Instituto Tecnológico y de Estudios Superiores de Monterrey

Campus Monterrey

School of Engineering and Sciences

Satellite-derived data and ground-based measurements relationships for assessing local and regional distributions of $PM_{2.5}$: model development and applications

A dissertation presented by

Johana Margarita Carmona García

Submitted to the
School of Engineering and Sciences
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy
in
Engineering Science

Major in Environmental Sciences

Monterrey Nuevo León, Jun 15th, 2020

Instituto Tecnológico y de Estudios Superiores de Monterrey
Campus Monterrey

School of Engineering and Sciences

Instituto Tecnológico y de Estudios Superiores de Monterrey

Campus Monterrey

School of Engineering and Sciences

The committee members, hereby, certify that have read the dissertation presented by **Johana Margarita Carmona García** and that it is fully adequate in scope and quality as a partial requirement for the degree of Doctor of Philosophy in Engineering Science, with a major in Environmental Sciences.

Dr. Alberto Mendoza Domínguez
Tecnológico de Monterrey
School of Engineering and Sciences
Principal Advisor

Dr. Diego Fabián Lozano García
Tecnológico de Monterrey
Co-advisor

Dr. Pawan Gupta
STI/USRA, NASA Marshall Space Flight Center
Committee Member

Dr. José Guadalupe Ríos Alejandro
Tecnológico de Monterrey
Committee Member

Dr. Gerardo Manuel Mejía Velázquez
Tecnológico de Monterrey
Committee Member

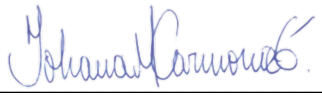
Dr. Rubén Morales Menéndez
Dean of Graduate Studies
School of Engineering and Sciences

Monterrey Nuevo León, Jun 15th, 2020

Declaration of Authorship

I, Johana Margarita Carmona García, declare that this dissertation titled, Satellite-derived data and ground-based measurements relationships for assessing local and regional distributions of $PM_{2.5}$: model development and applications and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this dissertation has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this dissertation is entirely my own work.
- I have acknowledged all main sources of help.
- Where the dissertation is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.



Johana Margarita Carmona García
Monterrey Nuevo León, Jun 15th, 2020

@2020 by Johana Margarita Carmona García
All rights reserved

Dedication

Thanks God for this moment. All the glory is for you. Thank you for my life that has been full of happy moments. For the gifts, talents and strengths with which you have provided me. Thank you for allowing me to fulfill this important academic goal. Thank you especially for blessing me with my family. I dedicate my thesis to them, and I highly appreciate their presence during the doctoral time:

To my husband, I feel so grateful to have your love, unconditional support and your motivating words. Thanks for being my fan, I am yours.

To my beloved mommy, in those moments when I didn't feel strong enough, your existence encouraged me to get ahead. I am blessed because you are here.

To my daddy, your love and your words of serenity are unique. I find in you an incomparable tenderness.

To my brothers Uri and Mauro, thank you for your listening to me whenever I need it, thanks for your advices, guidance and example of excellence.

To my sisters-in-law Andre and Estefy, my nieces Sofi and Fabi, you bring joy to our lives.

Mauro, Andre and Sofi, I really appreciate the way you take care of mommy, the love you profess for her is a source of tranquility for me.

Finally, thanks to family in law for loving me. My love and respect for you.

Acknowledgements

I thank my thesis committee for their teachings, support, and advices that allowed the development of this project.

This is a great opportunity to express all my acknowledgement to Alberto Mendoza for being an inspiring teacher, for allowing me work under your guidance, and for the resources managed to achieve different activities within the PhD program. This would not have been possible without your help.

I also want to thank to Fabián Lozano, José Ríos and Gerardo Mejía for guiding me and sharing with me your own experiences that encouraged me to continue working on my project.

I would like to show my gratitude to Pawan Gupta, for allowing me to do my research-stay under your guidance, for training, for agreeing to be my external advisor and for having the best attitude to teach me, guide me, and work with me.

Thanks to Tecnológico de Monterrey to support me on tuition and the Energy and Climate Change Research Group for the financial support to attend conferences, courses and research stay.

Thanks to CONACYT for the support for living and grant agreement PN 2014/247079 for computational resources and publication funds.

Thanks to COLCIENCIAS Colombia for fellowship through grant Conv-568.

Thanks to the Universities Space Research Association for the support for living and the NASA-Goddard Space Center for all the facilities during the research stay. Thanks to the MODIS Dark Target Algorithm team, mainly Robert Levy.

Thanks to the NASA-ARSET program for training me on remote sensing applications for health and air quality.

My acknowledgement to María Covarrubias, Gloria Ayala and Quetzalli Rocha for their support in completing the administrative procedures to complete this project.

To my colleagues and Friends: Anita, Edson, Hugo, Eder, Yasmany and Lizzete, Thanks for your friendship and support.

To my family and friends, thanks for your love. My deep love for you.

Satellite-derived data and ground-based measurements relationships for assessing local and regional distributions of $PM_{2.5}$: model development and applications

By

Johana Margarita Carmona García

Abstract

Estimating concentrations of ground-level $PM_{2.5}$ (particulate matter with an aerodynamic diameter less than or equal to 2.5 microns) from satellite-derived Aerosol Optical Depth (AOD) through statistical models is a promising method to evaluate the spatial and temporal distribution of $PM_{2.5.}$ Although PM concentrations are most accurately measured using ground-based instruments, the spatial coverage is often too sparse to determine local and regional variations in $PM_{2.5}$. AOD satellite data offers the opportunity to overcome the spatial limitation of ground-based measurements. Combining ground-based measurements with satellite and reanalysis data can be a robust tool to assess air pollution models. However, estimating $PM_{2.5}$ surface concentrations from AOD satellite data is challenging, since multiple factors can affect the relationship between the total-column of AOD and the surface-concentration of $PM_{2.5}$.

This study aims to establish a relationship between AOD satellite data and ground-based data that will allow designing relational models for the study of local $PM_{2.5}$ pollution and the regional $PM_{2.5}$ distributions in northeastern Mexico (NEM). First, an Ensemble Multiple Linear Regression Model (MLR) and a Neural Network Model (NN) were developed to estimate the relationship between the AOD and ground-concentrations of $PM_{2.5}$ within the MMA. The best performance of the models was obtained using a daily scheme, an AOD at 550 μm from the MYD04_3k product in combination with Temperature, Relative Humidity, Wind Speed and Wind Direction ground-based data. For the MLR developed, a correlation coefficient of $R \sim 0.57$ and mean percentage error ~ -6% were obtained. The NN showed a better performance than the MLR, with a correlation coefficient of $R \sim 0.73$ and mean percentage error ~ –3%. The results obtained confirmed that satellite-derived AOD in combination with meteorological fields may allow to estimate $PM_{2.5}$ local distributions.

