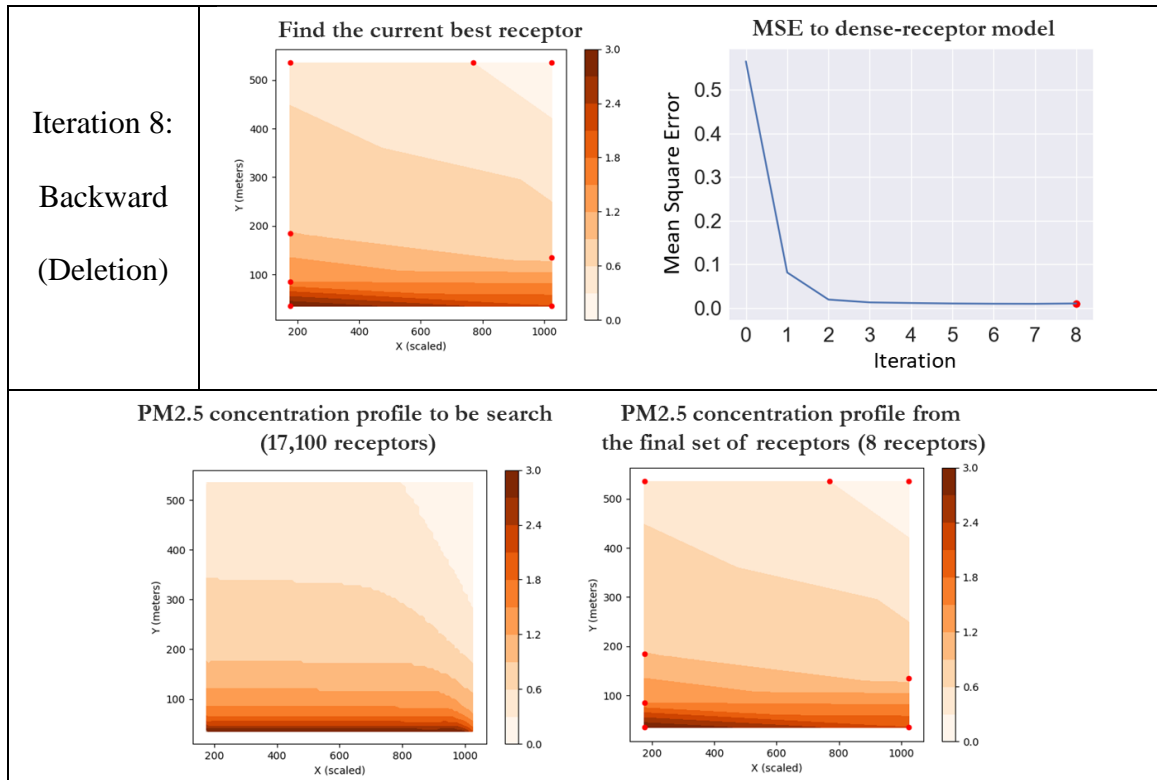


Figure 12 – The Example of Step-wise Search for Selecting Optimal Receptor Locations

4.5 Dynamic grid-receptor Model Development

The above section addressed optimal receptor placement and demonstrated that the step-wise receptor searching method can approximate the pollutant concentrations predicted by the dense receptor model with efficiency. The next stage in this study develops a model that selects the optimal set of receptors among candidate receptors using relevant explanatory variables. Hereafter, this process will be referred to as dynamic grid-receptor modeling. This model was developed based on the optimal receptor locations identified by the step-wise searching method, and corresponding R-LINE input parameters (Figure 13). The dynamic grid-receptor model consists of two sub-models: 1) the model that selects the number of optimal receptors, and 2) the model that identifies the optimal receptor locations.

The model is a step-by-step process that begins with selecting the number of optimal receptors, and then it predicts the optimal locations of receptors given the number of the optimal number of receptors.

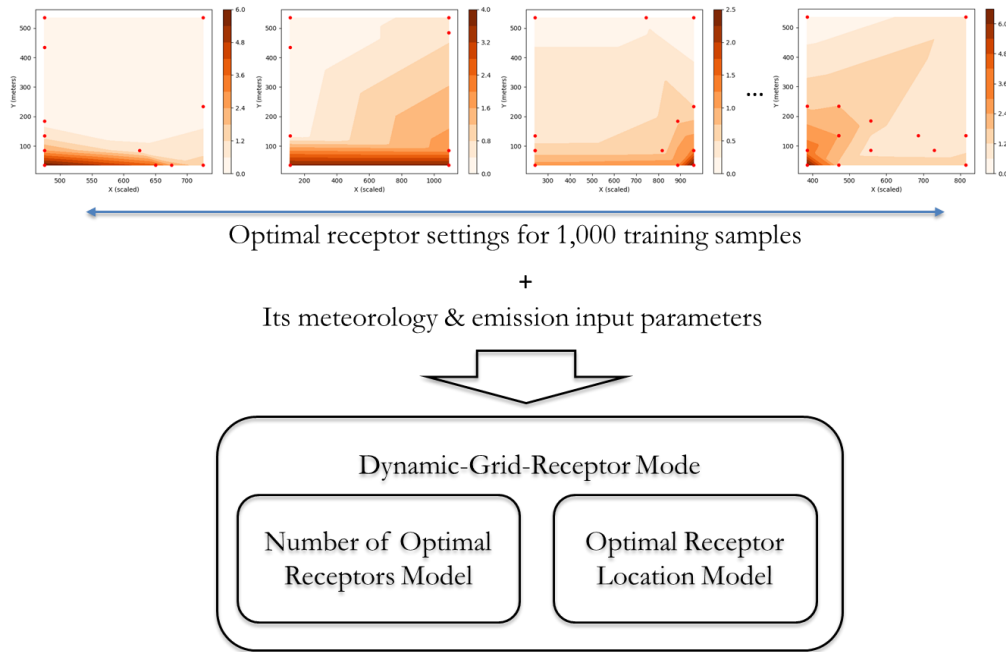


Figure 13 – Development of the Dynamic Grid-Receptor Model

4.5.1 Number of Optimal Receptor Model

The first model predicts the optimal number of receptors assigned for each link segment. The multiple linear regression model was developed using link emission rates and R-LINE meteorological parameters as potential explanatory variables (Equation (3)) (Greene, 2012).

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k + \varepsilon \quad (3)$$

Here, y is the optimal number of receptors to be estimated, x_k are the predictors, β_k is the regression coefficients, and ε is an error term.

The model results are shown in Table 3, suggesting that the difference in emission rates between the adjacent links and some of the meteorological parameters are important variables selecting the optimal number of receptors. The study first tested all variables in Table 3 (see the column labeled “Including all variables”). The results suggest that some of the variables (i.e., link emission, temperature, link length) are not of significant interest in identifying the number of optimal receptors, and thus they are excluded in the final model (the column “Excluding insignificant variables” in Table 3).

The signs of the model coefficients appear to be intuitively correct, and the trends appear to be related to the fundamental relationships identified in Chapter 4.3. For example, the large difference in emission rates between the adjacent links requires the placement of more receptors, potentially because the optimal receptor setting needs more receptors to control for the larger variation in pollutant concentration profile derived by the emission difference (Figure 11). Regarding wind speed, lower wind speeds appear to require more receptors. This can be also explained by the relationship between the wind speed and $PM_{2.5}$ concentration profile (Figure 10), in that the changes in $PM_{2.5}$ concentration profile with the distance from the link are greater under the lower wind speed condition, requiring more receptors to be placed close to the road. In terms of wind direction, under the condition that wind blows roughly parallel to the link (west or east winds) it requires a smaller number of receptors (as pollutants are carried along the roadway, and do not disperse very far north of the roadway). This may be because a larger variation in $PM_{2.5}$ concentration exists only on one side (either left-end or right-end of the link segment).

Table 3 – Linear Regression Model for Optimal Number of Receptors

Description	Including all variables		Excluding insignificant variables	
	Estimate	P-value	Estimate	P-value
Intercept	10.579	0.916	10.912	0.000
Link emission ($\mu\text{g}/\text{m}/\text{s}$)	-0.676	6.249	-	-
Difference in link emissions to that of adjacent link ($\mu\text{g}/\text{m}/\text{s}$)	24.039	5.270	24.942	0.000
Temperature ($^{\circ}\text{C}$)	0.018	0.012	-	-
Wind speed (m/s)	-0.129	0.038	-0.130	0.001
Wind direction (if 110° to 250°)	1.416	0.513	1.380	0.007
Link length (m)	0.000	0.001	-	-
Adjusted R-squared	0.163		0.171	

4.5.2 *Optimal Receptor Location Model*

Given the number of optimal receptors, a set of optimal receptor locations are chosen among the candidate receptor sites. To this end, this study developed a series of logistic regression models (equation (4)) to predict the probabilities of being chosen for the optimal location (Washington, et al., 2011; Greene, 2012). The study constructed a total of 121 logistic regression models for each candidate receptor site, with the explanatory variables related to link emission and meteorological conditions.

$$E\{y_i\} = \frac{\exp(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_k x_{ik})}{1 + \exp(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_k x_{ik})} \quad (4)$$

Here, y_i is an independent Bernoulli random variable for the i^{th} observation with an expected value $E\{y_i\}$, where y equals one if the location is chosen for the optimal location, and y equals zero otherwise. The logistic regression model is constructed with k predictor variables, known constant x and coefficients β to be estimated. Here, the probability of being chosen as the optimal location is expressed as equation (5):

$$P(X = 1) = \frac{1}{1 + \exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k)} \quad (5)$$

In terms of the explanatory variables, the same set of the variables used for the model for the optimal number of receptors was also applied to the logistic regression models with some variable modifications, while one more variable (the number of optimal receptors) is added to the variable set, assuming that the probability of each receptor being chosen for the optimal receptor depends on the total number of optimal receptors to be placed. In particular, the models considered two variables representing the emission differences in the adjacent links for both of the left-side (link 1 vs. link 2) and the right-side (link2 vs. link 3) (equation (6)).

$$emission\ difference = \frac{|emission_{link2} - emission_{link1\ or\ link3}|}{emission_{link2}} \quad (6)$$

Although a parsimonious variable selection process is preferred to increase the goodness-of-fit of the models, this study included all of the considered variables in the logistic regression models to estimate the relative probability of one receptor being chosen among the others under the same conditions (the probabilities of all of the candidate receptors being chosen are estimated with the same set of predictors).

As such, the dynamic grid-receptor model estimates the optimal number of receptors and the optimal receptor locations for each link segment. The modeling process is illustrated in Figure 14. As shown, the probabilities of being selected for the optimal receptor location are estimated for each candidate receptor site. Then, the model will select the receptor sites of the top N highest probabilities (where N is the optimal number of receptors).

The estimated model coefficients are shown in Table 14 in Appendix A, showing that most of the estimated coefficients are statistically significant at a p-value of 0.001. The model results showed that the signs and magnitudes of the coefficients that determine the probability of being chosen for the optimal receptor location vary depending on X and Y positions of receptor sites. For example, for the location X (scaled) = -0.5 and Y = 100 m, the probability of being chosen as the optimal receptor site increases as the emission difference on the left-side (link 1 vs. link 2) increases, while the difference in the emission difference on the right-side does not increase the probability of being chosen for the receptor site. As such, the model results suggest that the optimal receptor positions are quite dynamic depending on the link emission profiles and meteorological conditions.

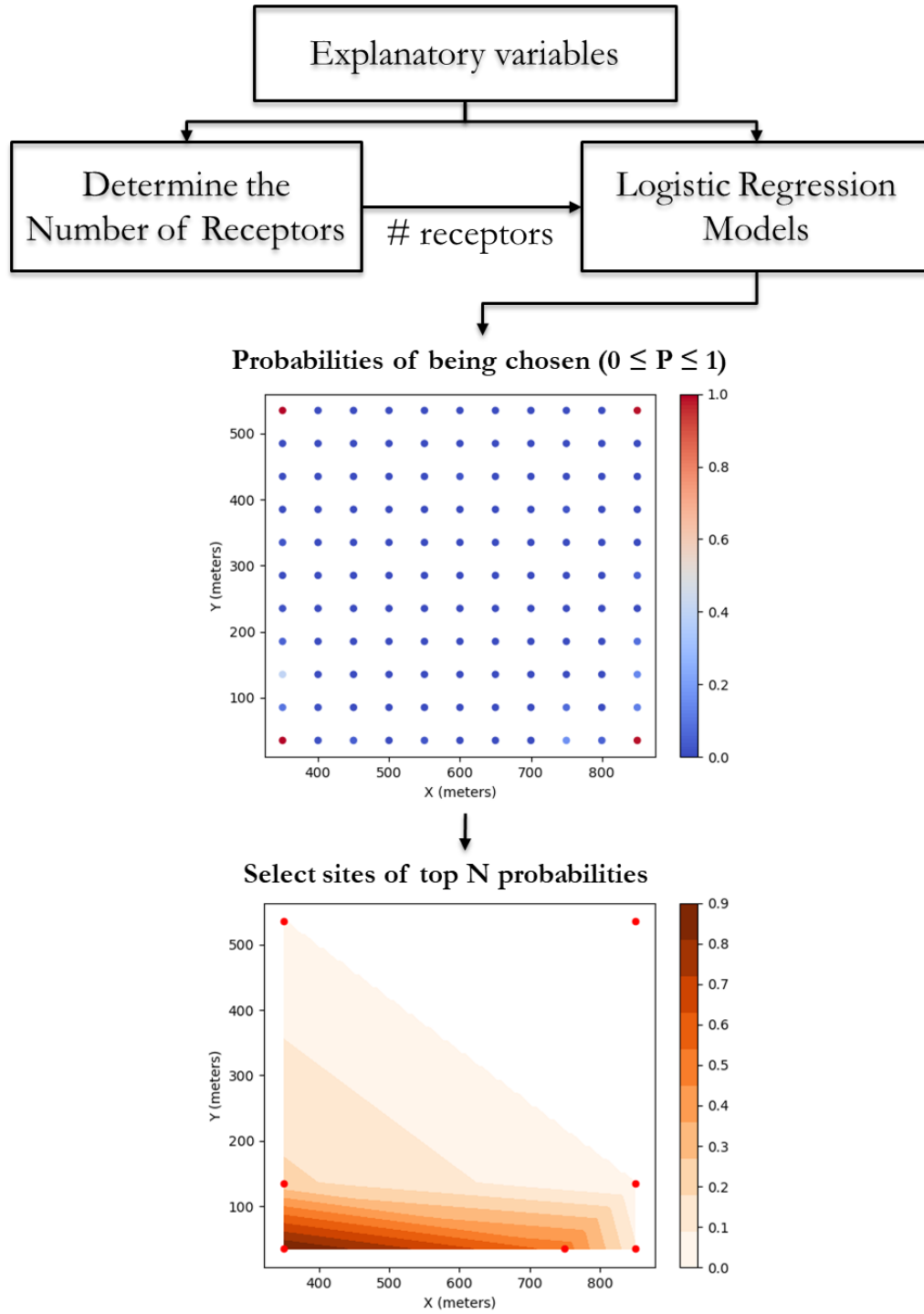


Figure 14 – Dynamic Grid-receptor Model Development

4.6 Model Verification

The dynamic grid-receptor model is verified using the 300 test sample results generated from the R-LINE simulations. Figure 15 illustrates the model verification process. In this process, the dynamic grid-receptor model is applied to the 300 test samples, by determining the optimal receptor locations based on its R-LINE input parameter values. Then, the $PM_{2.5}$ concentration profiles predicted by the dynamic grid-receptor model are compared with those predicted by the high-density receptor model (setting receptors every 5m), to measure how well the dynamic grid-receptor model approximates the $PM_{2.5}$ concentration profiles predicted by the dense receptor model using mean absolute error (MAE) (equation (7)). For each sample, MAE is calculated based on the $PM_{2.5}$ concentration profiles (on a 100 by 100 grid space) predicted by the high-density and dynamic-grid receptor models, and the average of MAEs for all test samples is calculated. Thereby, a small \overline{MAE} indicates that the dynamic grid-receptor well approximates the $PM_{2.5}$ concentration profiles as predicted by the high-density receptor model.

$$\overline{MAE} = \frac{1}{n} \sum_{i=1}^n \sum_{x,y} |PM_{2.5} (dense)_{i,x,y} - PM_{2.5} (dynamic-grid \text{ or } static-grid)_{i,x,y}| \quad (7)$$

Here, i refers to each test sample, and x, y refers to each grid area in the heatmaps.

To measure the performance of the dynamic grid-receptor model in terms of model prediction accuracy, the same processes are also applied to the static-grid-receptor model, where the number of receptors for the static-grid-receptor model is set to always higher

than or equal to that of the dynamic grid-receptor model (equation (8)). Then, \overline{MAE} for the static-grid-receptor model as compared to the high-density receptor model was also calculated (equation (7)).

$$N_2 = \lceil \sqrt{N_1} \rceil^2 \quad (8)$$

Here, N_1 is the number of dynamic grid-receptors, and N_2 is that for static-grid-receptors.

The model verification results, shown in Figure 15 and Table 4, suggest that the dynamic grid-receptor model can approximate the PM_{2.5} concentration profiles predicted by the high-density receptor model with a small number of receptors. For example, the dynamic grid-receptor model selected 13.2 receptors on average to approximate the PM_{2.5} concentration profiles as predicted by the dense-receptor model, which requires 11,541 receptors on average.

The comparison results for the \overline{MAE} s of the dynamic-grid and static-grid receptor models show that the dynamic grid-receptor model better approximates the PM_{2.5} concentration profile than the static-grid-receptor model (Table 4). For example, the \overline{MAE} for the dynamic grid-receptor model was 0.458 for the entire hypothetical network area ($Y = 0$ m to 500 m), which is smaller than that for the static-grid-receptor model (0.555). The trends become more obvious when it comes to the near-road areas ($Y = 0$ m to 200 m): 0.777 for the dynamic grid-receptor model and 1.277 for the static-grid-receptor model. However, the prediction power of the dynamic grid-receptor appears to be lower than the static-grid-receptor model for areas far from the link segment (e.g., $Y = 200$ m to 500 m). This may be because the dynamic grid-receptor model tends to place more receptors near

the roadway segment to capture the large variations in PM_{2.5} concentration in the areas (i.e., MAE is dominated by the near-road values). However, considering that the near-road areas are more vulnerable to high pollution concentrations, such a receptor placement strategy focusing on the better prediction on the near-road areas appears to be a better method in identifying hot-spots.

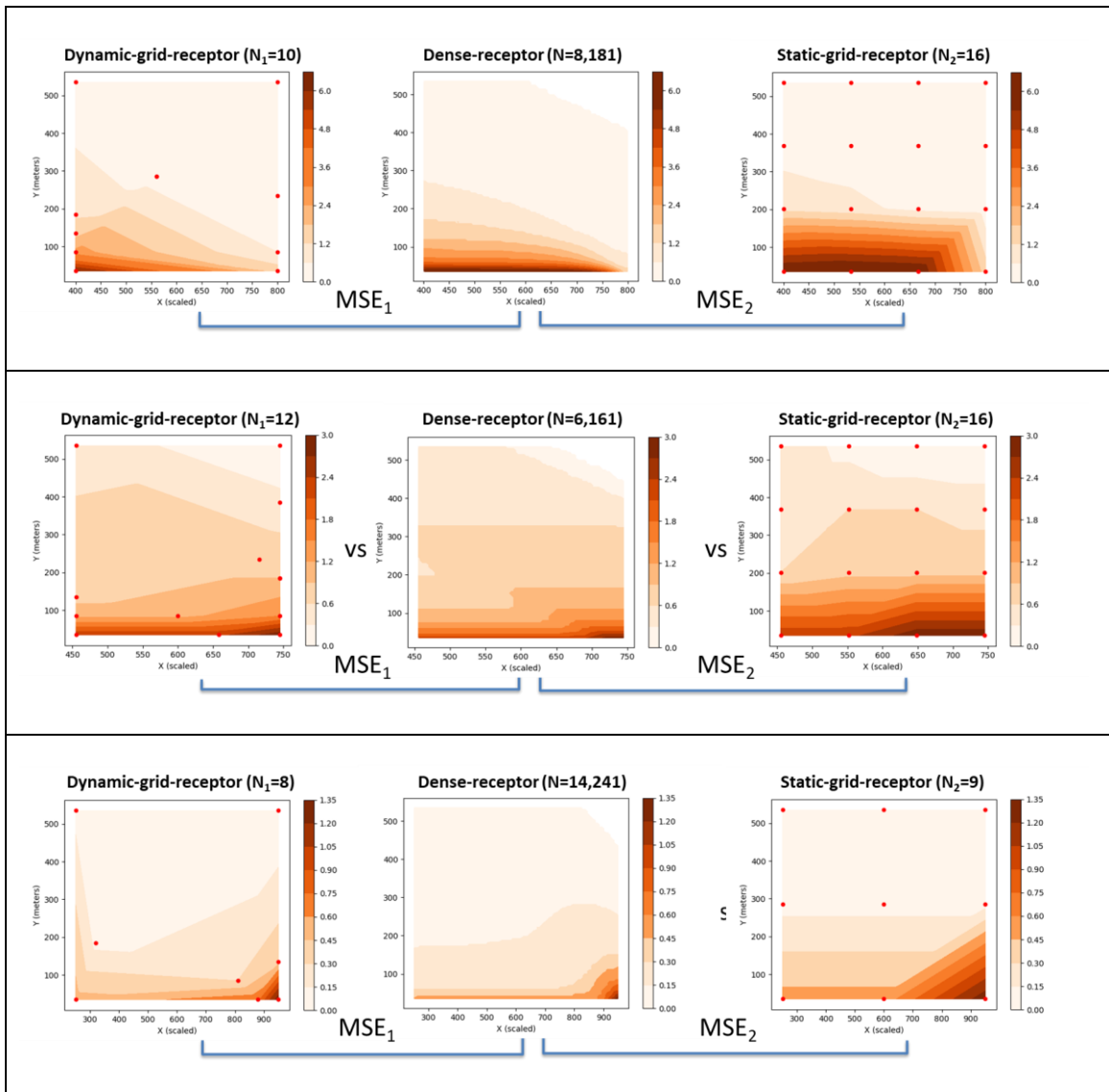


Figure 15 – Examples of Model Verification Processes

Table 4 – Comparison Analysis Results between Dense-receptor vs. Dynamic-grid/Static-grid Receptor Models

Receptor Area (distance from link area)	Dense-receptor	Dense-receptor vs. Static Grid-Receptor	Dense-receptor vs. Dynamic Grid-Receptor
Average number of receptors	11,541.1	16.3	13.2
\overline{MAE} (Y = 0m to 500m)	-	0.555	0.458
\overline{MAE} (Y = 0m to 100m)	-	1.920	1.010
\overline{MAE} (Y = 0m to 200m)	-	1.277	0.777
\overline{MAE} (Y = 100m to 500m)	-	0.214	0.319
\overline{MAE} (Y = 200m to 500m)	-	0.074	0.244

4.7 Case Studies

The case studies focus on evaluating the model prediction power of PM_{2.5} concentrations with the dynamic grid-receptor model, by applying the dynamic grid-receptor model to practical transportation networks. To this end, the annual average PM_{2.5} concentration estimation results are obtained using both the dynamic grid-receptor model and the static-grid-receptor model for comparison. The annual average PM_{2.5} concentrations for each of the case study areas were estimated with AERMOD, based on the Georgia meteorological processor data (Georgia EPD, 2019), and link emission rates estimated by ARC-ABM15 (see Chapter 6.8). Again, because each study area (16km²) requires many receptors to be

processed, Georgia Tech's PACE supercomputing cluster was used for the AERMOD simulations.

The study selected three subareas in the Metro Atlanta area (Figure 16): 1) the City of Decatur, GA; 2) Midtown Atlanta, GA; and 3) Downtown Atlanta, GA. Each study area represents different levels of road network densities. For example, the City of Decatur accounts for the areas of low roadway density, and the downtown Atlanta accounts for the high roadway density area. For each of the areas, four different models are applied to evaluate the dynamic grid-receptor models as follows.

- The City of Decatur, dynamic grid-receptor model (5,312 receptors, 325 links)
- The City of Decatur, static-grid-receptor model (5,329 receptors, 325 links)
- The City of Decatur, high-density receptor model (16,482 receptors, 325 links)
- The City of Decatur, low-density grid-receptor model (624 receptors, 325 links)
- Midtown Atlanta, dynamic grid-receptor model (10,441 receptors, 613 links)
- Midtown Atlanta, static-grid-receptor model (10,609 receptors, 613 links)
- Midtown Atlanta, high-density receptor model (8,282 receptors, 613 links)
- Midtown Atlanta, low-density grid-receptor model (624 receptors, 613 links)
- Downtown Atlanta, dynamic grid-receptor model (22,662 receptors, 1,632 links)
- Downtown Atlanta, static-grid-receptor model (22,801 receptors, 1,632 links)
- Downtown Atlanta, high-density receptor model (16,482 receptors, 1,632 links)
- Downtown Atlanta, low-density grid-receptor model (624 receptors, 1,632 links)

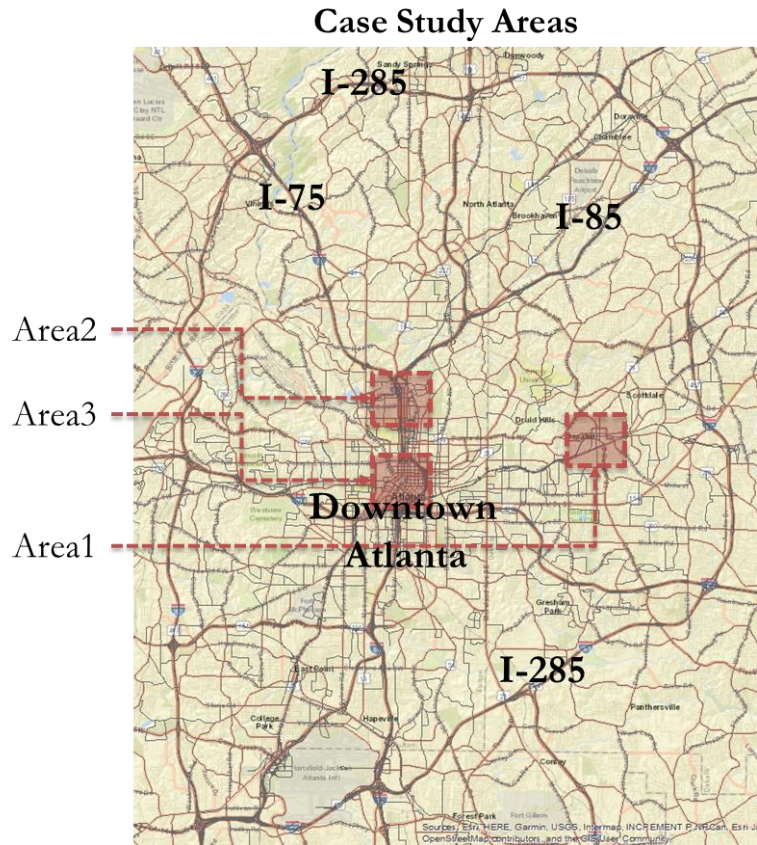


Figure 16 – Case Study Areas for Dynamic Grid-Receptor Model Evaluation

Figure 17, Figure 18, and Figure 19 illustrate the receptor settings for each of the cases. Figure 17 illustrates the receptor settings for the City of Decatur, applying the four different models. The figure shows that the dynamic grid-receptor model predicts the optimal receptor locations based on the related link emission rates and meteorology conditions, thereby setting a total of 5,312 receptors along the roadway network (the practical implementation of the dynamic grid-receptor model for the ARC-ABM15 network is described in more detail in the following section). As shown, the dynamic grid-receptor model tends to set relatively dense receptors near road segments (e.g., 5m away from the link segments), and set relatively lower-dense receptors with the distance from the link segments. For a fair comparison, the static-grid-receptor model set a similar

number of receptors (5,329 receptors, around 70m receptor resolution) to that of the dynamic grid-receptor model.

This study also applied two other models, high-density receptor and low-density grid-receptor models to evaluate the prediction of the dynamic grid-receptor model. Figure 17 illustrates how those models set the receptors. The high-density receptor model arbitrary selects a number of link segments in the study area, and sets 5m-resolution receptors along the line perpendicular to the link segment and within 200m from the link segment. This model was applied as a way to derive the elaborate near-road PM_{2.5} concentration profile, and the model results are compared with the results from the dynamic grid-receptor model. Lastly, the low-density grid-receptor model (200m resolution) was also applied to measure the difference in the prediction of the dynamic grid-receptor model as compared to the traditional modeling method.

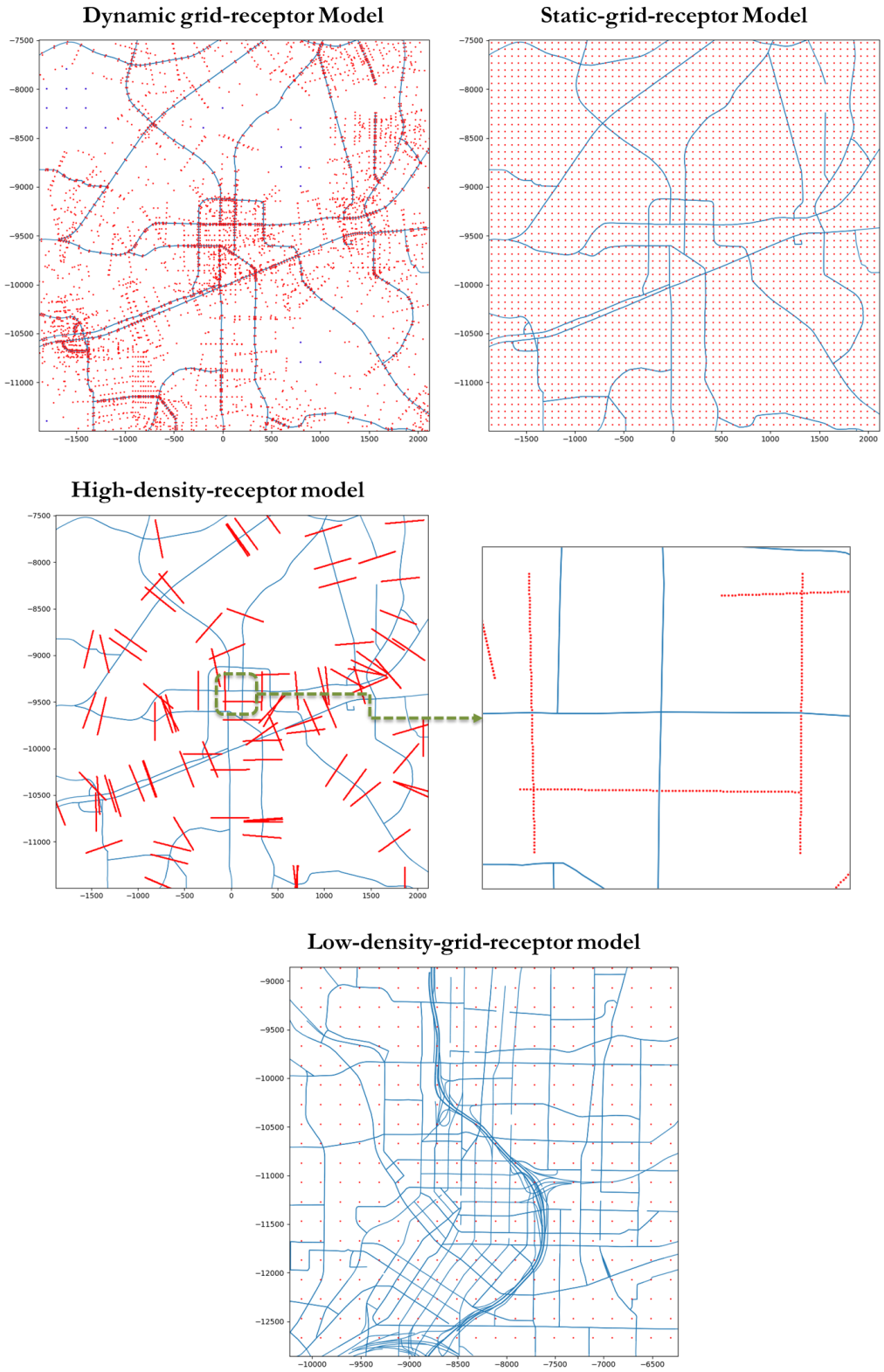


Figure 17 – Receptor Settings for the City of Decatur

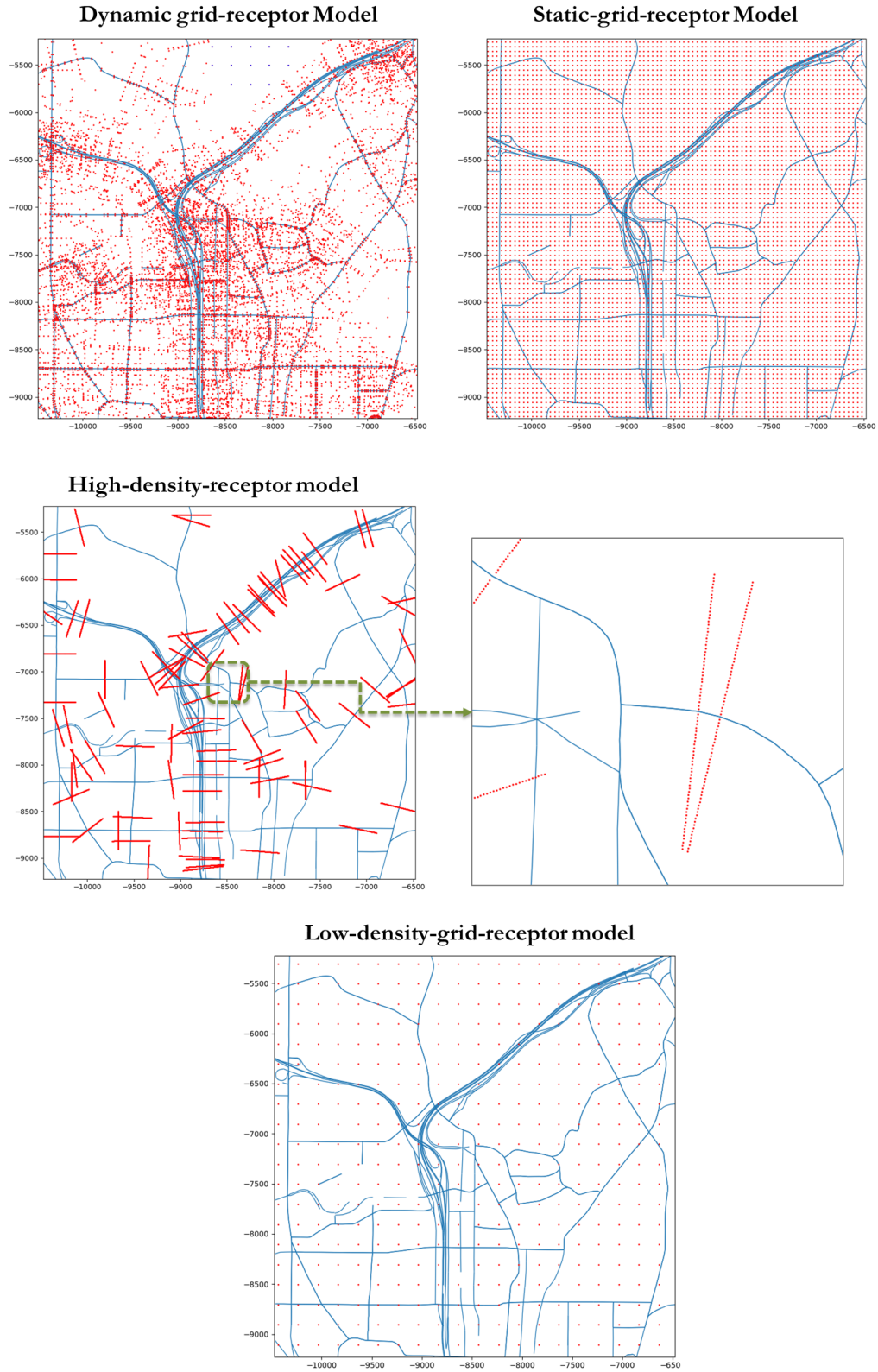


Figure 18 - Receptor Settings for Midtown Atlanta

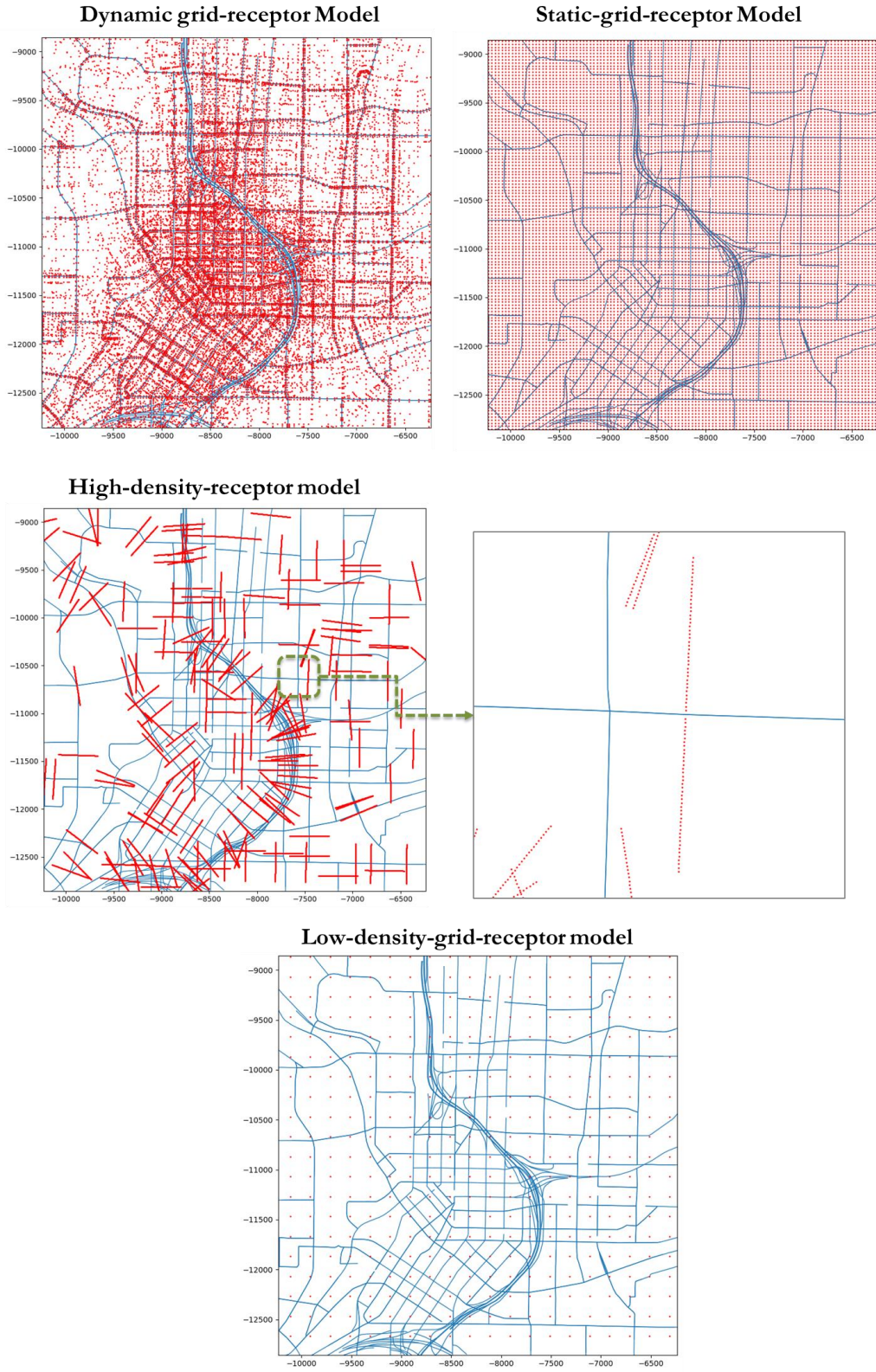


Figure 19 - Receptor Settings for Downtown Atlanta

4.7.1 Practical Implementation of Dynamic grid-receptor Model to Transportation Network

This study develops a streamlined PythonTM-based program to automate the processes of setting optimal receptors by applying the dynamic grid-receptor model coefficients to the link emission rates database and desired meteorology conditions. As such, setting the optimal receptors per link is easy to implement through the program, allowing user-specific input values. For example, the case study applied a wind speed of 1.0 m/s assuming for the worst-case scenarios in the dynamic grid-receptor model. For wind direction, the case study applied all wind directions (0° to 360°) at every 45° , by selecting any receptors which are identified by each of the wind direction values. For this, the program iteratively develops several dynamic grid-receptor models for each link segment, applying the wind directions at every 45° , and then the model extracts unique receptors identified by each of the modeling iterates.