Then, both the observed and model-estimated daily $PM_{2.5}$ concentrations were classified according to the categories of the Mexican Air Quality Index for $PM_{2.5}$

($PM_{2.5}$ MX-AQI). The observed $PM_{2.5}$ MX-AQI revealed a distinct seasonal variation: a decreasing trend was observed from spring to summer, but then concentrations increased from fall to winter, indicating that air quality across the region is worse in winter than in summer. The developed NN showed 90.4% overall accuracy, 2.5% overestimation and 7.2% underestimation. The results obtained confirmed that the developed model can be used to estimate $PM_{2.5}$ MX-AQI with high accuracy.

Finally, a method that uses neural networks (NN) was developed to combine five years of $PM_{2.5}$ data recorded at local monitoring sites with MERRA-2 reanalysis data. Based on these data, monthly pollution maps were generated to analyze the spatial and temporal patterns of $PM_{2.5}$ concentration fields. AOD, temperature, specific humidity, dust $PM_{2.5}$, sea salt $PM_{2.5}$, black carbon (BC), organic carbon (OC) and sulfate (SO_4^{2-}) reanalysis data were identified as factors that significantly influence the regional distribution of aerosol. The NN estimated a $PM_{2.5}$ monthly mean of 29.4 μg m^{-3}, with the following performance: correlation coefficient $R \sim 0.82$, root mean square error = 8.4 μg m^{-3} and mean percentage error = –2.7%. Significant spatial clustering of high $PM_{2.5}$ was observed, with the highest $PM_{2.5}$ levels located in the MMA, which is the major source to air pollution in this entire area. The estimated data indicates that $PM_{2.5}$ are not distributed uniformly throughout the region, varying spatially as well as temporally, as the mixed mountainous-and-valley topography complicates dispersion of local emissions.

In conclusion, the magnitude of air pollution varies with seasons and regions, and correlates with meteorological factors. The methodology developed here can be used to identify new monitoring sites and can help face information gaps. These findings can also help epidemiologists to better understand the adverse health effects related to $PM_{2.5}$, and serve as a stepping stone towards designing effective regional environmental planning and emissions control strategies.

Contents

Abstract vi
Contents viii
Figures xi
Tables xiii
CHAPTER 1 INTRODUCTION 1
1.1 Problem statement and context 2
1.2 Research questions 8
1.3 Research objectives 9
1.3.1 General objective 9
1.3.2 Specific objectives 9
1.4 Justification 9
1.5 Study area 11
1.5.1 Location 11
1.5.2 Climate 11
1.5.3 Air quality monitoring and $PM_{2.5}$ pollution in the MMA 13
CHAPTER 2 FUNDAMENTALS 15
2.1 Remote sensing fundamentals 16
2.1.1 Spectral regions and atmospheric windows 16
2.1.2 Radiation 17
2.1.3 Types of remote sensing instruments 17
2.1.4 Resolution of remote sensing 17
2.2 Radiative transfer 19
2.2.1 Radiative transfer (processes) 19
2.2.2 Radiative transfer equation - visible radiation 20
2.3 Aerosol properties 21
2.3.1 Aerosol Optical Depth 21
2.3.2 Single Scattering Albedo 21
2.3.3 Ångström Exponent 22
2.4 AOD - $PM_{2.5}$ relationship 22
2.5 MODIS aerosol modeling 22
2.5.1 MODIS - Dark Target Algorithm (10 km) 23

2.5.2 MODIS - Dark Target Algorithm (3 km product) 31
2.5.3 MODIS – Deep Blue Algorithm (10 km) 32
2.5.4 MODIS – Combined product (Merged DT and DB, 10 km) 37
2.6 MERRA-2 reanalysis model. 38
2.6.1 GEOS-5 model 39
2.6.2 GOCART aerosol module 40
2.7 Artificial Neural Network 41
2.7.1 Multi-layer perceptron artificial neural networks 41
2.7.2 Levenberg-Marquardt algorithm, generalization and overfitting 42
2.7.3 Artificial Neural Network to estimate $PM_{2.5}$ 43
CHAPTER 3 METHODOLOGY 44
3.1 AOD and ground-based data relationship 47
3.1.1 Model development and validation 47
3.1.2 Ground- based $PM_{2.5}$ and meteorological data 47
3.1.3 MODIS AOD data sets 49
3.1.4 $PM_{2.5}$ concentration and MODIS AOD spatio-temporal collocation 52
3.1.5 Model development and validation 54
3.1.6 Statistical performance 58
3.2 Developing a local application: Categorization of metropolitan air quality index based on satellite-derived data and meteorological ground-based measurements 59
3.2.1 Metropolitan Air Quality Index (MX-AQI) 59
3.2.2 Data and methods 59
3.2.3 Classification strategy 60
3.2.4 Performance evaluation 60
3.3 Developing a regional application: Analysis of spatial and temporal distribution of $PM_{2.5}$ pollution over Northeastern Mexico using MERRA-2 Reanalysis Datasets. 63
3.3.1 Data and methods 63
3.3.2 MERRA-2 datasets 64
3.3.3 Model development 65
CHAPTER 4 RESULTS 66
4.1 AOD and ground-based data relationship 67
4.1.1 Predictor variables and model selection 67

4.1.2 The spatio-temporal variability of the correlation among AOD, meteorological variables and $PM_{2.5}$ 71
4.1.3 Model development and validation 79
4.2 Developing a local application: Categorization of metropolitan air quality index based on satellite-derived data and meteorological ground-based measurements 87
4.2.1 Air quality index classification 87
4.2.2 Classification behavior 88
4.3 Developing a regional application: Analysis of spatial and temporal distribution of $PM_{2.5}$ pollution over Northeastern Mexico using MERRA-2 Reanalysis Datasets. 91
4.3.1 Distribution of data: $PM_{2.5}$ from SIMA and AOD from MERRA-2 91
4.3.2 $PM_{2.5}$ components from MERRA-2 93
4.3.3 Model performance 97
4.3.4 Regional $PM_{2.5}$ distribution 100
CHAPTER 5 CONCLUSIONS, SUMMARY AND FUTURE WORK 102
5.1 Conclusions 103
5.2 Summary 105
5.3 Future Work 106
REFERENCES 108
PUBLISHED PAPERS 123
CURRICULUM VITAE 125

Figures

Chapter 1. INTRODUCTION

Figure 1.1. Location and municipalities of the Monterrey Metropolitan Area (MMA). 12

Figure 1.2. Location of monitoring sites that form part of the SIMA network in the MMA. 14

Chapter 2. FUNDAMENTALS

Figure 2.1. Electromagnetic spectrum. 18

Figure 2.2. Flowchart of the Dark Target algorithm over land (modified from Levy et al. [2013]). 30

Figure 2.3. Flowchart of the Deep Blue algorithm over land (Hsu et al., 2013). 37

Chapter 3. METHODOLOGY

Figure 3.1. Research methodology framework. 46

Figure 3.2. Methodology framework for establishing the relationship between the AOD and ground-based data to estimate $PM_{2.5}$ within the MMA. 48

Figure 3.3. Ground-based data and MODIS AOD spatial collocation. 53

Figure 3.4. Selection and validation of the "*best dataset*". 55

Figure 3.5. A schematic of the NN model used in this study for estimation of $PM_{2.5}$. 57

Figure 3.6. Methodology framework for categorizing daily $PM_{2.5}$ AQI. 60

Figure 3.7. Scheme of the confusion matrix 63

Figure 3.8. Methodology framework for analyzing the spatial - temporal distribution of $PM_{2.5}$ 64

Chapter 4. RESULTS

Figure 4.1. Histograms of temperature and relative humidity data included in the daily scheme over domain 3 ($N = 694$). 68

Figure 4.2. Wind rose diagrams for the entire study episode. ($N = 694$). 70

Figure 4.3. Similarities and differences in the AOD distribution from the four SDS in domain 3. 73

Figure 4.4. Similarities and differences in the data distribution of the observed hourly $PM_{2.5}$ means and the observed daily $PM_{2.5}$ means. 77

Figure 4.5. Diurnal $PM_{2.5}$ variability. 78
Figure 4.6. NN model validation. 80
Figure 4.7. Statistical performance obtained using MLR and NN models to estimate daily $PM_{2.5}$ means, over the MMA. (all episode) 81
Figure 4.8. Statistical performance obtained by each season using MLR and NN models to estimate daily $PM_{2.5}$ means, over the MMA. 83
Figure 4.9. Statistical performance obtained using the daily $PM_{2.5}$ means estimated by the MLRs performed by seasons and the daily $PM_{2.5}$ means estimated by the NNs trained by seasons, over the MMA. (The fitted models results were merged by seasons). 84
Figure 4.10. Comparison of daily $PM_{2.5}$ means estimated by the statistical models developed in this study and the daily $PM_{2.5}$ means recorded by SIMA ($PM_{2.5}$ observed). 86
Figure 4.11. Frequencies of MX-AQI categories for $PM_{2.5}$. 87
Figure 4.12. Overall classification of MX-AQI categories for $PM_{2.5}$. 88
Figure 4.13. Overall performance by MX-AQI categories 89
Figure 4.14. Confusion Matrix, Evaluating "Good" – "Moderate" Level 90
Figure 4.15. Confusion Matrix, Evaluating "Moderate" – "Bad" Level 91
Figure 4.16. Distribution of meteorological variables recorded by SIMA that influenced the AOD-$PM_{2.5}$ relationship in the MMA 92
Figure 4.17. Comparison of temporal behavior of monthly AOD means obtained from MERRA-2 and monthly $PM_{2.5}$ means recorded by SIMA, over MMA for the period covering from January 2010 to December 2014. 93
Figure 4.18. Annual profile $PM_{2.5}$ components from MERRA-2, over MMA for the period covering from January 2010 to December 2014. 95
Figure 4.19. Comparison of monthly BC mean concentration estimated by INECC (2016) with monthly BC mean concentration estimated by MERRA-2 during 2015. 97
Figure 4.20. Statistical performance obtained using NN to estimate monthly $PM_{2.5}$ means, over the MMA for the period covering from January 2010 to December 2014. 98
Figure 4.21. Comparison among the monthly $PM_{2.5}$ means (µg m-3) estimated by the NN developed in this study, MERRA-2 derived data and SIMA data. 99
Figure 4.22. Regional distribution of monthly $PM_{2.5}$ means in NEM. 101

Tables

Chapter 2. FUNDAMENTALS

Table 2.1. The primary spectral regions used in earth remote sensing (Schowengerdt, 2007)..... 16

Table 2.2. Aerosol modes, characteristics and differences..... 28

Table 2.3. Lookup table used in DT algorithm (ATBD, 2013)..... 29

Table 2.4. Surface Reflectance Coefficients[1] for MODIS 0.47 and 0.65 μm (Hsu et al., 2013)..... 34

Chapter 3. METHODOLOGY

Table 3.1. AOD-SDS from MODIS aerosol products..... 50

Table 3.2. Categories based on $PM_{2.5}$ MX-AQI and relevant information (EPA, 2009; EPA, 2014; EPA, 2015; NADF-009-AMBT-2017. 61

Table 3.3. Standard evaluation metrics contained in the confusion matrix..... 62

Table 3.4. Meteorological SDS from MERRA-2..... 65

Table 3.5. Aerosol SDS from MERRA-2..... 65

Chapter 4. RESULTS

Table 4.1. Refined datasets developed to evaluate the relationship among AOD, meteorological variables and $PM_{2.5}$ over the MMA..... 67

Table 4.2. The descriptive statistics of the four AOD - SDS averaged in the three tested domains. 71

Table 4.3. Impact of various spatio-temporal average schemes on the correlation among AOD, meteorological variables and $PM_{2.5}$. 74

Table 4.4. Ensemble model development and comparison with the MLR using the original *best dataset* 79

Table 4.5. Summary of the descriptive statistics of the daily $PM_{2.5}$ (observed and estimated) and statistical performance obtained using the models developed in this study. 85

Table 4.6. Overall performance of the $PM_{2.5}$ MX-AQI classification obtained using the models developed in this study..... 89

Table 4.7. Comparison of $PM_{2.5}$ concentrations (μg m^{-3}) from components reported by MERRA-2 (M-2) with those reported by other MMA studies. 96

Chapter 1 INTRODUCTION

1.1 Problem statement and context

Atmospheric aerosols are composed of a mixture of liquid and solid particles suspended in the air. They are complex pollutants with particle diameters in the range of 10^{-9} - 10^{-4} m and with several physical, chemical, and biological features (Harrison et al., 2000; Vu et al., 2015). Atmospheric aerosols are released from a range of natural and anthropogenic sources. High temporal variations in contributions from natural sources are widely known (inter-annual and seasonal variability). Windblown soils, sea salt, pollen, spores, natural forest fires, volcanic eruptions, crustal weathering, seawater drops, and evaporation are natural sources (Alvesa et al., 2012; Vianaa et al., 2014). Meanwhile, anthropogenic sources include fossil fuel combustion, vehicle exhausts, industrial activities, incineration, and biomass burning, among others (Karagulian et al., 2015).