There are several programming rules embedded in the program that aim to increase the model prediction accuracy as well as the modeling efficiency. The rules embedded in the dynamic grid-receptor model are illustrated in Figure 20. First, the link segment-based receptor generation process tends to generate some unnecessary receptors at intersections or the sections where multiple links meet. For example, the dynamic grid-receptor model generates numerous receptors for each link connected to the intersection node, while some of the receptors may not be directly related to the specified link segment (rather, they are associated with other links belonging to the intersection). In this circumstance, the program sets a receptor area that is specified within the angles that are half of the angles defined by the two adjacent link segments (Rule 1 in Figure 20). From this, the program only assigns

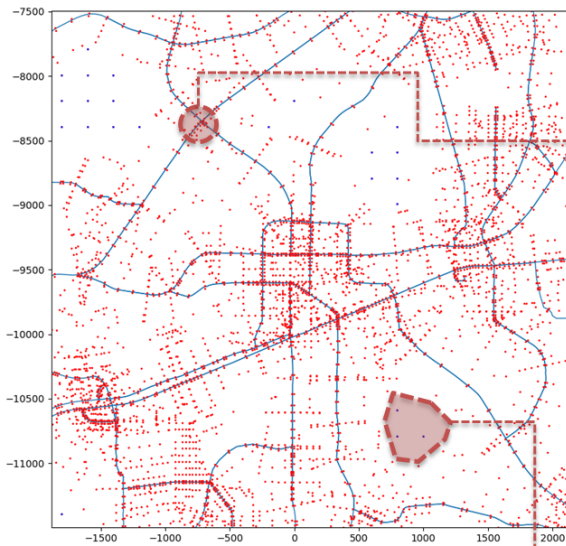
the receptors that are directly related to the specified link segment. (The program drops the receptors that are not included in the receptor area.)

Second, because the dynamic grid-receptor model is developed based on the hypothetical network that specifies the receptor areas within 500m away from the link segment, the dynamic grid-receptor model can only place receptors within the area. In this circumstance, some areas in the transportation network (i.e., areas more than 500m away from the transportation network) are not covered by the dynamic grid-receptor model. To make up for these areas, the program sets a minimum 200m-resolution receptor grid to infill in these areas (Rule 2 in Figure 20).

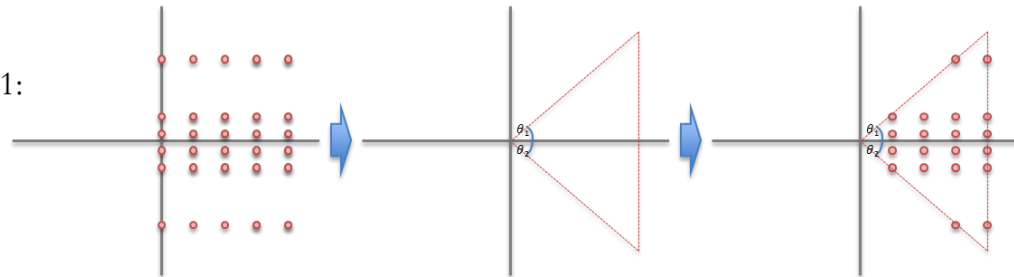
Third, because the dynamic grid-receptor model is designed for the near-road receptors, the on-road receptors (those placed in mixing zone areas) generated by the dynamic grid-receptor model are checked and removed in the last step of the program (a supplemental analysis may be required to integrate both of the over-the-road and near-road receptors in the same modeling framework for exposure assessment).

General rules:

- The dynamic-grid-receptor model is applied to every link segment
- Wind speed is assumed to be 1m/s for worst case
- Wind direction is applied for all direction at every 45°



Rule 1:



- Link segment-based receptor placements tends to produce a lot of receptors at intersections
- Remove the receptors that is more close to other link segments

Rule 2:

- The dynamic-grid-receptor model is applied for the areas within 500m away from the link segment
- Thus, after the model implementation, 200m-resolution receptors are placed for the rest of the areas

Rule 3:

- After all, on-road receptors are removed (above the research focus)
- For this, the buffer areas is set for the areas within link width + 5m

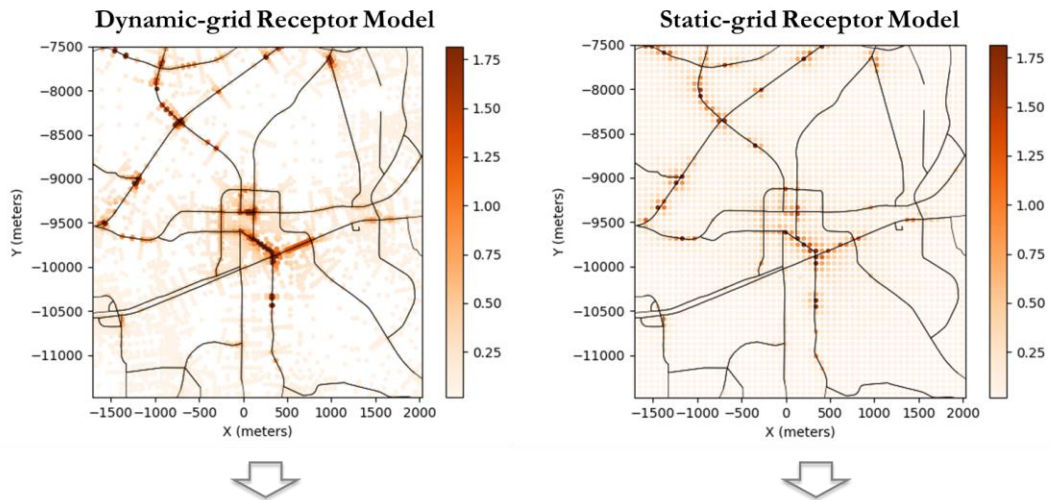
Figure 20 – Dynamic Grid-Receptor Model Implementation for the ARC-ABM15 Network

4.7.2 *Comparison of Dynamic grid-receptor Model and Static-grid-receptor Model*

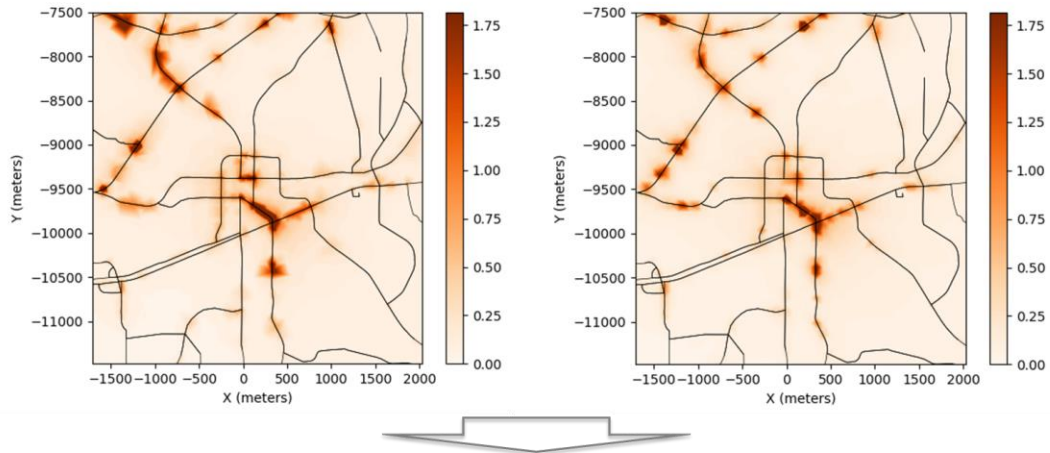
For each of the case study areas, the annual average PM_{2.5} concentrations at the receptors governed by both of the dynamic grid-receptor and the static-grid-receptor models are predicted (Figure 21, Figure 22, Figure 23). As previously discussed, a similar number of receptors were generated by the dynamic grid-receptor and the static-grid-receptor models for a fair comparison: for example, 5,312 receptors were generated by the dynamic grid-receptor model vs. 5,329 receptors were generated by the static grid-receptor model, for the City of Decatur. Then, the PM_{2.5} concentrations predicted for the receptors are used for generating PM_{2.5} concentration heatmaps onto 1,000 by 1,000 grid spaces. Overall, the two models appear to produce similar PM_{2.5} concentration profiles, identifying hot-spots at common geographic locations. The difference in PM_{2.5} concentration profiles is calculated by subtracting the PM_{2.5} concentration profile from the dynamic grid-receptor model by that from the static-grid-receptor model. The result shows that the PM_{2.5} concentration profiles predicted by both of the models are quite similar in most of the areas (the white-colored areas in Figure 21-(c)). The results also suggest that the dynamic grid-receptor model tends to predict higher PM_{2.5} concentrations, particularly near the roadways. This may be because the dynamic grid-receptor model tends to set at least some number of receptors near road areas, making it possible to consider the high PM_{2.5} concentration estimates near road areas in generating the PM_{2.5} concentration profile. However, because the static-grid-receptor pays less attention to near-road pollutant concentration, and the distance between the receptor and the link segment is random, the static-grid-receptor model is less able to capture the high PM_{2.5} concentration in some of the near-road areas. The same trends were found for the other case study areas. For example, the areas with

red-colored in Figure 22-(c) and Figure 23-(c) appear to be consistently larger than the areas with blue-colored, suggesting that the dynamic grid-receptor model has a better prediction for high concentration areas.

(a) Annual Average PM2.5 Concentrations at Receptors ($\mu\text{g}/\text{m}^3$)



(b) Annual Average PM2.5 Concentration Profile (1000 by 1000 grid map)



(c) Difference in the PM2.5 Concentration Profiles
PM2.5 (Dynamic-grid) – PM2.5 (Static-grid)

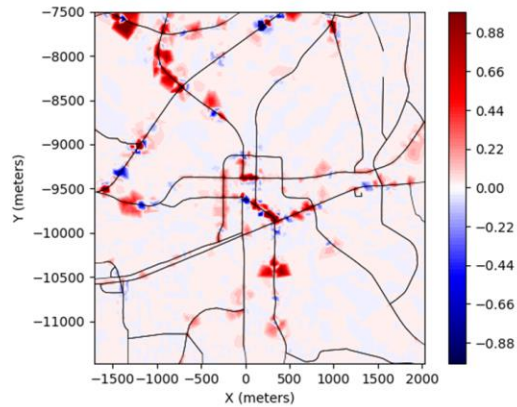
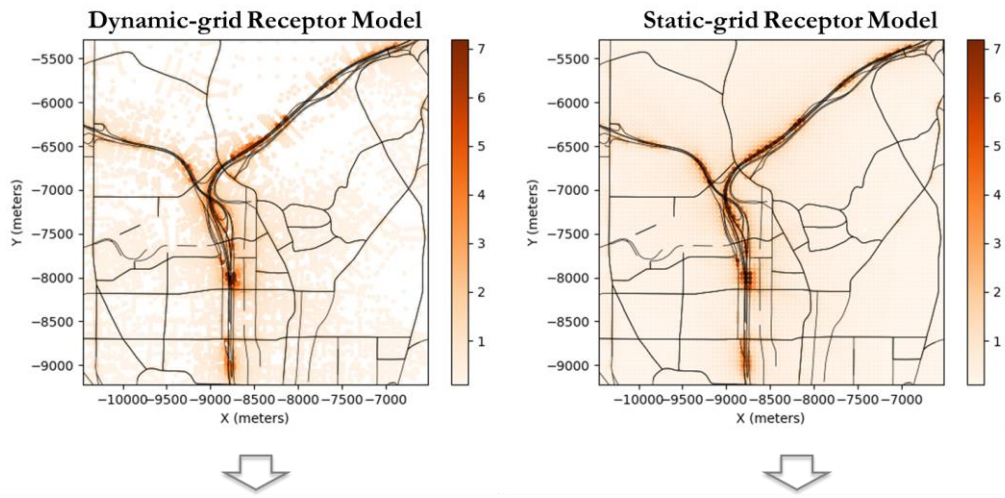
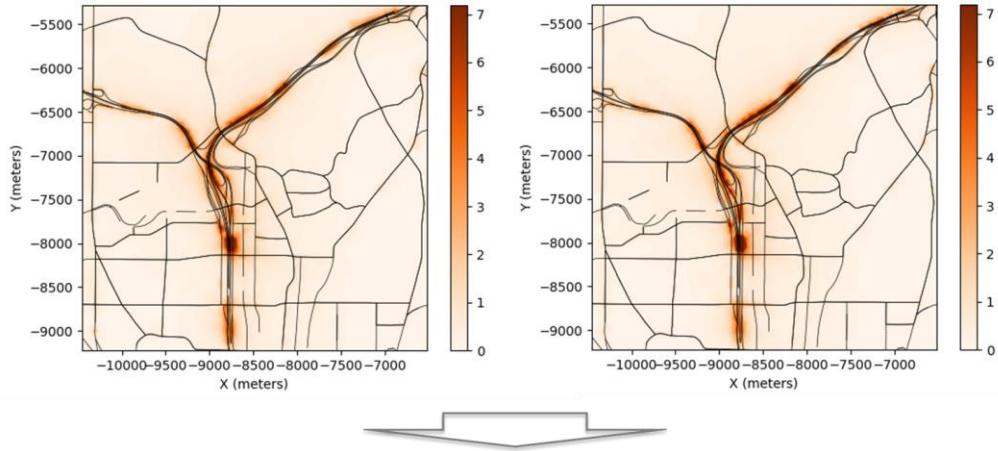


Figure 21 – Dynamic Grid-Receptor Model vs. Static Grid-Receptor Model (City of Decatur, GA)

(a) Annual Average PM2.5 Concentrations at Receptors ($\mu\text{g}/\text{m}^3$)



(b) Annual Average PM2.5 Concentration Profile (1000 by 1000 grid map)



(c) Difference in the PM2.5 Concentration Profiles

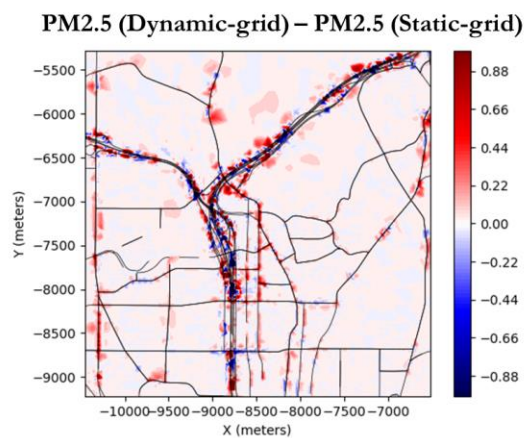
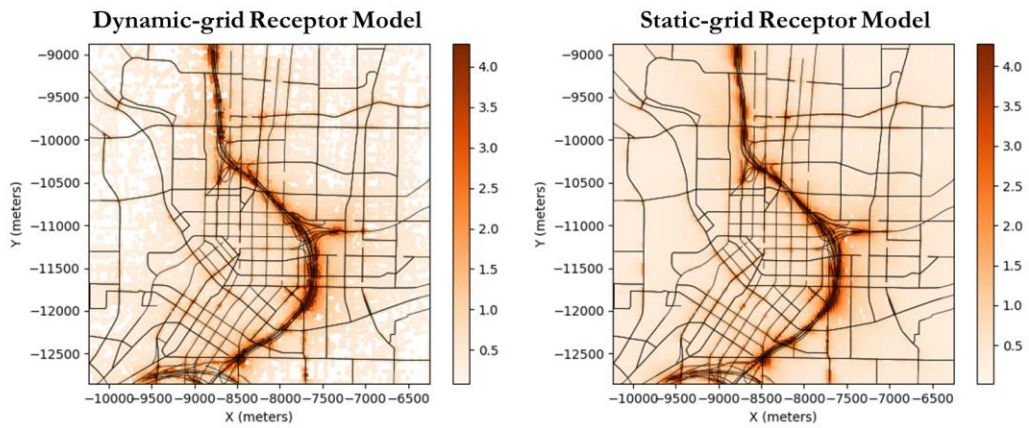
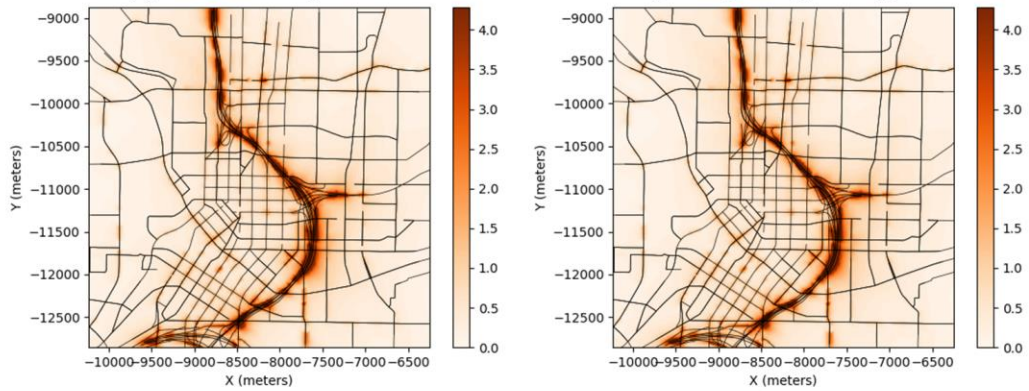


Figure 22 – Dynamic Grid-Receptor Model vs. Static Grid-Receptor Model (Midtown Atlanta, GA)

(a) Annual Average PM2.5 Concentrations at Receptors ($\mu\text{g}/\text{m}^3$)



(b) Annual Average PM2.5 Concentration Profile (1000 by 1000 grid map)



(c) Difference in the PM2.5 Concentration Profiles

PM2.5 (Dynamic-grid) – PM2.5 (Static-grid)

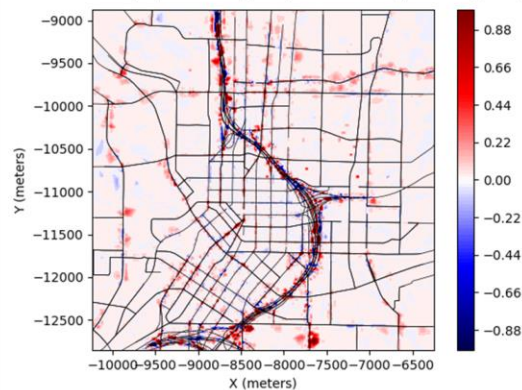
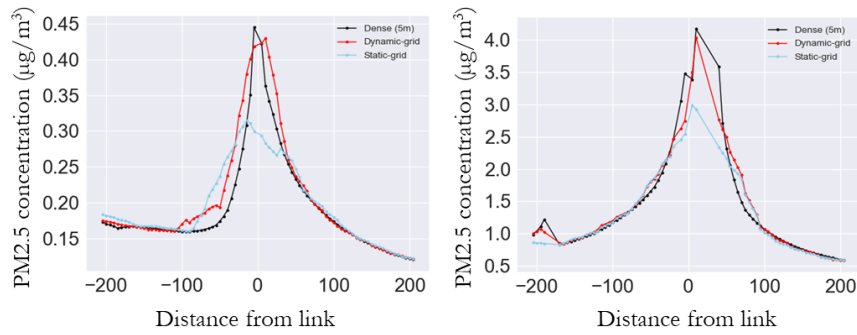


Figure 23 – Dynamic Grid-Receptor Model vs. Static Grid-Receptor Model (Downtown Atlanta, GA)

A supplemental analysis was conducted to evaluate the prediction of the dynamic grid-receptor model compared to the static-grid-receptor model, by comparing the $PM_{2.5}$ concentration profiles predicted by the two models with those predicted by the high-density receptor model. The analysis was conducted using the 5m-receptor site resolution set for the high-density receptor model. The $PM_{2.5}$ concentration profiles for the dynamic grid-receptor and static-grid-receptor models were retrieved by projecting the $PM_{2.5}$ concentration estimates onto the heatmaps generated for those models. Some examples of the $PM_{2.5}$ concentration profiles among the models are shown in Figure 24. The results suggest that the dynamic grid-receptor model better-predicts pollutant concentrations than the static-grid-receptor model for some areas, and vice versa, depending on the geographic locations. This suggests that the near-road pollutant concentration profiles can be biased depending on the receptor placements, raising a question about which type of model can overall predict unbiased pollutant concentration profiles, which will be addressed in the following statements.

(a) Examples that the dynamic-grid-receptor model produce a better prediction



(b) Examples that the static-grid-receptor model produce a better prediction

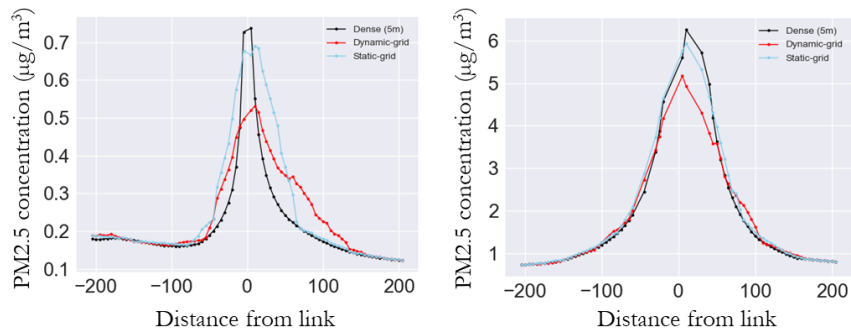


Figure 24 – PM_{2.5} Concentration Profiles for High-Density Receptor, Dynamic Grid-Receptor, and Static-Grid-Receptor Models

Based on the PM_{2.5} concentration profile charts in Figure 24, the difference in the PM_{2.5} concentrations predicted by these different models is measured by equation (9) and (10). Equation (9) explains the mean absolute error in the PM_{2.5} concentrations predicted by the high-density receptor model versus the dynamic grid-receptor or static-grid-receptor models. This measurement is adopted to identify which models better approximate the PM_{2.5} concentration profiles as predicted by the high-density receptor model. A similar measurement was also adopted that measures the mean absolute percentage error (equation (10)).

$$MAE = \frac{1}{n} \sum_{i=1, d=1}^n |PM2.5 (high\ dense)_{i,d} - PM2.5 (dynamic\ or\ static)_{i,d}| \quad (9)$$

$$MAPE = \frac{1}{n} \sum_{i=1, d=1}^n \left| \frac{PM2.5 (high\ dense)_{i,d} - PM2.5 (dynamic\ or\ static)_{i,d}}{PM2.5 (high\ dense)_{i,d}} \right| \quad (10)$$

Here, i refers to each link segment selected for high-density receptor model, and d is the distance from the link segment ($0 \leq d \leq 200$).

The MAE and MAPE results for each case study area are summarized in Table 5. The results show that the dynamic grid-receptor model better approximates the $PM_{2.5}$ concentration profiles as predicted by the high-density receptor model than the static-grid-receptor model across the three different regions, although the marginal differences between the dynamic grid-receptor and static-grid-receptor models are only $0.004 \mu\text{g}/\text{m}^3$ to $0.011 \mu\text{g}/\text{m}^3$ or 0.3% to 0.6%.

Summarizing the results addressed in this section, the case study results suggest that the dynamic grid-receptor model appears to produce better predictions of potential hot-spots than the static-grid-receptor model. The dynamic grid-receptor model also appears to better approximate the near-road pollutant concentration profiles as predicted by a high-density receptor model than (or at least similar to) the static-grid-receptor.

Table 5 – Comparison Results of PM_{2.5} Concentration Profiles among High-density Receptor, Dynamic grid-receptor, and Static-grid-receptor Models

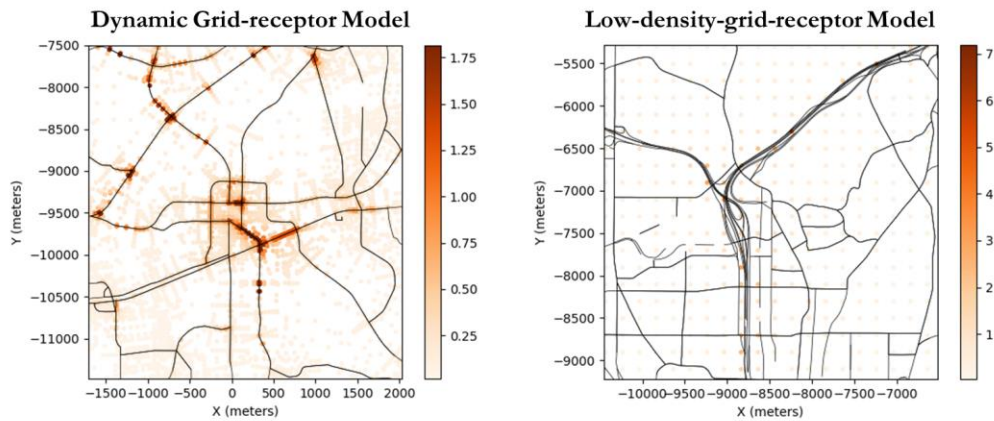
Classification	Description	Dense-receptor vs. Dynamic grid-receptor	Dense-receptor vs. Static-grid-receptor
Case 1: City of Decatur	MAE	0.053	0.060
	MAPE	0.049	0.050
Case 2: Midtown Atlanta	MAE	0.073	0.081
	MAPE	0.069	0.072
Case 3: Downtown Atlanta	MAE	0.057	0.068
	MAPE	0.047	0.053

4.7.3 Dynamic Grid-Receptor Model vs. Low-density Receptor Model

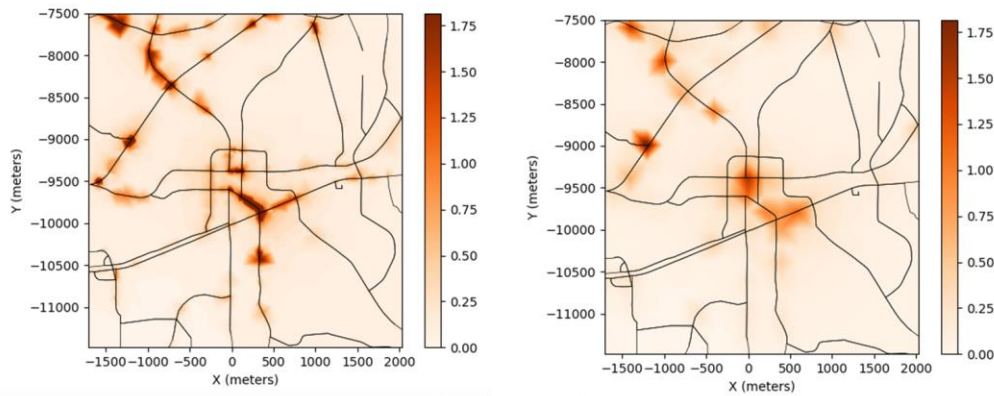
The above section suggests that the dynamic grid-receptor model produces comparable PM_{2.5} concentration profiles to the high-density receptor model. This section focuses on addressing whether the dynamic grid-receptor model can produce a better pollutant concentration profile than the traditional low-density grid-receptor model (e.g., 200m resolution receptor setting). The same approach was also applied for comparing the PM_{2.5} concentration profiles predicted by the dynamic grid-receptor and low-density grid-receptor models. Figure 25, Figure 26, and Figure 27 suggest that the difference in the PM_{2.5} concentration profiles predicted by the dynamic grid-receptor and low-density grid-

receptor model appears to be much greater than the difference between the dynamic grid-receptor and the static-grid-receptor models. In particular, a huge difference in the $PM_{2.5}$ concentration profiles was observed for downtown Atlanta. This suggests that the low-density grid-receptor model may produce a biased pollutant concentration profile, depending on the proximity between the link segments and the receptor location.

(a) Annual Average PM2.5 Concentrations at Receptors ($\mu\text{g}/\text{m}^3$)



(b) Annual Average PM2.5 Concentration Profile (1000 by 1000 grid map)



(c) Difference in the PM2.5 Concentration Profiles
PM2.5 (Dynamic-grid) – PM2.5 (Low-density-grid)

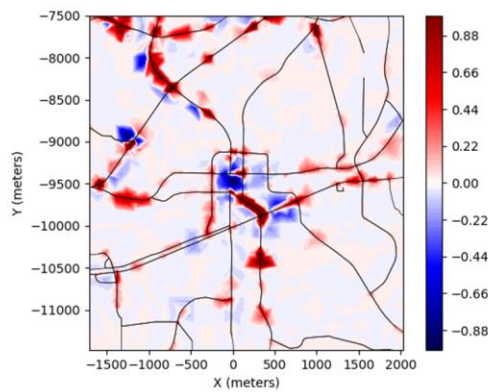
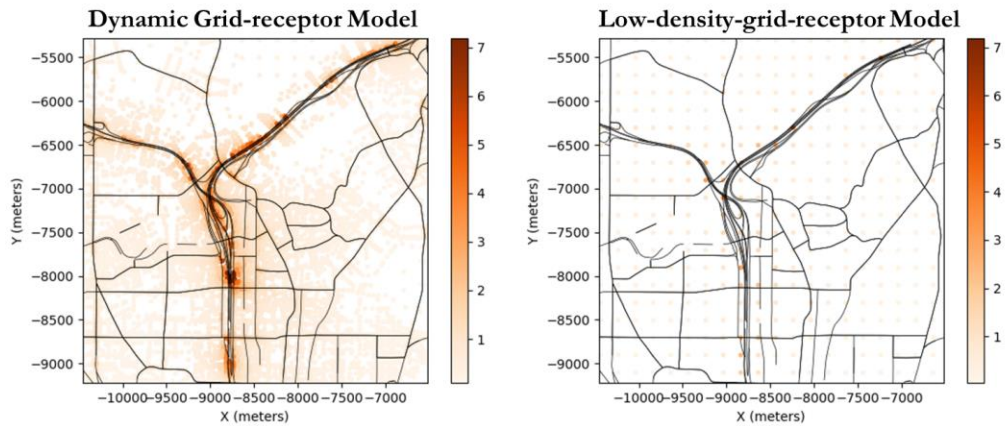
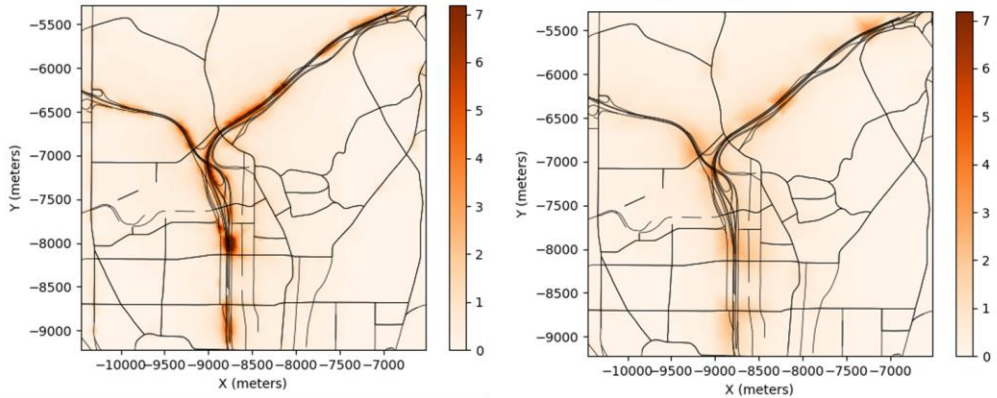


Figure 25 – Dynamic Grid Receptor Model vs. Low-density Grid-Receptor Model (City of Decatur, GA)

(a) Annual Average PM2.5 Concentrations at Receptors ($\mu\text{g}/\text{m}^3$)



(b) Annual Average PM2.5 Concentration Profile (1000 by 1000 grid map)



(c) Difference in the PM2.5 Concentration Profiles
PM2.5 (Dynamic-grid) – PM2.5 (Low-density-grid)

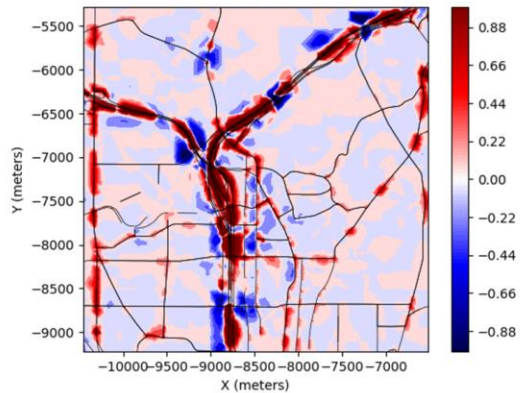
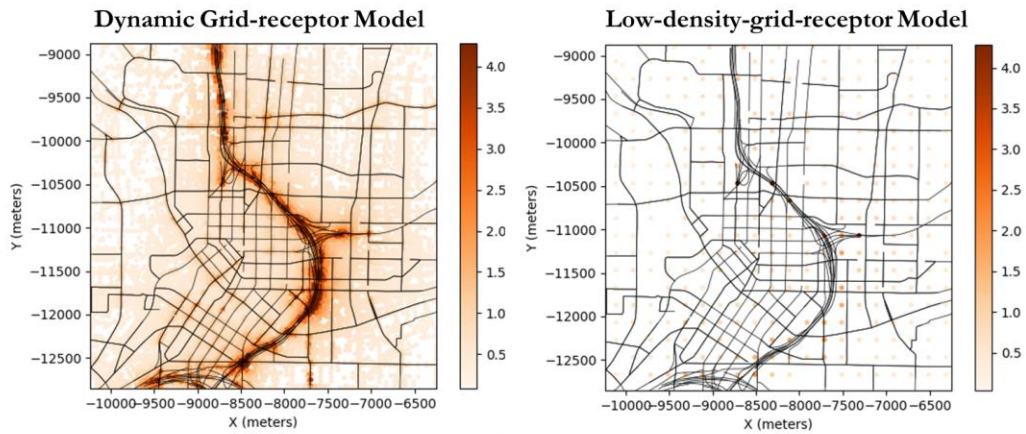
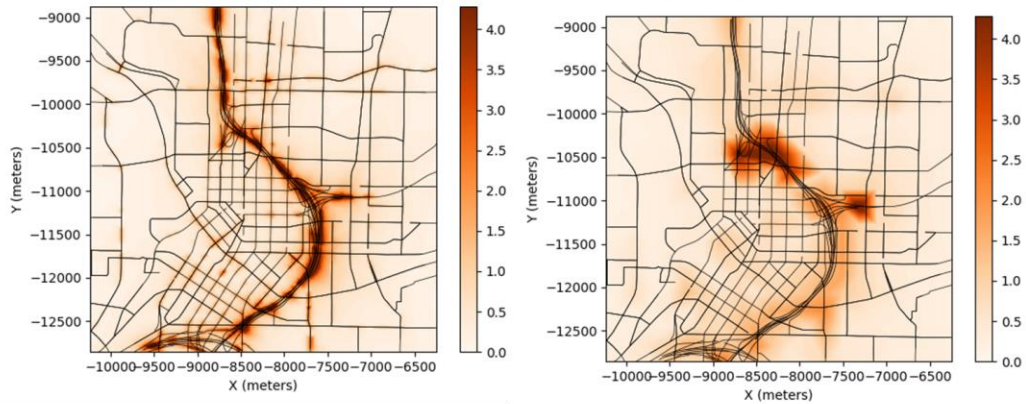


Figure 26 – Dynamic Grid-Receptor Model vs. Low-density Grid-Receptor Model (Midtown Atlanta, GA)

(a) Annual Average PM2.5 Concentrations at Receptors ($\mu\text{g}/\text{m}^3$)



(b) Annual Average PM2.5 Concentration Profile (1000 by 1000 grid map)



(c) Difference in the PM2.5 Concentration Profiles
 $\text{PM}_{2.5}(\text{Dynamic-grid}) - \text{PM}_{2.5}(\text{Low-density-grid})$

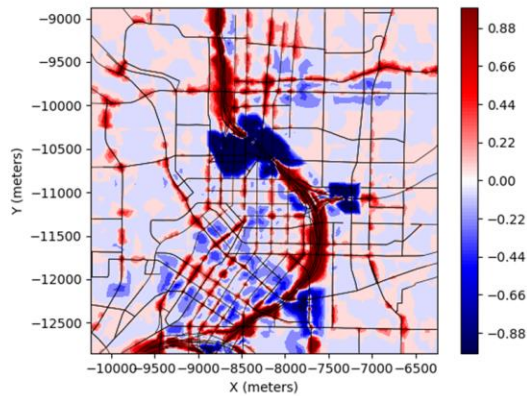


Figure 27 – Dynamic Grid-Receptor Model vs. Low-density Grid-Receptor Model (Downtown Atlanta)

Similar to the previous section, the difference in the $PM_{2.5}$ concentrations predicted by the high-density receptor model and low-density grid-receptor is measured by equation (11) and (12). The MAE and MAPE results for each case study area are summarized in Table 6. Comparing the MAE and MAPE results in Table 5 and Table 6, it also suggests that the $PM_{2.5}$ concentration profiles predicted by the low-density grid-receptor model are quite different from those predicted by the dynamic grid-receptor model. This also suggests that the dynamic grid-receptor model can produce a much more unbiased pollutant concentration profile than is obtained in traditional low-density grid modeling.

$$MAE = \frac{1}{n} \sum_{i=1, d=1}^n |PM_{2.5} (high\ dense)_{i,d} - PM_{2.5} (low\ dense\ grid)_{i,d}| \quad (11)$$

$$MAPE = \frac{1}{n} \sum_{i=1, d=1}^n \left| \frac{PM_{2.5} (high\ dense)_{i,d} - PM_{2.5} (low\ dense\ grid)_{i,d}}{PM_{2.5} (high\ dense)_{i,d}} \right| \quad (12)$$

Here, i refers to each link segment selected for high-density receptor model, and d is distance from the link segment ($0 \leq d \leq 200$).

Table 6 – Comparison Results of PM_{2.5} Concentration Profiles among High-density Receptor, Dynamic grid-receptor, and Low-density grid-receptor Models

	Description	Dense-receptor vs. Dynamic grid-receptor	Dense-receptor vs. Low-dense-grid-receptor
Case 1: City of Decatur	MAE	0.053	0.112
	MAPE	0.049	0.314
Case 2: Midtown Atlanta	MAE	0.073	0.230
	MAPE	0.069	0.169
Case 3: Downtown Atlanta	MAE	0.057	0.340
	MAPE	0.047	0.261

4.8 Chapter Summary

This chapter focused on developing a strategic receptor placement method that could minimize the number of receptors (and therefore receptor-link combinations used in modeling) without undermining the predictive capabilities compared to high-resolution concentration profiles generated by dense receptor placement. The fundamental relationship among dispersion model input parameters and predicted pollutant concentration profile was identified through a huge number of R-LINE sample runs. The results indicate that link emission rates, wind speed, and wind direction are important factors determining the PM_{2.5} concentration profile. Based on the 1,000 R-LINE simulation results with randomized input settings, this work demonstrated optimal receptor locations,

based on a step-wise searching method developed in this study, that minimizes the errors in the $PM_{2.5}$ concentration profiles predicted by the sub-set of receptor model as compared to those predicted by the high-density receptor model. The optimal receptor locations identified by the step-wise searching method suggests that with only a few numbers of receptors (around 13 receptors for each link segment), the optimal receptor model can approximate the $PM_{2.5}$ concentration profiles as predicted by the high-density receptor model.

The optimal receptor locations and its R-LINE input parameters were then used in developing a generalized receptor placement model (called dynamic grid-receptor model). The model consists of two sub-models: 1) the model for predicting the optimal number of receptors based on multiple linear regression, and 2) the model for predicting the optimal locations of receptors based on logistic regression. The dynamic grid-receptor model was verified with 300 R-LINE samples, by comparing the $PM_{2.5}$ concentration profiles predicted by the dynamic grid-receptor model with those predicted by the high-resolution-receptor model. The results suggest that the dynamic grid-receptor model can better approximate the $PM_{2.5}$ concentration profile than the static-grid-receptor modeling approach used in traditional modeling methods (where the distance between receptors is uniform, e.g., 200m resolution).

The dynamic grid-receptor model was then applied to three case study areas in metropolitan Atlanta to evaluate the model performance in the practical transportation network (e.g., ARC-ABM15 as in this study). The model results suggest that the dynamic grid-receptor model can also better approximate the $PM_{2.5}$ concentration profile predicted by the high-density receptor model for the practical transportation network than the static-

grid-receptor model. The greater prediction power of the dynamic grid-receptor model may be attributed to the strategic receptor placement that tends to set more receptors near road areas where higher pollutant concentrations are observed (which yields more accurate concentration fields). The better prediction of the high pollution areas helps to produce more unbiased pollutant concentration profiles.

For researchers and practitioners who implement air quality dispersion modeling of traffic-related pollution at a regional-scale, setting proper locations of receptors covering the entire project area has been a challenge. In this regard, this study could provide an idea about how many receptors are required and where those receptors need to be located. Thus, the dynamic grid-receptor model can help minimize the regional-scale dispersion modeling runtime, by removing unnecessary receptors in the analysis, while not undermining the pollutant concentration profile as predicted by high-density receptor model.

CHAPTER 5. DEVELOPMENT OF ROADWAY LINK SCREENING MODEL

This chapter is adopted from “Kim, D., Liu, H., Rodgers, M.O., and Guensler, R. (2020). Development of Roadway Link Screening Model for Regional-level Near-road Air Quality Analysis: A Case Study of Particulate Matter. Presentation at the 99th Transportation Research Board (TRB) Annual Meeting.”

5.1 Introduction

The impacts of traffic-generated emission sources on criteria pollution levels for project-level conformity assessments and the National Environmental Policy Act (NEPA) analyses are generally assessed through the application of computer models. The U.S. Environmental Protection Agency (USEPA) requires that the air quality impacts of mobile source emissions on the surrounding environment be assessed using its recommended mobile source emissions and dispersion modeling tools. For example, according to USEPA’s conformity guidance, MOVES (MOTOR Vehicle Emission Simulator) is designated as the official mobile source emission rate model for regulatory air quality analyses (USEPA, 2015a). When it comes to mobile source air dispersion modeling, several tools are recommended for transportation and air quality conformity assessment, including AERMOD and CAL3QHC/CAL3QHCR (USEPA, 2019a); CALINE4 can also be used in screening analysis, if permission is obtained from the USEPA Regional Office (40 CFR Appendix W). However, line source dispersion models employed in large-scale air quality impact assessments require significant computational resources. This is

especially true when it comes to the task of conducting air quality impact assessments for complex urban transportation projects consisting of numerous roadways, whose potential impacts on surrounding neighborhoods are affected by dynamic spatial and temporal traffic patterns and variable meteorological conditions. Therefore, enhancing the computational efficiency of microscale dispersion modeling applications makes large-scale assessments that employ higher-resolution receptor placement more feasible, and therefore also improves the reliability of assessment outcomes.

Computational efficiency has been a major concern in previous studies of near-road air quality impact assessment conducted over large geographic scales (Guensler, et al., 2000; Shafi, 2008; Guensler, et al., 2008; Kall, et al., 2008; Vallamsundar and Lin, 2012; D'Onofrio, et al., 2016; Wu, 2018; Zhai, et al., 2016; Zhai, et al., 2019; Liu, et al., 2017; Briant, et al., 2013). For example, Guensler, et al. (2000) and Vallamsundar and Lin (2012) proposed a streamlined modeling framework connecting emission rate models and dispersion models to help metropolitan planning organizations and practitioners implement conformity and NEPA processes within the same framework. Almost every one of the projects in the literature has called for future research focusing on an efficient modeling design to achieve high estimation precision while minimizing computational cost.

Some prior studies were able to use simplified sets of model assumptions, with fairly low-resolution receptor grids (e.g., greater than 200 m by 200 m resolution) to scale up air quality impact assessment of transportation projects to the metropolitan or other regional levels (D'Onofrio, et al., 2016; Wu, 2018; Zhai, et al., 2016; Zhai, et al., 2019). While this regular, but sparse, receptor placement strategy helps reduce computational complexity and model run time, sparse receptor grids can cause estimation bias. Predicted

concentrations at individual receptors must properly consider the spatial distribution of all nearby roadways, and concentration estimates are very sensitive to the distance from receptor to the nearest roadway links (Wu and Niemeier, 2016) and each link's mass flux (Shafi, et al., 2008). Hence, low-density grids can lead to biased results (over-prediction or under-prediction) depending upon the mass flux from the roadways and the distance separating individual link-receptor combinations. Liu, et al. (2017) and Kim, et al. (2019c) used the state-of-the-practice methodology recommended by USEPA (USEPA, 2015a; USEPA, 2018) to assess near-road pollutant concentration profiles by spatially adjusting the density and location of receptors in proximity to link emission sources in all directions, based upon the spatial layout of the roadway network. In these analyses, a distributed computing cluster supported the calculations for the larger number of receptors. However, if distributed computing technology is not available for an analysis, which is currently true for most public agencies and consulting firms, modeling efforts that use dense receptor grids encounter significant computational challenges.

Most existing dispersion modeling studies attempt to account for all major arterial and freeway links included within, or located near, the transportation project to compute pollutant concentration (Batterman, et al., 2014; Zhai, et al., 2016; Liu, et al., 2017; Kim, et al., 2019c), hence, extensive computational resources are generally required when high-resolution concentration fields are desired. The methodology usually become even more infeasible when applied to a large metropolitan area, because the number of roadway links that need to be considered for each receptor concentration computation increases dramatically with a wider coverage area. In a case study in Metro Atlanta, Kim, et al. (2019c) showed that AERMOD processing time increases with the number of links (Figure

28) and the number of receptors (Figure 29). In particular, Figure 28 shows that AERMOD processing time increases with the number of links considered, taking about 11.5 hours for 161,188 links and a single receptor. Figure 29 shows that AERMOD processing time increases with the number of receptors considered, taking about 2.5 hours for a single link and 10,000 receptors. Combining those conditions, the model run time for AERMOD to compute PM_{2.5} concentration considering the entire Atlanta roadway network (161,188 links) iterated across a high-resolution grid of receptors for the Atlanta Metro area (1,163,958 receptors) would take longer than one year when performed on a typical desktop computer.

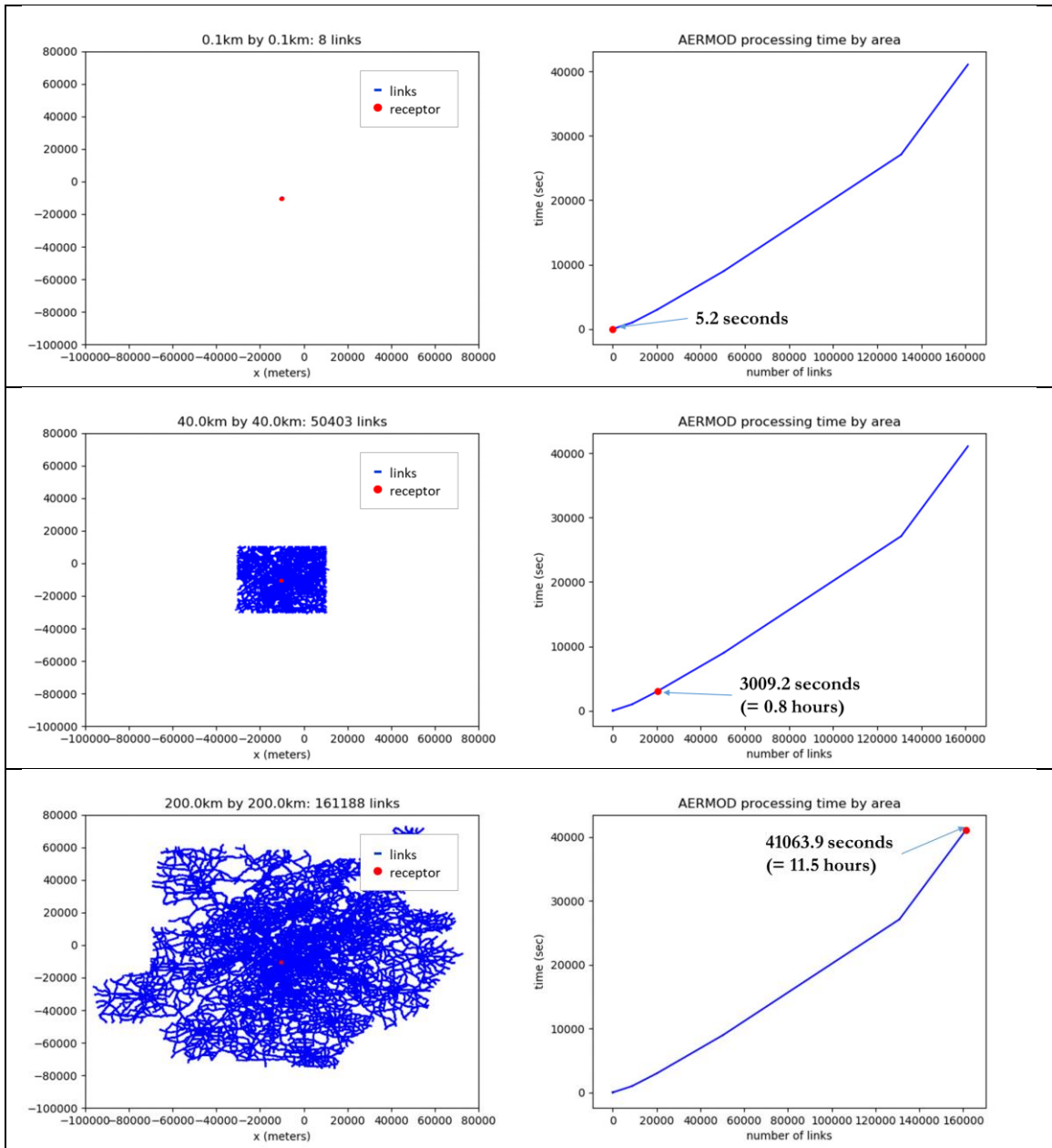


Figure 28 – Examples of AERMOD Processing Time by Number of Links

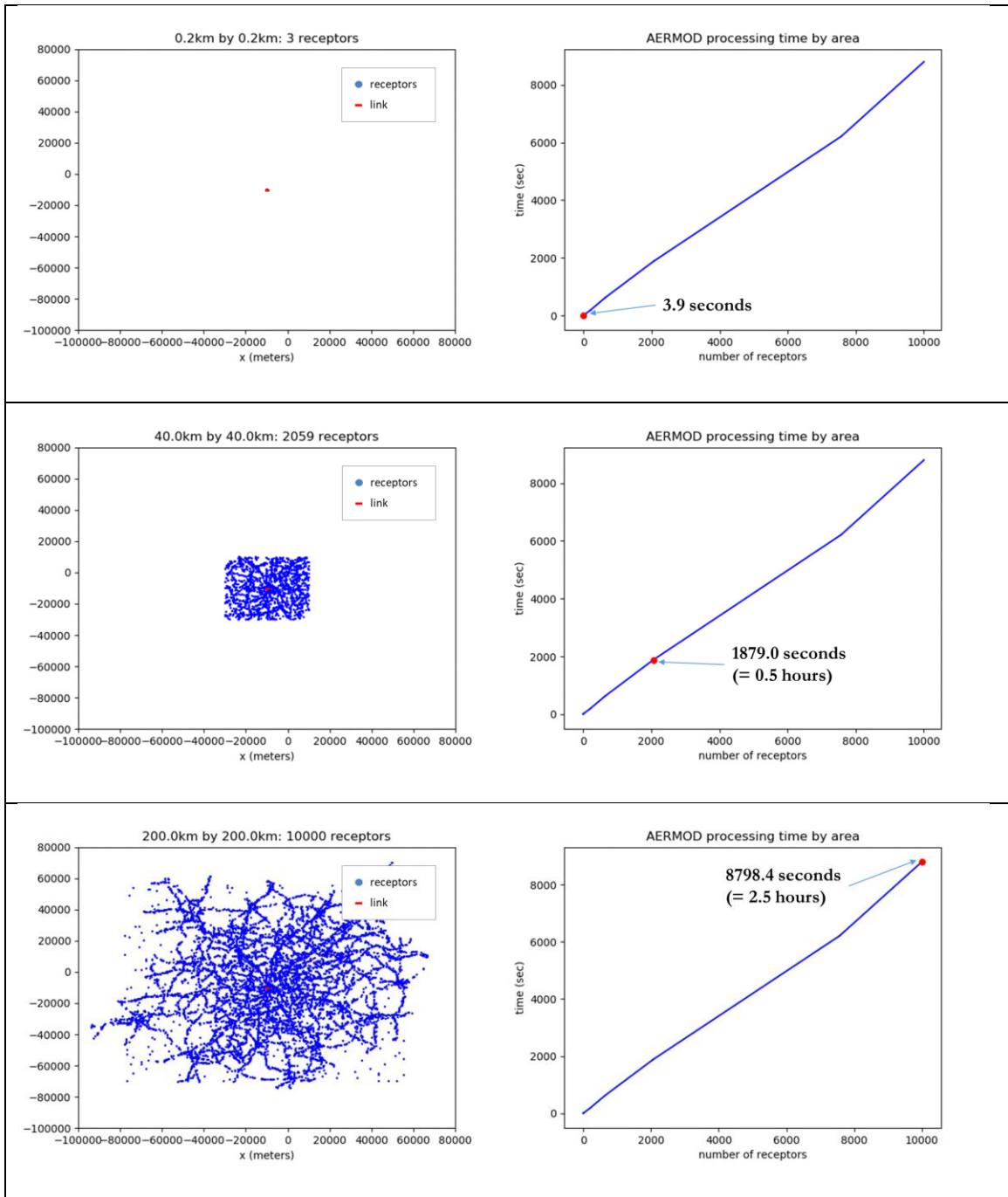


Figure 29 – Examples of AERMOD Processing Time by Number of Receptors

To address the significant computational challenge of large-scale applications of dispersion models, this study proposes an innovative approach, hereafter referred to as supervised link screening (SLS), to reduce the number of link-receptor pairs employed in

modeling efforts. The concept of link screening has been previously applied in CALINE4 modeling with MOBILE model emission rate (Shafi; 2008; Guensler, et al. 2008), using a decision tree-based method to identify and eliminate roadway links with zero contribution (i.e., the receptor is upwind of the link), and minimal contribution to receptor concentration, using roadway mass flux and distance separation between source and receptor (Guensler, et al., 2008). With the advent of enhanced computational and machine learning techniques, and the introduction of more advanced emission models (MOVES), and more advanced dispersion models (e.g., AERMOD) that require much more complicated meteorological inputs (Georgia EPD, 2019; Texas Commission on Environmental Quality, 2019), the development of new link screening methods is a timely endeavor. In this Chapter, the machine learning classification algorithms are applied to link screening. A random forest (RF) classifier is used to identify and remove roadway links with negligible concentration contributions for each receptor with high precision. This study then verifies a significant increase in the efficiency of link elimination using direct comparison tests. The new link screening method is applied to the 20-county metropolitan Atlanta area to demonstrate the promising performance of SLS for regional-scale applications of microscale dispersion modeling.

5.2 Model Development

This study develops a supervised binary link screening model that classifies roadway links as either significant or insignificant, based upon the predicted contribution of the link to the pollutant concentration at a particular receptor. Several supervised machine-learning models were developed and applied in this modeling effort, and the random forest (RF) classifier was selected for its superior predictive capability. The RF classifier was

developed using a set of classification variables that are calculated based upon the hourly mass emissions from the roadway link, distance to receptor, and meteorological parameters. The variables that influence predicted pollutant concentration at a specific location are identified and used in the classifier. Furthermore, this study develops different link-screening classifiers for six criteria air pollutants. As a demonstration, the classifier developed for fine particulate matter (PM_{2.5}) is described in detail and applied to Metro Atlanta. However, the modeling structure and preparation of input datasets for PM_{2.5} can be applied to all pollutants of interest.

5.2.1 Data

The data used to the roadway classifier include pollutant concentration estimates derived from the dispersion model (used as the screening variable), link emissions (grams/hour), roadway geometry, and meteorological parameters as predictive variables. First, the on-road vehicle emissions profile for each link is generated based on traffic operating conditions and fleet composition to generate fleet emission rates in grams/vehicle-hour, and then multiplied by hourly traffic volumes to generate the resulting hourly mass emissions (grams/hour) for each roadway link. The emissions are then used to calculate the PM_{2.5} mass flux for each roadway link (given the roadway length and width) as dispersion model inputs. Spatial data include roadway geometry, receptor locations, and surface terrain roughness. The road and receptor geometry are used to calculate the distance between a given receptor and roadway segment. Meteorological data include wind speed and direction, mixing height, temperature, etc. The input data are used to train and verify an RF link-screening classifier for the 20-county Metropolitan Atlanta Region. This approach allows the prediction results of the classifier to be assessed for a large

metropolitan area. Because a metropolitan region as large as Atlanta includes a diversity of urban, suburban, and exurban settings, with a variety of transportation infrastructure development, traffic operating conditions, and local meteorological conditions, sufficient environmental complexity is available in the input data and concentration outcomes. Hence, the application helps ensure that the link-screening model is transferrable to other regions.

5.2.1.1 Link Emissions

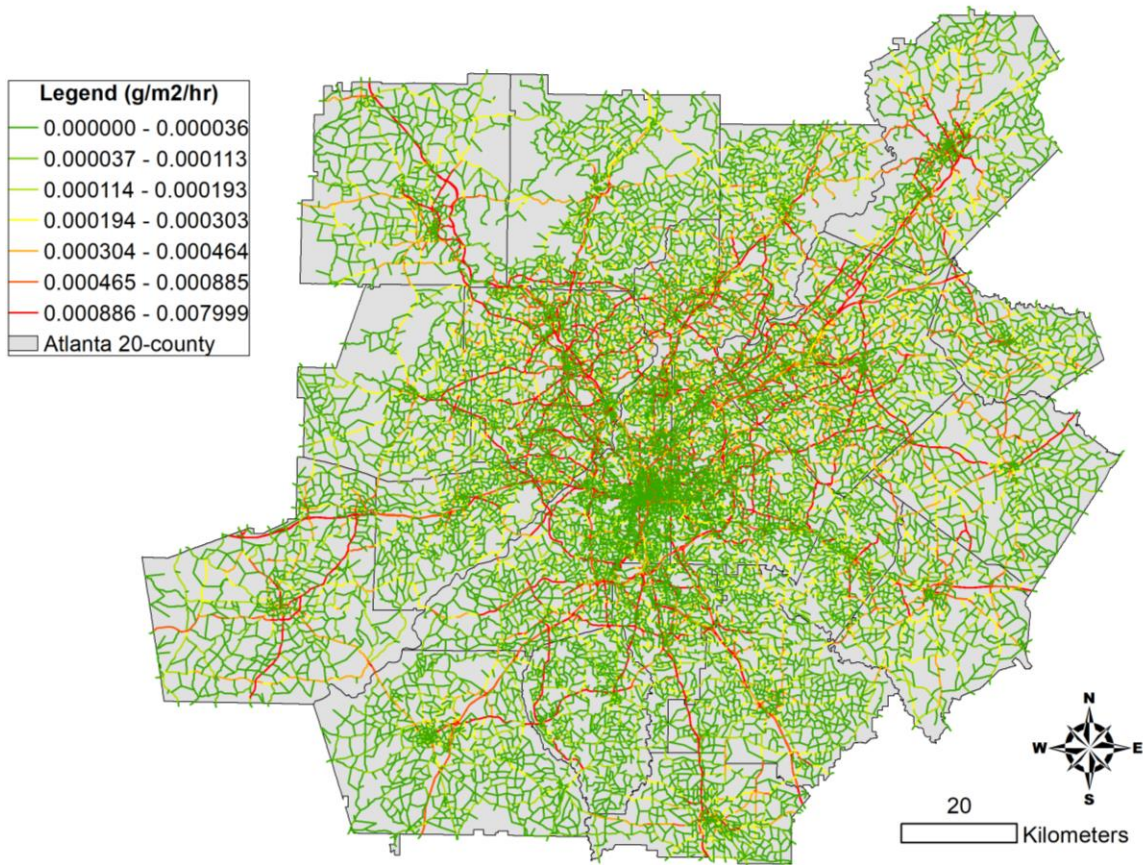
The on-road PM_{2.5} emission rates are calculated by integrating vehicle activity data from the Atlanta Regional Commission's activity-based travel demand model (ARC-ABM), used in regional air quality management planning (ARC, 2017), with applicable emission rates from MOVES-Matrix (Liu, et al., 2019b). The ARC-ABM outputs contain vehicle activity data in vehicle-miles traveled (VMT) and average speed for each roadway segment within Metro Atlanta. Based on this information, the PM_{2.5} emission rate for each roadway segment is queried from MOVES-Matrix, a high-performance emission rate lookup system (Liu, et al., 2019b). MOVES-Matrix is a multidimensional array containing emission rate outputs from numerous MOVES model runs enabled by the Partnership for an Advanced Computing Environment (PACE) high-performance computing (HPC) cluster (PACE, 2017). MOVES-Matrix generates exactly the same emission rates as running MOVES on a case-by-case basis, while using the same structure of input variables (e.g., source type, drive cycle, road grade) and algorithms used by MOVES. However, MOVES-Matrix calculates link-based emission rates more than 200 times faster than a traditional MOVES interface run (Guensler et al., 2016; Liu et al., 2019) and MOVES input files never need to be prepared because MOVES-matrix already contain the results of all model input

combinations. As such, using MOVES-Matrix in emissions calculations for large-scale networks (such as the Metro Atlanta network) helps increase the speed of generating input data used to calculate link emissions and mass flux. The link emissions calculations are performed using equation (13), in which each vehicle source type (vehicle class), by model year for that vehicle class, by operating mode condition for that vehicle class and model year, is multiplied by its applicable MOVES emission rate.

$$Activity_{fleet} \sum_{ST} \sum_{MY} \sum_{FTS} (ST\% \times MY\%_{ST} \times FTS\%_{ST,MY}) \times ER_{ST,MY,FTS} \quad (13)$$

Here: $Activity_{fleet}$ is on-road vehicle activity of a link (in vehicle-miles or vehicle-seconds), $ST\%$ is the source type distribution on a link, $MY\%_{ST}$ is the model year distribution within a source type on a link, and $FTS\%_{ST,MY}$ is the distribution of on-road operations (facility type by average speed) on the link within each source type and model year pair, and ER is the corresponding emission rate for each source type by model year by on-road operating condition combination (which can be queried from MOVES-Matrix). In particular, this study performed the link emissions calculations for 12 months and 24 hours within a day (weekday and weekend).

Figure 30 depicts the example of the estimated $PM_{2.5}$ mass flux (grams/m²/hour) by link for the Metro Atlanta road network (161,188 links) in the morning peak hour (8:00 AM to 8:59 AM) for January 2015.



Note: the figure illustrates the example emission rates estimated for 8:00 AM to 8:59 AM in January 2015.

Figure 30 – Estimated Hourly Average PM_{2.5} Mass Flux (g/m²/hour) for the Metropolitan Atlanta Area

5.2.1.2 Meteorological Conditions

The meteorology data used in this study were obtained from the Georgia Department of Natural Resources (DNR) Environmental Protection Division (Georgia EPD, 2019). Georgia EPD provides five years (2014-2018) of meteorological data for 15 sub-regions in the state of Georgia. The meteorological data are processed by using AERMET, the meteorological processor approved by USEPA, based on data from 15 Automated Surface

Observing Systems (ASOS) surface- and upper-air stations (Nadolski, 1998). The resulting meteorological database generated by AERMET contains the post-processed meteorological conditions for each sub-region and hour of the year. Therefore, the meteorology database contains a total of 657,000 meteorological conditions (15 sub-regions \times 5 years \times 365 days \times 24 hours). The database provides meteorological parameter values for each hour, and the list of the parameters are summarized in Table 7, along with the descriptive statistics for each parameter in the database. For most of the variables, the processed meteorological conditions in Georgia vary significantly across the hours of the year and sub-regions. For example, the surface temperature ranges from $-9.0\text{ }^{\circ}\text{C}$ (264.2 K) to $38.3\text{ }^{\circ}\text{C}$ (311.4 K), with a mean of $19.8\text{ }^{\circ}\text{C}$ (292.9 K) and a standard deviation of around $8.4\text{ }^{\circ}\text{C}$ (8.4 K), indicating that the data include extremely cold and hot weather conditions. Yet, a few variables may not represent a wide range of meteorological conditions. For example, the Bowen ratio ranges from 0.470 to 0.730, which implies a wet surface over most of the Atlanta region. This surface condition is likely reasonable due to the abundance of temperate forests and grasslands in the Atlanta metropolitan area (Nowak and Greenfiled, 2012; Park and Guldmann, 2020) as compared to other metro areas in the U.S. Hence, the link screening model developed based on the meteorological processor data for Georgia may need to be revised before being applied to regions with significantly different meteorological conditions.

Table 7 – Descriptive Statistics of Meteorological Variables

Variab les	Description	Min.	Max.	Mean	Standard Deviation
H	Sensible heat flux (W/m ²)	-64.0	228.1	34.0	63.5
temp	Reference temperature on ground surface (K)	264.2	311.4	292.9	8.4
u*	Surface friction velocity (m/s)	0.033	1.515	0.203	0.114
w*	Convective velocity scale (m/s)	0.015	2.285	1.169	0.497
Z _{ic}	Height of convectively-generated boundary layer, or mixing height (m)	2.0	2,664.0	843.5	545.8
Z _{im}	Height of mechanically-generated boundary layer (m)	14.0	3,999.0	245.8	204.0
VPTG	Vertical potential temperature gradient above the convective mixing height (K/m)	0.005	0.034	0.007	0.004
L	Monin-Obukhov length (m)	-8,888.0	8,888.0	-9.5	318.6
z ₀	Surface roughness length (m)	0.012	0.091	0.031	0.017
B ₀	Bowen ratio	0.470	0.730	0.599	0.130
R	Albedo	0.150	1.000	0.624	0.388
W _s	Reference wind speed (m/s)	0.0	22.7	2.8	1.6
W _d	Reference wind direction (degrees)	0.0	360.0	184.0	106.7
Z _{ref}	Reference height for wind (m)	10.1	10.1	10.1	0.0
Z _{temp}	Reference height for temperature (m)	2.0	2.0	2.0	0.0
ipcode	Precipitation type code (0=none, 11=liquid, 22=frozen)	-	-	-	-
p _{amt}	Precipitation rate (mm/hour)	0.0	70.6	0.1	1.2
rh	Relative humidity (percent)	7.0	100.0	70.9	20.7
pres	Station atmospheric pressure (mb)	971.0	1,034.0	1,011.9	5.1
ccvr	Cloud cover (tenths)	0.0	10.0	3.4	4.1

5.2.1.3 Sample Selection

A large number of random samples are needed to train and test the supervised link-screening (SLS) classifier. These random samples should include the dependent variable (i.e., classification of each link as significant versus insignificant pollutant contributor for each link-receptor pair) and all relevant classification variables expected to affect model predictions. A spatially random sample of roadway links is first selected from the Atlanta network. A random set of receptors is generated over x- and y-coordinate space using a random point generator within a geographic bound. A random pairing process was then used to generate 79,328 link-receptor pairs for analysis. About 70% of the sample sets (i.e., 55,530 sample sets) are used to train the classifier, and the remaining 30% is used for model verification.

5.2.1.4 Pollutant Concentrations

AERMOD predicts the PM_{2.5} receptor concentration for each link-receptor pair in the random sample. The screening option embedded in AERMOD is used for this dispersion modeling, so that AERMOD forces the model calculations to represent values for the plume centerline, regardless of the source-receptor-wind direction orientation (USEPA, 2016). In other words, the AERMOD screening option internally adjusts the source-receptor-wind direction orientation to derive the highest predicted air pollutant concentrations for the link-receptor pair, and thereby returns the highest concentration based on the worst hourly meteorological condition for the year. This worst-case hourly meteorological condition identified by AERMOD is used as the pollutant concentration value for developing the link screening classifier. Because the AERMOD screening option

is likely to overestimate the contribution of roadway links to any receptors' predicted pollutant concentrations, links screened as non-significant under the worst-case meteorological scenario should have even lesser impact under more favorable meteorological scenarios (i.e., scenarios with greater dispersion). The urban population parameter is required in AERMOD to select the population-dependent urban boundary layer height. During the AERMOD dispersion modeling process, the urban population parameter is randomly varied, ranging from 50,000 (small city) to 1,000,000 (large city), uniformly distributed in one-person unit increments. Finally, based on the link and receptor geometry data, the Euclidean distance between a specific link (geometric center) and a receptor is calculated. In summary, the output from the AERMOD-assisted dispersion modeling contains the worst-case hourly PM_{2.5} concentration for each sample set as a function of its corresponding link emission rates, meteorological conditions, and urban population.

5.2.1.5 Link Screening Thresholds

The binary dependent variable, or 'class' (i.e., identification of a link as significant versus insignificant for a given receptor), is generated from the resulting PM_{2.5} concentrations that are computed for each link-receptor pair. The link is defined as significant (and the dependent variable is set to 1) if the concentration contribution from the link at the receptor is greater than a specific pollutant concentration threshold (δ). The link is defined as insignificant (and the dependent variable is set to 0) if the concentration contribution from the link at the receptor is less than the pollutant concentration threshold (δ). In this study, the threshold value for link contribution to concentration is set to 0.1 $\mu\text{g}/\text{m}^3$, and then to 0.01 $\mu\text{g}/\text{m}^3$, both of which are relatively small contributions at a receptor, given the

predictive capability of microscale dispersion models. The former threshold value ($\delta = 0.1 \mu\text{g}/\text{m}^3$) was used in a previous study conducted by Shafi (2008), which employed CALINE4 (Benson, 1984) and set worst-case wind speed of one meter per second (concentration outputs are automatically rounded by CALINE4 to one decimal place). The previous screening (Shafi, 2008) removed all links that contributed to a reported concentration impact of $0.0 \mu\text{g}/\text{m}^3$. The initial threshold value for this study was also set to $0.1 \mu\text{g}/\text{m}^3$. A more stringent threshold value ($\delta = 0.01 \mu\text{g}/\text{m}^3$) was then considered in this study as a more conservative rule for identifying insignificant links, given that AERMOD reports predicted concentrations beyond the first decimal place.

5.2.2 Variable Selection

In developing a parsimonious model, variable selection should be designed to obtain good performance (i.e., explanatory power) without including excess variables and overfitting the model. The variable selection for link significance classification was conducted through two main steps: 1) finding candidate variables that potentially affect $\text{PM}_{2.5}$ concentrations, and 2) identifying a final set of variables through an iterative model calibration process.

In the first step, three different approaches are considered in selecting candidate variables: logistic regression (LR) coefficients, feature importance rankings provided by a random forest (RF) classifier, and significances of variables identified in the literature. The binary LR model is generally used to predict the probability of a certain class of binary case (Greene, 2012; Washington, et al., 2011); in this study, whether the predicted concentration contribution is greater than the pre-defined threshold δ . The statistical significance of coefficients of the explanatory variables of the LR model is used as criteria

for variable selection. The LR model was estimated using the routines contained in the R statistics software 'glm' library (R Core Team, 2017). The results in Table 8 show that link emission rate, distance between link and receptor, surface friction velocity, and mechanical mixing height parameters are statistically significant in predicting $PM_{2.5}$ concentrations at the 90% confidence level. The signs of most of the significant coefficients are intuitively correct. For example, as the link source produces more $PM_{2.5}$ emissions, the resulting $PM_{2.5}$ concentration impact increases. The $PM_{2.5}$ concentration at the receptor decreases as the link source's distance to the receptor increases. An increase in surface roughness length decreases $PM_{2.5}$ concentration. However, the signs of some of the variables appeared to be counter-intuitive. For example, an increase in wind speed increased $PM_{2.5}$ concentration. This trend may be attributed to the fact that some of the meteorological variables are highly correlated. In particular, the team found that the correlations among surface friction velocity, Monin-Obukhov length, mechanical mixing height, and wind speed were high (Pearson's correlation coefficients among these variables are either less than -0.5 or greater than 0.5), as these variables are related to each other in AERMOD formulation (USEPA, 2019b). Thus, these correlated variables need to be further investigated through a parsimonious variable selection process.

The feature importance rankings provided by the RF classification were also used to assess the usefulness of each variable in identifying link significance, calculated by the Gini index from classification results (Archer and Kimes, 2008). Scikit-Learn, a machine learning library in Python (Pedregosa, et al., 2011), is used to apply the RF classifier. The results in Table 8 (in the "Variable Importance (from RF Classifier)" column) show that the link emission rate and the distance between link and receptor are the most important

features in estimating the PM_{2.5} concentration rates. The urban population (which impacts the AERMOD boundary layer height), surface roughness length, and wind speed are also identified as important in the feature importance rankings. These results are in line with the logistic regression results.

An extensive literature review identified potential relationships between PM_{2.5} concentrations and explanatory variables. The results are summarized in the “Literature Review” column in Table 8. In particular, the column “Expected Coefficient Signs” in Table 8 summarizes the impacts of explanatory variables on PM_{2.5} concentrations described in the literature: “positive” indicates that an increase in the variable value is expected to increase PM_{2.5} concentration, and “negative” indicates that an increase in the variable value is expected to decrease the PM_{2.5} concentration. For example, an increase in hourly PM_{2.5} emission from the link source tends to increase PM_{2.5} concentration (Liu and Kim, 2019a; Zhang, et al., 2018). On the contrary, an increase in the distance between the link and receptor should decrease the predicted PM_{2.5} concentration (Liu, et al., 2017; Wu and Niemeier, 2016; Igri, et al., 2011). The urban population input variable is used in AERMOD to account for the urban nighttime heat island effect, which generally increases the urban nighttime boundary layer mixing height and thus decreases observed concentrations (USEPA, 2019d). Note that the AERMOD urban roughness option was not considered in this study because the option is not allowed for regulatory analysis (USEPA, 2019d).

The literature also revealed relationships between some of the meteorological variables and PM_{2.5} concentration. Igri, et al. (2011) investigated the variations in pollutant concentration patterns depending on the three surface parameters (surface roughness

length, albedo, and Bowen ratio). The research revealed that an increase in albedo or surface roughness length decreases pollutant concentration, while an increase in the Bowen ratio increases pollutant concentration. Some research emphasized the important role of sensible heat flux on pollutants concentration. Gamo, et al. (1994) concluded that an increase in sensible heat flux is related to an increase in convective mixing height, thereby lowering pollutants concentration, while USEPA (2019d) and Igri, et al. (2011) suggested that because the sensible heat flux is widely used in processing various surface parameters (e.g., albedo and Bowen ratio) as well as meteorological parameters (e.g., surface friction velocity, mixing height, and Monin-Obukhov length), the impact of sensible heat flux on $PM_{2.5}$ concentrations may vary. Many journal articles suggested that surface temperature and wind speed are among the most important factors affecting pollutant concentration levels (Zhang, et al., 2015; Tecer, et al., 2008; Akpinar, et al., 2008; Lin, et al., 2015; Zhang, et al., 2018), suggesting that increases in surface temperature and wind speed tend to reduce pollutant concentrations. In addition, research revealed that some variables (including vertical potential temperature gradient, relative humidity, station pressure, and cloud cover) have a positive impact on $PM_{2.5}$ concentration, while mixing height and precipitation rate have a negative impact (Hien, et al., 2002; Tecer, et al., 2008; Lin, et al., 2015; Zhang, et al., 2015; Akpinar, et al., 2008; Zhang, et al., 2018).

An extensive literature review was conducted to provide a comprehensive overview of the fundamental relationship between various dispersion parameters and pollutant concentrations. The effect and significance of the introduced variables are summarized in Table 8. This overview helps limit meaningful variables for link screening. At the same time, an appropriate variable selection for link screening also requires an understanding of

the technical formulation of AERMOD. In the AERMOD formulation, some of the variables are interconnected and derived from each other (Cimorelli et. al., 2005; USEPA, 2019d). For example, sensible heat flux depends on surface characteristics such as Bowen ratio (indicator for surface moisture), implying that the increase in Bowen ratio (i.e., dry condition) leads to the increase in sensible heat flux. When the variables are correlated, one of those variables with an overriding effect may be selected over the other variables.

In the second step of parameter screening, the candidate variables identified in the first step are evaluated in light of the literature review and the technical formulation of AERMOD to select a final set of explanatory variables. This process iteratively developed candidate LR models using multiple combinations of the candidate variables until a final model that only included those coefficients that were statistically significant with correct signs (corresponding to the column “Expected Coefficient Signs” in Table 8) were identified. The final LR model result is summarized in the column labeled “Logistic Regression for Selected Variables” in Table 8. This process helped resolve the potential multi-collinearity problem noted for some of the variables (e.g., wind speed, surface friction velocity, sensible heat flux, and Monin-Obukhov length), by selecting the most representative variable among them. For example, the final variable set only includes wind speed, excluding its correlated variables, and the coefficient sign was negative as expected. Based on this parsimonious variable selection processes, six variables were selected for use in the final link-classification model, as listed in the column “Selected in Final Model” in Table 8. The final variables include: 1) link emission rate, 2) distance between link and receptor, 3) urban population, 4) sensible heat flux, 5) surface roughness length, and 6) reference wind speed.

Table 8 – Results of Variable Selection Models

Variables	Step 1: Identifying Candidate Variables					Step 2: Selection of Final Variable Set			
	Logistic Regression for All Considered Variables		Variable Importance (from RF Classifier)	Literature Review		Selected as Candidate Variables	Logistic Regression for Selected Variables		Selected for Final Model
	Coefficients	Z-score		Expected Coefficient Signs	References		Coefficients	Z-score	
(Intercept)	33.280	0.668	-	-	-	-	1.538	10.869	-
Link emission rate (grams/hour) ¹	1.005	50.319	0.318	Positive	Liu and Kim (2019); Zhang, et al. (2018)	Yes	0.992	50.555	Yes
Distance between link and receptor (km)	-8.557	-46.826	0.396	Negative	Liu, et al. (2017); Wu and Niemeier (2016); Igri, et al. (2011)	Yes	-8.427	-47.103	Yes
Urban population (million) ²	-0.168	-1.428	0.034	Positive or negative	USEPA (2018b)	Yes	-0.151	-1.296	Yes

Sensible heat flux (W/m ²)	0.023	1.039	0.021	Positive or negative	Gamo, et al. (1994); Igri, et al. (2011); USEPA (2019b)	Yes	0.062	3.921	Yes
Surface friction velocity (m/s)	-13.780	-2.602	0.021	Positive	Donateo and Contini (2014)	Yes	-	-	No
Convective velocity scale (m/s)	-31.280	-0.698	0.001	-	-	No	-	-	No
Vertical potential temperature gradient above the convective mixing height (K/m)	35.510	0.657	0.000	Positive	Hien, et al. (2002)	No	-	-	No
Height of convectively generated mixing height (m)	-0.015	-0.365	0.000	Negative	Liu and Kim (2019)	No	-	-	No
Height of mechanically generated mixing height (m)	-0.008	-2.190	0.022	Negative	Liu and Kim (2019)	Yes	-	-	No
Monin-Obukhov length (m)	0.000	1.284	0.025	-	-	Yes	-	-	No
Surface roughness length (m)	-1.101	-1.580	0.024	Negative	Igri, et al. (2011)	Yes	-4.414	-8.549	Yes

Bowen ratio	-0.532	-2.969	0.017	Positive	Igri, et al. (2011)	Yes	-	-	No
Albedo	0.249	1.640	0.017	Negative	Igri, et al. (2011)	Yes	-	-	No
Reference wind speed (m/s)	1.149	5.367	0.022	Negative	Zhang, et al. (2015); Tecer, et al. (2008); Akpınar, et al. (2008); Zhang, et al. (2018)	Yes	-0.241	-2.022	Yes
Reference height for wind (m)	0.029	0.393	0.001	-	-	No	-	-	No
Reference temperature on ground surface (K)	0.005	0.974	0.024	Negative	Zhang, et al. (2015); Tecer, et al. (2008); Akpınar, et al. (2008); Zhang, et al. (2018); Lin, et al. (2015)	Yes	-	-	No
Precipitation type 11=liquid	-0.409	-2.019	0.001	-	-	Yes	-	-	No
Precipitation type 22=frozen	0.060	0.050	0.000	-	-	No	-	-	No
Precipitation amount (mm/hour)	0.059	0.816	0.001	Negative	Tecer, et al. (2008); Lin, et al. (2015)	Yes	-	-	No

Relative humidity (%)	0.000	0.010	0.021	Positive	Zhang, et al. (2015); Tecer, et al. (2008); Akpinar, et al. (2008)	Yes	-	-	No
Station pressure (mb)	-0.010	-2.999	0.022	Positive	Akpinar, et al. (2008); Zhang, et al. (2018)	Yes	-	-	No
Cloud cover (tenths)	-0.011	-1.313	0.008	Positive	Tecer, et al. (2008)	Yes	-	-	No

¹ Note that the variable for link emission rate is a mass emission rate from a link source (grams/hour) rather than mass flux rate (grams/meter²/hour). Although the mass flux rate may better explain predicted concentrations when wind is perpendicular to the road, this study adopted the mass emission rate because the mass emission rate may be better for explaining predicted concentrations in connection with road geometry and wind direction.

² Note that AERMOD processes a continuous value for the urban population to calculate urban-rural temperature difference and nocturnal urban boundary layer height (USEPA, 2019d).

5.2.3 Model Development

Several supervised machine-learning models are applied to identify significant and insignificant links as a function of the six variables selected for model development as outlined above. The applied models include logistic regression (LR), regression tree (RT), random forest (RF), support vector machine (SVM), adaptive boosting (AdaBoost), and neural network (NN). All of these models are widely used for classification in computer science and industry. LR is estimated by the R statistics software (R Core Team, 2017), and the rest of the models are estimated by Scikit-Learn in Python (Pedregosa, et al., 2011).

The theoretical background of LR was described earlier. Here, LR is used to classify the links with only the selected variables. The probability of being a significant link, P , is modeled by LR, and the predicted label is determined by P : significant link if $P \geq 0.5$, and insignificant link if otherwise.