Airborne Particulate Matter (PM) forms part of atmospheric aerosols (Taiwo et al., 2014). Besides the varying emission sources, PM also differs in size and composition. PM size is classified on the basis of aerodynamic diameter, which is defined as the diameter of a sphere of unit density $\left(\rho_p = 1\ g\ cm^{-3}\right)$ that has the same terminal gravitational settling velocity in still air as the actual real particle. Current air quality standards for PM are based on total mass of $PM_{2.5}$ (PM with aerodynamic diameter of 2.5 μm and less) and PM_{10} (PM with aerodynamic diameter 10 μm and less). Both PM size ranges are indicators of air quality (WHO, 2005; EPA, 2004). In regards to its chemical composition, it is complex and may include several crustal elements (e.g., Si, Ca), metals (e.g., V, Al, Fe, Ti, Cr, Ni, Cu, Zn, Pb), cations (e.g., Na^+, NH_4^+, K^+), anions (e.g., Cl^-, NO_3^-, SO_4^2), Organic Carbon (OC), and Elemental Carbon (EC) (Liang et al., 2016). Strong seasonal and geographic variations in chemical composition have been identified (Martins and Carrilho da Graça, 2018).

One major concern with the presence of high levels of PM in the atmosphere is related to health issues. Different health effects and exposure-response relationships are associated with exposure to PM. Several disruptions in respiratory and cardiovascular systems functions may be provoked by PM. These particles can penetrate deep into the respiratory system and have been associated with various disturbances, including increased risk in asthma, short-term decreases in lung function, lung inflammation and obstructive pulmonary disease (Balmes et al., 2014; Feng et al., 2016; Xing et al., 2016). Concerning the cardiovascular

system, there is an increased risk of affecting cardiac function and heart rate variability, inducing systemic inflammatory changes and development of atherosclerosis as well as coronary and aortic lesions (Atkinson et al., 2015; Du et al., 2016; Chen et al., 2018). In addition, PM inhalation can increase deep vein thrombosis and coagulation (Robertson and Miller, 2018). Short-term elevations in $PM_{2.5}$ may also increase the risk of ischemic and/or hemorrhagic stroke morbidity and mortality (Wang et al., 2014). Neurodegenerative diseases and neurodevelopmental disorders including dementia, Alzheimer's disease, autism spectrum disorder, Parkinson's disease, and mild cognitive impairment (MCI) are associated to $PM_{2.5}$ exposure (Fu et al., 2019). Initiation and progression of diabetes mellitus and eliciting adverse birth outcome are associated to $PM_{2.5}$ exposure (Feng et al., 2016).

Another major concern related to PM is its properties as a radiation-forcing agent. Atmospheric aerosols influence the radiation budget of the earth-climate system. Studies of the optical properties of aerosols are crucial to fully understand the aerosol radiative effects. The Intergovernmental Panel on Climate Change (IPCC; IPCC, 2014) has categorized aerosol radiative forcing effects as follows. First, the radiative effect due to aerosol–radiation interactions, formerly known as direct effect, is the change in radiative flux caused by the combined scattering and absorption of radiation by atmospheric aerosols. Second, the aerosol–cloud interactions including: i. the first indirect effect or cloud albedo effect: hygroscopic aerosol particles act as cloud condensation nuclei, aerosols modify the cloud microphysics (changes in clouds' radiative properties, their frequency and their lifetimes), as such clouds are more highly reflecting (increasing the cloud albedo), this induces a negative radiative forcing; ii. the second indirect effect or cloud lifetime effect: the effect of aerosols on clouds' height, lifetime and water content (related to the amount of water required before precipitation occurs); and, iii. associated cloud changes, often referred to as the semi-direct aerosol effect: absorbing aerosols modify atmospheric stability in the boundary layer (some aerosols modify the air temperature, its humidity and the vertical stability of the air column) and free troposphere, cloud cover is expected to decrease if absorbing aerosol is embedded in the cloud layer. Absorbing aerosols embedded in cloud drops enhance their absorption, which can affect the dissipation of clouds. Third, the impact of black carbon (BC, sometimes interchangeably referred also as elemental carbon or EC) on snow and ice surface albedo: BC in the snow or ice can lead to a decrease of the surface albedo. Because absorption by ice

is very weak at visible and ultraviolet (UV) wavelengths, BC in snow makes the snow darker and increases absorption.

In addition to climate studies (Agrawal et al., 2014), characterization of the aerosol optical properties is also important for a number of applications (Kahnert et al., 2014), including air pollution (Fishman et al., 2008, Prud'homme et al., 2013, Chang et al., 2014), atmospheric visibility (Hadjimitsis et al., 2010), and air quality monitoring from space and from ground-based instruments (Duncan et al., 2014; Lee et al., 2014). However, the understanding of these properties is complex, because they are determined by chemical composition, concentration, size, shape, and internal structure of particles suspended in air. All these characteristics are highly variable in space and time (Kokhanovsky, 2008).

In PM monitoring applications, typically, PM concentration is measured using ground-based devices. These ground-based observations represent in situ measurements of mass concentration. They are sparse and not numerous. For example, in Latin America most of the monitoring stations are located only in a given country's capital city and a few major cities (Riojas et al., 2016). $PM_{2.5}$ monitors are frequently located at specific sites that represent neighborhood (500 m to 4 km) or urban scales (4 km to 100 km) to determine compliance with the annual standard and, at maximum, population-oriented locations for comparison with the 24-hour standard (EPA, 1997). Neighborhood and urban-scale monitors do not have the required coverage to estimate regional and global distributions of PM. Meanwhile, satellite data have tremendous potential for deriving indirect estimates of ground PM pollution from aerosol optical properties, overcoming the well-known spatial limitation of traditional measurements, and could serve as a complementary cost-effective method for PM monitoring on local and regional scales where ground-based observations are scarce or lacking altogether (Gupta et al., 2006; Péré et al., 2009).

Aerosol data products (processed data sets) are developed from a subset of satellite sensors, such as MODIS, MISR, AATSR, CALIOP, POLDER, OMI, VIIRS, OCM that provide important observations of atmospheric aerosols properties (Mhawish et al., 2018). Aerosol data products derived from MODIS sensor (MODerate resolution Imaging Spectroradiometer) aboard both the NASA's Aqua and Terra Satellites, with nearly global coverage at high/moderate spatial resolutions, calibration accuracy and radiometric stability,

are being widely used to estimate ground-level particle characteristics, even over water and complex land surfaces (Kahna, 2011; Streets, 2013).

One of the most robust aerosol parameters retrieved in MODIS aerosol products is the Aerosol Optical Depth (AOD), also known as Aerosol Optical Thickness (AOT). AOD is the integral of aerosol extinction along the vertical atmospheric column from earth's surface to the top of the atmosphere (TOA). Promising empirical correlations were reported by Hoff and Christopher (2009), they reviewed and discussed over 30 papers that have explored the relationship between surface PM and AOD retrieved from satellite. However, it is not straightforward to compare AOD and PM directly, because AOD derived from satellite is an optical quantity representing columnar information, whereas PM describes near-surface mass concentration. In addition, variations in local conditions, aerosol type, hygroscopicity, and variations in aerosol chemical composition play an important role in determining the strengths of such correlations. AOD - PM relationship also varies in space and time, which needs to be assessed specifically for different regions and seasons (Li et al., 2015). Therefore, the relationship between AOD and PM should be determined regionally to account for its specific conditions (Schaap et al., 2009).

Chu et al. (2016) reviewed 116 articles focused on the prediction of PM from satellite AOD and noted that the correlations between AOD and PM ground-based measurements have shown considerable variation depending on the season, location, temporal and spatial scales. Engel-Cox et al. (2004) performed the first extended –both in terms of temporal and spatial coverage- study of correlations between MODIS-AOD and PM ground-based concentration data, by applying the Single Linear Regression (SLR) Model at a national scale, in the United States. They observed that the Eastern and Midwest portions of the US showed better correlation coefficients than other areas, particularly, the Western US, which showed little or no correlation among the variables of interest.

Different statistical techniques have been used in an effort to understand and/or improve the correlation between AOD and PM. Some of these techniques consider additional variables and have been used for climate and air quality purposes. For example, for an application in the St. Louis Area, Liu et al. (2007) employed General Linear Regression Models (GLM) that included meteorological fields as covariates, obtaining $R = 0.5$. Gupta and Christopher

(2008) used Multiple Linear Regression (MLR) and Neural Network Techniques (NNT), considering Planet Boundary Layer (PBL) height, location, temperature (T), and Relative Humidity (RH), in an analysis performed for the Southeastern US. Their study achieved an *R* of 0.7 using MLR and *R* of 0.83 using the NNT, improving the *R* of 0.41 obtained through a SLR model.

Strawa et al. (2013) combined observations from multiple satellites using Generalized Additive Model (GAM) in order to improve estimates of local PM by using only satellite observations (without surface measurements or input from meteorological models). Their application resulted in an *R* of 0.77, when compared to the *R* of 0.17 obtained with the SLR model. Li et al. (2015) used AOD retrieved by different satellite sensors to derive $PM_{2.5}$ focused on monthly means applying a developed combined maximum covariance analysis. Comparison was focused on large-scale spatial and temporal variability, and both seasonal and interannual scales were examined. Their spatial and temporal analysis showed that AOD - $PM_{2.5}$ relationship can be highly variable for different regions and on different time scales.

Van Donkelaar et al. (2016) estimated global $PM_{2.5}$ concentrations using information from satellite, simulation and ground-based sources by applying a Geographically Weighted Regression (GWR) to global geophysically based satellite-derived $PM_{2.5}$ estimates. This approach demonstrates that the addition of even sparse ground-based measurements to more globally continuous $PM_{2.5}$ data sources can yield valuable improvements to $PM_{2.5}$ characterization on a global scale. Zheng et al. (2017) used both ground and satellite observations in Beijing. This study suggests that aerosol type, relative humidity, planetary boundary layer height, wind speed and direction, and the vertical structure of aerosol distribution are influential factors to consider when the AOD–$PM_{2.5}$ relationships are investigated.

Soni et al. (2018) evaluated the relationship between PM_{10}, AOD and meteorological parameters (T, RH and WS) over Jaipur, a semi-arid region in India using both multivariable regression ($R = 0.75$) and log-linear regression models ($R = 0.76$). The models were validated over Jodhpur (another arid region) extending the analysis using the same coefficients obtained from models trained over Jaipur. $R = 0.75$ and $R = 0.80$ were obtained using multivariable regression and log-linear regression models, respectively. Stirnberg et al.

(2018) investigated the PM_{10} - AOD relationship on an hourly basis for Berlin, Germany. They found that the relationship between AOD and PM_{10} was rarely linear and strongly influenced by ambient relative humidity, boundary layer height, wind direction and wind speed.

For epidemiological studies, Kloog et al. (2011, 2014) applied a mixed-effects models approach to AOD retrievals in a combination with Land Use Regression (LUR) models, in order to study the temporally and spatially resolved short-term and long-term exposures to $PM_{2.5}$ and PM_{10}. Also, Van Donkelaar et al. (2010) combined AOD data with a chemical transport model for mapping global ground-level $PM_{2.5}$ aimed to facilitate studies of chronic exposure to particulate matter in regions of the world without extensive ground-based monitoring networks. Van Donkelaar et al. (2015) combined satellite-derived $PM_{2.5}$ sources to produce improved global estimates of long-term $PM_{2.5}$ exposure and changes over time. Lee et al. (2016) estimated daily ground-level $PM_{2.5}$ concentrations combining AOD data with land use regression in California, United States, using mixed effects model with AOD, air basin adjustment, land use parameters, and meteorological data. The $PM_{2.5}$ predictive capability was improved by expanding the spatial coverage of estimated $PM_{2.5}$. This may reduce exposure errors for health effect studies in California.

In Mexico, the research about AOD - PM relationship is limited. Focused on the health effects, Just et al. (2015) developed the first high spatially and temporally resolved $PM_{2.5}$ exposure model for the Mexico City Metropolitan Area (MCMA). They predicted daily $PM_{2.5}$ and analyzed the geographical distribution of $PM_{2.5}$. The developed daily $PM_{2.5}$ exposure model can yield a great impact in studying acute and chronic associations of air pollution with consequent morbidity and mortality in the MCMA. The model obtained, which is based on mixed effects, allows daily estimates even in areas within the city that are not adjacent to the few stationary municipal monitors and showed that precipitation and height of the boundary layer are both important factors influencing the relationship between AOD and $PM_{2.5}$ in the area.