RT is one of the most widely used classification techniques that generally take a binary classification outcome to construct a ‘decision tree’. The tree-structured classification system consists of a set of attributes that splits the data through a binary partition, thus generating two resultant regions (Quinlan, 1986). This approach is attractive because it provides a symbolic representation that helps human interpretation (Camdeviren, et al., 2007). It should be noted that the RT model in this study utilizes a different set of meteorological variables through a more parsimonious variable selection process, compared to the older model presented in Shafi (2008).

RF is a supervised ensemble classifier that is a collection of individual tree predictors. Each tree depends on the values of a random vector sampled independently and

with the same distribution for all trees in the forest (Breiman, 2001). Each tree in the ensemble is trained on a random subset of the training sample. This bootstrap aggregation (or “bagging”) approach makes the RF classifier less sensitive to the quality of training samples and to overfitting than other machine learning models (Breiman, 2001; Belgiu and Drăguț, 2016).

SVM is also one of the widely used classification algorithms that seek a separating hyperplane that is a maximal margin classifier with respect to training data (Yuan and Cheu, 2003; Hastie et al., 2009a). In SVM, to prevent the model from overfitting the data, and to make the model less error-prone in the training process, it is common to adjust two parameters: cost parameter C and a kernel parameter γ (Hsu, et al., 2003). In this study, the optimal C and γ are determined by a k-fold cross-validation procedure, and the procedure was conducted using ‘Optunity,’ a Python library supporting SVM modeling in the Scikit-Learn (Claesen, et al., 2014). The optimal C was estimated to be 83.8 and 96.7 for $\delta = 0.1 \mu\text{g}/\text{m}^3$ and $\delta = 0.01 \mu\text{g}/\text{m}^3$, respectively, and the optimal γ was estimated to be 0.181 and 0.419 for $\delta = 0.1 \mu\text{g}/\text{m}^3$ and $\delta = 0.01 \mu\text{g}/\text{m}^3$, respectively.

AdaBoost is an algorithm introduced to remedy binary class problems identified by Freund and Schapire (1997). The AdaBoost algorithm is an iterative procedure that tries to approximate a Bayes classifier, by combining many weak classifiers (Hastie, et al., 2009b; Shafique and Hato 2015). AdaBoost has been particularly successful when applied to binary classification problems (Hastie, et al., 2009b). While the AdaBoost algorithm can be implemented within various machine learning models, this study used a decision tree-based AdaBoost algorithm, known as the “best off-the-shelf classifier” (Breiman, 1996).

Neural Networks (NNs) employ a supervised learning algorithm based on the concept of extracting linear combinations of inputs as derived features, and then modeling the target as a nonlinear function of these features (Hastie, et al., 2009a). There are several implementations of NNs; the Multi-Layer Perceptron (MLP), Convolutional Neural Networks, and Recurrent Neural Networks are among the most commonly used NNs. This study used the MLP that is suitable for classification prediction problems where inputs are assigned to a class or label (Montavon, et al., 2012). The NN with MLP is modeled with an input layer, a hidden layer, and an output layer of neurons. The input data are scaled and transformed to a zero to one range using $\frac{x_i - \min(x)}{\max(x) - \min(x)}$, where x is equal to one when $\max(x) = \min(x)$. From the empirical analysis, the study used two hidden layers with twenty neurons for the first layer and five for the second layer.

5.2.4 Model Results

The performance and prediction accuracy of the classification models for link screening (i.e., properly identifying the links significant or insignificant) are compared. Twelve classification models (the six models for each threshold value, $\delta = 0.1 \mu\text{g}/\text{m}^3$ and $\delta = 0.01 \mu\text{g}/\text{m}^3$) are trained and tested for prediction accuracy using the 79,328 receptor-link pairs that were randomly selected from the Atlanta Metropolitan area. Table 9 summarizes the overall accuracy (OA) of each of the link screening models. The results demonstrate that the developed classification models are successful at identifying significant and insignificant links, with an overall accuracy of greater than 95% in all cases. Moreover, the results show that the tree-based models (RF and AdaBoost), outperformed all other classifiers by yielding an overall prediction accuracy of over 97% for both threshold values.

While the difference in OA between RF and AdaBoost is probably still acceptable (97.7% - 99.0% versus 97.4% - 98.9%), the RF showed slightly better performance in identifying significant links than AdaBoost (88.2% - 95.4% versus 80.3% - 91.9%). Considering that misclassifying significant links as insignificant may result in an underestimation of pollutant concentrations, the RF model seems to be a better alternative than the AdaBoost model. It is noteworthy that the OA of the RT model presented in Shafi (2008) was only about 88%, which is much lower than the models that are developed and presented in this dissertation. This is primarily because the previous model used an arbitrary set of worst-case meteorological variables for dispersion modeling, limited wind speeds to one meter per second, and used CALINE 4 model outputs that were necessarily constrained to one decimal place by the model outputs. Hence, the older model presented in Shafi (2008) may have excluded many potentially significant links that the newer models include. On the other hand, as a screening tool used to identify whether a violation of an ambient air quality standard is likely to occur, the older screening tool may perform acceptably (with much lower computational cost).

Table 9 – Overall Prediction Accuracy of Link Screening Models

Classification Model		Actual and Predicted Classes				Prediction Accuracy		
		S _a ; S _p (A)	S _a ; IS _p (B)	IS _a ; S _p (C)	IS _a ; IS _p (D)	Significant Links A/(A+B)	Insignificant Links D/(C+D)	Overall (A+D)/(A+B+C+D)
δ = 0.1 μg/m³	Logistic Regression	619	280	74	22,825	68.9%	99.7%	98.5%
	Regression Tree	618	284	81	22,815	68.5%	99.6%	98.5%
	Random Forest	745	100	133	22,820	88.2%	99.4%	99.0%
	Support Vector Machine	574	291	52	22,881	66.4%	99.8%	98.6%
	Adaptive Boosting	682	167	101	22,848	80.3%	99.6%	98.9%
	Neural Network	730	136	106	22,826	84.3%	99.5%	99.0%
δ = 0.01 μg/m³	Logistic Regression	4,071	596	262	18,869	87.2%	98.6%	96.4%
	Regression Tree	4,108	497	551	18,642	89.2%	97.1%	95.6%

Random Forest	4,573	222	332	18,671	95.4%	98.3%	97.7%
Support Vector Machine	3,601	1049	129	19,019	77.4%	99.3%	95.1%
Adaptive Boosting	4,272	375	246	18,905	91.9%	98.7%	97.4%
Neural Network	4,053	711	265	18,769	85.1%	98.6%	95.9%

Here:

S_a refers to an actual significant link; S_p refers to a predicted significant link;

IS_a refers to an actual insignificant link; IS_p refers to a predicted insignificant link.

A: number of significant links with a prediction as significant (good prediction).

B: number of significant links with a prediction as insignificant (bad prediction).

C: number of insignificant links with a prediction as significant (bad prediction).

D: number of insignificant links with a prediction as insignificant (good prediction).

Figure 31 illustrates the example of roadway links that are identified by the RF model as significantly contributing to the $PM_{2.5}$ concentration at one receptor site in downtown Atlanta. Without link screening, the whole set of roadway links (about 161,188 links), as shown in Figure 31(a), needs to be run to calculate the pollutant concentrations at a single receptor. However, after applying the link-screening model, only those links that have an impact greater than the threshold on the receptor are selected and considered in the concentration calculations, as shown in Figure 31(b). Refining the threshold value from $\delta = 0.1 \mu\text{g}/\text{m}^3$ to $\delta = 0.01 \mu\text{g}/\text{m}^3$ increases the number of significant link sources for each receptor, and these links spatially surround the receptor without a clear pattern, as the link's contribution is a function of many variables (Figure 31(c)). In the case of Metro Atlanta's road network, on average, the RF model with a threshold of $\delta = 0.1 \mu\text{g}/\text{m}^3$ selected 23 links (0.014% of all links), and with a threshold of $\delta = 0.01 \mu\text{g}/\text{m}^3$, RF selected 175 links (0.108% of all links).

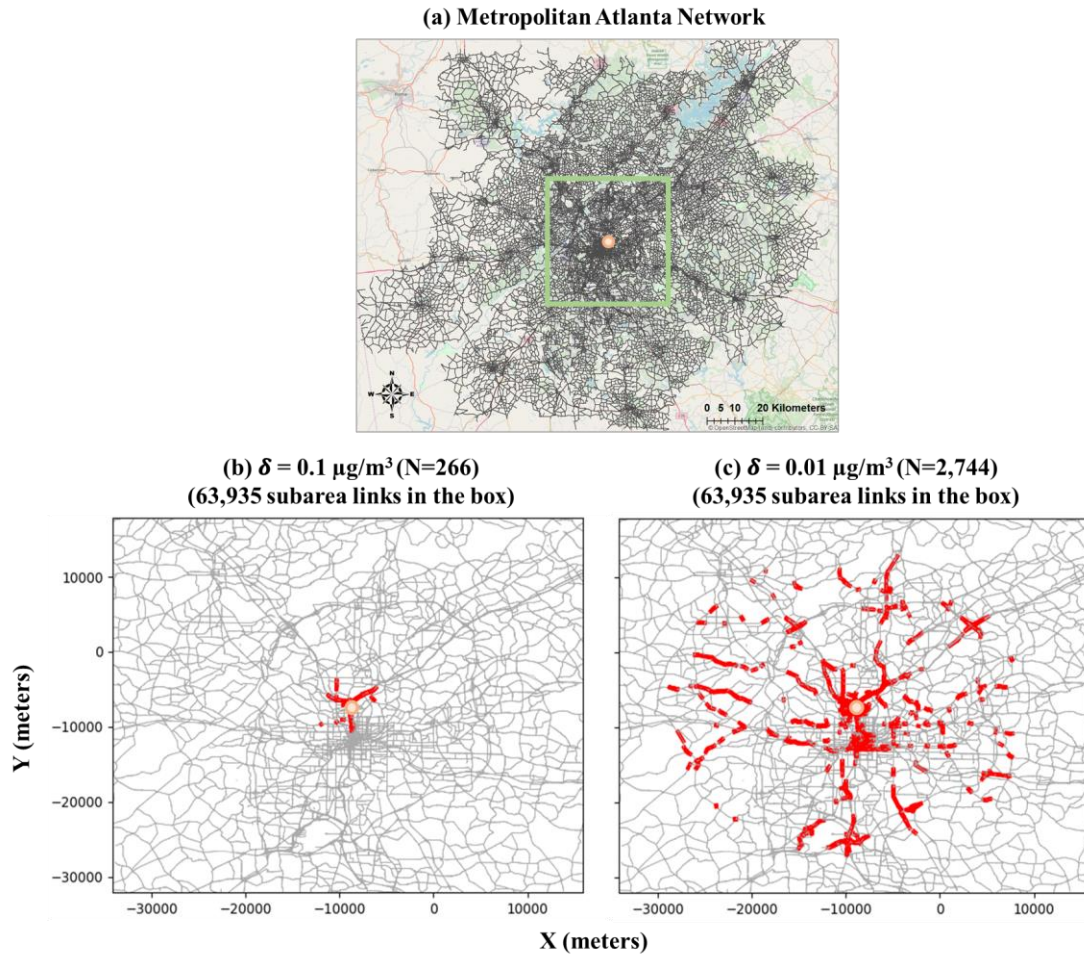


Figure 31 – Location of the Significant Links for a Receptor as Identified by Random Forest

5.3 Case Study

The case study focuses on the estimation of $\text{PM}_{2.5}$ concentrations in the Metro Atlanta area. The concentration estimation results that are obtained with and without the link-screening model integrated with AERMOD are compared with regard to total computation time and similarity in concentration outcomes. As expected, conducting mobile source dispersion modeling for the entire Metro Atlanta area roadway network is computationally formidable. Therefore, this study selected two subareas in Metro Atlanta (Figure 33(a)):

downtown Atlanta and northwest Atlanta. Downtown Atlanta has a dense highway and street network (35,234 links in an area of 878 km²) and heavy traffic volumes. Whereas, northwest Atlanta has less-congested traffic on fewer links (20,893 links in an area of 1,193 km²). The air quality dispersion model is first applied in the two case study areas using all existing links in the subarea (hereafter referred to as a ‘whole-link’ model), and then with the subset of links selected by the link screening model as being significant (‘reduced-link’ model). A set of 2,000 receptors is randomly placed in x- and y-coordinate space for Downtown Atlanta (2,000 receptors and 35,2234 links) and Northwest Atlanta (2,000 receptors and 20,893 links), respectively. As a result, six different model specifications are created:

- Downtown Atlanta, whole-link model
- Downtown Atlanta, reduced-link model, $\delta = 0.1 \mu\text{g}/\text{m}^3$
- Downtown Atlanta, reduced-link model, $\delta = 0.01 \mu\text{g}/\text{m}^3$
- Northwest Atlanta, whole-link model
- Northwest Atlanta, reduced-link model, $\delta = 0.1 \mu\text{g}/\text{m}^3$
- Northwest Atlanta, reduced-link model, $\delta = 0.01 \mu\text{g}/\text{m}^3$

The PM_{2.5} link emissions for the links in each subarea are first extracted from the Metro Atlanta network (see the earlier section “Link Emissions”). The links are coded for use with AERMOD and link emissions are converted to mass flux rates (grams/m²/s). The reduced-link model selects the significant links associated with each of the 2,000 receptors, and stores the link-receptor pairs in the database for modeling. To speed up AERMOD dispersion modeling, input files are separately generated for each link-receptor

combination so that the AERMOD modeling can be implemented with Georgia Tech's PACE supercomputing cluster (submitting multiple jobs simultaneously). The AERMOD input files prepared for the whole-link model contain the information about the concentration impact of all of the links (e.g., emission rates, geometry) on each of 2,000 receptors. In contrast, the AERMOD input files created for the reduced-link model only contain the information of the selected links for each receptor.

Figure 32 shows the correlation of the predicted annual average $PM_{2.5}$ concentration results for the reduced-link and whole-link models, where δ denotes a threshold value used in the reduced link models and MSE denotes mean squared error. The results show that their correlations are extremely high, with R^2 values ranging from 95.8% to 97.3%, with the fitted slope ranging from 0.9472 to 0.9999, indicating that the reduced-link models can approximate the $PM_{2.5}$ concentrations predicted by the whole-link models. The mean squared errors (MSEs) between the $PM_{2.5}$ concentrations predicted by the reduced-link and whole-link models only ranged from $0.084 \mu\text{g}/\text{m}^3$ to $0.272 \mu\text{g}/\text{m}^3$, on a case-by-case basis. This finding suggests that the $PM_{2.5}$ emissions from some significant link sources predominantly determine the modeled $PM_{2.5}$ concentrations at the receptor.

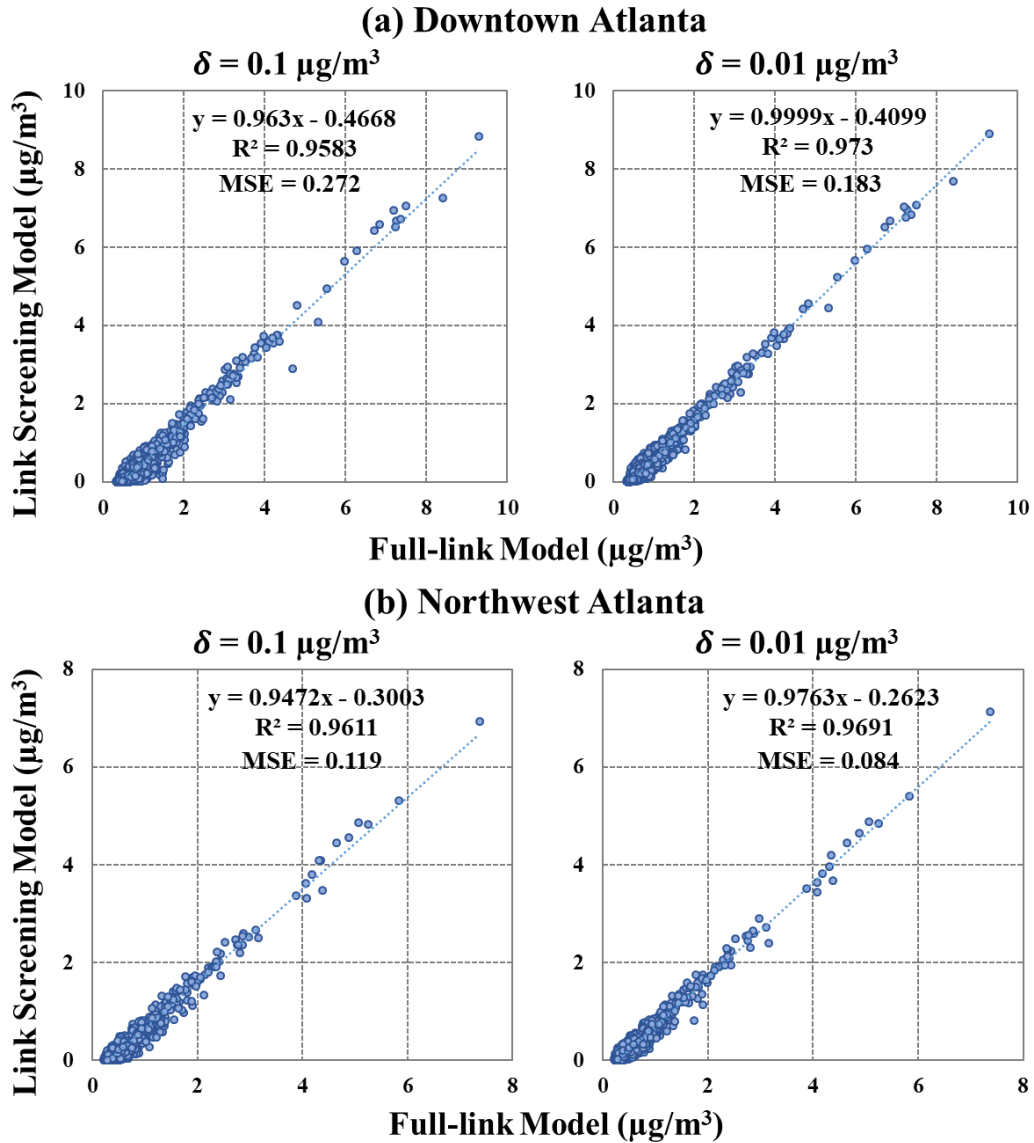


Figure 32 – Comparison of Annual Average PM_{2.5} Concentrations by Different Model Specifications

The results also show that the prediction power of the reduced-link models is high and results are comparable to those of the whole-link models, regardless of the parameters considered. As expected, the analyses also indicate that a further increase in estimation

precision can be achieved by using a lower threshold value in identifying insignificant links, which comes with an accompanying cost of increased modeling time. When more links are selected as being significant and included in AERMOD modeling due to the lower threshold ($\delta = 0.01$), the predicted $PM_{2.5}$ concentrations become more closely align with the predictions of whole-link models. Furthermore, Figure 33 shows that the spatial distributions of the estimated annual average $PM_{2.5}$ concentration profiles across the two study areas are very similar across the reduced-link and whole-link models. Figure 33(b) and (c) display contour plots representing the $PM_{2.5}$ concentrations predicted for the 2,000 receptor sites. The proposed reduced-link model is especially effective in predicting high pollutant areas in that both model settings identified the same $PM_{2.5}$ hot spots.

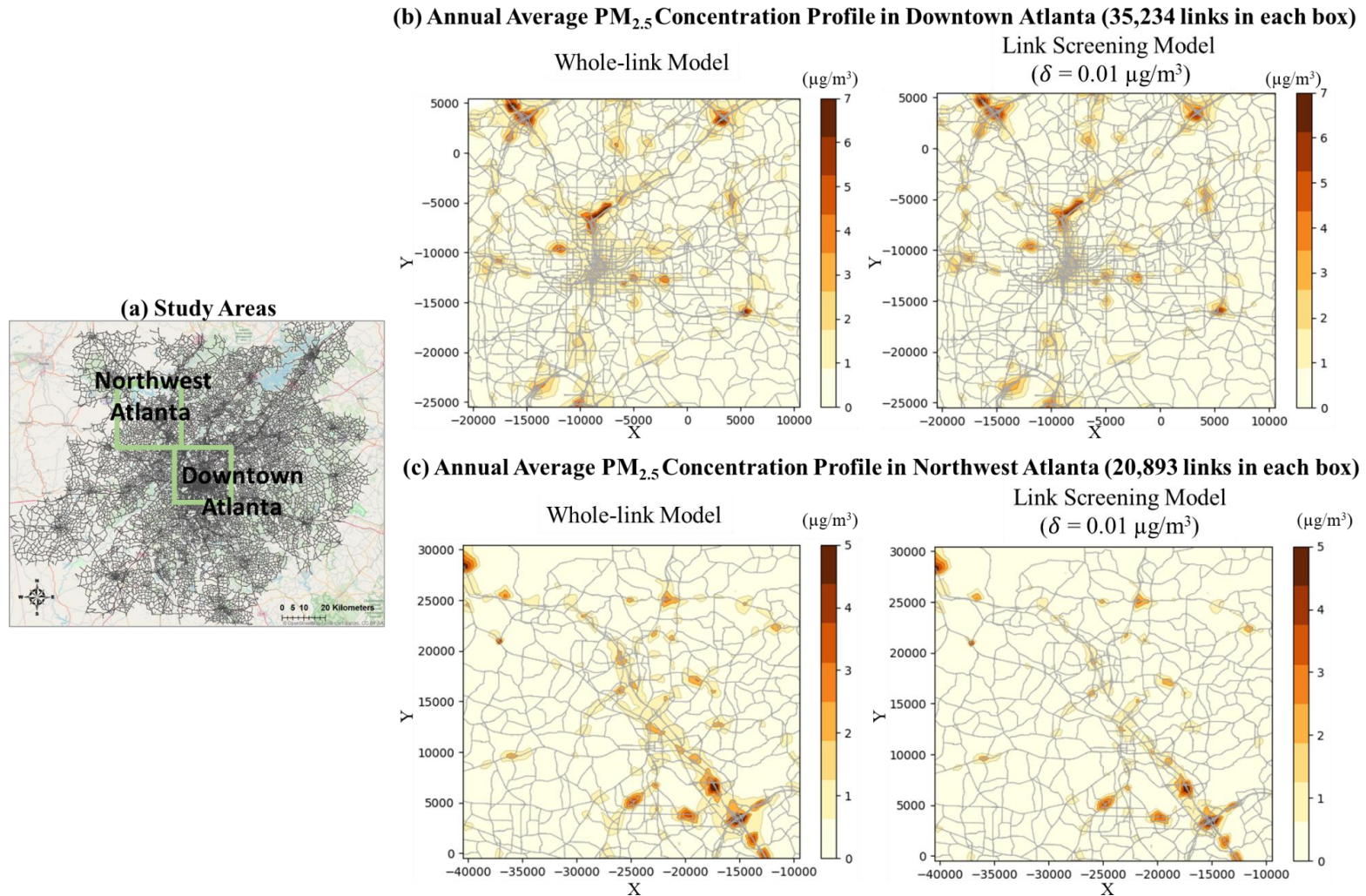


Figure 33 – Predicted Annual Average PM_{2.5} Concentration Profile of Study Area

Table 10 summarizes the AERMOD simulation run-time for each model scenario. The results show that the AERMOD run-time of the reduced-link models is dramatically reduced compared to that of the whole-link model. The speed increase results from the significant reduction in link-receptor pairs that need to be processed in AERMOD. Note that the time for identifying significant links based on the RF model is much smaller than the AERMOD execution time. For example, in the case of the reduced-link model with $\delta = 0.1 \mu\text{g}/\text{m}^3$ for downtown Atlanta, the link screening takes an average 57 milliseconds, while the AERMOD execution takes about 32 seconds (57 milliseconds:32 seconds = 1:561). The significant reduction in dispersion modeling run-time is therefore due significantly to the decrease in AERMOD simulation run-time. Specifically, the AERMOD run-time of the reduced-link model with $\delta = 0.1 \mu\text{g}/\text{m}^3$ for downtown Atlanta is 511 times faster than the whole-link model (16,354 seconds vs. 32 seconds). The reduced-link model processed 1,258 times fewer links (35,234 links vs. 28 links) during the dispersion modeling than the whole-link model.

The results also show that the AERMOD run-time reductions from the reduced-link models for downtown Atlanta are much greater than those for northwest Atlanta. The AERMOD run-times for the reduced-link models for downtown Atlanta were 113 to 511 times faster than those for the whole-link models, while the AERMOD run-times for the models for northwest Atlanta (where roadway density is lower) were only 92 to 372 times faster. This finding suggests that the reduced-link model may decrease AERMOD run-time more for areas with dense road networks, than for suburban road networks. For example, the reduced-link model with $\delta = 0.01 \mu\text{g}/\text{m}^3$ selected 0.605% of the links existing in

downtown Atlanta (213 links among 35,234 links), while the same model for northwest Atlanta selected 0.637% of the available links (133 links among 20,893 links).

In addition, link screening reduces data storage requirements by decreasing AERMOD input file sizes. As presented in Table 10, the AERMOD input file size for the whole-link model for downtown Atlanta is 26,273 kilobytes for one receptor case, leading to 52 gigabytes for 2,000 receptors. In contrast, the average input file size for the reduced-link model with $\delta = 0.1 \mu\text{g}/\text{m}^3$ for the same area is only 26 kilobytes, implying that the data storage requirement is reduced by a factor of 1,000. The significant decreases in input file size leads to an additional reduction in dispersion modeling execution time. A dispersion modeling covering a large geographic area often necessitates the use of distributed computing, as is the case in this research, and smaller file sizes help reduce the data transferring time significantly between local computer and distributed computing clusters.

Table 10 – AERMOD Simulation Results

Area	Models	Average Number of Links	Average Link Screening Time (milliseconds)	Average AERMOD input file size (kilobytes)	Average AERMOD Run-time (seconds)	Run-time Ratio (whole-link/RF)
Downtown Atlanta	Whole-link	35,234	10	26,273	16,354	-
	RF ($\delta = 0.1 \mu\text{g}/\text{m}^3$)	28	57	26	32	511
	RF ($\delta = 0.01 \mu\text{g}/\text{m}^3$)	213	61	163	145	113
Downtown Atlanta	Whole-link	20,893	9	15,586	11,526	-
	RF ($\delta = 0.1 \mu\text{g}/\text{m}^3$)	17	49	18	31	372
	RF ($\delta = 0.01 \mu\text{g}/\text{m}^3$)	133	54	105	125	92

Note: δ denotes a threshold of the reduced-link model. The average link screening time includes the time for identifying significant links through the developed RF models and generating the resulting AERMOD input files.

5.4 Summary of Findings

Enhancing the computational efficiency of regional-scale near-road air quality analysis is critical, if these models are to become feasible for assessing the air quality impact assessment of large and complex transportation projects. Not all roadway links contribute significantly to the pollutant concentration level at every receptor. The number and location of links that significantly influence receptor concentrations vary in space and time as a function of roadway geometry, mobile source emissions, and meteorological conditions. Prior regional modeling applications of microscale dispersion models have typically included all link-receptor combinations in dispersion model iterations; however, this brute-force approach results in high computational costs. As an innovative way to significantly improve modeling efficiency, this study developed a supervised link screening (SLS) approach to eliminate roadway links that do not contribute significantly to predicted receptor concentration from these analyses. The SLS presented in this paper and applied to an Atlanta regional case study was derived from a supervised machine learning Random Forest (RF) classification model. The final variable set includes six variables used for the link-screening model, including link emission rate, distance between link and receptor, urban population, and some meteorological variables. The classification results showed that the developed RF model yielded a high classification accuracy (greater than 95%). Using refined AERMOD model variables to develop an RF screening model provided much higher classification accuracy in identifying significant link-receptor pairs than did a previous link screening method that employed regression tree analysis with CALINE 4.

To demonstrate the methodology and its computational performance in detail, the RF classifier is applied to AERMOD case studies for PM_{2.5} concentrations in downtown

Atlanta (urban) and northwest Atlanta (suburban). The precision and efficiency of the dispersion modeling integrating the developed RF classifier ('reduced-link' model) are compared to the dispersion modeling conducted without a link-screening process ('whole-link' model). The results showed that the AERMOD run-time of the reduced-link models is greatly decreased (a 98.9% - 99.8% reduction in model run time) compared to that of the whole-link model, primarily because the number of links processed during the AERMOD simulation decreased significantly.

In terms of estimation precision, the results also showed that the $PM_{2.5}$ concentration estimates obtained with the reduced-link models are close to those obtained with the whole-link models, with correlations ranging from 95% to 97%. The reduced-link model is especially efficient for identifying areas with potentially high pollutant concentrations that may require more detailed hot-spot analysis. A lower contribution cut-off threshold value for labeling significant versus insignificant links ($\delta = 0.1 \mu\text{g}/\text{m}^3$ vs. $\delta = 0.01 \mu\text{g}/\text{m}^3$) turned out to be very helpful in increasing the accuracy of predicted pollutant concentrations. Future research may need to assess the trade-off between estimation precision and computational cost with respect to the selection of the threshold parameter.

One of the caveats in this research is that the link screening model was developed using State of Georgia meteorological data, and this screening model may need to be re-derived for use in other regions (as meteorological factors do affect the link screening rules). A careful selection of meteorological data that provide reasonable ranges for individual variables for each study area is also recommended. The basic modeling framework and approaches of the link screening model presented in this paper remain applicable to other traffic-induced primary pollutants, such as NO_2 , with appropriate model parameter

development. While the link screening model presented in this study proved to be effective in selecting a compact set of significant links and thereby reducing computation time, the series of modeling procedures required to implement the model may be difficult for many practitioners to reproduce. Future research may consider developing simpler link screening rules/algorithms or data requirements (e.g., traffic volumes and speeds, weather observation data, etc.). For example, land-use regression (LUR) modeling, a statistical modeling approach might be used to estimate air pollution levels or ambient temperature measured at a set of geographic points, might be used as a substitute for screening models, particularly when available input data are limited.

The link screening methodology presented in this paper has significant potential to help researchers and practitioners implement air quality dispersion modeling for environmental assessment of regional-scale transportation projects at a much lower computational cost. The methodology is also expected to make it much more feasible to compare the air quality impacts across complex project scenarios and for transportation development alternatives over large geographic areas. All these aspects may be of interest to a broad readership including metropolitan planning organizations and practitioners engaged in near-road air quality modeling for transportation and air quality conformity and for environmental analysis under the National Environmental Policy Act.

CHAPTER 6. STREAMLINED DATA PROCESSING FOR REGIONAL-LEVEL MICROSCALE DISPERSION MODELING

This chapter is adopted from “Kim, D., Liu, H., Xu, X., Lu, H., Wayson, R., Rodgers, M.O., and Guensler, R. (2020). Streamlined Data Processing for Regional Scale Applications of Line Source Dispersion Modeling via Distributed Computing. Presentation at the 99th Transportation Research Board (TRB) Annual Meeting.”

6.1 Introduction

Motivated by the challenges encountered in the previous efforts, this study develops an advanced modeling framework for a regional-level microscale dispersion modeling that integrates a high-performance emission rate lookup system known as MOVES-Matrix, employs innovative link screening and receptor locating methods that further accelerating model implementation, and provides flexibility for system use with multiple dispersion models. The research team has developed linkages for all of USEPA’s recommended line source dispersion models, including AERMOD Version 19191 (USEPA, 2019b), CALINE4 (Benson, 1984), and R-LINE (Snyder and Heist, 2013b). This paper demonstrates the application with AERMOD and R-LINE.

MOVES-Matrix operates with a multidimensional array containing 90 billion energy use and emission rate outputs for each modeling region (where a region is defined by a specific fuel program and inspection and maintenance program) generated from more than 146,000 MOVES model runs per region. MOVES-Matrix systematically iterates across specified ranges for all input variables, such that every combination of input variables that

affect energy use and emission rates is modeled. The comprehensive preprocessing of MOVES results allows users to query matrices to obtain applicable emission rates for any analysis, rather than having to run MOVES again. MOVES-Matrix queries produce exactly the same emission rates as running MOVES on a case-by-case basis, but these queries run more than 200 times faster than running MOVES to generate emission rates and project-specific MOVES input files do not need to be prepared (Liu, et al., 2019b; Guensler, et al., 2016). Using MOVES-Matrix boosts the computational speed of the model processes in conducting regional-level mobile source emissions calculations. The modeling system also integrates efficient receptor locating algorithms and link screening (to remove non-significant source-receptor combinations from the modeling work) to provide additional savings in computational time associated with line source pollutant dispersion modeling. Finally, the system features a streamlined processor that automates individual modeling steps ranging from data preparation, travel demand model (TDM) connectivity, emissions calculations, and dispersion modeling, using a user-friendly interface with minimal data input requirements. As an application, this paper presents the model's performance when applied to the 20-county Metropolitan Atlanta Region (hereafter referred to as Metro Atlanta), using MOVES-Matrix emission rates coupled with outputs from the Atlanta Regional Commission's (ARC) activity-based travel demand model (ARC, 2017).

6.2 Modeling Framework

As illustrated in Figure 34, the streamlined structure of the system for regional-scale line source dispersion analysis consists of four main sections: scenario set-up, input data preparation, emissions calculations, and dispersion modeling. This modeling framework utilizes various data sources including the regional travel demand model (TDM) database,

meteorology, terrain, road geometry, local fleet operation data, and background concentrations, as shown in Figure 34.

First, the regional TDM database is processed to produce link-level vehicle activity data (i.e., speed, volume, fleet composition) in the proper formats to be used in link-level emissions calculations. Second, the link-level vehicle activity data are linked to MOVES-Matrix to obtain the applicable emission rates for each link. Finally, the estimated emission rates are converted to the dispersion model inputs (in the model's applicable formats), and additional input parameters (e.g., meteorology, terrain, receptor locations) are also prepared. Because fine-resolution dispersion modeling for a large geographic area still requires substantial computational resources, the program is designed to use a distributed computing cluster (cloud computing resources).

The modeling framework supports two types of dispersion modeling: screening dispersion modeling and standard dispersion modeling. Screening dispersion modeling can be used for predicting the worst-case pollutant concentrations. To this end, screening dispersion modeling assumes the worst-case link emission rate and meteorological scenarios. In contrast, standard dispersion modeling allows seasonal and hourly variations in link emission rates and meteorological conditions to predict the annual average concentrations conforming to NAAQS.

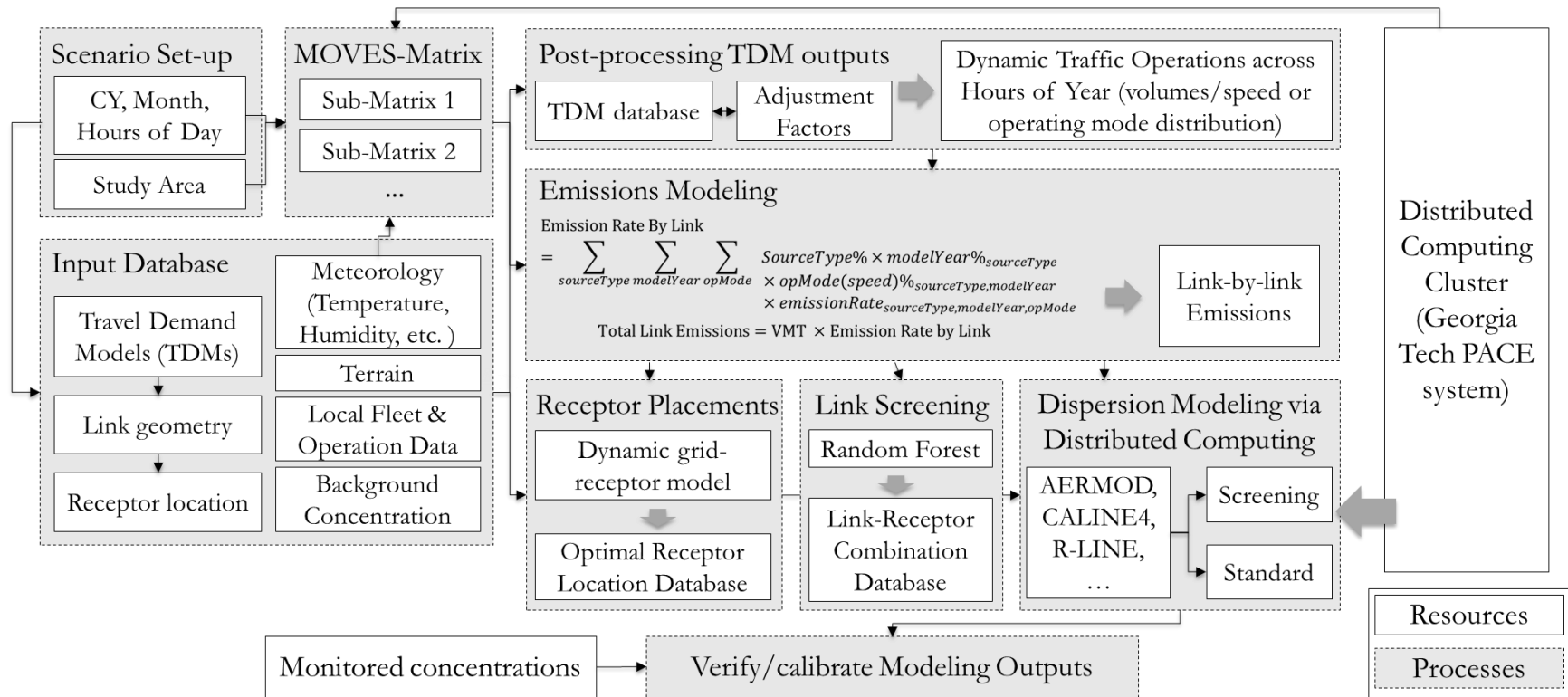


Figure 34 – Proposed Modeling Framework for Regional-level Microscale Dispersion Analysis

6.3 Data Preparation

Regional-scale applications of microscale pollutant dispersion modeling require a variety of inputs for comprehensive analysis. The inputs include: regional travel demand model (TDM) outputs, link and receptor geometry, meteorology and terrain profile, local fleet and operation data, background pollutant concentrations, etc. Because many of these input data are related, and linking one dataset to another is often complex, the streamlined linkages among these diverse input datasets are of importance for efficient modeling are described in this section.

6.3.1 Scenario Set-up

Dispersion model input data sets can be prepared in advance for each scenario analysis. Each scenario considers the temporal and spatial scales of the project. For example, the temporal scale of analysis may consider the present and future calendar years in which the on-road fleet and traffic conditions, and therefore vehicle emissions, are affected by planned transportation projects. The pre-set temporal scenario is then used to prepare the required input datasets for subsequent analyses. For example, transportation modelers will query TDM model outputs for a scenario, obtain traffic volumes and operating conditions, specify on-road fleet compositions, prepare meteorology data, querying sub-sets of corresponding MOVES-Matrix emission rates, etc. In addition, the spatial scale for project analyses needs to be defined to consider the geographical areas where traffic conditions are expected to change over time due to implementing transportation projects. Similarly, the pre-set spatial scale is used to prepare various input dataset for the project area.

6.3.2 *MOVES-Matrix*

MOVES-Matrix is a multidimensional array containing emission rate outputs obtained from the iterative MOVES model runs for all combinations of MOVES input variables that affect emission rate predictions. This process requires more than 146,000 MOVES runs to represent all emission and energy use rates for the Atlanta-region and was facilitated by the high-performance computing (HPC) cluster at Georgia Tech (Partnership for an Advanced Computing Environment (PACE)). The final MOVES-Matrix array for each modeling region contains more than 90 billion energy use and emission rates, which can be quickly queried and combined to generate on-road emission rates for any scenario analysis. MOVES-Matrix has proven its capability in estimating sub-regional and project-level emission impacts at more than 200-times faster than traditional methods (Guensler, et al. 2016; Liu, et al. 2017; Xu, et al. 2016; Xu, et al. 2018a; Xu, et al. 2018b; Kim, et al. 2019a; Kim, et al. 2019b). MOVES-Matrix is highly-desirable for regional-scale dispersion analysis, as it is capable of dealing with emissions calculations for numerous link sources under a variety of scenarios for seasonal and hourly variations in traffic operations and various meteorology conditions, fleet compositions, etc., without needing to create additional input files or to launch MOVES itself. This simple lookup system also helps minimize potential human errors in running MOVES for thousands of transportation links at a time.