A preliminary study to explore the influential factors on the AOD-PM relationship within the Monterrey Metropolitan Area (MMA), in Northeastern Mexico (NEM), was conducted by Carmona et al. (2014), for a period covering January to March of 2013. Single (SLR) and

multiple regression models (MLR) were applied in order to explore the relationship between PM_{10} concentrations based on ground-based measurements and MODIS-AOD. Meteorological and air quality data were obtained from the routine air quality monitoring stations (AQMS) of the MMA, and satellite data were acquired from the MODIS sensor on board of Aqua satellite. The MODIS product used was MYD04 - MODIS Aerosol product-tested in a 3 km × 3 km resolution. A *R* of 0.60 was obtained using MLR, improving the *R* of 0.45 from the SLR model. AOD was the main predictor for estimating PM_{10} levels followed by Relative Humidity.

In conclusion, further research and tool development that can enhance the performance of models attempting to relate satellite-based to ground-based $PM_{2.5}$ measurements are needed. Establishing column AOD-to-ground PM relationships are imperative for assessing the feasibility of using satellite aerosol data as a complement in air quality monitoring, and thus overcoming the spatial limitation of ground-based, fixed, air quality measurement sites. Combining ground-based measurements with satellite and reanalysis data can be an excellent tool to assess air pollution models. The growing records of both aerosol and meteorological remote sensing data yields a new view on urban and regional air quality distribution.

1.2 Research questions

- May AOD satellite retrievals in combination with statistical models and ground-based data be useful for assessing the local $PM_{2.5}$ concentrations over the MMA?
- What are the most important factors (i.e., meteorological data, spatio-temporal averaging scheme) that influence the models' performance and would be better to improve the correlation models?
- How accurately do the models predict the Air Quality Index (AQI) categories for $PM_{2.5}$ over MMA?
- May data from global reanalysis datasets in combination with statistical models be useful for assessing the regional $PM_{2.5}$ distributions over Northeastern Mexico?

1.3 Research objectives

1.3.1 General objective

To establish a relationship between AOD satellite data and ground-based data that will allow designing relational models for the study of local $PM_{2.5}$ pollution and the regional $PM_{2.5}$ distributions in northeastern Mexico.

1.3.2 Specific objectives

1. To develop and validate statistical models aimed to estimate $PM_{2.5}$ concentrations based on AOD satellite data and meteorological ground-based data.
2. To evaluate the ability of the validated models to accurately predict the AQI categories for $PM_{2.5}$ in a local application over MMA.
3. To analyze the spatial and temporal distribution of $PM_{2.5}$ pollution developing a regional application over NEM using data from the validated model dataset and data from global reanalysis datasets.

1.4 Justification

Estimating $PM_{2.5}$ using AOD data from satellites is promising, but poses several challenges. $PM_{2.5}$ and MODIS AOD represent two different atmospheric loadings of pollutants. On one hand, $PM_{2.5}$ is the dry mass of aerosols (with a specific particle size) measured at ground level and represents the particle mass concentration near the surface. On the other hand, MODIS AOD represents total columnar loading of all aerosol particles from the surface to the TOA averaged over a large spatial area. Hence, several factors influence the AOD and $PM_{2.5}$ relationship (Li et al., 2015). For example, aerosol vertical distribution has to be considered to understand the AOD and $PM_{2.5}$ relationship, since AOD is related to the amount of aerosol in the vertical column of atmosphere and $PM_{2.5}$ represents observations at surface level; elevated aerosol layers increase the AOD but are not observed by surface measurements (Schafer et al., 2008; He et al., 2008). Humidity conditions also has to be considered to establish AOD and $PM_{2.5}$ relationship. $PM_{2.5}$ is typically measured at relatively dry conditions (Collaud et al., 2013) while satellites provide information under all humidity conditions (Li et al., 2015). In the presence of water vapor, hydrophilic particles increase their size. Because the additional mass, AOD will increase and may also result in

discrepancies between AOD and $PM_{2.5}$ (Crumeyrolle et al., 2014). Additionally, cloud-free conditions are required to report satellite retrievals, reported in several meteorological conditions (including clouds).

The use of satellite data has tremendous potential in the advance of knowledge regarding $PM_{2.5}$ spatial distribution, especially in regions where the routine monitoring $PM_{2.5}$ sites are scarce. Satellite data is produced covering wide spatial areas and may provide a broad $PM_{2.5}$ regional perspective. Using satellite data has the potential to support environmental agencies in their air quality monitoring efforts, as satellite data may be used as a means to address the lack of spatial coverage of monitoring ground-based sites. Furthermore, these kinds of data may be used in studies about climate impacts of aerosols because they contribute to a better understanding of the solar attenuation and regionally radiative forcing. Moreover, long-term and short-term $PM_{2.5}$ exposure-response relationships studies can be addressed using these data.

While promising results have been achieved in using satellite AOD for surface air pollution research, sources of uncertainty have to be considered on this kind of studies including calibration, cloud contamination, surface parameterization and the assumptions made in the aerosol retrieval models (Li et al., 2015). Particularly, some limitations are faced when MODIS data is used for estimating $PM_{2.5,}$ such as: lack of information on vertical distribution, chemical composition and size distribution of the aerosols. Limitation on temporal resolution which is not appropriate for understanding the diurnal variability of $PM_{2.5}$. Gaps in spatial coverage due to clouds and limitation on spatial resolution of the products which could be insufficient for analysis at fine spatial resolution.

AOD - $PM_{2.5}$ relationship is highly variable in space and time and needs to be assessed specifically for different regions and seasons. In the current research, we demonstrate the viability of using MODIS data in order to establish a quantitative relationship between the MODIS AOD and ground-based observations for assessing -temporally and spatially- local and regional distributions of $PM_{2.5}$ in Northeastern Mexico.

1.5 Study area

1.5.1 Location

This study was conducted within the Monterrey Metropolitan Area (MMA) in Mexico, for a period covering January 2010 to December 2017. The MMA (25°40'N, 100°20'W) is located in the State of Nuevo León, in the northeast of Mexico, around 720 km N of Mexico City, and some 230 km S of the US border. The MMA is composed of eighteen municipalities (**Figure 1.1**) with an area of 7,657 km^2 (CONAPO, 2018). It is considered the largest urban area in Northeastern Mexico and the third most populous in the country with 4.7 million inhabitants (CONAPO, 2018). The MMA is one of the urban areas with the worst air quality problems in the country (Wakamatsu et al., 2017).

The study area is located on the boundary of the province of the Sierra Madre Oriental, which is characterized by a mountainous relief with alternations of mountain ranges and valleys (López-Ramos, 2008). Vertical drops of over 1400 meters are recognized, with minimum and maximum elevations from 542 to 1993 meters above sea level, respectively. The areas of lower altitude correspond to valleys where the population of the MMA is settled (Eguiluz et al., 2000).