Because emissions are a complex function of many locally dependent variables, and because the MOVES model interface is somewhat complex and requires numerous inputs for a specific emissions scenario, a significant amount of time and effort (i.e., human labor) is normally required to pre-process MOVES and prepare MOVES input files for

traditional scenario analyses. All of this pre-processing time is eliminated in using MOVES-matrix, as MOVES has already been run for all combinations of input variables. In addition, running the MOVES model is time-consuming because calculations of emissions always begin with base emissions rates, and for each run, the data are internally adjusted for aspects such as temperature, humidity, and fuel property. This adjustment makes it difficult to use the MOVES model to assess large-scale transportation networks that experience dynamic changes in on-road fleets and operating conditions. Thus, this study uses the MOVES-Matrix modeling approach to take advantage of the operating mode bin approach employed inside the MOVES model (see Figure 35), and the fact that the model allows users to specify the fleet composition down to a single-vehicle source type. For example, the CO₂ emission rate for the passenger trucks under the driving condition that operating mode bin is 25 and the vehicle speed ranges between 25 and 50 mph can be found through the MOVES-Matrix, which is approximately 4 grams per second. In this circumstance, because MOVES-Matrix contains all possible model input combinations for the selected regions and is well-structured based on MOVES input specifications, the users can easily query the desired MOVES emission rates (desirably using regular language scripts, e.g. Python, Java, C, etc.) from MOVES-Matrix arrays and obtain the exactly the same emission outputs that MOVES provides without ever having to launch MOVES. More details on the setup, implementation, and application of MOVES-Matrix can be found in Guensler, et al. (2016).

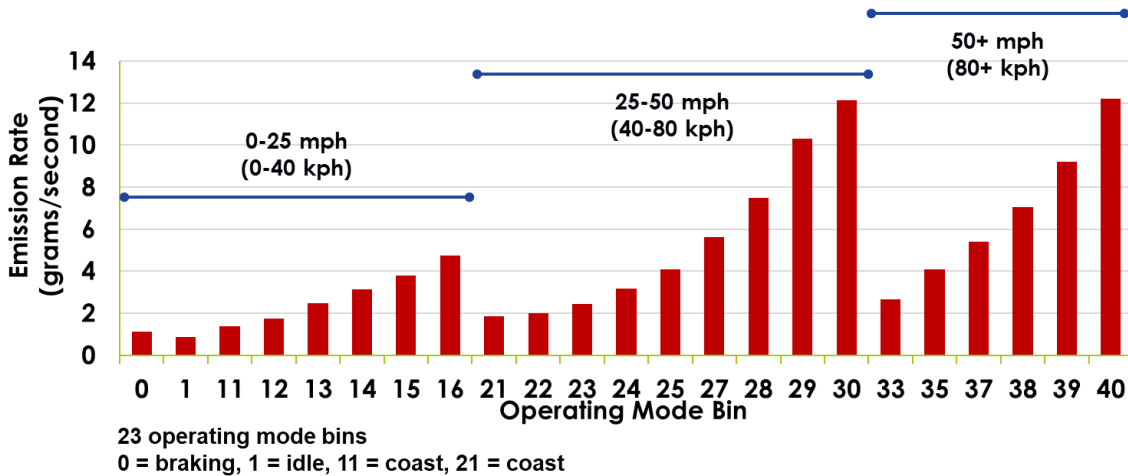


Figure 35 – Example of MOVES CO₂ Emission Rates by VSP Bin for Passenger Trucks (Model Year of 2016 in 2016)

6.3.3 Geographic Input Data

The spatial and geometric characteristics of roadway networks and receptors are a key factor in estimating the pollutant concentration profile. Because the proximity between road segments and receptor significantly influences the concentrations at a specific site, an optimal setting for receptor distribution is crucial for producing unbiased concentration estimates over the area of interest (Liu, et al., 2017; Wu and Niemeier, 2016; Guensler, et al., 2000). The modeling system incorporates a model to generate link geometry input files and proper receptor locations associated with the link geometry information. The program also integrates road grade information into the constructed road geometry as an elevation attribute to better estimate road grade contributions to link emissions, which has rarely been applied in previous regional-level microscale dispersion modeling (Liu, et al., 2019c). Finally, a built-in model feature converts the link-geometry and receptor location outputs to a format suitable for use in subsequent line source dispersion modeling.

6.3.3.1 Road Geometry

The dispersion models (e.g., AERMOD, CALINE4, R-LINE) typically allow either Universal Transverse Mercator (UTM) coordinates or X- and Y-coordinates in meters, specified relative to any user-specified coordinate origin. Most state-government and city-government agencies provide road geometry shapefiles, where each roadway link is represented by the coordinate information of a GIS-polyline. These road geometry data are defined in a specific geospatial coordinate reference system and usually need to be converted to a metric coordinate system suitable for each of the dispersion models. The conversion is made according to the input format requirements of each dispersion model. For example, R-LINE and CALINE4 require every link segment to be a straight-line, with coordinates representing the two link endpoints (network nodes). Therefore, the original road shapefile, which often includes polylines representing curved road segments with multiple nodes, must be transformed into a new shapefile by dividing each polyline into a set of straight lines. Furthermore, because the public release of CALINE4 only handles 20 links in a single simulation run, multiple CALINE4 input files with each file containing up to 20 link segments are generated by automatically decomposing the entire road network. AERMOD supports various road geometry formats including line, polygon, and different shapes of area sources. The USEPA's PM hot-spot guidance allows any of these options to be used for transportation projects (USEPA, 2015a). The modeling system developed as a part of this study utilizes the polygon option (Figure 36). Preparing roadway network geometries in proper formats for regional-scale dispersion modeling is usually a labor-intensive and time-consuming process, and because each dispersion model requires

different input formats, the streamlining program developed for pre-processing of these networks significantly reduces modeling labor.

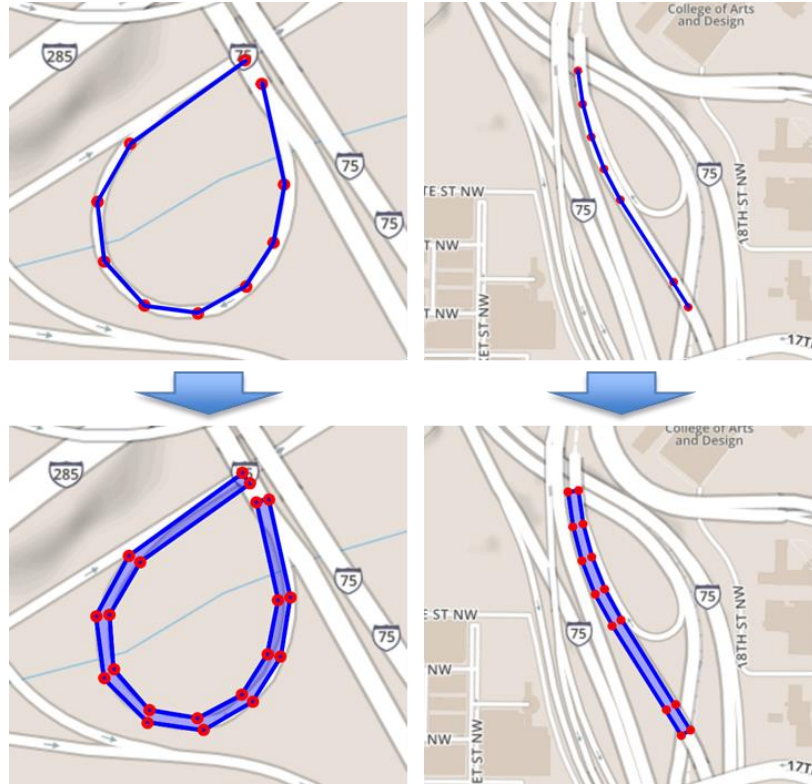
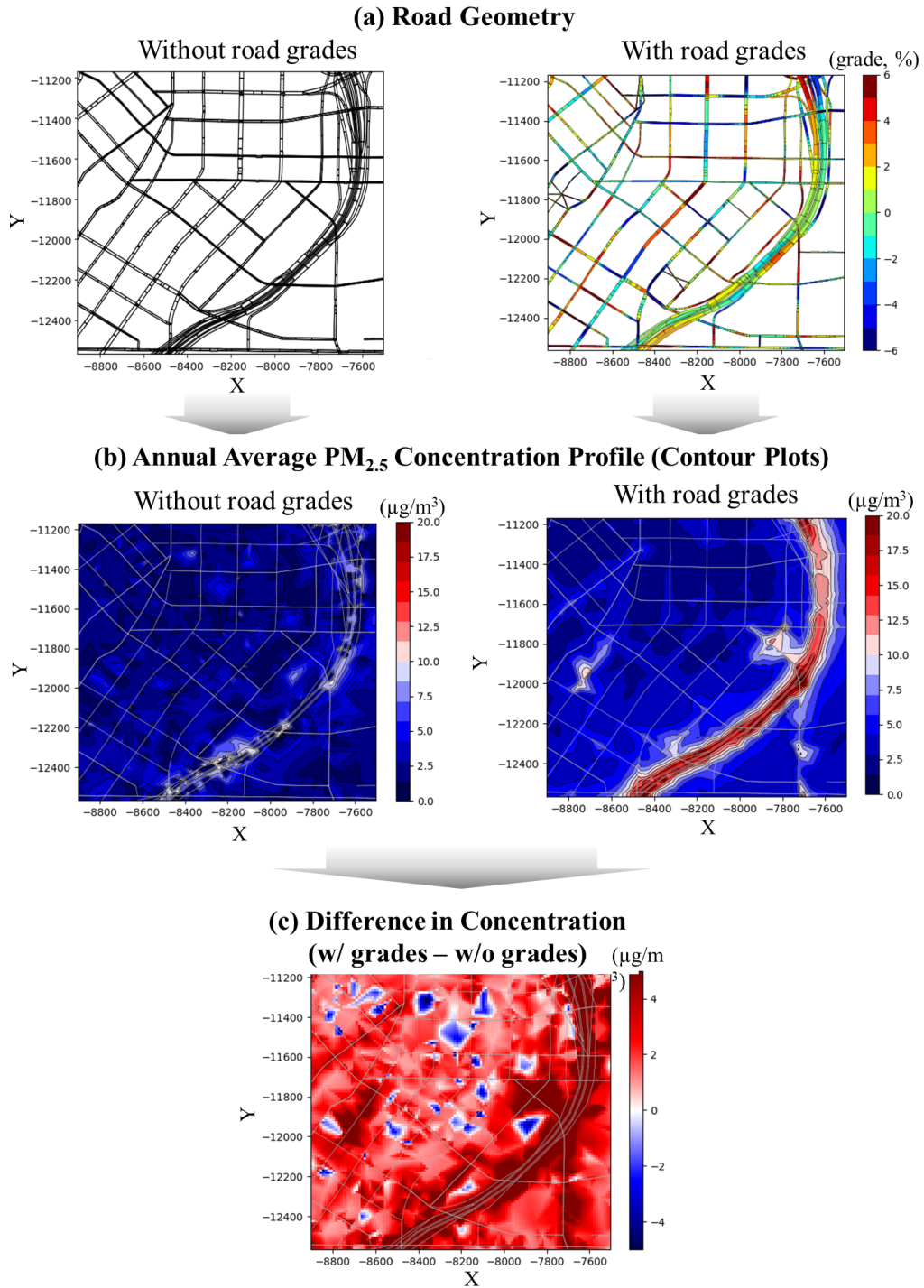


Figure 36 – Example of Generating Link Geometry Data using GIS-Polyline

6.3.3.2 Road Grade

Based upon the previous empirical analysis, the integration of road grade with link geometry data (as described in the preceding section) is important in line source dispersion model implementation (Liu, et al., 2019c). Roadway elevation profiles and road grade distribution are produced at high-spatial-resolution from the Digital Elevation Model (DEM), obtained from the US Geological Survey (Liu, et al., 2018). The resulting road grade profile is a layer consisting of points containing road grade values placed 10 m apart along the roadway network. This point layer is then used to further subdivide modeled

roadway links into sub-link for every change in road grade by 1% or greater, from -15% to 15%, as shown in Figure 37-(a). For example, the continuous portion of a road link where grade remains within a predefined 1% range (e.g., $1.5\% \geq \text{Grade} < 2.5\%$) becomes a new road segment (i.e., 2%). In the case of Metro Atlanta, each roadway link in the 203,000-link ARC-ABM network is divided into an average of about six new modeling links, based upon road grade. Preliminary analysis of the impact of road grade on predicted $\text{PM}_{2.5}$ concentrations in downtown Atlanta showed that ignoring road grade is likely to result in biased $\text{PM}_{2.5}$ concentrations across the area, particularly for potential concentration hot spots along the urban highways. The mean absolute error (MAE) in $\text{PM}_{2.5}$ concentrations were estimated to be $2.2 \mu\text{g}/\text{m}^3$ in downtown Atlanta when compared to those estimated with road grade and those without, with a 200 m by 200 m grid space. Figure 37(a) illustrates the road geometry and grade distribution; Figure 37(b) shows the annual average $\text{PM}_{2.5}$ concentration profile; and Figure 37(c) illustrates the difference in annual average $\text{PM}_{2.5}$ concentration profiles between estimates with and without road grade information.



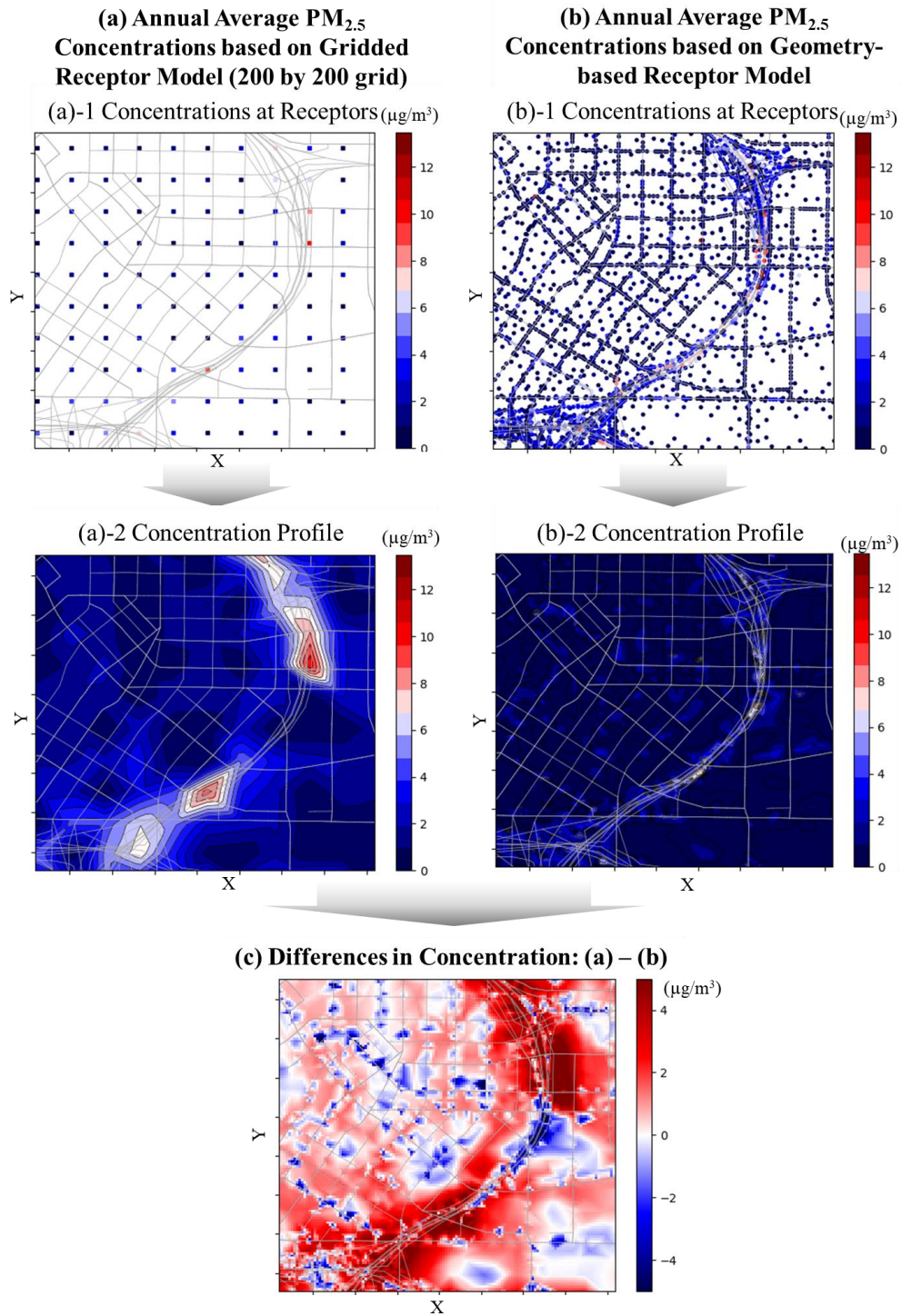
Note: PM_{2.5} concentrations were estimated based on ARC-ABM data using AERMOD.

Figure 37 – Impact of Road Grade on Pollutant Concentrations (Downtown Atlanta)

6.3.3.3 Receptor Location

Most previous studies for dispersion modeling have utilized gridded receptors with relatively low (e.g., 200 m by 200 m) resolution (see Zhai, et al., 2016; Wu, 2018; D'Onofrio, et al., 2016). The program presented in this study generates receptors that directly consider the road geometry across a metropolitan region at much finer spatial resolution. This approach places an increased density of receptors in areas of high road density to provide better predictions of pollutant sources and local area concentrations (all receptors are placed at a “human receptor” height of 1.5 m). The program also allows users to specify receptor resolutions for receptor installation at a specific distance interval along and adjacent to the roadway. Figure 38(b)-1 displays the receptors generated for downtown Atlanta. Receptors were placed at intervals of 100m along roadways, at 5-meters and 50-meters from the roadsides. Receptors are precluded from falling on top of any roadway polygons. Another group of receptors is then placed at intervals of 200m, at 100-, 200-, 400-, and 800-meters from the roadsides, again these receptors are precluded from falling onto roadway polygons. Finally, 1 km by 1 km gridded receptors are placed over the entire region. This receptor assignment mechanism, when used in dispersion modeling, yielded a much more precise understanding of the local PM_{2.5} concentrations (Figure 38(b)-2). Comparing the concentration profiles estimated with road geometry-based and gridded receptors shows that gridded receptors are likely to overestimate local exposure concentrations when receptors fall on top of (or immediately adjacent to) roadway polygons where high concentrations are predicted because these high values are spatially (linearly) averaged across the uniform low-density grid (proper spatial averaging requires application of non-linear weighting, based upon dispersion-related pollutant concentration

drop-offs with distance). Similarly, the models underestimate concentration fields when no receptors fall near roadway edges. The mean absolute difference between the concentrations from geometry-based and gridded receptor models was estimated as $1.01\mu\text{g}/\text{m}^3$ for downtown Atlanta (Figure 38(c)). Because an increased number of receptors results in longer total processing time, identifying an optimal setting of receptor resolution and placement is warranted for predicting concentration profiles at desired precision while minimizing the total computational time.

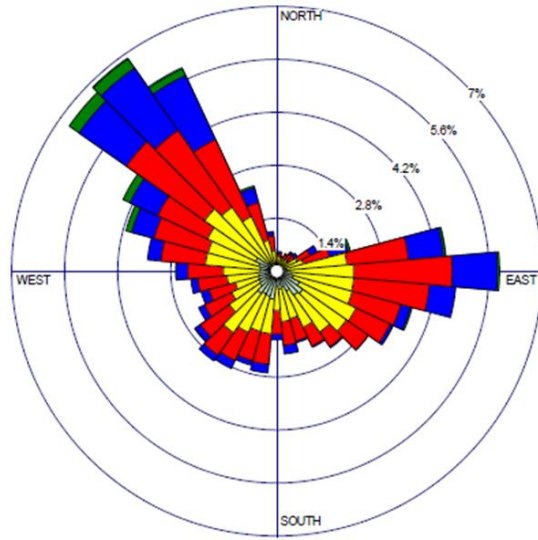


Note: PM_{2.5} concentrations were estimated based on ARC-ABM data using AERMOD.

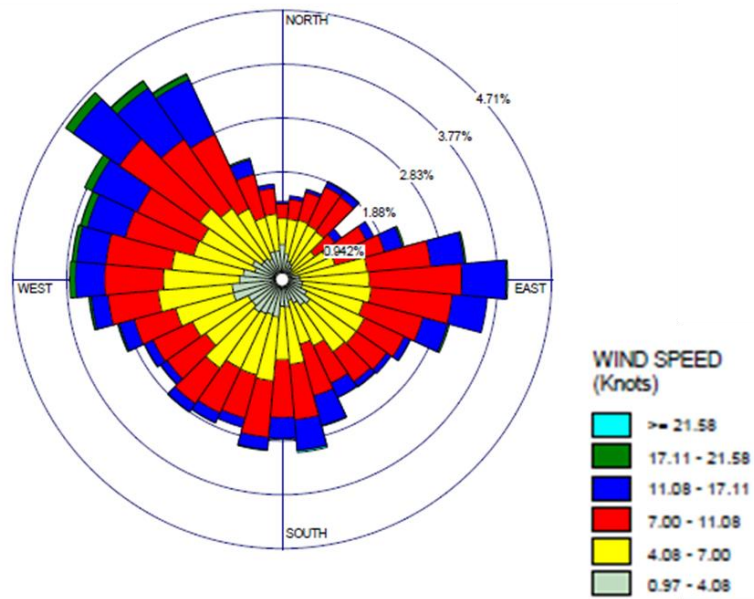
Figure 38 – Annual Average PM_{2.5} Concentrations from Geometry-based and Gridded Receptor Models

6.3.4 *Meteorology Data*

Meteorological condition data is one of the most important inputs in dispersion models. This meteorology data are commonly developed by state air agencies for use in AERMOD dispersion modeling (see an example of Georgia EPD, 2019). The meteorological files are normally processed using AERMET, AERSURFACE, and AERMINUTE, which are meteorological processors developed by USEPA, by taking surface and upper-air data provided by National Weather Service. The meteorological files contain a number of parameters explaining meteorological conditions across hours of the year (8,760 hours for 365 days), including wind profile (Figure 39), temperature, humidity, etc. In general, analyses take at least one full year of meteorology data to determine annual average pollutant concentrations conforming to the NAAQS standard. Simplified meteorology data are also prepared by categorizing the full year of data into a number of representative meteorological hours (D'Onofrio, et al., 2016; Zhai, et al., 2016). In particular, this study developed a modeling tool that matches hourly meteorological information in the meteorological files with hourly travel activity data (e.g., speed and volume) to extract appropriate emissions rates.



(a) Hartsfield-Jackson Atlanta Airport



(b) Savannah Airport

Data source: Georgia EPD (2019)

Figure 39 – Examples of Wind Rose Diagrams in Two Locations in Georgia, USA: Calendar Year of 2018

6.3.5 *Others*

Previous research has demonstrated that vehicle fleet composition can have a significant impact on emission results, and these inputs should represent local conditions (Granell, et al., 2002; Liu, et al., 2017). MOVES represents vehicle fleet features using 13 vehicle source type and 31 applicable model years (age 0 - 30 from investigated calendar year). Agencies can prepare the vehicle composition with data from the MOVES default database, while a more detailed data can be retrieved from state motor vehicle registration data, or license-plate data transcribed from the video and matched to the motor vehicle registration database (Liu, et al., 2017). In this regard, the modeling system developed in this study is capable of translating diverse local fleet data to be matched with appropriate emissions rates in the MOVES-Matrix platform.

Background concentrations are those emissions not from the project that also affect the project area. Under PM Hot-spot Guidance (USEPA, 2015a), background concentrations are combined with air quality modeling results to generate design values and determine project conformity. Ideally, background concentrations are determined by interagency consultation with state and local air quality agencies (USEPA, 2018), while air quality monitoring data can also be selected from USEPA's AirData website (www.epa.gov/airdata/).

Terrain conditions are also related to urban pollutant concentrations (Saide, et al., 2011). For some of the dispersion models (e.g., AERMOD and R-LINE), the urban terrain conditions are input in the form of surface roughness length. The selection of surface roughness length can be made with Grimmond and Oke (1999) (Table 11). However,

AERMOD set a default value of 1.0-meter surface roughness length, and USEPA addressed that any value for the urban roughness length other than 1.0 meter will be treated as a non-regulatory option. Thus, caution should be used when specifying a non-default urban roughness length, and the use of a non-default value should be clearly documented and justified (USEPA, 2018).

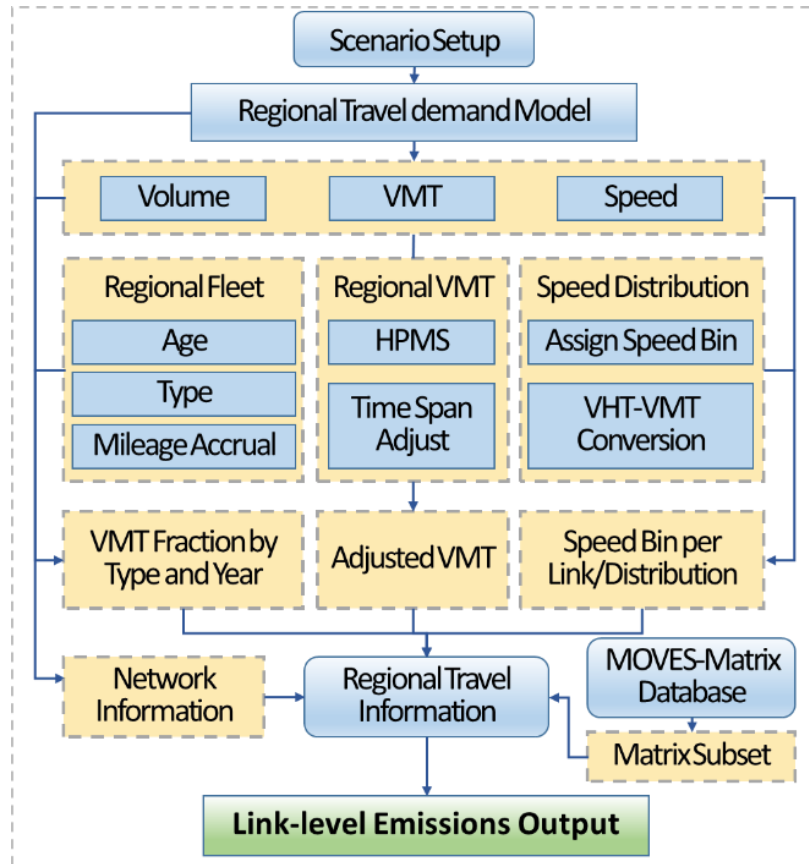
Table 11 – Typical Urban Boundary Layer Parameters

Urban surface form	Mean building height (m)	Displacement height (m)	Surface roughness length (m)
Low height and density	5 - 8	2 - 4	0.3 - 0.8
Medium height and density	7 - 14	3.5 - 8.0	0.7 - 1.5
Tall height and density	11 - 20	7 - 15	0.8 - 1.5
High rise	> 20	> 12	> 2.0

Source: Grimmond and Oke (1999)

6.4 Emissions Inventories

The modeling system employs an advanced emissions modeling method for regional emission analysis using MOVES-Matrix. The method is suitable for processing the outputs from the majority of TDM developed and provided by state and city planning agencies or traffic simulation model (Xu, et al., 2018a; Xu, et al., 2016) to generate regional mobile source emissions inventories (by summing link-by-link emissions). Typical TDMs predict static or dynamic traffic operating conditions (e.g., traffic volumes, link speeds) assigned to individual roadway links by time-period. These traffic data are then post-processed and linked to MOVES-Matrix emission rates. The process includes three underlying steps: 1) MOVES-Matrix preparation; 2) integration of MOVES-Matrix and the regional TDM outputs; and 3) emissions calculations (Figure 40).



Note: the figure has been modified based on Xu, et al. (2018a)

Figure 40 – Emissions Calculation Process

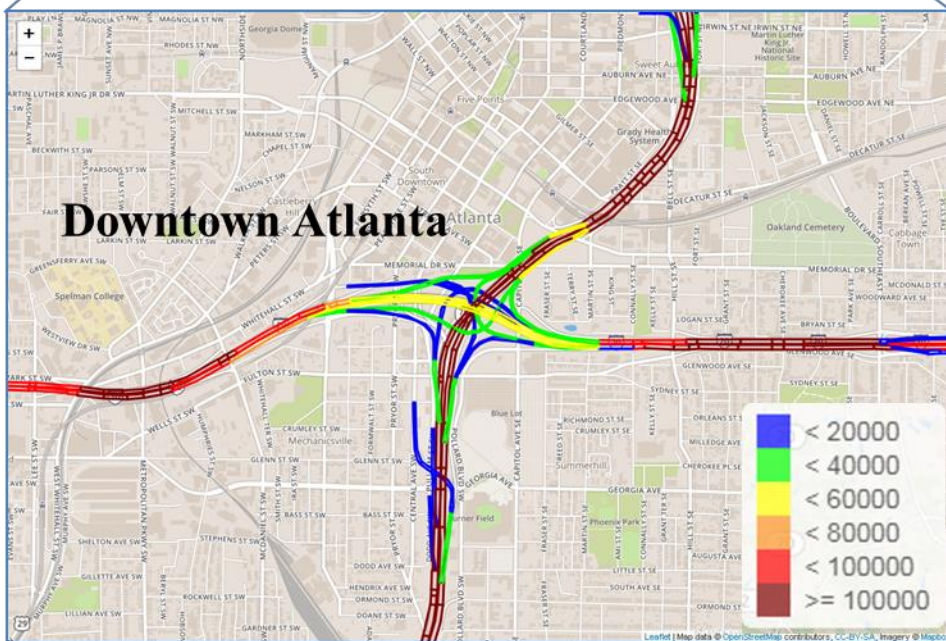
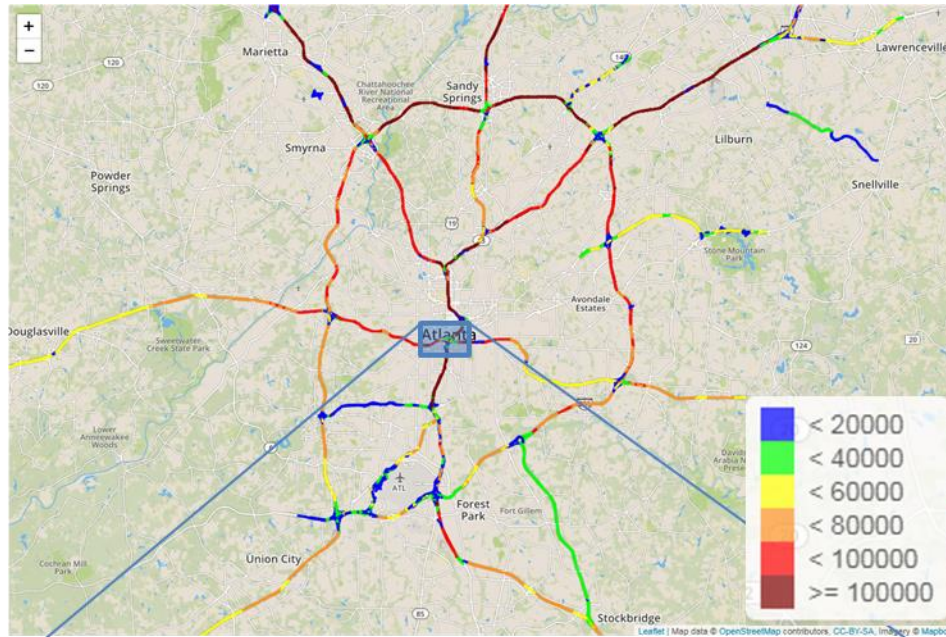
6.4.1 Integration of MOVES-Matrix and Regional Travel Demand Model

A series of Python™-script-based programs were created to process and convert regional TDM outputs into specific formats to link these TDM outputs with corresponding MOVES-Matrix emission rates. Typical TDM outputs include traffic volumes (or vehicle miles traveled, VMT), average speeds, and road types for each link (as examples in Figure 41 and Figure 42). The TDM outputs need to be further post-processed because the TDM is generally run for a certain traffic period (e.g., morning peak, afternoon peak) and reflects annual average traffic conditions. VMT adjustment factors representing the amount of vehicle activities during a specific time period are applied to the VMT outputs generated

by the TDM (see Equation (14)). Average link speeds included in the TDM outputs can be updated from field observation data, where applicable, to obtain more realistic on-road operating conditions. The roadway facility type within the TDM outputs may need to be matched with MOVES road types, where applicable. For example, vehicle operations on arterials are different from those on freeway segments. In sum, each link contains an array of the required information including VMT, average speed, road type, and VMT adjustment factors representative of variations across seasons (or months) and hours within the day.

$$\begin{aligned}
 VMT_{l,m,d,h} = & VMT_l \times \textit{Month adjustment factor}_m \\
 & \times \textit{Day adjustment factor}_d \times \textit{Hour adjustment factor}_h
 \end{aligned}
 \tag{14}$$

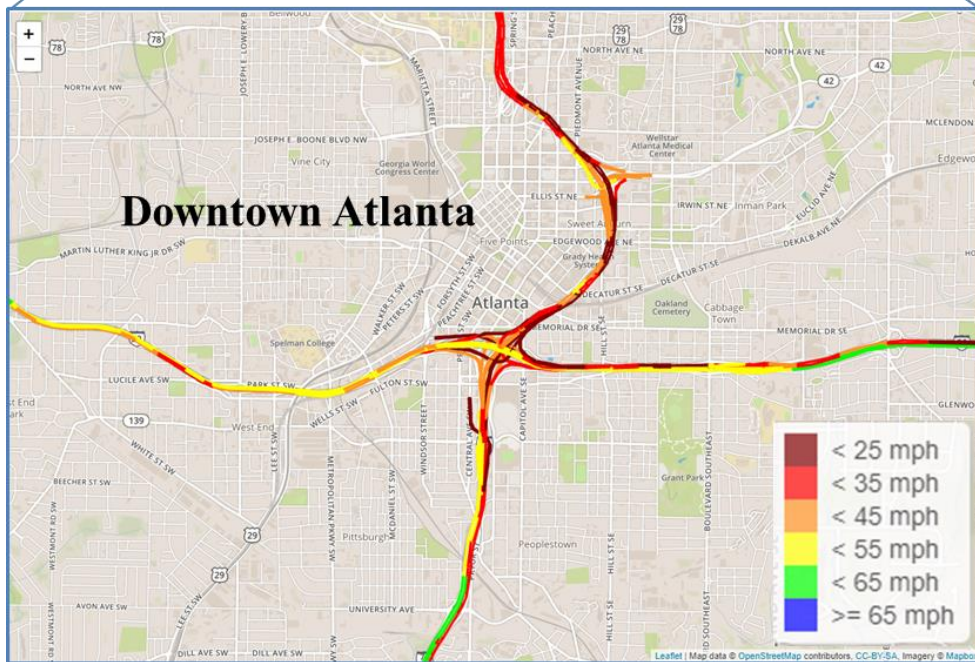
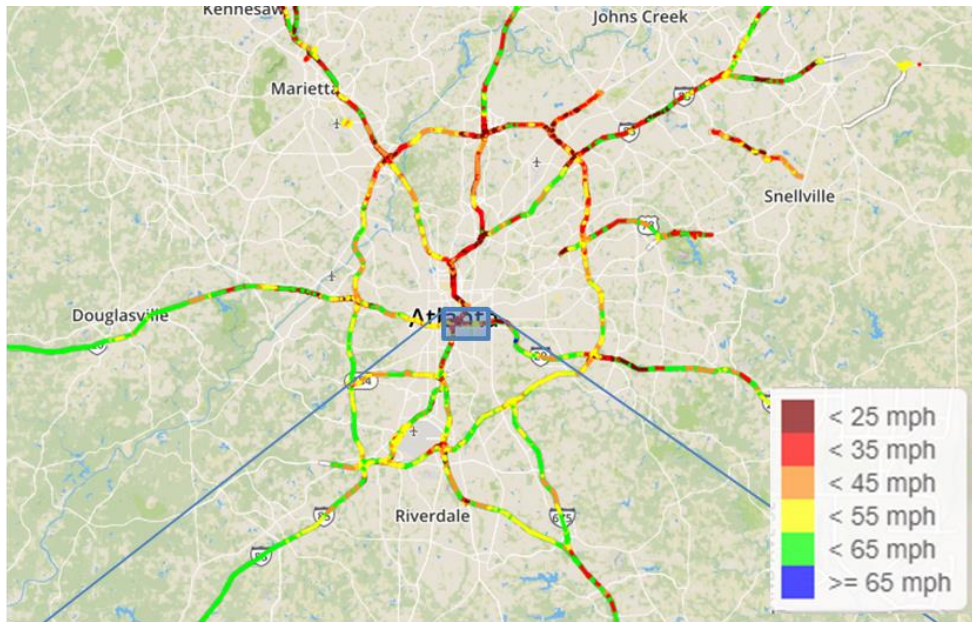
Here, l is link, m is month, d is day, h is hour,



Source: Kim, et al. (2019c)

Note: the figure shows the example of metro Atlanta major interstate highways.

Figure 41 – ARC-ABM’s Daily Link-by-Link Total Traffic Volumes in 2015



Source: Kim, et al. (2019c)

Note: the figure shows the example of metro Atlanta major interstate highways at 8:00 AM to 8:59 AM.

Figure 42 – ARC-ABM’s Link-by-Link Congested Traffic Speeds in 2015

MOVES classifies vehicle fleet features by 13 different vehicle source types (Table 12) and 31 applicable model years (the age of 0 to 30 from the target calendar year). By default, the program developed in this study uses the fraction of regional fleet composition by vehicle source type and age distribution in MOVES. Often, state vehicle registration data describing the proportion of each vehicle type registered in the state are used by impact assessment modelers in traditional MOVES runs. However, the program is also capable of integrating fleet composition based upon field observation (e.g., license plate studies) by allowing the user to allocate total VMT estimated by the TDM to MOVES-specific source types. Because MOVES only allows only one set of vehicle model years per run, users have to run MOVES separately for each set of links if the model year distributions are different. However, MOVES-Matrix allows users to specify any vehicle fleet composition data for road links (source type and model year distributions) without running MOVES. That is, users can specify different fleet compositions for each transportation link, represented by a three-dimensional array, with a fraction of population by source type, model year, and road type.

Table 12 – MOVES2014 On-road Source Types

Source Type ID	Source Type Name
11	Motorcycles
21	Passenger Cars
31	Passenger Trucks (primarily personal use)
32	Light Commercial Trucks (primarily non-personal use)
41	Intercity Buses (non-school, non-transit)
42	Transit Buses
43	School Buses
51	Refuse Trucks
52	Single Unit Short-Haul Trucks
53	Single Unit Long-Haul Trucks
54	Motor Homes
61	Combination Short-Haul Trucks
62	Combination Long-Haul Trucks

6.4.2 Emissions Calculation for Dispersion Modelling

Using MOVES-Matrix, emissions for each link are calculated by Equation (15); the same emissions calculation algorithm as used in MOVES. Corresponding vehicle activities, fleet composition, and on-road operating conditions are prepared for each link and used to assemble a composite emission rate for the link (the summation term in equation (15)). The composite emission rates are then multiplied by the amount of vehicle activity on the link to estimate the total emissions for the specific link segment.

Total emission for each link

$$= Activity_{fleet} \sum_{ST} \sum_{MY} \sum_{FTS} (ST\% \times MY\%_{ST} \times FTS\%_{ST,MY}) \quad (15)$$
$$\times ER_{ST,MY,FTS}$$

Here, $Activity_{fleet}$ is on-road vehicle activity of a link (in vehicle-miles or vehicle-seconds), $ST\%$, $MY\%$, and $FTS\%$ are source type distribution, model year distribution, and the distribution of on-road operations of the link in the form of a percentage of activity by facility type and average speed, and ER is the corresponding emission rates determined by operating speed, road type and fleet composition from MOVES-Matrix (Figure 43). As noted in previous studies, this process obtains exactly the same fleet composite emission rate for each link as is obtained by running MOVES separately for each link (Liu, et al., 2019b).

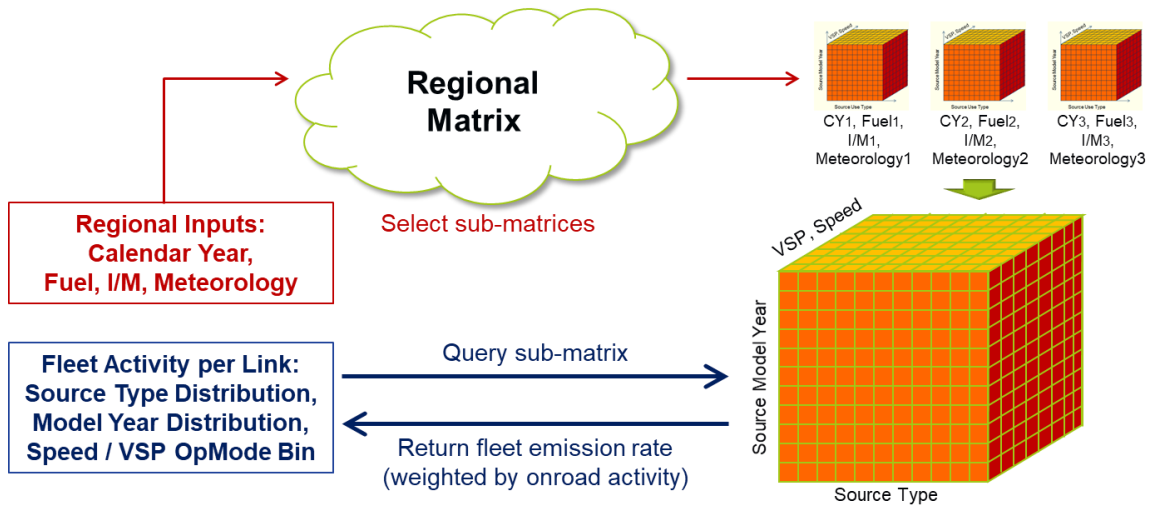


Figure 43 – MOVES-Matrix Run Module: Developing On-Road Fleet Emission Rates

6.5 Link Screening

In general, modeling traffic-induced pollutant dispersion for a large geographic area demands substantial computational resources and processing time. A preliminary study conducted in Metro Atlanta showed that the run time for AERMOD to process spatially-dense receptors installed in roadside and non-roadside areas (161,188 links and 1,163,958 receptors in Atlanta model) would require more than a year with a normal desktop computer (Kim, et al., 2019). To reduce computation time, previous research has implemented link screening, where roadway link and receptor combinations that do not impact concentrations at that receptor are excluded from the analysis (Guensler, et al., 2008; Shafi, 2008). However, link screening may result in under-prediction of concentrations. The contribution of concentration from a road to a receptor is a function of the mass flux from the roadway to the receptor, the distance between roadway and receptor, and a set of meteorological parameters. Building upon this methodology, this study created an updated objective version of the link screening method using a classification algorithm

based upon supervised random forests (RF) machine learning. Detailed documentation describing the development of this new link screening model will be addressed in Chapter 5.

6.6 Dispersion Modeling

Based on the database of link emissions, road geometry, receptors, and meteorological parameters, dispersion modeling is conducted for the project area. The modeling system supports screening modeling as well as standard modeling. Screening models generally predict the impacts of a “worst-case” or near worst-case analysis based upon conservative meteorological assumptions and on-road operating conditions. This approach can save much labor in obtaining high-resolution meteorology and on-road operating condition data, allowing the users to quickly evaluate alternative scenarios and identify potential hot spots where further investigation is needed. Standard modeling generally estimates the pollutant concentrations to meet regulatory modeling guidelines for planning activities and identifies likely concentrations in comparative scenario analysis (USEPA, 2018).

6.6.1 Screening Dispersion Modeling

The screening model approach essentially searches for worst conditions, based on mass flux calculations from mobile source emissions under a specific meteorological scenario. To derive the worst-case conditions for given on-road vehicle emissions, the database containing seasonal (or monthly) and hourly emission profiles can be used to identify the time periods that yield the highest emission rates among the seasons and hours considered. This process outputs the highest concentrations for each receptor during the year, for a set of given meteorological conditions.

The process also involves the adjustment of the meteorological parameters (in the dispersion modeling), such as atmospheric stability, wind speed, and wind direction, so that the dispersion model yields the highest concentration rates with the given link emission rates. It is noteworthy that CALINE4 and AERMOD, the dispersion modeling tools used in the modeling system also have internal screening options.

For CALINE4, the default worst-case meteorological parameter values are generally set to 1.0 m/s for wind speed, atmospheric stability G (7), and 30 to 100 m for the mixing height so that the model will simulate the worst-case concentration likely to be experienced in the field. Then, CALINE4 searches the worst-case wind angle for each source, by cycling through wind directions (Benson, 1984).

In the case of AERMOD, it finds the worst-case meteorological scenario among those conditions specified in a meteorological processor file (AERMET format), which is generally provided by the state or local agency responsible for implementing environmental regulations (USEPA, 2018). For example, the AERMET meteorology files for Atlanta are managed by the Georgia Department of Natural Resources (Georgia DNR), Environmental Protection Division (EPD). The internal screening mode in AERMOD forces the model calculations to represent values for the plume centerline, regardless of the source-receptor-wind direction orientation (USEPA, 2016).

Because these screening models do not necessarily perform the calculations considering dynamic meteorology conditions across the project area, they are naturally computationally efficient in identifying potential hot spots while minimizing the total model run time.

6.6.2 *Standard Dispersion Modeling*

The program database can also be used in traditional, or ‘standard’ modeling. For the standard models, link emission profiles that vary by season and hour of day are included within the model input files. CALINE4 uses fixed link emission rates for a specific date and hour in each model run. In other words, CALINE4 does not consider seasonal and hourly variations in link emission rates within a single model run. Therefore, a number of input files must be created that contain link emission rates by season and hour combination (e.g., 96 input files are created when four seasons and 24 hours are considered). Within each input file, appropriate meteorological conditions are coupled with corresponding season and hour. The modeling system can post-process the meteorological file in AERMET format to identify meteorological conditions for each season and hour. Although R-LINE accepts the AERMET output file, it has fixed link emission rates assigned for a specific date and hour for a single model run. Therefore, as with CALINE4, multiple input files need to be generated for R-LINE that contain the link emission rates for each season and hour. Then a subset of meteorological elements is extracted from AERMET for season and hour. In the case of AERMOD, because of its integrated model specification, standard modeling generates a single file that contains varying link emission profiles across different seasons and hours. After CALINE4 and R-LINE are run separately for each season and hour, the concentration results obtained from each model need to be aggregated across the year to assess whether the outcome conforms to the NAAQS. Those results are aggregated differently for each pollutant type (e.g., 8-hour average for CO, annual average for primary PM_{2.5}); whereas, AERMOD handles this aggregation process internally. Another

difference is that R-LINE and AERMOD deal with all NAAQS pollutants, whereas CALINE4 only handles CO, NO₂, and particulate matter.

6.7 Distributed Computing Strategy

Even with the efficiencies associated with MOVES-matrix, estimating regional-scale pollutant dispersion incorporates numerous sources and receptors under dynamic traffic and meteorological conditions, which requires extensive computational resources. The research team has priority access to the Partnership for an Advanced Computing Environment (PACE) high-performance computing (HPC) cluster at Georgia Tech. PACE was established for the primary purpose of providing an environment for distributed high-performance computing. For this study, the computational challenges will be met using this distributed computing resource. For implementation, dispersion model input data were split into individual dispersion model input files, where each file included data for a single receptor and its significant link sources selected by the RF classifier (Figure 44). Now that these numerous small-sized input files can be simultaneously processed through the distributed computing cluster (Figure 44), the total processing time will substantially decrease. Up to 500 PACE processors were available to the team through Georgia Tech's PACE allocation, meaning that 500 individual AERMOD input files are can be processed simultaneously, producing an (almost) corresponding savings in wall-clock processing time. The research team is in the process of expanding the system so that any user can take advantage of similar cloud-computing platforms, such as Amazon Web Services (AWS) or Microsoft Azure. Note that the modeling system is designed to assign one core to one processor for modeling convenience. However, this strategy may decrease the distributed computing efficiency to some extent because this strategy is not likely to use the maximum

computing capacity of each processor. This calls for future research regarding optimal core-processor allocation strategy that maximizes the modeling efficiency.

For each receptor,

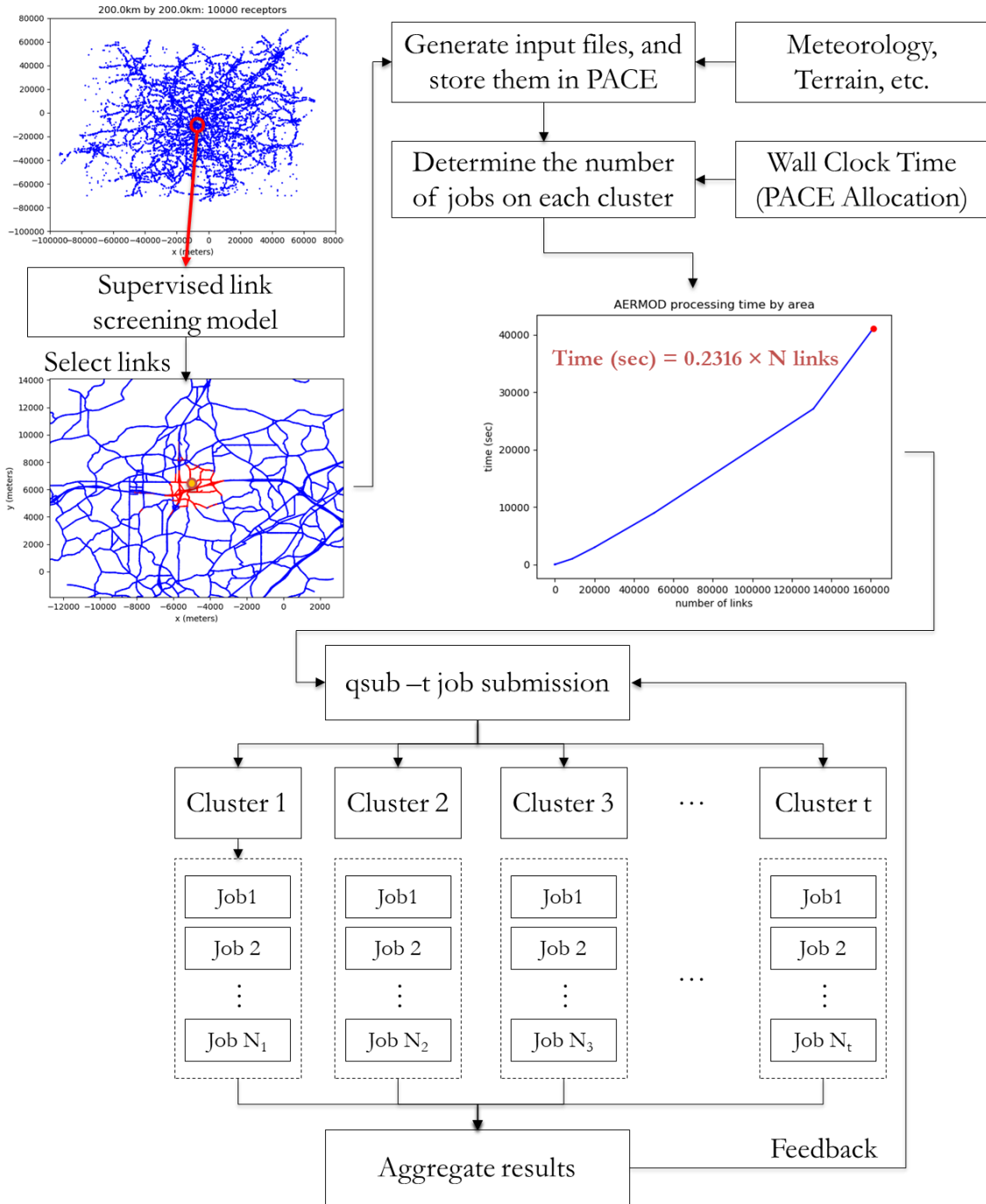


Figure 44 – Distributed Modeling Process for Dispersion Modeling

6.8 Case Study

This section describes the results of applying the modeling system to the 20-county metropolitan Atlanta area as a case study, focusing on the estimation of $PM_{2.5}$ concentrations in the area. The 20-county metropolitan Atlanta area includes Cherokee, Clayton, Cobb, DeKalb, Douglas, Fayette, Fulton, Gwinnett, Henry, Rockdale, Barrow, Bartow, Carroll, Coweta, Forsyth, Hall, Newton, Paulding, Spalding, and Walton counties. The number of roadway links considered in this application, including existing freeway, arterial, and local roads, excluding centroid connectors, was 161,188. Following the spatial distribution of these roadway links, receptors are placed across the area at fine spatial resolution. As explained earlier, receptors are installed based on the dynamic grid-receptor model. As a result, 878,731 receptors are created in Metro Atlanta. Figure 47 shows the geographic layout of these receptors.

Vehicle activity and link geometry data are obtained from the ARC-ABM model, the regional TDM developed for the regional transportation plan and the Atlanta Region's Plan (ARC, 2017). The ARC-ABM output files contain link-level vehicle activity data, including traffic volumes and average speeds for five different time periods (early AM, AM, midday, PM, evening/late night periods). These traffic data are further elaborated by using month, day, and hour VMT adjustment factors to create variations in link-level VMT (Xu, et al., 2018a). The monthly and daily adjustment factors are provided by MOVES as default, and the hourly adjustment factor is derived from post-processing of ARC-ABM outputs (Figure 45-(a) and (b)). To account for varying meteorological conditions affecting vehicle emissions rates (e.g., more incomplete combustion at low ambient temperature), the meteorological processor data are provided by the Georgia EPD (2019) (upper air

station ID: 53819; surface station ID: 3813; latitude: 33.355278; longitude: 84.566944; station height: 243.2 m). Combining the lowest, highest, and average surface temperatures and humidity, nine (3×3 matrix) different meteorological conditions are considered for each of the months (Figure 45). Therefore, the case study considers 5,184 alternative scenarios (12 months × 2 types of days (weekday and weekend) × 24 hours × 3 temperatures × 3 humidities) in the emissions calculation. Notably, estimating emissions under such a variety of conditions for a regional-scale dispersion analysis has been rarely conducted in prior studies, possibly because of the complexity of using the MOVES GUI in processing numerous link sources. In this study, the modeling was feasible because the system could take advantage of the MOVES-Matrix platform. For each of the 5,184 scenarios, the vehicle emissions rates corresponding to the fleet, on-road operating conditions, and meteorological conditions are retrieved from MOVES-Matrix.

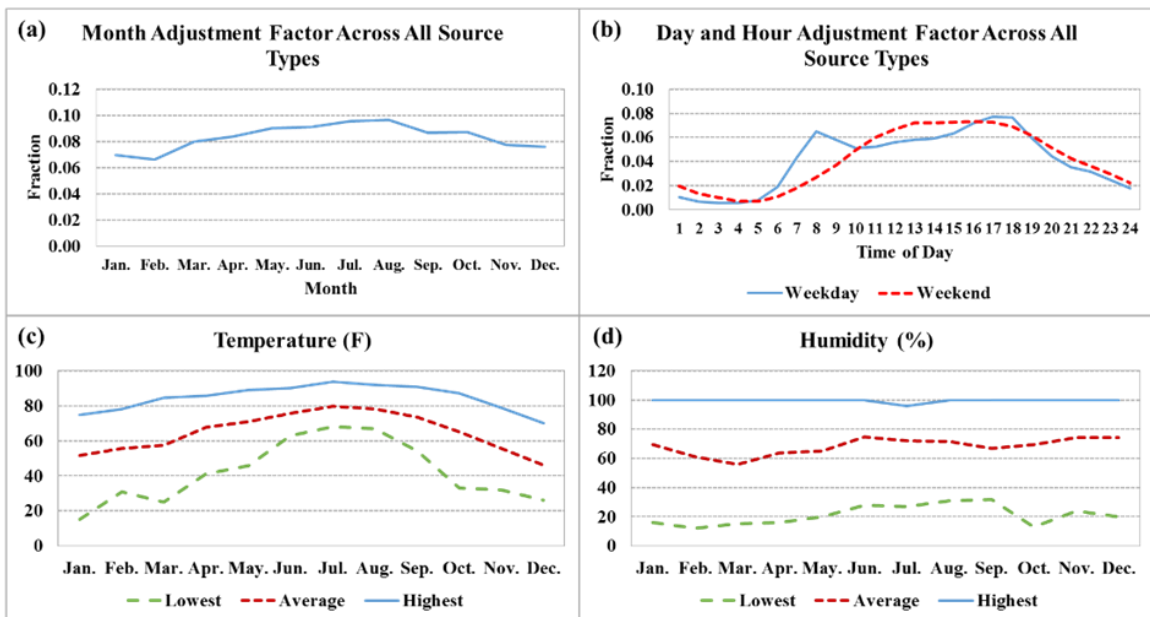


Figure 45 – Temporal and Meteorological Factors Considered in Emissions Calculations

Figure 46 displays the sensitivity analysis results of the estimated total PM_{2.5} link emissions, indicating that the total PM_{2.5} emissions produced in Metro Atlanta vary by traffic operating condition, by month (Figure 46(a)), day of week (Figure 46(b)), temperature (Figure 46(c)), humidity (Figure 46(d)), hour of day (Figure 46(e)). In each sub-chart, the gray circle represents the sample mean and the green error bar represents the 99% confidence limit on the mean (that is, we are 99% confident that the true mean falls somewhere within the error bar). These estimated link emissions are then converted to the dispersion model format; for example, link emission rates were converted into the mass flux rates (g/s/m²) using the polygon-shaped link geometry data. The raw link geometry data were obtained from ARC-ABM network data.

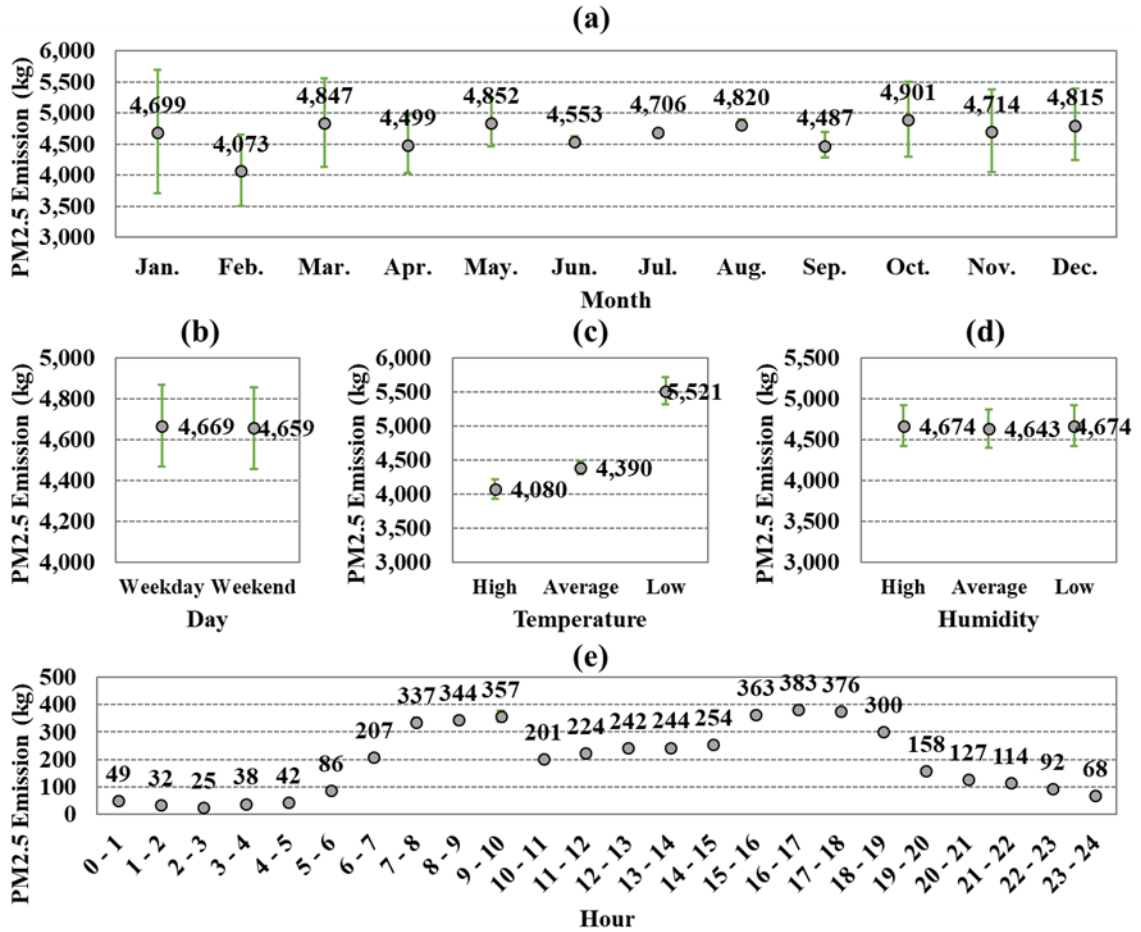
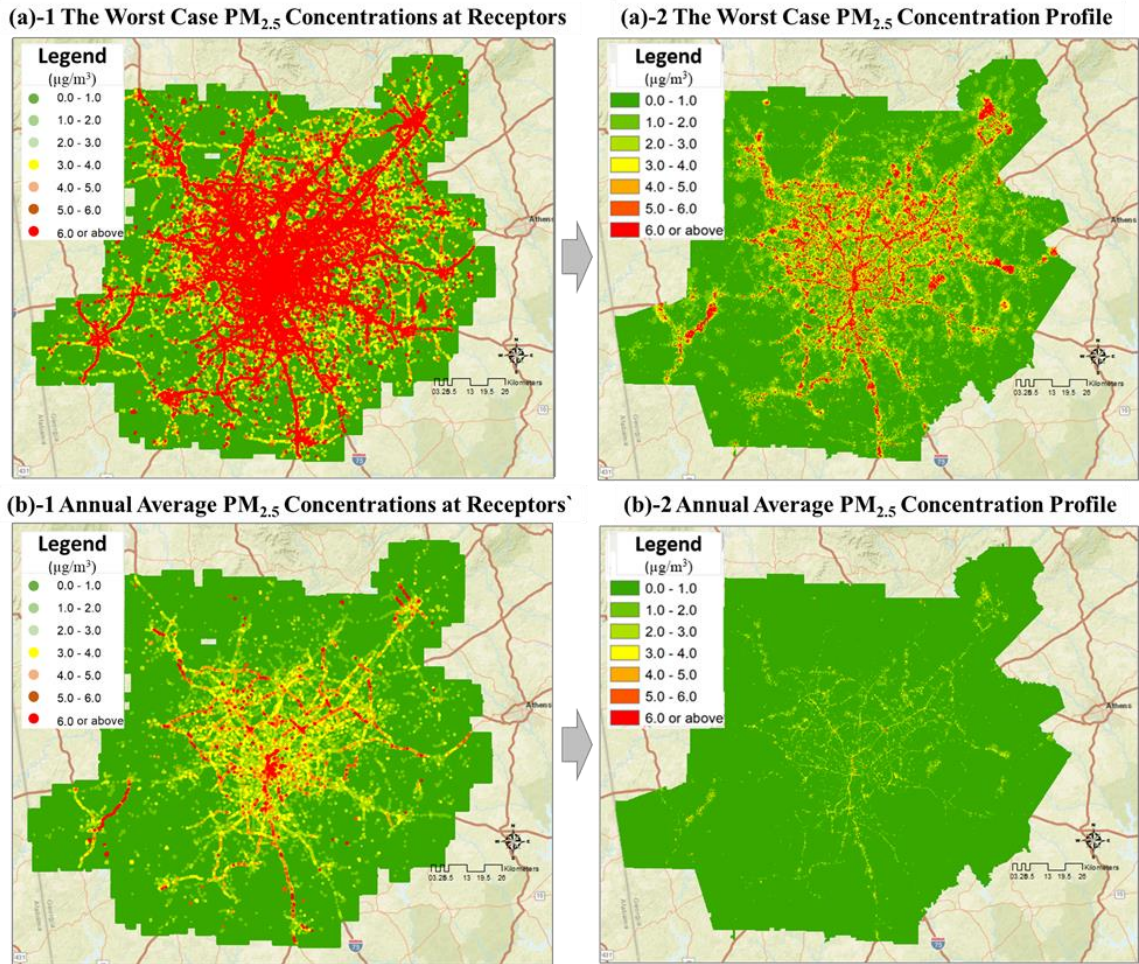


Figure 46 – Sensitivity Analysis of the Total PM_{2.5} Emissions in Metro Atlanta

The PM_{2.5} dispersion modeling was conducted in AERMOD due to its efficiency in terms of incorporating a multitude of link emission rates that vary by season and hour of day. Furthermore, the streamlined modeling system can be further enhanced if new models can be applied to significantly reduce the number of link-receptor combinations that are considered when calculating receptor concentrations. For screening dispersion modeling, the highest link emission rate among the 5,184 link emission scenarios was selected for the AERMOD link emission rate. For standard dispersion modeling, the average seasonal and hourly link emission rates (96 rates = 4 seasons × 24 hours) were

estimated by aggregating the 5,184 link emission rates into the corresponding season and hour, and they were used for link emission rates in AERMOD. In this way, the standard dispersion modeling connects the seasonal and hourly link emission rates to the corresponding meteorological conditions through the AERMOD processing.

Figure 47 shows the results of estimating PM_{2.5} concentrations using both screening and standard dispersion models for Metro Atlanta. The screening dispersion model results provide the estimated PM_{2.5} concentrations for the worst-case total emission scenario among the alternative scenarios (Figure 47(a)). The emission rates for the weekday in January with the lowest temperature and humidity conditions turned out to be the worst-case situation. For the standard model, 24-hour link emission rates for four months (January, April, July, and October, representing the four seasons) were used in the AERMOD input file. As expected, the standard dispersion model produced consistently lower PM_{2.5} concentrations than the screening model. The overall model results seem to be in line with the common patterns in prior studies in that higher PM_{2.5} concentrations are clearly observed along the major Interstate highways (e.g., I-85, I-75, and I-285) in Metro Atlanta (Kim, et al., 2019c). The case study also demonstrated a greater performance of the streamlined modeling system, compared to existing approaches, in terms of total processing time. Processing a single AERMOD input file took around 3.5 minutes for each screening model run (about three days in total) and 12.7 minutes for each standard model run (about ten days in total). Traditional methods would require many years to perform the same analyses on a single machine.



Note: The concentration profile graphs are generated based on the receptor concentrations ((a)-1, (b)-1) using the Kriging tool in ArcGIS.

Figure 47 – Estimated PM_{2.5} Concentration Results: (a) Screening model, (b) Standard Model

The dramatic reduction in computational time in implementing the regional-scale dispersion modeling framework resulted primarily from the innovative methods for reducing the number of receptor-link combinations required by using the supervised machine learning classifier, and employing the PACE distributed computing platform. For example, Table 13 shows that the dynamic grid-receptor model reduced the number of

receptors by 92%, and the link screening model reduced the number of links by 99%, respectively. In addition, the dispersion modeling through PACE boosted the AERMOD simulation speed by 100 times as compared to the AERMOD modeling with a normal desktop computer. As such, the availability of the streamlined modeling program developed in this study is expected to make it more feasible to assess regional-level air quality impacts of transportation projects while taking various project scenarios into account.

Table 13 – Expected and Actual AERMOD Runtimes for Metro Atlanta (Standard Modeling)

Classification	Average number of links	Receptors	Total runtime
Traditional modeling	161,188	10,626,677 ¹	many years ^{2,3}
After applying dynamic grid-receptor model	161,188	878,731	many years ^{2,3}
After applying link screening	23	878,731	1,220 days ^{2,3}
Dispersion modeling through PACE	23	878,731	10 days ³

¹ The number of receptors is determined based on USEPA’s hot-spot guidance (25-m resolution near roads and 100-m resolution for other areas)

² Expected runtime

³ The runtime does not include ARC-ABM model runtime. The ARC-ABM traffic assignment runs around 1.5 days (or 4 days with path retention)

6.9 Chapter Summary

This study illustrated in this chapter introduced a streamlined modeling framework to implement regional-scale line source dispersion analysis with high spatial resolution. The modeling system provided a significant improvement not only in model prediction resolution, but also in computational efficiency and simplicity of the user interface. The modeling system augments roadway link attribute data by directly integrating road grade information with the link geometry information. As demonstrated in previous work, the road grade contributed to the emissions calculations and should not be ignored in the identification of potential concentration hot stops (e.g., continuous uphill roadway segments with a heavy mixture of passenger cars and trucks). Including grade yields more realistic estimates. The program also used an increased number of receptors whose density and locations are spatially dependent on the configuration of roadways. This approach can produce a more complete concentration profile than simple, sparsely-gridded receptors, and thereby help better capture the spatial variability of pollutant concentrations profile. The program allows variation in the temporal and meteorological conditions for dispersion analysis, comparing 5,184 alternative scenarios in the emissions calculations as presented in the case study.

Despite the increase in input data precision and spatiotemporal resolution, a huge decrease in computational time was also achieved by integrating some innovative features (i.e., AERMOD processing time for metro Atlanta decreased from years to 10 days). The program utilizes the MOVES-Matrix platform to rapidly produce link emissions rates for mobile source emissions under a variety of scenarios combining varying traffic operating conditions by different time periods. Running MOVES-Matrix is more than 200 times

faster than running the MOVES interface to achieve the same output. MOVES-Matrix outputs can also be automatically processed to obtain the appropriate fleet average emission rates for each link, by season, hour of day, and meteorological condition. The modeling system also employs a link screening feature that objectively identifies all irrelevant link sources that do not significantly affect the pollutant concentration of each receptor, and excludes the irrelevant link-receptor combinations from the modeling. The verification results showed that while each project is unique, 0.1 - 3.0% of the selected link sources could closely approximate most pollutant concentrations estimated using all of the links. The presented program also employed a distributed computing cluster to perform a regional-level microscale dispersion analysis. Multiple input files created for the number of receptors were split and simultaneously processed, significantly reducing the total run-time.

The case study based on the 20-county metropolitan Atlanta area showed that the modeling system generated reliable concentration estimates with high computational efficiency. The total processing time for running dispersion models was around 3 days and 10 days, for the screening model and standard model, respectively, which seem reasonable for regional-scale analysis involving a huge number of link sources and receptors and performing comparative analysis across thousands of scenarios (especially considering that comparable analysis conducted with traditional methods on a regular desktop would take years to complete). The streamlined modeling system can also be interfaced with most of the travel demand models, dynamic traffic assignment models, and traffic simulation models for use in air quality impact assessment with regional and sub-regional analyses. The authors anticipate that once the modeling system is deployed on conventional cloud

computing platforms (such as AWS or Microsoft Azure), federal, state, and local agencies will be able to compare numerous project scenarios at the regional level very rapidly, reliably, and automatically. The efficiency and precision of the modeling system should be of interest to a broad group of stakeholders who are in vehicle emission modeling, transportation conformity analysis, near-road air quality impact assessment, and health impact assessment.

CHAPTER 7. CONCLUSIONS AND FUTURE RESEARCH

Chapter 7 summarizes the conclusions and contributions of the research presented in this dissertation, including the proposed regional-level microscale dispersion modeling process, theoretical contributions, and practical contributions. Some limitations of the dissertation research are listed and future research needs are discussed.

7.1 Regional-Level Microscale Dispersion Modeling Processes

Traditional efforts in region-wide line source dispersion modeling have faced major concerns regarding modeling complexity, data preparation, and computational efficiency. To address these issues, the work developed a streamlined data processing tool for efficiently preparing extensive input datasets needed for regional-scale line source dispersion modeling. In particular, the modeling system integrated several advanced modeling features, to generate reliable pollutant concentration estimates with high computational efficiency, including MOVES-Matrix integration, supervised link screening, dynamic grid-receptor modeling, and distributed computing.

The modeling system integrates MOVES-Matrix (a high-performance emission rate lookup system) for efficient on-road emissions calculation that accounts for dynamic traffic operations. Because USEPA's MOVES model requires a complex model set-up and huge computational resources for large-scaled on-road emissions calculations, most of the traditional modeling methods that used the MOVES model have tried to simplify the dynamic traffic operating conditions into the averaged conditions. Such simplifications reduce the total number of MOVES runs and the total model run-times for MOVES;

however, these approaches produce model output bias (e.g., unsuitable in identifying seasonal peak emissions). The integration of MOVES-Matrix for on-road emissions calculations (as in this dissertation) helps reduce model bias, by supporting calculations under dynamic traffic conditions with reduced modeling efforts and reasonable model run-times. This is because the MOVES-Matrix can be integrated using Java, Python, Perl or any similar program, MOVES-Matrix can finish the emissions computation tasks 200 times faster than using the MOVES GUI and the generated results are exactly the same.

The integration of MOVES-Matrix also allows users to consider road grade information in the on-road emissions calculations. The modeling system augments roadway link attribute data by directly integrating road grade information with the link geometry information. In the process, emission rates for desired road grade intervals (e.g., every change in road grade by 1%) can be prepared in the MOVES-Matrix platform so that they can be easily queried for each link segment. To this end, generating road grade profiles can be conducted from the USGS's DEM within the modeling framework. As such, the emissions calculation considering road grade profile became much more efficient. Considering that previous modeling systems have ignored road grade profiles in emissions calculations due to the complexities of generating road grade profiles for large-scale networks (and the increased model run-times for emissions calculations with road grades through MOVES GUI), this work helps practitioners obtain more realistic emissions estimates (e.g., potential concentration hot-spots near continuous uphill roadway segments with a heavy mixture of passenger cars and trucks).

The modeling system also integrated a distributed computing cluster to perform a regional-level line source dispersion analysis, which has been barely employed in previous

modeling efforts. The distributed computing processing multiple dispersion modeling jobs could reduce the total dispersion model run-time with the number of distributed computing clusters. For example, the distributed computing with 500 PACE clusters (available to the team through Georgia Tech's PACE allocation) produced corresponding savings in wall-clock processing time. With a lack of previous modeling efforts that implemented distributed computing in dispersion modeling, the significant reduction in dispersion modeling run-time through the distributed computing as in this work is unique in this field.

A supervised link screening methodology is also incorporated into the streamlined modelling framework as an innovative way to significantly improve modeling efficiency. The link screening models are developed based on a supervised machine learning Random Forest classification algorithm, with a large number of training and testing datasets predicted by AERMOD. The link screening model is capable of eliminating roadway links that do not contribute significantly to receptor concentration with a high classification accuracy (greater than 95%). The case study results based on two sub-regions in the metropolitan Atlanta suggest that the link screening model could significantly reduce the total dispersion modeling run-time (a 98.9% - 99.8% reduction in model run-time), while preserving the pollutant concentrations as predicted by whole link models.

The work addressed in Chapter 4 proposed a strategic receptor placement method (called dynamic grid-receptor model) that minimizes the number of receptors without undermining the pollutant concentration profiles generated by high-density receptor model. The dynamic grid-receptor approach is proposed in the placement of receptor locations with respect to link geometry and meteorological parameters. As a result, the work suggests that the optimal receptor placement from the dynamic grid-receptor models readily

approximate the PM_{2.5} concentration profiles predicted by high-density receptor models. The model results also suggested that the dynamic grid-receptor model helps to reduce biased pollutant concentration profiles predicted by static grid-receptor models that most of previous studies have applied.

The case study applying the developed modeling framework to the 20-county metropolitan Atlanta area showed that the modeling system generated reliable concentration estimates with high computational efficiency. The total processing time for running the dispersion models was around 10 days, which seem to be affordable for regional-scale analysis involving a huge number of link sources and receptors and for performing comparative analysis across thousands of scenarios (especially considering that comparable analysis conducted with traditional methods on a regular desktop would take years to complete).

7.2 Theoretical Contributions

This dissertation has identified some theoretical findings that could be added to the literature on traffic-related pollutant dispersion research. First, some influential factors that significantly influence receptor concentrations were identified. In the absence of literature that summarized the influential factors that affect traffic-related pollutant concentrations, the thorough literature review results as well as the estimated coefficients from the logistic regression models summarized in Table 8 provide a better understanding of the relationships between the influential factors and relative contributions to predicted concentrations. For example, link emission rate can increase concentration, while concentration decreases with the distance between the link segment and the receptor site.

In particular, the dissertation suggested that six variables including link emission rate, distance between link and receptor, urban population, sensible heat flux, surface roughness length, and wind speed are the significant factors determining traffic-related pollutant concentrations. As such, the model results also suggested that the predicted concentration at receptor site should be interpreted in relation to those influential parameters.

The work also identified the fundamental relationships among some dispersion model input parameters and predicted near-road pollutant concentration profile. The results suggested that link emissions, wind speed, and wind direction are important factors that determine concentration profile. For example, steep gradients in concentrations at near-road areas were found when the near-road link emission rate is high and the wind speed low. In addition, high concentrations were found in the areas where the wind runs along roadway sections. Such trends helped identify the optimal locations of receptor sites (i.e., dynamic grid-receptor approach as in this work) that can produce unbiased concentration profiles, suggesting that the unbiased concentration profiles can be predicted with only a few numbers of receptors. It also shows that the optimal receptor model can approximate the concentration profiles as predicted by the high-density receptor model than standard gridding methods. Considering that traditional modeling methods have mostly applied standard gridding methods (where the distance between receptors are uniform in all directions, e.g., 200 m resolution), the model results identified in this work imply that concentration profiles predicted by the previous modeling efforts might have produced model biases to some extent because predicted concentrations are highly sensitive to the distance between emission source and receptor site. For researchers and practitioners who implement air quality dispersion modeling of traffic-related pollution, setting proper

locations of receptors has always been a challenge. In this regard, this work could provide an idea about how many receptors are required and where those receptors need to be located. In particular, the dynamic grid-receptor model can help minimize the regional-scale dispersion modeling run-time, by removing unnecessary receptors in the analysis, while not undermining the pollutant concentration profile as predicted by high-density receptor model.

7.3 Practical Contributions

This dissertation work also produced some important contributions to practical modeling fields. First, the modeling system is streamlined through the entire modeling processes including input preparation to dispersion modeling, and developed using PythonTM-based tools (PythonTM is currently one of the most popular programming languages). The open source program can be made available immediately to agencies and consulting firms. Once the modeling system is deployed on conventional cloud computing platforms (such as AWS or Microsoft Azure) any user could conduct the same modeling processes featured in this dissertation for their own analyses (see Appendix B, for more details about the model deployment on AWS).

Second, the modeling system developed in this work provides a user-friendly interface through the developed PythonTM program. Considering that current air quality impact assessment efforts at this scale are labor-intensive because of the complex input data preparation procedures, the modeling system can significantly reduce the labor of practitioners and modelers and help minimize analytical errors that can be produced through the complex procedures. For example, the modeling system provides a tool for

generating source geometry used for dispersion models, by converting transportation network files (typically, containing GPS coordinates of both link-ends and mid-points for curved link geometry) into source geometry formats in dispersion models. In general, converting the road geometry of general transportation network files into dispersion model formats are complex (e.g., AERMOD requires to calculate the area of emission source, and R-LINE and CALINE requires to calculate link length). Thus, manual conversion requires a huge amount of time and effort. In addition, the conversion generally requires coordinate transformation (e.g., GPS coordinates to coordinates in meters). In this regard, the PythonTM program developed based on GeoPandas (PythonTM-based geo-spatial analysis tool) allows to quickly generate the source geometry required for dispersion modeling. Transferring data through subsequent modeling steps is also challenging (e.g., converting travel demand model outputs to be used for emissions calculation, and converting link emissions outputs to dispersion model inputs). In this regard, the streamlined program can also provide a way to efficiently connect each of the modeling steps.

Third, the modeling system can be interfaced with transportation models at any scale. For example, the modeling system can be linked to most of travel demand models, dynamic traffic assignment models, traffic simulation models, and real-time traffic monitoring data for use in air quality impact assessment with regional and sub-regional analyses. Modeling requirements may differ, depending on the scope of air quality impact assessments, but connectivity can occur at any scale. For example, the project-level analysis may require microscopic traffic simulation model (e.g., VISSIM) for predicting individual vehicle operations. Some projects may conduct analysis based on observational data. Although the different traffic data inputs may require different model connections to

be developed in Python, the modeling system is capable of conducting dispersion modeling within that modeling framework. For example, MOVES-Matrix allows to process any traffic data (from individual- to aggregated-level data) for emissions calculation. In addition, because the modeling system processes the modeling on a step-by-step basis, once the users manipulate their own data corresponding to the data formats required for each step in the developed modeling framework, the same modeling processes can be performed by others. Specifically, real-time pollutant concentration predictions can be performed by automatically connecting real-time traffic monitoring station data to the modeling system deployed on cloud-computing platforms. The linkage between real-time traffic data and the modeling system may also help predict potential hot-spots in connection with forecasted traffic and weather information. To this end, the input formats in the developed modeling system is simple (e.g., comma separated values with its headers) and easy-to-understand with detailed variable descriptions. In this regard, the modeling system should be of interest to a broad group of stakeholders who are involved in various fields including vehicle emission modeling, transportation conformity analysis, near-road air quality impact assessment, and health-impact assessment.

7.4 Limitations and Future Research

There are some limitations and future research remained for further improving the model prediction and efficiency. First, this modeling demonstration focused on modeling PM_{2.5}, which is one of the NAAQS criteria air pollutants (i.e., ground-level ozone, particulate matter, carbon monoxide, lead, sulfur dioxide, and nitrogen dioxide), for the particular concerns about the increasing severity of PM_{2.5} concentrations and related public health problems in countries with significant air pollution issue, including China, India, and South

Korea, in the last several years. Governments and environmental and health authorities concerned about traffic-induced air pollution impacts will benefit from the research presented on $PM_{2.5}$. The same basic modeling framework and approaches presented in this paper should be applicable to other traffic-induced primary pollutants. To this end, future research is recommended to expand the modeling framework for other pollutants and to develop the appropriate modeling parameters for the various pollutions. Some of the caveats and discussions regarding expanding the modeling system for other pollutants are addressed throughout this section.