1.5.2 Climate

The average daily temperature varies greatly depending on the season: reaches temperatures above 40 °C during the summer and minimum temperatures of 0 ºC in the winter (Aguilar-Barajas et al., 2015). The average annual rainfall is 300 to 500 mm. The maximum incidence of rains occurs in September with 80 to 90 mm and the minimum, from 5 to 10 mm, occurs in the months of January and December. The number of days with appreciable rainfall varies between 55 to 60 days a year, being cloudless the rest of the days of the year (Quispe-Medina, 2005). In this area, the climate has a heat wave condition; that is, a small season less rainy, within the rainy season (also called mid-summer drought). Frosts occur for 1 to 8 days on average in the months of November, December, January and February. Hailstorms are of irregular distribution and do not keep a definite behavior pattern. These occur, in general, in a range of 0 to 2 days throughout the year. Their incidence is associated with the first months of the rainy season (April, May and June) (Aguilar-Barajas et al., 2015).

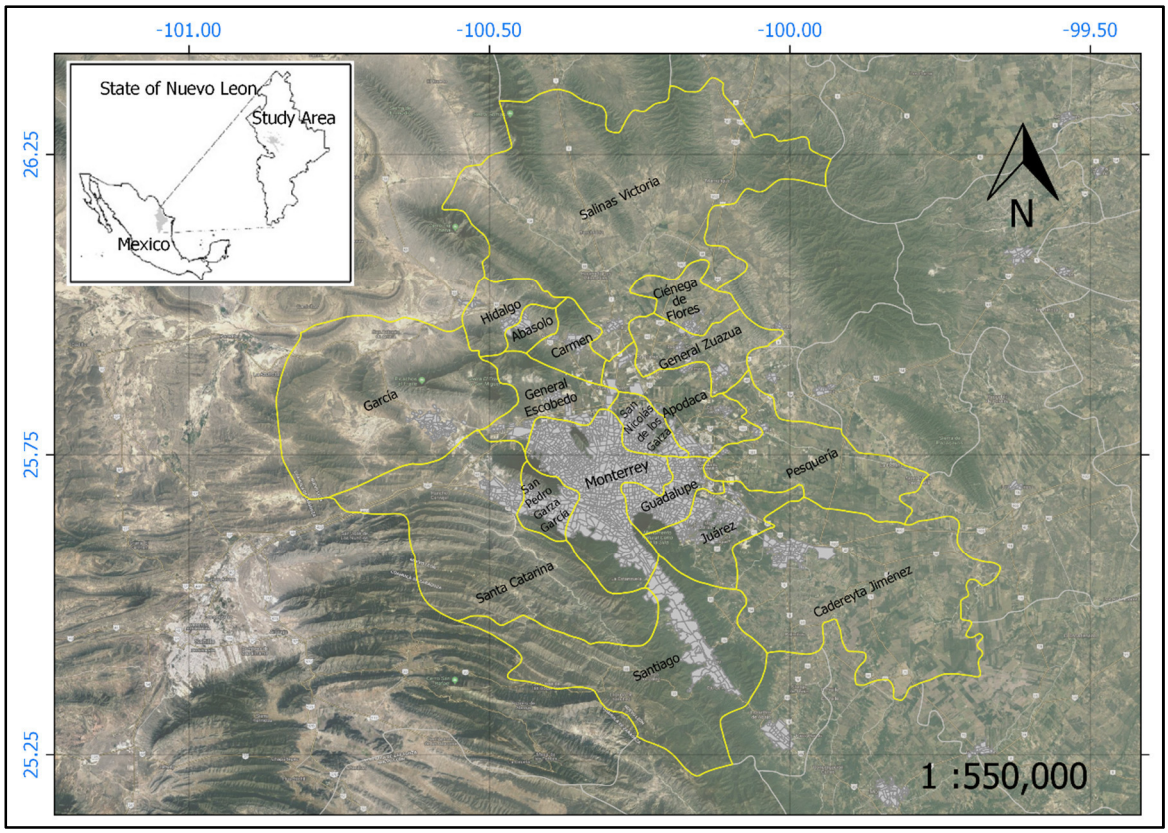


Figure 1.1. Location and municipalities of the Monterrey Metropolitan Area (MMA).

The MMA is located within the high-pressure subtropical zone along with regional orographic conformation influenced by continental maritime air masses and modified by the upper circulation of the atmosphere. The prevailing winds for the MMA are the trade winds coming from the northeast and in second place are those coming from the southeast and are directed westward between N 270 ° E and N 292.5 ° E. For the central zone, the prevailing winds go towards the southwest. For the northwest and the northeast zones almost, a quarter of the winds are directed towards directions between the west and the northwest, although for the northwest there are strong winds heading southeast. Most of the winds captured in the southeast and southwest are directed from east to west (Quispe-Medina, 2005). When the trade winds collide with the Sierra Madre Oriental, they cause abundant rainfall that reaches the Monterrey Valley during the month of June. In September, the rains are more intense, with more precipitation and deeper penetration. In winter, the westerly winds, which are less humid than the trade winds, cause dryness that lasts until spring, presenting masses of cold air that decrease the temperature.

1.5.3 Air quality monitoring and $PM_{2.5}$ pollution in the MMA

The MMA has a network of air quality monitoring stations known locally as the Sistema Integral de Monitoreo Ambiental de Nuevo León (SIMA) composed currently of thirteen monitoring sites (**Figure 1.2**). Criteria pollutants (O_3, SO_2, NO_x, CO, $PM_{2.5}$ and PM_{10}) and meteorological variables (temperature, relative humidity, wind speed, wind direction, pressure, precipitation and solar radiation) have been monitored continuously, with data summarized as hourly averages. The Mexican standards NOM-036-SEMARNAT-1993 and NOM-156- SEMARNAT-2012 are followed to accomplish the calibration, maintenance procedures, and quality assurance/quality control (QA/QC) protocols (Hernández Paniagua et al., 2017).

High concentrations of $PM_{2.5}$ have been measured in the MMA such that it is in nonattainment with respect to Mexican Air Quality Standard for $PM_{2.5}$ (Wakamatsu et al., 2017). In 2015, the MMA exhibited annual averages in its routine monitoring network in the range of 20–34 μg m^{-3} for $PM_{2.5}$; exceeding the $PM_{2.5}$ national air quality standards of 12 μg m^{-3} (Mancilla et al., 2019). Annual seasonal variation in the concentration of $PM_{2.5}$ has been observed, on which winter showed the highest concentration while the lowest concentration was observed on the first half of fall (Blanco et al., 2015).