Second, future research should be to conduct model verification and calibration by comparing predictions with observational data collected from monitoring stations. As addressed in some previous research, the predicted concentrations tend to overestimate high concentrations and underestimate the lower range of concentrations. Because this work is designed to consider dynamic traffic operations and meteorological conditions in the analysis, the modeling framework is expected to reduce model prediction bias. Model verification against measured concentrations is an essential future activity to confirm the reduction in prediction bias.

To this end, future research also needs to be conducted to better understand the overall limitations of dispersion models. As addressed, dispersion models tend to overestimate the gradient of pollutant concentrations near roadway due to the technical formulations, embedded in dispersion models, that deal with mechanical mixing over the roadway (and near roadway within mixing zone). However, the modeling system presented in this dissertation does not address this overestimation in concentrations (i.e., the predicted concentrations are directly retrieved from the dispersion models without

verification/calibration). Additional attention needs to be paid to verifying/calibrating the predicted concentrations over the roads or near roads. In addition, future research needs to understand the characteristics of field-monitored concentrations for model verifications/calibrations. For example, PM_{2.5} concentrations measured from monitoring stations are attributed to: 1) primary PM_{2.5} emissions emitted from vehicle tailpipes, and 2) secondary PM_{2.5} formation from precursor emissions such as sulfur dioxide, nitrogen oxides, volatile organic compounds, and ammonia. In practice, the primary PM_{2.5} emissions can be modeled by dispersion models; however, most of the dispersion models do not account for the secondary PM_{2.5} emissions in the modeling. Understanding the potential gaps between modeled and monitored pollutant concentrations is essential for proper processes of verifying/calibrating the predicted concentrations modeled by dispersion models.

Another caveat in this research is that the modeling approach does not account for any issues associated with dispersion parameters. In other words, the modeling approach uses the parameters that are embedded in the dispersion models without verifying/calibrating the parameter values. That is, dispersion parameters still have their own limitations, and therefore future research needs a further improvement on dispersion parameters through a sort of model calibration process. For example, the modeling system is designed to use the meteorological processor data that are already developed and provided by state agencies (e.g., Georgia EPD's database). The data from state agencies tend to represent macroscopic meteorological patterns of a large-scale geographic area (e.g., Fulton County in Georgia). As such, the modeling system that applies the data from state agencies does not account for the microscale meteorological conditions of a specific project

area. This may result in biased model predictions for some areas. For example, Georgia EPD provides meteorological processor data for 15 sub-regions for the State of Georgia, and as such, each sub-region data tends to cover a large-scale area. Meanwhile, it is also possible that dynamic surface meteorological patterns are dependent upon geographic locations within the sub-region. For example, the Fulton County dataset is used to reflect the whole Fulton county area, but the area consists of a variety of land uses and surface characteristics (e.g., high-rise and paved areas in downtown Atlanta, medium building heights and medium-levels of pavement area in Midtown Atlanta, and residential area with low building heights and primarily vegetated areas in suburban Atlanta). The diverse surface characteristics may affect the local meteorological conditions, and thereby affect local pollutant concentrations. This calls for future research that focuses on developing a modeling framework that accounts for microscale meteorological data processing at a local scale. Similarly, the modeling system also does not account for area-specific terrain profiles. The modeling regime simply applies macroscopic terrain profiles (surface roughness length from the meteorological processor data provided by state agencies) that represent the terrain profile for a large-scale area, such as Fulton County. As such, future research may also need to further reduce model prediction bias by processing area-specific terrain profiles considering microscopic terrain characteristics (e.g., high-rise building area in downtown Atlanta). In the same manner, future research also needs to incorporate roadway-specific terrain profiles (e.g., depressed roadway, elevated roadways, ridgelines). To this end, the surface elevation profile data from the USGS DEM may help identify area-specific terrain profiles for use in dispersion modeling.

Third, future research may need to assess more efficient distributed computing strategies that account for the optimal link-receptor pairs for each model run. For modeling convenience, this work applied a brute-force approach where each model run considers a pair of multiple-links and a single-receptor. However, a better model performance comparison may be achieved by optimal job allocation strategies with a combination of multiple-links and multiple-receptors. In this way, the optimal performance can be achieved by reducing modeling spin-up time related to input file generation for each job, job allocation to cluster, and dispersion modeling spin-up and writing output files. In addition, the modeling system employed that assigns one core to one processor may actually have decreased the distributed computing efficiency because this strategy is not likely to use the maximum computing capacity of each processor. Thus, future research needs to consider optimal core-processor allocation strategy that maximizes the computing capacity of each processor, considering the distributed computing specifications (e.g., number of cores in each processor, memory, CPU, maximum number of processors and cores allowed for project).

Fourth, the dynamic grid-receptor model developed based on linear interpolations for pollutant concentration profiles need to be further evaluated by comparing the model results with those generated by non-linear smoothing methods such as Kriging. Although the linear interpolation method was very efficient in developing the dynamic grid-receptor model by reducing the computational resources and time, this approach may produce somewhat biased model predictions because of weak performances in smoothing profiles by the linear interpolation. More advanced non-linear interpolation methods would likely improve model prediction accuracy (at a computational time cost). To this end, future

research needs to consider developing optimal receptor placement models based on non-linear interpolation methods when computational resources are available. Future research should compare the concentration profiles produced by both of the linear and non-linear interpolation methods to identify the improvement in model predictions by the non-linear interpolation method over the linear interpolation method.

Fifth, future research regarding the supervised link screening model can help to assess the trade-off between estimation precision and computational cost with respect to the selection of the threshold parameter (cut-off threshold value for labeling significant versus insignificant links). As demonstrated in the dissertation work, model accuracy and run-time are highly sensitive to the threshold parameter. For example, a lower threshold value turned out to be very helpful in increasing the accuracy of predicted pollutant concentrations while it increases the model run-time. On the contrary, a higher threshold value decreases the model accuracy while reducing the model run-time. For modeling convenience, this dissertation only tested two threshold parameters of $0.1 \mu\text{g}/\text{m}^3$ and $0.01 \mu\text{g}/\text{m}^3$ for the case of $\text{PM}_{2.5}$ concentrations. However, an optimal threshold value can be identified by conducting sensitivity analysis on the threshold values. In addition, threshold values for other primary pollutants also need to be evaluated. To this end, future research may need to consider the reasonable searching ranges of threshold parameters, considering the relative difference of the predicted concentrations to NAAQS pollution levels. For example, the threshold parameter values for $\text{PM}_{2.5}$ were set between $0.1 \mu\text{g}/\text{m}^3$ and $0.01 \mu\text{g}/\text{m}^3$ in this study, considering that the NAAQS for $\text{PM}_{2.5}$ is $35 \mu\text{g}/\text{m}^3$ (24-hour average). Contributions lower than $0.1 \mu\text{g}/\text{m}^3$ have a marginal impact on the NAAQS $\text{PM}_{2.5}$ concentration level, but the threshold value of $0.1 \mu\text{g}/\text{m}^3$ helps to identify only a small set

of significant links. However, in the case of CO, higher threshold parameters may be considered in that the NAAQS for CO is 9 ppm (8-hour average) which is equivalent to 1,164.2 $\mu\text{g}/\text{m}^3$ (under the condition of 20°C and 1,013 millibars). In this case, 1 $\mu\text{g}/\text{m}^3$ to 10 $\mu\text{g}/\text{m}^3$ may be proper selections for link screening threshold parameters for CO.

Sixth, sensitivity analysis on modeling inputs should be considered in future research. This dissertation proposed conducting dispersion modeling by integrating high temporal and spatial resolution data to improve the model predictions; however, the work did not test the model predictions as compared to the low-resolution data. For example, the modeling system proposed predicting concentrations under dynamic traffic conditions, while the analysis on how much the model predictions were improved as compared to the concentrations predicted by the annual average traffic conditions. In the same manner, sensitivity analysis on pollutant concentrations with and without road grades or with simplified and hourly meteorological conditions need to be conducted. Because the model prediction and model run-time depending on the data resolution should be of interest for practitioners and modelers, the comparison analysis needs to be conducted to provide a better understanding of the model performance. To this end, several parameters that potentially affect predicted pollutant concentrations should be considered in future research, including:

- Link screening threshold parameters, δ
- Annual average traffic conditions vs. dynamic traffic conditions
- Simplified meteorological conditions vs. hourly meteorological conditions
- With road grades vs. without road grades

- Breaking roadway links into smaller sections

Lastly, the modeling system needs to be expanded for use in human exposure analysis. To this aim, the concentrations predicted by the developed modeling system can be linked to human activity data. For example, the predicted concentration results can be linked to American time use survey (ATUS) data that surveys on the amount of time that people spend doing various activities, such as paid work, childcare, volunteering, and socializing (U.S. Bureau of Labor Statistics, 2020), although the further data manipulation may be required because ATUS data does not include the activity location data. In this regard, U.S. census data where population data is tracked into small survey area can be considered. Using census data, the overall human exposure to traffic-related pollutant concentrations could be estimated for each census tract; however, data accounting for dynamic human activities still needs to be incorporated to assess the different level of human exposure by time of day. In particular, vehicle trajectory data (e.g., path retention data from TDM, observational data such as license plate matching data) can be used to evaluate the on-road human exposure.

7.5 Publication Outline

The research articles listed below have been conducted in part of this dissertation work.

Guensler, R., Liu, H., Xu, Y., Akanser, A., Kim, D., Hunter, M. P., and Rodgers, M.O. (2017) Energy Consumption and Emissions Modeling of Individual Vehicles. *Transportation Research Record*, 2627(1), 93-102. <https://doi.org/10.3141/2627-11>

Kim, D., Ko, J., Xu, X., Liu, H., Rodgers, M.O., and Guensler, R. (2019a). Evaluating the Environmental Benefits of Median Bus Lanes: A Microscopic Simulation Approach, *Transportation Research Record: Journal of the Transportation Research Board*, <https://doi.org/10.1177/0361198119836982>.

- Kim, D., Guin, A., Ko, J., Rodgers, M., and Guensler, R. (2019b) Energy and Air Quality Impacts of Truck-Only Lanes: A Case Study of Interstate 75 Between Macon and McDonough, Georgia. Presented at 98th Annual Meeting of the Transportation Research Board, Washington, D.C. <https://trid.trb.org/view/1573113>.
- Kim, D., Liu, H., Xu, X., Lu, H., Wayson, R., Rodgers, M.O., and Guensler, R. (2019c) A Regional Air Quality Impact Assessment Screening Tool based upon MOVES-Matrix and AERMOD. Presented at Guideline on Air Quality Models: Planning Ahead Conference, Air & Waste Management Association, March 12, Durham, North Carolina, USA.
- Kim, D., H. Liu, M.O. Rodgers, and R. Guensler. (2020a) Development of Roadway Link Screening Model for Regional-level Near-road Air Quality Analysis: A Case Study of Particulate Matter. Atmospheric Environment. Under Revision.
- Kim, D., H. Liu, X. Xu, H. Lu., R. Wayson, M.O. Rodgers, R. Guensler. (2020b) Streamlined Data Processing for Regional Scale Applications of Line Source Dispersion Modeling via Distributed Computing. Computer-aided Civil and Infrastructure Engineering. Under Revision.

APPENDIX A. LOGISTIC REGRESSION MODEL RESULTS FOR DYNAMIC GRID-RECEPTOR MODEL

The section summarized the logistic regression model results described in section 4.5.2. As indicated, a number of logistic regression models are developed for predicting the probability of each candidate receptor site being selected as the optimal receptor site. Table 14 particularly summarized the model coefficients and its p-values. As a total of 121 candidate receptor sites are considered, 121 logistic regression model results are summarized in Table 14, showing that each row shows the model coefficients for each model. For the reference receptor sites (set for the corners of the receptor area), the model results are non-applicable because those receptor sites are selected in all cases. For practical implementation, they are set to be included in the optimal receptor set in all cases.

Table 14 - Logistic Regression Model Results for Each Candidate Receptor Site

Model classification		Model coefficients (values in parenthesis indicates p-values)						
x (scaled) ¹	y	Number of receptors	Emission rate for focus link (link2)	Percentage difference in emission on the left-side $(\frac{ emi_{link2}-emi_{link1} }{emi_{link2}})$	Percentage difference in emission on the right-side $(\frac{ emi_{link2}-emi_{link3} }{emi_{link2}})$	Wind speed	Wind direction	Link length
-0.5	0	n.a. (n.a.)	n.a. (n.a.)	n.a. (n.a.)	n.a. (n.a.)	n.a. (n.a.)	n.a. (n.a.)	n.a. (n.a.)
-0.5	50	-0.00780 (0.208)	0.00000 (0.000)	-0.04260 (0.000)	-0.02520 (0.000)	-0.12080 (0.000)	0.03360 (0.000)	-0.00040 (0.000)
-0.5	100	0.03400 (0.000)	0.00020 (0.000)	0.03760 (0.002)	-0.04570 (0.000)	-0.04610 (0.000)	0.00280 (0.038)	0.00020 (0.062)
-0.5	150	0.02990 (0.000)	0.00010 (0.000)	-0.03040 (0.008)	0.02110 (0.004)	-0.02160 (0.000)	0.00200 (0.130)	-0.00040 (0.002)
-0.5	200	0.06700 (0.000)	0.00020 (0.000)	-0.13980 (0.000)	-0.00680 (0.270)	-0.04220 (0.000)	0.00430 (0.000)	-0.00080 (0.000)
-0.5	250	0.04660 (0.000)	0.00000 (0.000)	0.09690 (0.000)	-0.05360 (0.000)	0.00550 (0.186)	-0.00050 (0.590)	-0.00030 (0.000)
-0.5	300	0.08870 (0.000)	-0.00020 (0.000)	0.00390 (0.588)	-0.36440 (0.000)	0.00990 (0.009)	-0.00730 (0.000)	-0.00070 (0.000)
-0.5	350	0.05450 (0.000)	0.00010 (0.000)	-0.04750 (0.000)	0.00240 (0.477)	-0.05190 (0.000)	0.01020 (0.000)	-0.00090 (0.000)
-0.5	400	-0.01570 (0.000)	-0.00020 (0.000)	-0.05450 (0.000)	0.01810 (0.000)	-0.01760 (0.000)	0.01070 (0.000)	-0.00010 (0.169)
-0.5	450	0.05860 (0.000)	0.00000 (0.000)	-0.86180 (0.000)	-1.05250 (0.000)	-0.10690 (0.000)	0.00090 (0.001)	-0.00300 (0.000)
-0.5	500	n.a. (n.a.)	n.a. (n.a.)	n.a. (n.a.)	n.a. (n.a.)	n.a. (n.a.)	n.a. (n.a.)	n.a. (n.a.)
-0.4	0	0.14760 (0.000)	-0.00040 (0.000)	0.03480 (0.000)	0.02010 (0.001)	-0.00730 (0.153)	-0.02000 (0.000)	-0.00080 (0.000)

Model classification		Model coefficients (values in parenthesis indicates p-values)						
x (scaled) ¹	y	Number of receptors	Emission rate for focus link (link2)	Percentage difference in emission on the left-side $\left(\frac{ emi_{link2}-emi_{link1} }{emi_{link2}}\right)$	Percentage difference in emission on the right-side $\left(\frac{ emi_{link2}-emi_{link3} }{emi_{link2}}\right)$	Wind speed	Wind direction	Link length
-0.4	50	0.13060 (0.000)	-0.00010 (0.000)	0.05930 (0.000)	-0.02010 (0.000)	-0.01050 (0.007)	-0.01500 (0.000)	-0.00010 (0.298)
-0.4	100	0.08830 (0.000)	-0.00010 (0.000)	0.21090 (0.000)	-0.31680 (0.000)	-0.00650 (0.033)	-0.03320 (0.000)	0.00090 (0.000)
-0.4	150	0.07280 (0.000)	-0.00020 (0.000)	-0.28490 (0.000)	0.05180 (0.000)	-0.02680 (0.000)	-0.02560 (0.000)	-0.00010 (0.014)
-0.4	200	-0.03150 (0.000)	-0.00020 (0.000)	-0.02810 (0.000)	0.03710 (0.000)	-0.02090 (0.000)	-0.03400 (0.000)	0.00110 (0.000)
-0.4	250	0.06520 (0.000)	-0.00020 (0.000)	0.03800 (0.000)	-0.01540 (0.000)	-0.02060 (0.000)	-0.00900 (0.000)	-0.00110 (0.000)
-0.4	300	-0.08570 (0.000)	0.00000 (0.000)	-0.02330 (0.000)	-0.04350 (0.000)	-0.14270 (0.000)	-0.01600 (0.000)	0.00140 (0.000)
-0.4	350	-0.18180 (0.000)	-0.00020 (0.000)	-0.00400 (0.122)	-0.04300 (0.000)	-0.16740 (0.000)	-0.01460 (0.000)	0.00310 (0.000)
-0.4	400	0.06600 (0.000)	-0.00020 (0.000)	-0.70950 (0.000)	-0.53060 (0.000)	-0.09360 (0.000)	-0.02630 (0.000)	0.00110 (0.000)
-0.4	450	-0.05490 (0.000)	-0.00010 (0.000)	-0.21550 (0.000)	-0.34470 (0.000)	-0.04330 (0.000)	-0.03610 (0.000)	-0.00090 (0.000)
-0.4	500	0.01600 (0.000)	-0.00010 (0.000)	-0.54590 (0.000)	-0.13400 (0.000)	-0.11560 (0.000)	-0.02470 (0.000)	0.00170 (0.000)
-0.3	0	0.18030 (0.000)	-0.00020 (0.000)	0.26420 (0.000)	-0.27050 (0.000)	-0.05650 (0.000)	-0.01110 (0.000)	-0.00460 (0.000)
-0.3	50	0.11810 (0.000)	-0.00030 (0.000)	0.00140 (0.790)	-0.02800 (0.000)	-0.01410 (0.000)	-0.01020 (0.000)	-0.00120 (0.000)
-0.3	100	0.14490 (0.000)	-0.00010 (0.000)	-0.02040 (0.000)	0.01570 (0.000)	-0.10040 (0.000)	-0.02090 (0.000)	-0.00130 (0.000)
-0.3	150	0.08200 (0.000)	-0.00030 (0.000)	-0.11990 (0.000)	0.01320 (0.000)	-0.06180 (0.000)	-0.01140 (0.000)	-0.00120 (0.000)

Model classification		Model coefficients (values in parenthesis indicates p-values)						
x (scaled) ¹	y	Number of receptors	Emission rate for focus link (link2)	Percentage difference in emission on the left-side $\left(\frac{ emi_{link2}-emi_{link1} }{emi_{link2}}\right)$	Percentage difference in emission on the right-side $\left(\frac{ emi_{link2}-emi_{link3} }{emi_{link2}}\right)$	Wind speed	Wind direction	Link length
-0.3	200	0.02650 (0.000)	-0.00030 (0.893)	0.11280 (0.000)	-0.09820 (0.000)	-0.02470 (0.000)	-0.00860 (0.000)	-0.00140 (0.000)
-0.3	250	0.04990 (0.000)	-0.00010 (0.000)	-0.86710 (0.000)	0.04880 (0.000)	-0.09150 (0.000)	-0.02590 (0.000)	-0.00120 (0.000)
-0.3	300	0.04510 (0.000)	-0.00010 (0.000)	-0.02490 (0.000)	-0.14500 (0.000)	-0.02760 (0.000)	-0.01160 (0.000)	-0.00050 (0.000)
-0.3	350	0.08230 (0.000)	-0.00030 (0.000)	-0.28000 (0.000)	-0.09890 (0.000)	-0.11140 (0.000)	-0.00950 (0.000)	-0.00100 (0.000)
-0.3	400	0.09470 (0.000)	-0.00010 (0.000)	-0.69110 (0.000)	0.06450 (0.000)	-0.09920 (0.000)	-0.01010 (0.000)	-0.00150 (0.000)
-0.3	450	-0.09700 (0.000)	-0.00010 (0.000)	-0.27470 (0.000)	-0.32890 (0.000)	-0.23140 (0.000)	-0.00900 (0.000)	0.00210 (0.000)
-0.3	500	0.02030 (0.000)	0.00000 (0.000)	-0.35600 (0.000)	-0.80990 (0.000)	-0.10610 (0.000)	-0.02470 (0.000)	0.00100 (0.000)
-0.2	0	0.05670 (0.000)	-0.00030 (0.000)	0.00910 (0.022)	-0.00780 (0.002)	-0.01680 (0.000)	0.01250 (0.000)	-0.00410 (0.000)
-0.2	50	0.16900 (0.000)	-0.00030 (0.000)	-0.02630 (0.000)	-0.02060 (0.000)	-0.01860 (0.000)	-0.00170 (0.001)	-0.00330 (0.000)
-0.2	100	0.07060 (0.000)	-0.00040 (0.000)	0.09880 (0.000)	-0.10070 (0.000)	0.04050 (0.000)	0.00400 (0.000)	-0.00240 (0.000)
-0.2	150	0.12930 (0.000)	-0.00020 (0.000)	0.07540 (0.000)	-0.06920 (0.000)	0.01370 (0.000)	-0.00850 (0.000)	-0.00320 (0.000)
-0.2	200	0.05210 (0.000)	-0.00020 (0.000)	0.19040 (0.000)	-0.22070 (0.000)	-0.00630 (0.000)	-0.00790 (0.000)	-0.00220 (0.000)
-0.2	250	0.03110 (0.000)	-0.00020 (0.000)	-0.05890 (0.000)	-0.00440 (0.002)	-0.05930 (0.000)	-0.02150 (0.000)	-0.00240 (0.000)
-0.2	300	0.09240 (0.000)	-0.00010 (0.000)	-0.73140 (0.000)	-0.19680 (0.000)	-0.09980 (0.000)	-0.03780 (0.000)	-0.00190 (0.000)

Model classification		Model coefficients (values in parenthesis indicates p-values)						
x (scaled) ¹	y	Number of receptors	Emission rate for focus link (link2)	Percentage difference in emission on the left-side $\left(\frac{ emi_{link2} - emi_{link1} }{emi_{link2}}\right)$	Percentage difference in emission on the right-side $\left(\frac{ emi_{link2} - emi_{link3} }{emi_{link2}}\right)$	Wind speed	Wind direction	Link length
-0.2	350	0.11140 (0.000)	-0.00010 (0.000)	-0.37810 (0.000)	-0.68440 (0.000)	-0.12500 (0.000)	-0.03820 (0.000)	-0.00260 (0.000)
-0.2	400	-0.09210 (0.000)	0.00000 (0.000)	0.12760 (0.000)	-0.64860 (0.000)	-0.19330 (0.000)	-0.00930 (0.000)	-0.00330 (0.000)
-0.2	450	-0.02390 (0.000)	-0.00010 (0.000)	0.00110 (0.568)	0.00810 (0.000)	-0.05320 (0.000)	-0.02440 (0.000)	-0.00260 (0.000)
-0.2	500	0.10560 (0.000)	-0.00030 (0.000)	-0.31680 (0.000)	0.02370 (0.000)	-0.06500 (0.000)	-0.03120 (0.000)	-0.00030 (0.000)
-0.1	0	0.15480 (0.000)	-0.00040 (0.000)	0.06760 (0.000)	-0.06710 (0.000)	-0.15190 (0.000)	0.00860 (0.000)	-0.00500 (0.000)
-0.1	50	0.19500 (0.000)	-0.00040 (0.000)	0.14720 (0.000)	-0.18600 (0.000)	-0.03160 (0.000)	0.00290 (0.000)	-0.00420 (0.000)
-0.1	100	0.17490 (0.000)	-0.00030 (0.000)	-0.00750 (0.097)	0.01760 (0.000)	-0.02320 (0.000)	-0.01280 (0.000)	-0.00370 (0.000)
-0.1	150	0.17720 (0.000)	-0.00010 (0.000)	0.01360 (0.000)	-0.24520 (0.000)	-0.05620 (0.000)	-0.01990 (0.000)	-0.00250 (0.000)
-0.1	200	0.08100 (0.000)	0.00000 (0.000)	-0.02690 (0.000)	0.00630 (0.000)	-0.03370 (0.000)	-0.02510 (0.000)	-0.00320 (0.000)
-0.1	250	0.13190 (0.000)	-0.00010 (0.000)	-0.04030 (0.000)	-0.03270 (0.000)	-0.06300 (0.000)	-0.01910 (0.000)	-0.00560 (0.000)
-0.1	300	0.05880 (0.000)	-0.00010 (0.000)	-0.12420 (0.000)	-0.10830 (0.000)	-0.02720 (0.000)	-0.02540 (0.000)	-0.00160 (0.000)
-0.1	350	0.12290 (0.000)	-0.00010 (0.055)	0.03040 (0.000)	-0.24850 (0.000)	-0.07890 (0.000)	-0.04440 (0.000)	-0.00370 (0.000)
-0.1	400	0.02650 (0.000)	-0.00020 (0.000)	0.16510 (0.000)	-0.09920 (0.000)	-0.18480 (0.000)	-0.02490 (0.000)	-0.00140 (0.000)
-0.1	450	-0.01630 (0.000)	0.00010 (0.000)	0.14660 (0.000)	-0.10840 (0.000)	-0.10940 (0.000)	-0.03680 (0.000)	-0.00090 (0.000)

Model classification		Model coefficients (values in parenthesis indicates p-values)						
x (scaled) ¹	y	Number of receptors	Emission rate for focus link (link2)	Percentage difference in emission on the left-side $\left(\frac{ emi_{link2} - emi_{link1} }{emi_{link2}}\right)$	Percentage difference in emission on the right-side $\left(\frac{ emi_{link2} - emi_{link3} }{emi_{link2}}\right)$	Wind speed	Wind direction	Link length
-0.1	500	-0.00060 (0.777)	-0.00020 (0.000)	0.02280 (0.000)	0.03380 (0.000)	-0.08140 (0.000)	-0.02290 (0.000)	0.00020 (0.000)
0.0	0	-0.07820 (0.000)	-0.00020 (0.000)	0.08050 (0.000)	-0.01680 (0.000)	-0.07810 (0.000)	0.00760 (0.000)	-0.00480 (0.000)
0.0	50	0.15850 (0.000)	-0.00040 (0.000)	-0.02160 (0.000)	0.00900 (0.000)	-0.12970 (0.000)	0.00260 (0.000)	-0.00260 (0.000)
0.0	100	0.14120 (0.000)	-0.00020 (0.000)	-0.18480 (0.000)	0.02370 (0.000)	-0.16360 (0.000)	-0.01820 (0.000)	-0.00390 (0.000)
0.0	150	0.18940 (0.000)	-0.00030 (0.000)	-0.19930 (0.000)	0.01700 (0.000)	-0.00800 (0.000)	-0.01140 (0.000)	-0.00300 (0.000)
0.0	200	0.08450 (0.000)	-0.00030 (0.000)	-0.22570 (0.000)	-0.02780 (0.000)	-0.03480 (0.000)	-0.02270 (0.000)	-0.00210 (0.000)
0.0	250	0.10790 (0.000)	-0.00020 (0.000)	-0.25260 (0.000)	0.01620 (0.000)	-0.05100 (0.000)	-0.01720 (0.000)	-0.00400 (0.000)
0.0	300	0.12700 (0.000)	0.00000 (0.000)	0.10190 (0.000)	-0.05360 (0.000)	-0.06730 (0.000)	-0.02450 (0.000)	-0.00390 (0.000)
0.0	350	0.18080 (0.000)	-0.00010 (0.174)	-0.69960 (0.000)	-0.08350 (0.000)	-0.38840 (0.000)	-0.05940 (0.000)	-0.00380 (0.000)
0.0	400	-0.04810 (0.000)	-0.00010 (0.672)	-0.18930 (0.000)	0.02150 (0.000)	-0.06850 (0.000)	-0.04150 (0.000)	-0.00120 (0.000)
0.0	450	-0.19410 (0.000)	-0.00010 (0.000)	-0.09720 (0.000)	0.02120 (0.000)	-0.08440 (0.000)	-0.01960 (0.000)	-0.00220 (0.000)
0.0	500	0.11260 (0.000)	-0.00020 (0.723)	0.04530 (0.000)	-0.01500 (0.000)	-0.07750 (0.000)	-0.03450 (0.000)	-0.00130 (0.000)
0.1	0	0.06880 (0.000)	-0.00030 (0.000)	0.06980 (0.000)	0.00540 (0.004)	-0.09290 (0.000)	0.00010 (0.752)	-0.00350 (0.000)
0.1	50	0.16670 (0.000)	-0.00030 (0.000)	-0.02810 (0.000)	0.01990 (0.000)	-0.06370 (0.000)	-0.01330 (0.000)	-0.00240 (0.000)

Model classification		Model coefficients (values in parenthesis indicates p-values)						
x (scaled) ¹	y	Number of receptors	Emission rate for focus link (link2)	Percentage difference in emission on the left-side $\left(\frac{ emi_{link2}-emi_{link1} }{emi_{link2}}\right)$	Percentage difference in emission on the right-side $\left(\frac{ emi_{link2}-emi_{link3} }{emi_{link2}}\right)$	Wind speed	Wind direction	Link length
0.1	100	0.16980 (0.000)	-0.00020 (0.000)	-0.01960 (0.000)	0.00540 (0.016)	-0.08590 (0.000)	-0.01630 (0.000)	-0.00150 (0.000)
0.1	150	0.09030 (0.000)	-0.00030 (0.000)	0.18470 (0.000)	-0.24520 (0.000)	-0.06870 (0.000)	-0.01900 (0.000)	-0.00210 (0.000)
0.1	200	0.15280 (0.000)	-0.00030 (0.000)	-0.15390 (0.000)	-0.12420 (0.000)	-0.09750 (0.000)	-0.01400 (0.000)	-0.00370 (0.000)
0.1	250	0.12900 (0.000)	-0.00010 (0.000)	0.16230 (0.000)	-0.09950 (0.000)	-0.19640 (0.000)	-0.02970 (0.000)	-0.00170 (0.000)
0.1	300	0.10090 (0.000)	-0.00010 (0.000)	-0.66970 (0.000)	0.03160 (0.000)	-0.13830 (0.000)	-0.03310 (0.000)	-0.00230 (0.000)
0.1	350	0.04130 (0.000)	-0.00010 (0.000)	-0.60000 (0.000)	0.02700 (0.000)	-0.15850 (0.000)	-0.02880 (0.000)	-0.00230 (0.000)
0.1	400	0.00540 (0.001)	-0.00030 (0.000)	-0.31980 (0.000)	0.02550 (0.000)	-0.08900 (0.000)	-0.01000 (0.000)	-0.00100 (0.000)
0.1	450	-0.02640 (0.000)	-0.00010 (0.000)	0.22370 (0.000)	-0.12890 (0.000)	-0.04970 (0.000)	-0.07170 (0.000)	-0.00170 (0.000)
0.1	500	0.13110 (0.000)	-0.00010 (0.000)	-0.01060 (0.009)	0.04430 (0.000)	-0.02950 (0.000)	-0.04820 (0.000)	-0.00120 (0.000)
0.2	0	0.11730 (0.000)	-0.00030 (0.000)	-0.18530 (0.000)	0.16060 (0.000)	-0.18480 (0.000)	0.00550 (0.000)	-0.00300 (0.000)
0.2	50	0.15840 (0.000)	-0.00040 (0.000)	-0.02700 (0.000)	0.03090 (0.000)	-0.07630 (0.000)	-0.01000 (0.000)	-0.00120 (0.000)
0.2	100	0.10240 (0.000)	-0.00020 (0.000)	-0.04360 (0.000)	0.00370 (0.133)	-0.00630 (0.003)	-0.00610 (0.000)	-0.00230 (0.000)
0.2	150	0.06900 (0.000)	-0.00010 (0.000)	-0.85180 (0.000)	0.00570 (0.001)	-0.03570 (0.000)	-0.01160 (0.000)	-0.00160 (0.000)
0.2	200	0.07580 (0.000)	-0.00010 (0.000)	-0.16460 (0.000)	-0.38360 (0.000)	-0.01160 (0.000)	-0.02800 (0.000)	-0.00180 (0.000)

Model classification		Model coefficients (values in parenthesis indicates p-values)						
x (scaled) ¹	y	Number of receptors	Emission rate for focus link (link2)	Percentage difference in emission on the left-side $\left(\frac{ emi_{link2} - emi_{link1} }{emi_{link2}}\right)$	Percentage difference in emission on the right-side $\left(\frac{ emi_{link2} - emi_{link3} }{emi_{link2}}\right)$	Wind speed	Wind direction	Link length
0.2	250	-0.17700 (0.000)	0.00000 (0.000)	0.02640 (0.000)	0.02570 (0.000)	-0.05000 (0.000)	-0.03630 (0.000)	-0.00070 (0.000)
0.2	300	0.04680 (0.000)	-0.00010 (0.000)	-0.07690 (0.000)	-0.06060 (0.000)	-0.10770 (0.000)	-0.02240 (0.000)	-0.00060 (0.000)
0.2	350	0.13790 (0.000)	0.00000 (0.000)	-0.99550 (0.000)	-0.50340 (0.000)	-0.12720 (0.000)	-0.02640 (0.000)	-0.00330 (0.000)
0.2	400	-0.06290 (0.000)	-0.00010 (0.000)	-0.00410 (0.063)	0.04610 (0.000)	-0.03330 (0.000)	-0.02360 (0.000)	0.00000 (0.066)
0.2	450	-0.01660 (0.000)	-0.00010 (0.000)	-0.79690 (0.000)	0.04590 (0.000)	-0.15840 (0.000)	-0.03200 (0.000)	-0.00080 (0.000)
0.2	500	0.06170 (0.000)	-0.00010 (0.000)	-0.01350 (0.001)	0.00160 (0.521)	-0.01170 (0.000)	-0.02420 (0.000)	0.00110 (0.000)
0.3	0	0.14840 (0.000)	-0.00040 (0.000)	0.07220 (0.000)	-0.05760 (0.000)	-0.08260 (0.000)	-0.00470 (0.000)	-0.00230 (0.000)
0.3	50	0.16450 (0.000)	-0.00020 (0.000)	-0.04390 (0.000)	0.01590 (0.000)	-0.02150 (0.000)	-0.00490 (0.000)	-0.00120 (0.000)
0.3	100	0.08640 (0.000)	-0.00030 (0.000)	-0.02790 (0.000)	0.03190 (0.000)	-0.08550 (0.000)	-0.01190 (0.000)	-0.00090 (0.000)
0.3	150	0.08470 (0.000)	-0.00020 (0.001)	0.24770 (0.000)	-0.20340 (0.000)	-0.04550 (0.000)	-0.01420 (0.000)	-0.00130 (0.000)
0.3	200	0.04030 (0.000)	-0.00010 (0.000)	-0.01980 (0.000)	0.03370 (0.000)	-0.11500 (0.000)	-0.02350 (0.000)	-0.00100 (0.000)
0.3	250	0.07580 (0.000)	-0.00020 (0.000)	-0.33860 (0.000)	-0.05090 (0.000)	-0.03800 (0.000)	-0.01610 (0.000)	-0.00070 (0.000)
0.3	300	-0.04100 (0.000)	-0.00020 (0.000)	-0.55890 (0.000)	0.02090 (0.000)	0.00530 (0.000)	-0.01730 (0.000)	-0.00080 (0.000)
0.3	350	0.07090 (0.000)	0.00000 (0.000)	-0.04050 (0.000)	0.05710 (0.000)	-0.18130 (0.000)	-0.03520 (0.000)	-0.00140 (0.000)

Model classification		Model coefficients (values in parenthesis indicates p-values)						
x (scaled) ¹	y	Number of receptors	Emission rate for focus link (link2)	Percentage difference in emission on the left-side $\left(\frac{ emi_{link2}-emi_{link1} }{emi_{link2}}\right)$	Percentage difference in emission on the right-side $\left(\frac{ emi_{link2}-emi_{link3} }{emi_{link2}}\right)$	Wind speed	Wind direction	Link length
0.3	400	0.08500 (0.000)	-0.00020 (0.000)	-0.64500 (0.000)	-0.21940 (0.000)	-0.16320 (0.000)	-0.01010 (0.000)	-0.00140 (0.000)
0.3	450	0.09090 (0.000)	-0.00010 (0.000)	0.05940 (0.000)	-0.04980 (0.000)	-0.03310 (0.000)	-0.02140 (0.000)	-0.00080 (0.000)
0.3	500	0.02760 (0.000)	-0.00020 (0.000)	-0.02160 (0.000)	0.01870 (0.000)	-0.01330 (0.000)	-0.02390 (0.000)	0.00110 (0.000)
0.4	0	0.18810 (0.000)	-0.00050 (0.000)	-0.01500 (0.116)	0.03430 (0.000)	0.00260 (0.603)	-0.02830 (0.000)	0.00050 (0.000)
0.4	50	0.11800 (0.000)	-0.00010 (0.001)	-0.06360 (0.000)	0.01450 (0.004)	-0.01930 (0.000)	-0.02590 (0.000)	0.00050 (0.000)
0.4	100	0.10630 (0.000)	-0.00010 (0.000)	-0.25960 (0.000)	0.03120 (0.000)	-0.00090 (0.770)	-0.02420 (0.000)	0.00040 (0.000)
0.4	150	0.08050 (0.000)	-0.00020 (0.000)	-0.45730 (0.000)	0.05390 (0.000)	-0.03160 (0.000)	-0.03400 (0.000)	0.00090 (0.000)
0.4	200	0.11370 (0.000)	-0.00010 (0.000)	-0.57120 (0.000)	-0.21250 (0.000)	-0.04800 (0.000)	-0.01680 (0.000)	-0.00030 (0.000)
0.4	250	0.03330 (0.000)	-0.00030 (0.000)	-0.23770 (0.000)	-0.21510 (0.000)	-0.02720 (0.000)	-0.00780 (0.000)	-0.00080 (0.000)
0.4	300	0.07860 (0.000)	-0.00010 (0.000)	-0.03260 (0.000)	-0.01460 (0.000)	-0.09170 (0.000)	-0.01980 (0.000)	-0.00110 (0.000)
0.4	350	-0.05410 (0.000)	-0.00010 (0.000)	-0.05090 (0.000)	0.03770 (0.000)	-0.13770 (0.000)	-0.01020 (0.000)	0.00150 (0.000)
0.4	400	0.07170 (0.000)	-0.00010 (0.000)	-0.41600 (0.000)	-1.22860 (0.000)	-0.05600 (0.000)	-0.01720 (0.000)	-0.00280 (0.000)
0.4	450	0.06760 (0.000)	-0.00010 (0.000)	-0.39930 (0.000)	-1.10530 (0.000)	-0.12020 (0.000)	-0.01770 (0.000)	-0.00150 (0.000)
0.4	500	-0.04570 (0.000)	-0.00010 (0.000)	0.02860 (0.000)	-0.01360 (0.000)	-0.07170 (0.000)	-0.02060 (0.000)	0.00100 (0.000)

Model classification		Model coefficients (values in parenthesis indicates p-values)						
x (scaled) ¹	y	Number of receptors	Emission rate for focus link (link2)	Percentage difference in emission on the left-side $\left(\frac{ emi_{link2} - emi_{link1} }{emi_{link2}}\right)$	Percentage difference in emission on the right-side $\left(\frac{ emi_{link2} - emi_{link3} }{emi_{link2}}\right)$	Wind speed	Wind direction	Link length
0.5	0	n.a. (n.a.)	n.a. (n.a.)	n.a. (n.a.)	n.a. (n.a.)	n.a. (n.a.)	n.a. (n.a.)	n.a. (n.a.)
0.5	50	0.02560 (0.000)	0.00010 (0.000)	0.01540 (0.123)	-0.02790 (0.000)	-0.12530 (0.000)	0.01870 (0.000)	-0.00060 (0.000)
0.5	100	-0.01800 (0.031)	0.00040 (0.000)	-0.09270 (0.000)	0.06600 (0.000)	-0.04600 (0.000)	0.01460 (0.000)	0.00020 (0.097)
0.5	150	0.03470 (0.000)	-0.00010 (0.000)	0.03020 (0.008)	-0.07860 (0.000)	-0.01950 (0.001)	0.00760 (0.000)	0.00000 (0.801)
0.5	200	0.02830 (0.000)	0.00010 (0.000)	0.04070 (0.000)	-0.03340 (0.000)	-0.01710 (0.000)	-0.00030 (0.780)	-0.00080 (0.000)
0.5	250	0.02870 (0.000)	0.00000 (0.000)	0.06830 (0.000)	-0.06850 (0.000)	-0.06450 (0.000)	-0.00240 (0.012)	-0.00030 (0.004)
0.5	300	0.02900 (0.000)	0.00010 (0.000)	0.01870 (0.001)	-0.07700 (0.000)	-0.07840 (0.000)	0.00060 (0.395)	-0.00010 (0.025)
0.5	350	0.02240 (0.000)	-0.00010 (0.000)	0.09060 (0.000)	-0.11850 (0.000)	0.01190 (0.000)	0.01500 (0.000)	-0.00150 (0.000)
0.5	400	0.04420 (0.000)	-0.00020 (0.000)	-0.16620 (0.000)	-0.04380 (0.000)	-0.03230 (0.000)	0.00430 (0.000)	0.00020 (0.007)
0.5	450	-0.19580 (0.000)	-0.00010 (0.339)	0.26000 (0.000)	-0.19640 (0.000)	0.07780 (0.000)	-0.02420 (0.000)	-0.00080 (0.000)
0.5	500	n.a. (n.a.)	n.a. (n.a.)	n.a. (n.a.)	n.a. (n.a.)	n.a. (n.a.)	n.a. (n.a.)	n.a. (n.a.)

¹ The column values indicate the X positions in scale. -0.5 indicates the left-most position of Link 2, 0.0 indicates the middle, and 0.5 indicates the right-most position.

APPENDIX B. THE DEPLOYMENT STRATEGY OF MODELING SYSTEM ON AMAZON WEB SERVICES

This section suggests a strategy to deploy the developed modeling system on Amazon Web Services (AWS), and provides the expected deployment cost. AWS provides a variety of cloud computing products for diverse user demands, including products specialized for computing, storage, database, internet service, blockchain, etc. In addition, the AWS pricing varies depending on the type of products and the specification of the desired cloud computing system.³ Therefore, the proper selection of AWS product and pricing plan is important to provide affordable modeling system through AWS to the users.

Based on the review of AWS product website (<https://aws.amazon.com/>), this study suggested some suitable options among various AWS products and pricing plans that may match with the developed modeling system. Basically, this study assumed that the similar modeling structure addressed in this dissertation will be deployed on AWS. That is, each job consisting of multi-links and a single receptor will be assigned to each processor and multiple jobs are processed simultaneously, and as such, this strategy inevitably requires many AWS processors and cores to be deployed on AWS. In addition, the AWS modeling system will be used per client's request to reduce the total AWS ownership cost. (This option is best for occasional use.) Considering these assumptions, the study suggested the following AWS product and pricing plan.

- Amazon EC2 product (one of the common types of AWS cloud computing system)

³ AWS cost estimation was based on <https://aws.amazon.com/ec2/pricing/on-demand/>.

- On-Demand service (pay for compute capacity by the hour with no long-term commitments)
- “Compute optimized” and Linux based operating systems

Table 15 summarized the cost of AWS service usages that adopts the above AWS service type. As shown, Amazon EC2 On-Demand service provides a variety of system specifications with increasing unit prices by CPU, EC2 compute unit (ECU), memory, and storage. In this case, because the developed modeling system particularly require more compute units for multiple dispersion modeling processes, some powerful Amazon EC2 classes such as “c5d.metal” may be preferred. Note that “c5d.metal” class provides up to 375 ECU (AWS compute unit: one ECU is similar in power to a 1.0 - 1.2 GHz 2007 Intel Xeon or AMD Opteron processor).

Considering these desired AWS specifications, the study estimated the expected cost of dispersion modeling for the metro Atlanta area through AWS. As mentioned in chapter 6.8, the standard dispersion modeling time for metro Atlanta took around 10 days (= 240 hours) with 500 PACE processors. The Amazon EC2 c5d.metal provides up to 375 processors, thus the dispersion modeling for metro Atlanta is expected to take 13.3 days (10 days \times 500 processors / 375 processors, which is 320 hours) through AWS. Thus, the total cost of dispersion modeling for metro Atlanta is expected to be \$1,475 (320 hours \times \$4.608 per Hour). Because the modeling cost can be scalable by the size of network (as mentioned, the dispersion modeling time is related to the number of links and receptors), the estimated cost in this study may help agencies to estimate their expected cost of their analysis.

Table 15 - Amazon EC2 On-Demand Pricing as of April, 2020: Compute Optimized & Linux/Unix Services

Class	Virtual CPU (vCPU)	ECU ¹	Memory (GiB)	Instance Storage (GB)	Linux/UNIX Usage
c5.large	2	10	4	EBS Only	\$0.085 per Hour
c5d.large	2	10	4	1 x 50 NVMe SSD	\$0.096 per Hour
c4.large	2	8	3.75	EBS Only	\$0.10 per Hour
c5n.large	2	10	5.25	EBS Only	\$0.108 per Hour
c5.xlarge	4	20	8	EBS Only	\$0.17 per Hour
c5d.xlarge	4	20	8	1 x 100 NVMe SSD	\$0.192 per Hour
c4.xlarge	4	16	7.5	EBS Only	\$0.199 per Hour
c5n.xlarge	4	20	10.5	EBS Only	\$0.216 per Hour
c5.2xlarge	8	39	16	EBS Only	\$0.34 per Hour
c5d.2xlarge	8	39	16	1 x 200 NVMe SSD	\$0.384 per Hour
c4.2xlarge	8	31	15	EBS Only	\$0.398 per Hour
c5n.2xlarge	8	39	21	EBS Only	\$0.432 per Hour
c5.4xlarge	16	73	32	EBS Only	\$0.68 per Hour
c5d.4xlarge	16	73	32	1 x 400 NVMe SSD	\$0.768 per Hour
c4.4xlarge	16	62	30	EBS Only	\$0.796 per Hour
c5n.4xlarge	16	73	42	EBS Only	\$0.864 per Hour
c5.9xlarge	36	139	72	EBS Only	\$1.53 per Hour
c4.8xlarge	36	132	60	EBS Only	\$1.591 per Hour
c5d.9xlarge	36	139	72	1 x 900 NVMe SSD	\$1.728 per Hour
c5n.9xlarge	36	139	96	EBS Only	\$1.944 per Hour
c5.12xlarge	48	188	96	EBS Only	\$2.04 per Hour
c5d.12xlarge	48	188	96	2 x 900 NVMe SSD	\$2.304 per Hour
c5.18xlarge	72	281	144	EBS Only	\$3.06 per Hour
c5d.18xlarge	72	281	144	2 x 900 NVMe SSD	\$3.456 per Hour
c5n.18xlarge	72	281	192	EBS Only	\$3.888 per Hour
c5n.metal	72	N/A	192	EBS Only	\$3.888 per Hour
c5.24xlarge	96	375	192	EBS Only	\$4.08 per Hour
c5.metal	96	375	192	EBS Only	\$4.08 per Hour
c5d.24xlarge	96	375	192	4 x 900 NVMe SSD	\$4.608 per Hour
c5d.metal	96	375	192	4 x 900 NVMe SSD	\$4.608 per Hour

REFERENCES

- Akpinar, S., Oztop, H. F., and Akpinar, E. K. (2008) Evaluation of Relationship between Meteorological Parameters and Air Pollutant Concentrations during Winter Season in Elazığ, Turkey. *Environmental Monitoring and Assessment*, 146(1-3), 211-224. <https://doi.org/10.1007/s10661-007-0073-9>.
- ARC, Atlanta Regional Commission. (2017) Activity Based Modeling. <https://atlantaregional.org/transportation-mobility/modeling/modeling/>.
- Archer, K., and Kimes, R. (2008) Empirical Characterization of Random Forest Variable Importance Measures. *Computational Statistics & Data Analysis*, 52(4), 2249-2260. <https://doi.org/10.1016/j.csda.2007.08.015>.
- Aziz, H. A., and Ukkusuri, S. V. (2012) Integration of Environmental Objectives in A System Optimal Dynamic Traffic Assignment Model. *Computer-Aided Civil and Infrastructure Engineering*. 27(7), 494-511. <https://doi.org/10.1111/j.1467-8667.2012.00756.x>.
- Bachman, W., Granell, J., Guensler, R., and Leonard, J. (1998) Research Needs in Determining Spatially Resolved Sub-fleet Characteristics. *Transportation Research Record*. 1625. 139-146. Transportation Research Board. Washington, DC. 1998.
- Bachman, W., Sarasua, W., Hallmark, S., and Guensler, R. (2000) Modeling Regional Mobile Source Emissions in a Geographic Information System Framework. *Transportation Research Part C: Emerging Technologies*, 8(1-6), 205-229. [https://doi.org/10.1016/S0968-090X\(00\)00005-X](https://doi.org/10.1016/S0968-090X(00)00005-X).
- Barad, M. L. (1958). Project Prairie Grass, A Field Program in Diffusion. Volume 1 (No. Grp-59-Vol-1). Air Force Cambridge Research Labs Hanscom AFB MA.
- Barzyk, T. M., George, B. J., Vette, A. F. Williams, R. W., Croghan, C. W., and Stevens, C. D. (2009) Development of a Distance-to-roadway Proximity Metric to Compare Near-road Pollutant Levels to a Central Site Monitor. *Atmospheric Environment*, 43(4), 787-797. <https://doi.org/10.1016/j.atmosenv.2008.11.002>.

- Batterman, S., Chambliss, S. and Isakov, V. (2014) Spatial Resolution Requirements for Traffic-related Air Pollutant Exposure Evaluations. *Atmospheric Environment*, 94, 518-528. <https://doi.org/10.1016/j.atmosenv.2014.05.065>.
- Belgiu, M., and Drăguț, L. (2016) Random Forest in Remote Sensing: A Review of Applications and Future Directions. *ISPRS Journal of Photogrammetry and Remote Sensing*, 114, 24–31. <https://doi.org/10.1016/j.isprsjprs.2016.01.011>.
- Bellander, T., Berglind, N., Gustavsson, P., Jonson, T., Nyberg, F., Pershagen, G., and Järup, L. (2001) Using Geographic Information Systems to Assess Individual Historical Exposure to Air Pollution from Traffic and House Heating in Stockholm. *Environmental Health Perspectives*, 109(6), 633-639. <https://doi.org/10.1289/ehp.01109633>.
- Benson, P.E. (1979) CALINE-3. A Versatile Dispersion Model for Predicting Air Pollution Levels Near Highways and Arterial Streets FHWA/CA/TL-79/23, California Department of Transportation, Sacramento, CA
- Benson, P. E. (1982). Modifications to The Gaussian Vertical Dispersion Parameter, Σ_z , Near Roadways. *Atmospheric Environment* (1967), 16(6), 1399-1405. [https://doi.org/10.1016/0004-6981\(82\)90060-9](https://doi.org/10.1016/0004-6981(82)90060-9).
- Benson, P. E. (1984) CALINE 4-A Dispersion Model for Predicting Air Pollutant Concentrations Near Roadways. No. FHWA-CA-TL-84-15 Final Report. <https://trid.trb.org/view/215944>.
- Benson, P. E. (1992). A Review of The Development and Application of The CALINE3 and 4 Models. *Atmospheric Environment. Part B. Urban Atmosphere*, 26(3), 379-390. [https://doi.org/10.1016/0957-1272\(92\)90013-I](https://doi.org/10.1016/0957-1272(92)90013-I).
- Bernstein, J. A., N. Alexis, C. Barnes, I. L. Bernstein, A. Nel, D. Peden, D. Diaz-Sanchez, S. M. Tarlo, and P. B. Williams. (2004) Health effects of air pollution. *Journal of Allergy and Clinical Immunology*. 114(5), 1116–1123: <https://doi.org/10.1016/j.jaci.2004.08.030>.
- Breiman, L. (1996) Bagging Predictors. *Machine Learning*, 24, 123–140. <https://doi.org/10.1007/BF00058655>.
- Breiman, L. (2001) Random Forest. *Machine Learning*, 45(1), 5-32. <https://doi.org/10.1023/A:1010933404324>.

- Briant, R., Seigneur, C., Gadrat, M., and Bugajny, C. (2013) Evaluation of Roadway Gaussian Plume Models with Large-scale Measurement Campaigns. *Geoscientific Model Development*, 6(2), 445-456. <https://doi.org/10.5194/gmd-6-445-2013>.
- Brook, R. D., S. Rajagopalan, C. A. Pope III, J. R. Brook, A. Bhatnagar, A. V. Diez-Roux, F. Holguin, Y. Hong, R. V. Luepker, M. A. Mittleman, and A. Peters. (2010) Particulate Matter Air Pollution and Cardiovascular Disease: An Update to the Scientific Statement from the American Heart Association. *Circulation*. 121(21), pp.2331-2378: <https://doi.org/10.1161/CIR.0b013e3181dbee1>.
- Caltrans, California Department of Transportation (2019) Project-Level Air Quality Analysis. <https://dot.ca.gov/programs/environmental-analysis/air-quality/project-level-air-quality-analysis>.
- Camdeviren, H., Yazici, A., Akkus, Z., Bugdayci, R., and Sungur, M. (2007) Comparison of Logistic Regression Model and Classification Tree: An Application to Postpartum Depression Data. *Expert Systems with Applications*, 32(4), 987-994. <https://doi.org/10.1016/j.eswa.2006.02.022>.
- Carruthers, D.J., Holroyd, R.J., Hunt, J.C.R., Weng, W.S., Robins, A.G., Apsley, D.D., Thompson, D.J. and Smith, F.B. (1994) UK-ADMS: A New Approach to Modelling Dispersion in The Earth's Atmospheric Boundary Layer. *Journal of Wind Engineering and Industrial Aerodynamics*, 52, pp.139-153. <https://doi.org/10.1016/j.atmosenv.2003.10.052>
- CERC, Cambridge Environmental Research Consultants. (2019) ADMS-Roads [.https://www.cerc.co.uk/environmental-software/ADMS-Roads-model.html](https://www.cerc.co.uk/environmental-software/ADMS-Roads-model.html).
- Cimorelli, A.J., Wilson, R.B., Perry, S.G., Venkatram, A., Weil, J.C., Paine, R.J., Lee, R.F. and Peters, W.D., (1998) Minimum Meteorological Data Requirements for AERMOD—study and Recommendations. USEPA Version, 98314, p.98022. <https://www3.epa.gov/ttn/scram/7thconf/aermod/degrade.pdf>.
- Cimorelli, A.J., Perry, S.G., Venkatram, A., Weil, J.C., Paine, R.J., Wilson, R.B., Lee, R.F., Peters, W.D. and Brode, R.W. (2005) AERMOD: A Dispersion Model for Industrial Source Applications. Part I: General Model Formulation and Boundary Layer Characterization. *Journal of Applied Meteorology*, 44(5), 682-693. <https://doi.org/10.1175/JAM2227.1>.
- Claesen, M., Simm, J., Popovic, D., and Moor, B. D. (2014) Hyperparameter Tuning in Python using Optunity. In *Proceedings of the International Workshop on Technical*

Computing for Machine Learning and Mathematical Engineering, 1, 3.
<https://github.com/claesenm/optunity>.

Donateo, A., and Contini, D. (2014) Correlation of Dry Deposition Velocity and Friction Velocity over Different Surfaces for PM_{2.5} and Particle Number Concentrations. *Advances in Meteorology*. <http://dx.doi.org/10.1155/2014/760393>.

D'Onofrio, D., Kim, B., Kim, Y., and Kim, K. (2016) Atlanta Roadside Emissions Exposure Study-Methodology & Project Overview. <http://atlantaregional.org/wp-content/uploads/2017/03/arees-documentation.pdf>

Draxler, R. R. (1976). Determination of Atmospheric Diffusion Parameters. *Atmospheric Environment* (1967), 10(2), 99-105. [https://doi.org/10.1016/0004-6981\(76\)90226-2](https://doi.org/10.1016/0004-6981(76)90226-2).

Eckhoff, P. A., & Braverman, T. N. (1995). Addendum to the User's Guide to CAL3QHC Version 2.0 (CAL3QHCR User's Guide).

Finn, D., Clawson, K. L., Carter, R. G., Rich, J. D., Eckman, R. M., Perry, S. G., Isakov, V., Heist, D. K. (2010) Tracer Studies to Characterize the Effects of Roadside Noise Barriers on Near-Road Pollutant Dispersion Under Varying Atmospheric Stability Conditions. *Atmospheric Environment* 44, 204-214.

Freund, Y., Schapire, R. E. (1997) A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. *Journal of Computer and System Sciences*, 55(1), 119–139. <https://doi.org/10.1006/jcss.1997.1504>.

Gamo, M., Goyal, P., Kumari, M., Mohanty, U. C., and Singh, M. P. (1994) Mixed-layer Characteristics as Related to the Monsoon Climate of New Delhi, India. *Boundary-Layer Meteorology*, 67(3), 213-227. <https://doi.org/10.1007/BF00713142>.

Georgia EPD, Georgia Environmental Protection Division. (2019) Georgia AERMET Meteorological Data. <https://epd.georgia.gov/air/georgia-aermet-meteorological-data>.

Gifford, F. A. (1961) Use of Routine Meteorological Observations for Estimating Atmospheric Dispersion. *Nuclear Safety*, 2, 47-51.

Granell, J., Guensler, R., and Bachman, W. H. (2002) Using Locality-Specific Fleet Distributions in Emissions Inventories: Current Practice, Problems and Alternatives.

Published in the CD-ROM Proceedings of the 79th Annual Meeting of the Transportation Research Board. Washington, DC. January 2002.

- Greene, W. (2012) *Econometric Analysis* (7th ed.). Upper Saddle River, New Jersey: Prentice Hall.
- Grimmond, C.S.B., and Oke, T.R. (1999) Aerodynamic Properties of Urban Areas Derived from Analysis of Surface Form. *Journal of Applied Meteorology.*, 38, 1262-1292. [https://doi.org/10.1175/1520-0450\(1999\)038<1262:APOUAD>2.0.CO;2](https://doi.org/10.1175/1520-0450(1999)038<1262:APOUAD>2.0.CO;2)
- Guensler, R., Rodgers, M.O., Leonard II, J., and Bachman, W. (2000) A Large Scale Gridded Application of the CALINE4 Dispersion Model. *Transportation Planning and Air Quality IV*. A. Chatterjee, Ed. American Society of Civil Engineers. New York, NY. <https://cedb.asce.org/CEDBsearch/record.jsp?dockkey=0123540>.
- Guensler, R., Pandey, V., Kall, D., Shafi, G., Blaiklock, P., Rodgers, M.O., and Hunter, M., (2008) MOBILE-Matrix and CALINE-Grid: Project-Level Conformity Screening and Microscale Air Quality Impact Assessment Tools. Prepared for the Georgia Department of Transportation, Atlanta, GA. Georgia Institute of Technology. Atlanta, GA. June 2008.
- Guensler, R., Liu, H., Xu, X., Xu, Y., and Rodgers, M.O. (2016) MOVES-Matrix: Setup, Implementation, and Application (16-6362). Presented at 95th Annual Meeting of the Transportation Research Board, Washington, D.C. <https://trid.trb.org/view/1394387>.
- Guensler, R., Liu, H., Xu, Y., Akanser, A., Kim, D., Hunter, M. P., and Rodgers, M.O. (2017) Energy Consumption and Emissions Modeling of Individual Vehicles. *Transportation Research Record*, 2627(1), 93-102. <https://doi.org/10.3141/2627-11>
- Hagler, G. S. W., Baldauf, R. W., Thoma, E. D., Long, T. R., Snow, R. F., Kinsey, J. S., Oudejans, L., and Gullett, B. K. (2009) Ultrafine Particles near a Major Roadway in Raleigh, North Carolina: Downwind Attenuation and Correlation with Traffic-related Pollutants. *Atmospheric Environment*, 43(6), 1229-1234. <https://doi.org/10.1016/j.atmosenv.2008.11.024>.
- Hanna, S. R., Briggs, G. A., and Hosker Jr, R. P. (1982). *Handbook on Atmospheric Diffusion* (No. DOE/TIC-11223). National Oceanic and Atmospheric Administration, Oak Ridge, TN (USA). Atmospheric Turbulence and Diffusion Lab. <https://www.osti.gov/biblio/5591108>.

- Hanna, S. R. (1984) Applications in Air Pollution Modeling. In Atmospheric Turbulence and Air Pollution Modelling (pp. 275-310). Springer, Dordrecht.
- Hartley, S., Rosenbaum, A., Holder, C., Cohen, J., Graham, S., Brode, R., Thurman, J., Langstaff, J., and Fox, T. (2009) Application of AERMOD to Region-wide Emissions in an Urban Setting in Support of the NO₂ NAAQS. https://www.researchgate.net/publication/325269688_Application_of_AERMOD_to_region-wide_emissions_in_an_urban_setting_in_support_of_the_NO2_NAAQS_review
- Hastie, T., Tibshirani, R., Friedman, J. H. (2009a) The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer Series in Statistics second ed., corrected 7th printing. <http://statweb.stanford.edu/~tibs/ElemStatLearn/>.
- Hastie, T., Rosset, S., Zhu, J., and Zou, H. (2009b) Multi-class Adaboost. Statistics and its Interface, 2(3), 349-360. <http://dx.doi.org/10.4310/SII.2009.v2.n3.a8>.
- HEI, Health Effects Institute. (2010) Traffic-related air pollution: A Critical review of the literature on emissions, exposure, and health effect. Boston, MA: HEI. <https://www.healtheffects.org/publication/traffic-related-air-pollution-critical-review-literature-emissions-exposure-and-health>.
- Heist, D. K., Perry, S. G., Brixey, L. A., (2009) A Wind Tunnel Study of The Effect of Roadway Configurations on The Dispersion of Traffic-Related Pollution. Atmospheric Environment 43, 5101-5111.
- Heist, D., Isakov, V., Perry, S., Snyder, M., Venkatram, A., Hood, C., Stocker, J., Carruthers, D., Arunachalam, S., and Owen, R.C., (2013) Estimating Near-Road Pollutant Dispersion: A Model Inter-Comparison. Transportation Research Part D: Transport and Environment, 25, pp.93-105. <https://doi.org/10.1016/j.trd.2013.09.003>.
- Hien, P. D., Bac, V. T., Tham, H. C., Nhan, D. D., and Vinh, L. D. (2002) Influence of Meteorological Conditions on PM_{2.5} and PM_{2.5} – 10 Concentrations during the Monsoon Season in Hanoi, Vietnam. Atmospheric Environment, 36(21), 3473-3484. [https://doi.org/10.1016/S1352-2310\(02\)00295-9](https://doi.org/10.1016/S1352-2310(02)00295-9).
- Hodan, W. M., and Barnard, W. R. (2004) Evaluating the Contribution of PM_{2.5} Precursor Gases and Re-Entrained Road Emissions to Mobile Source PM_{2.5} Particulate Matter Emissions. MACTEC Federal Programs, Research Triangle Park, NC.

<http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.593.2783&rep=rep1&type=pdf>.

- Hsu, C. W., Chang, C.C., and Lin, C. J. (2003) A Practical Guide to Support Vector Classification. Department of Computer Science, National Taiwan University. https://www.researchgate.net/profile/Chenghai_Yang/publication/272039161_Evaluating_unsupervised_and_supervised_image_classification_methods_for_mapping_cotton_root_rot/links/55f2c57408ae0960a3897985/Evaluating-unsupervised-and-supervised-image-classification-methods-for-mapping-cotton-root-rot.pdf.
- Hu, S., Fruin, S., Kozawa, K., Mara, S., Paulson, S. E., and Winer, A. M. (2009) A Wide Area of Air Pollutant Impact Downwind of a Freeway during Pre-sunrise Hours. *Atmospheric Environment*, 43(16), 2541-2549. <https://doi.org/10.1016/j.atmosenv.2009.02.033>.
- Igri, P., Vondou, D., and Kamga, F. (2011) Case Study of Pollutants Concentration Sensitivity to Meteorological Fields and Land Use Parameters over Douala (Cameroon) using AERMOD Dispersion Model. *Atmosphere*, 2(4), 715-741. <https://doi.org/10.3390/atmos2040715>.
- Kall, D., V. Pandey, J., and Guensler, R. (2008) MOBILE-Matrix and CALINE-Grid: Project-Level Conformity Screening and Microscale Air Quality Impact Assessment Tools. 18th Annual On-Road Vehicle Emissions Workshop, San Diego, CA. Coordinating Research Council. Atlanta, GA. March 2008.
- Kampa, M., and Castanas, E. (2007) Human Health Effects of Air Pollution. *Environmental pollution*. 151(2), pp.362-367: <https://doi.org/10.1016/j.envpol.2007.06.012>
- Kim, D., Ko, J., Xu, X., Liu, H., Rodgers, M.O., and Guensler, R. (2019a). Evaluating the Environmental Benefits of Median Bus Lanes: A Microscopic Simulation Approach, *Transportation Research Record: Journal of the Transportation Research Board*, <https://doi.org/10.1177/0361198119836982>.
- Kim, D., Guin, A., Ko, J., Rodgers, M., and Guensler, R. (2019b) Energy and Air Quality Impacts of Truck-Only Lanes: A Case Study of Interstate 75 Between Macon and McDonough, Georgia. Presented at 98th Annual Meeting of the Transportation Research Board, Washington, D.C. <https://trid.trb.org/view/1573113>.
- Kim, D., Liu, H., Xu, X., Lu, H., Wayson, R., Rodgers, M.O., and Guensler, R. (2019c) A Regional Air Quality Impact Assessment Screening Tool based upon MOVES-

Matrix and AERMOD. Presented at Guideline on Air Quality Models: Planning Ahead Conference, Air & Waste Management Association, March 12, Durham, North Carolina, USA.

Kim, D., H. Liu, M.O. Rodgers, and R. Guensler. (2020a) Development of Roadway Link Screening Model for Regional-level Near-road Air Quality Analysis: A Case Study of Particulate Matter. Presentation at the 99th Transportation Research Board (TRB) Annual Meeting.

Kim, D., H. Liu, X. Xu, H. Lu., R. Wayson, M.O. Rodgers, R. Guensler. (2020b) Streamlined Data Processing for Regional Scale Applications of Line Source Dispersion Modeling via Distributed Computing. Presentation at the 99th Transportation Research Board (TRB) Annual Meeting.

Lin, G., Fu, J., Jiang, D., Wang, J., Wang, Q., and Dong, D. (2015) Spatial Variation of the Relationship between PM_{2.5} Concentrations and Meteorological Parameters in China. *BioMed Research International*. <http://dx.doi.org/10.1155/2015/684618>.

Liu, H., Xu, Y., Rodgers, M. O., and Guensler, R. (2015). Developing Vehicle Classification Inputs for Project-Level MOVES Analysis. *Transportation Research Record: Journal of the Transportation Research Board*, No. 2503, 2015, pp. 81-90. <https://doi.org/10.3141/2503-09>.

Liu, H., Xu, X., Rodgers, M.O., Xu, Y., and Guensler, R. (2017) MOVES-Matrix and Distributed Computing for Microscale Line Source Dispersion Analysis. *Journal of the Air & Waste Management Association*, 67(7). <https://doi.org/10.1080/10962247.2017.1287788>.

Liu, H. (2018) Modeling the Impact of Road Grade on Vehicle Operation, Vehicle Energy Consumption, and Emissions. Doctoral Dissertation, Georgia Institute of Technology.

Liu, H., and Kim, D. (2019a) Simulating the Uncertain Environmental Impact of Freight Truck Shifting Programs. *Atmospheric Environment*, 214, 116847. <https://doi.org/10.1016/j.atmosenv.2019.116847>.

Liu, H., Guensler, R., Xu, X., Xu, Y., and Rodgers, M. (2019b) MOVES-Matrix for High-Performance On-road Energy and Emission Rate Modeling Applications. *Journal of the Air and Waste Management Association*, 69(12), 1415-1428. <https://doi.org/10.1080/10962247.2019.1640806>.

- Liu, H., M.O. Rodgers, and R. Guensler (2019c). Impact of Road Grade on Vehicle Speed-Acceleration Distribution, Emissions, and Line Source Dispersion Modeling on Freeways. *Transportation Research Part D: Transport and Environment*. Volume 69, pp: 107-122. <https://doi.org/10.1016/j.trd.2019.01.028>.
- McHugh, C. A., Carruthers, D. J., and Edmunds, H. A. (1997) ADMS and ADMS–urban. *International Journal of Environment and Pollution*, 8(3-6), 438-440. https://www.cerc.co.uk/environmental-software/assets/data/doc_presentations/CERC_2011_ADMS-Urban_Developments_in_modelling_dispersion_from_the_city_scale_to_the_local_scale_StockerJ.pdf
- Montavon, G., Orr, G., and Müller, K. (2012) *Lecture Notes in Computer Science Series. LNCS: Vol. 7700. Neural networks: tricks of the trade*. Springer Verlag. <https://doi.org/10.1007/978-3-642-35289-8>.
- Nadolski, V. L. (1998) *Automated Surface Observing System (ASOS) User's Guide*. National Oceanic and Atmospheric Administration, Department of Defense, Federal Aviation Administration, United States Navy.
- Nowak, D. J., Greenfield, E. J., 2012. Tree and Impervious Cover Change in US cities. *Urban Forestry & Urban Greening*, 11(1), 21-30. <https://doi.org/10.1016/j.ufug.2011.11.005>.
- PACE (2017) *Partnership for an Advanced Computing Environment (PACE)*. <http://www.pace.gatech.edu>.
- Park, Y., Guldmann, J. M. (2020) Understanding Disparities in Community Green Accessibility under Alternative Green Measures: A Metropolitan-wide Analysis of Columbus, Ohio, and Atlanta, Georgia. *Landscape and Urban Planning*, 103806. <https://doi.org/10.1016/j.landurbplan.2020.103806>.
- Pasquill, F. (1974) *Atmospheric Diffusion*. John Wiley & Sons.
- Pasquill, F. (1976) *Atmospheric Dispersion Parameters in Gaussian Plume Modeling. Part 2. Possible Requirements for Change in the Turner Workbook Values*.
- Pasquill, F., and Michael, P. (1977) *Atmospheric Diffusion*. *Physics Today* 30, 55. <https://doi.org/10.1063/1.3037599>.

- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., and Vanderplas, J. (2011) Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12, 2825-2830. <http://www.jmlr.org/papers/v12/pedregosa11a>.
- Quinlan, J. (1986) Induction of Decision Trees. *Machine learning*, (1), 81-106. <https://doi.org/10.1007/BF00116251>.
- R Core Team. (2017) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0: <http://www.R-project.org/>.
- Saide, P. E., Carmichael, G. R., Spak, S. N., Gallardo, L., Osses, A. E., Mena-Carrasco, M. A., and Pagowski, M. (2011). Forecasting Urban PM10 And PM2. 5 Pollution Episodes in Very Stable Nocturnal Conditions and Complex Terrain Using WRF–Chem CO Tracer Model. *Atmospheric Environment*, 45(16), 2769-2780. <https://doi.org/10.1016/j.atmosenv.2011.02.001>.
- Samaranayake, S., Glaser, S., Holstius, D., Monteil, J., Tracton, K., Seto, E., Bayen, A. (2014) Real-Time Estimation of Pollution Emissions and Dispersion from Highway Traffic. *Computer-Aided Civil and Infrastructure Engineering*. 29(7), 546-558. <https://doi.org/10.1111/mice.12078>
- Shafi, G. (2008) Development of Roadway Link Screening Criteria for Microscale Carbon Monoxide and Particulate Matter Conformity Analyses through Application of Classification Tree Model. Master's Thesis, Georgia Institute of Technology. <http://hdl.handle.net/1853/28222>.
- Shafique, M. A., and Hato, E. (2015) Use of Acceleration Data for Transportation Mode Prediction. *Transportation*, 42(1), 163-188. <https://doi.org/10.1007/s11116-014-9541-6>.
- Snyder, M. G., Venkatram, A., Heist, D. K., Perry, S. G., Petersen, W. B., and Isakov, V. (2013a). RLINE: A Line Source Dispersion Model for Near-Surface Releases. *Atmospheric Environment*, 77, 748-756. <https://doi.org/10.1016/j.atmosenv.2013.05.074>
- Snyder, M. G., and Heist, D. K. (2013b). User's Guide for R-LINE Model Version 1.2 A Research LINE Source Model for Near-Surface Releases. Atmospheric Modeling and Analysis Division. https://www.cmascenter.org/r-line/documentation/1.2/RLINE_UserGuide_11-13-2013.pdf

- Tainio, M., de Nazelle, A. J., Götschi, T., Kahlmeier, S., Rojas-Rueda, D., Nieuwenhuijsen, M. J., de Sá, T. H., Kelly, P., and Woodcock, J. (2016) Can Air Pollution Negate the Health Benefits of Cycling and Walking? *Preventive Medicine*. 87, pp.233-236: <https://doi.org/10.1016/j.ypmed.2016.02.002>.
- Tecer, L. H., Süren, P., Alagha, O., Karaca, F., and Tuncel, G. (2008) Effect of Meteorological Parameters on Fine and Coarse Particulate Matter Mass Concentration in a Coal-Mining Area in Zonguldak, Turkey. *Journal of the Air & Waste Management Association*, 58(4), 543-552.
<https://www.tandfonline.com/doi/pdf/10.3155/1047-3289.58.4.543?needAccess=true>.
- Texas Commission on Environmental Quality. (2019) Meteorological Data for Refined Screening with AERMOD.
<https://www.tceq.texas.gov/permitting/air/modeling/aermod-datasets.html>.
- Turner, D. B. (1994). *Workbook of Atmospheric Dispersion Estimates: An Introduction to Dispersion Modeling*. CRC press.
- U.S. Bureau of Labor Statistics. (2020) American Time Use Survey.
<https://www.bls.gov/tus/>.
- USEPA, U.S. Environmental Protection Agency. (1995) User's Guide for the Industrial Source Complex (ISC3) Dispersion Models: Volume II.
<https://nepis.epa.gov/Exe/ZyPURL.cgi?Dockey=000031VJ.TXT>.
- USEPA, U.S. Environmental Protection Agency (2004) User's guide for the AERMOD Meteorological Preprocessor (AERMET). Research Triangle Park, NC, Office of Air Quality Planning and Standards.
https://www3.epa.gov/ttn/scram/7thconf/aermod/aermet_userguide.pdf.
- USEPA, U.S. Environmental Protection Agency. (2006) National ambient air quality standards (NAAQS).
- USEPA, U.S. Environmental Protection Agency. (2008) Integrated Science Assessment for Oxides of Nitrogen – Health Criteria. Research Triangle Park, NC: National Center for Environmental Assessment, Office of Research and Development.
<https://www.epa.gov/isa/integrated-science-assessment-isa-nitrogen-dioxide-health-criteria>

- USEPA, U.S. Environmental Protection Agency (2014) Official Release of the MOVES2014 Motor Vehicle Emissions Model for SIPs and Transportation Conformity. <https://www.federalregister.gov/documents/2014/10/07/2014-23258/official-release-of-the-moves2014-motor-vehicle-emissions-model-for-sips-and-transportation>.
- USEPA, U.S. Environmental Protection Agency (2015a) Project-Level Conformity and Hot-Spot Analyses. <https://www.epa.gov/state-and-local-transportation/project-level-conformity-and-hot-spot-analyses>.
- USEPA, U.S. Environmental Protection Agency (2015b) Technical Support Document (TSD) for Replacement of CALINE3 with AERMOD for Transportation Related Air Quality Analyses. https://www3.epa.gov/ttn/scram/11thmodconf/CAL3_AERMOD_Replacement_TSD.pdf.
- USEPA, U.S. Environmental Protection Agency (2015c) AERMINUTE User's Guide. https://www3.epa.gov/ttn/scram/7thconf/aermod/aerminute_userguide.pdf.
- USEPA, U.S. Environmental Protection Agency (2016) AERSCREEN User's Guide. https://www3.epa.gov/scram001/models/screen/aerscreen_userguide.pdf.
- USEPA, U.S. Environmental Protection Agency. (2018). 3-Day PM Hot-spot Training Course, Held at USEPA, Ann Arbor, MI, June 26-28.
- USEPA, U.S. Environmental Protection Agency. (2019a) Air Quality Dispersion Modeling - Preferred and Recommended Models. <https://www.epa.gov/scram/air-quality-dispersion-modeling-preferred-and-recommended-models>.
- USEPA, U.S. Environmental Protection Agency. (2019b). User's Guide for the AMS/EPA Regulatory Model (AERMOD). https://www3.epa.gov/ttn/scram/models/aermod/aermod_userguide.pdf.
- USEPA, U.S. Environmental Protection Agency. (2019c) MOVES2014b: Latest Version of Motor Vehicle Emission Simulator. <https://www.epa.gov/moves/latest-version-motor-vehicle-emission-simulator-moves>.
- USEPA, U.S. Environmental Protection Agency. (2019d) AERMOD Model Formulation and Evaluation. https://www3.epa.gov/ttn/scram/models/aermod/aermod_mfed.pdf.

- Vallamsundar, S., and Lin, J. (2012) MOVES and AERMOD used for PM2.5 Conformity Hot Spot Air Quality Modeling. *Transportation Research Record*, 2270, 39-48. <https://doi.org/10.3141/2270-06>.
- Vallero, D. A. (2014) *Fundamentals of Air Pollution*. Academic Press.
- Venkatram, A., Isakov, V., Yuan, J., and Pankratz, D. (2004) Modeling Dispersion at Distances of Meters from Urban Sources. *Atmospheric Environment*, 38(28), pp.4633-4641. <https://doi.org/10.1016/j.atmosenv.2004.05.018>.
- Venkatram, A., Snyder, M. G., Heist, D. K., Perry, S. G., Petersen, W. B., and Isakov, V. (2013) Re-Formulation of Plume Spread for Near-Surface Dispersion. *Atmospheric Environment*, 77, 846-855. <https://doi.org/10.1016/j.atmosenv.2013.05.073>
- Washington, S., Karlaftis, M., and Mannering, F. (2011) *Statistical and Econometric Methods for Transportation Data Analysis*. Boca Raton, FL: CRC Press.
- Wu, Y., and Niemeier, D. (2016) Strategy of AERMOD Configuration for Transportation Conformity Hot-spot Analysis. Presented at 95th Annual Meeting of the Transportation Research Board, Washington, D.C.
- Wu, Y. (2018) *Integrated Assessment for Health Effects of Sustainable Transportation Strategies*. Doctoral Dissertation, University of California, Davis. <https://search.proquest.com/docview/2135254695?pq-origsite=gscholar>.
- Xu, X., Liu, H., Anderson, J. M., Xu, Y., Rodgers, M.O., and Guensler, R. (2016) Estimating Project-Level Vehicle Emissions with Vissim and MOVES-Matrix. *Transportation Research Record: Journal of the Transportation Research Board*, 2016. 107-117. doi:10.3141/2570-12.
- Xu, X., Liu, H., Guin, A., Rodgers, M. O., and Guensler, R. (2018a) Regional Emission Analysis using Travel Demand Models and MOVES-Matrix. Presented at 97th Annual Meeting of the Transportation Research Board, Washington, D.C.
- Xu, X., Liu, H., Li, H., Rodgers, M.O., Guensler, R. (2018b) Integrating Engine Start, Soak, Evaporative, and Truck Hoteling Emissions into MOVES-Matrix. DOI: 10.1177/0361198118797208. *Transportation Research Record*. Washington, DC. 2018.

- Xu, Y., Jiang, S., Li, R., Zhang, J., Zhao, J., Abbar, S. and González, M.C. (2019) Unraveling Environmental Justice in Ambient PM_{2.5} Exposure in Beijing: A Big Data Approach. *Computers, Environment and Urban Systems*, 75, pp.12-21. <https://doi.org/10.1016/j.compenvurbsys.2018.12.006>
- Yuan, F., and Cheu, R. L., (2003) Incident Detection using Support Vector Machines. *Transportation Research Part C: Emerging Technologies*, 11(3-4), 309-328. [https://doi.org/10.1016/S0968-090X\(03\)00020-2](https://doi.org/10.1016/S0968-090X(03)00020-2).
- Zhai, X., Russell, A. G., Sampath, P., Mulholland, J. A., Kim, B. U., Kim, Y., and D'Onofrio, D. (2016) Calibrating R-LINE Model Results with Observational Data to Develop Annual Mobile Source Air Pollutant Fields at Fine Spatial Resolution: Application in Atlanta. *Atmospheric Environment*. 147, 446-457: <https://doi.org/10.1016/j.atmosenv.2016.10.015>.
- Zhai, X., Mulholland, J. A., Friberg, M. D., Holmes, H. A., Russell, A. G., and Hu, Y. (2019) Spatial PM_{2.5} Mobile Source Impacts Using a Calibrated Indicator Method. *Journal of the Air & Waste Management Association*, 2019. 1-13: <https://doi.org/10.1080/10962247.2018.1532468>.
- Zhang, H., Wang, Y., Hu, J., Ying, Q., and Hu, X. M. (2015) Relationships between Meteorological Parameters and Criteria Air Pollutants in Three Megacities in China. *Environmental Research*, 140, 242-254. http://www.caps.ou.edu/~xhu/files/Zhangetal_ER_accepted.pdf.
- Zhang, N., Huang, H., Duan, X., Zhao, J., and Su, B. (2018) Quantitative Association Analysis between PM_{2.5} Concentration and Factors on Industry, Energy, Agriculture, and Transportation. *Scientific Reports*, 8(1), 9461. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6013430/>.