$PM_{2.5}$ source apportionment studies based on a chemical characterization approach have been conducted in MMA (Martinez et al., 2012; Mancilla et al., 2015). Chemical profiles for $PM_{2.5}$ were determined for vehicular emissions (Mancilla and Mendoza, 2012). With these profiles and the use of organic molecular markers, it was found that anthropogenic sources had a major contribution to $PM_{2.5}$, and about 64% of the $PM_{2.5}$ were produced from vehicle exhausts. Another important source was meat cooking operations and biomass burning (Mancilla et al., 2016). Other sources that had been identified within the MMA are industrial activities, garbage burning, biogenic emissions and resuspended dust (Medina et al., 2015).

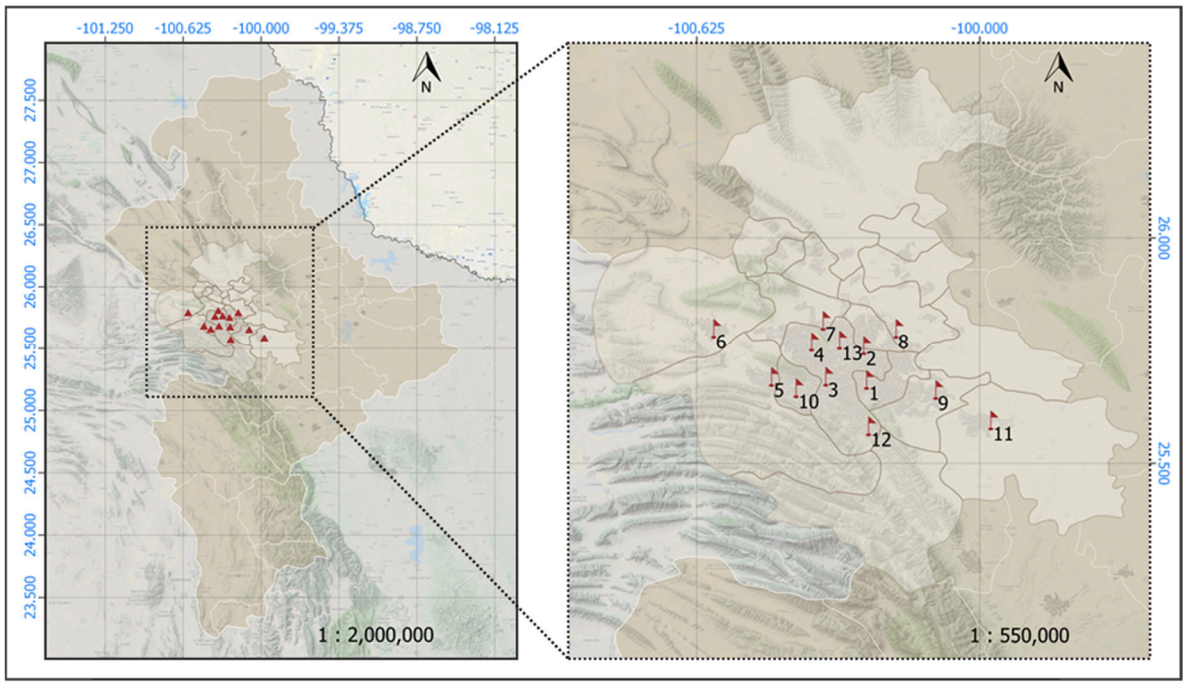


Figure 1.2. Location of monitoring sites that form part of the SIMA network in the MMA.

1. Southeast (La Pastora)
2. Northeast (San Nicolás)
3. Downtown (Obispado)
4. Northwest (San Bernabé)
5. Southwest (Santa Catarina)
6. Northwest 2 (García)
7. North (Escobedo)
8. Northeast 2 (Apodaca)
9. Southeast 2 (Juárez)
10. Southwest 2 (San Pedro)
11. Southeast 3 (Cadereyta)
12. South (Pueblo Serena)
13. Northeast 2 (C.U)

Chapter 2 FUNDAMENTALS

2.1 Remote sensing fundamentals

Remote sensing is the acquisition of information about an object or phenomenon without making direct physical contact with the object of interest (Conway, 1997). Remote sensing is defined, for the purpose of this thesis, as the measurement of aerosol properties using data acquired from satellites. The discussion to remote sensing in the current document will be limited of the earth's surface using optical signals.

Schowengerdt (2007) presents some concepts to understand the capabilities and limitations of remote sensing discipline:

2.1.1 Spectral regions and atmospheric windows

The major optical spectral regions used for earth remote sensing are shown in **Table 2.1.** They contain relatively transparent atmospheric "windows" through which (barring clouds in the non-microwave regions) the ground can be seen from space, and because effective radiation detectors are in them. Remote sensing of the earth's surface is limited to spectral transmission windows in the atmosphere.

Table 2.1. The primary spectral regions used in earth remote sensing (Schowengerdt, 2007).

name	wavelength range	radiation source	surface property of interest
Visible (VIS)	0.4–0.7 µm	solar	reflectance
Near InfraRed (NIR)	0.7–1.1 µm	solar	reflectance
Short Wave InfraRed (SWIR)	1.1–1.35 µm 1.4–1.8 µm 2–2.5 µm	solar	reflectance
MidWave InfraRed (MWIR)	3–4 µm 4.5–5 µm	solar, thermal	reflectance, temperature
Thermal or LongWave InfraRed (TIR or LWIR)	8–9.5 µm 10–14 µm	thermal	temperature
microwave, radar	1mm–1m	thermal (passive) artificial (active)	temperature (passive) roughness (active)

2.1.2 Radiation

There are two possible viewpoints for describing electromagnetic radiation. The particles that compose radiation are the *photons*. Similarly, it can be viewed as a wave propagating at the speed of light ($c \approx 3\times10^8$ ms^{-1} in a vacuum, a similar value in air). It is then characterized by its frequency (v), in s^{-1} or in hertz; or, equivalently, by its wavelength $\lambda = c/v$ (nm or μm) (Sportisse, 2010).

2.1.3 Types of remote sensing instruments

Passive remote sensors measure radiation naturally (sun as a source of energy or radiation) reflected or emitted from the ground, atmosphere, and clouds. The Visible, NIR, and SWIR regions (from 0.4 μm to about 3 μm) are the solar-reflective spectral range because the energy supplied by the sun at the earth's surface exceeds that emitted by the earth itself.
Active remote sensors employ an artificial source of radiation as a probe. The resulting signal that scatters back to the sensor characterizes either the atmosphere or the earth.

2.1.4 Resolution of remote sensing

Spatial Resolution: it is a measure of the area or size of the smallest dimension on the Earth's surface over which an independent measurement can be made by the sensor (size of the pixel). A measure of size of pixel is provided by the Instantaneous Field of View (IFOV), given among other by the technical specifications of the sensor system. The IFOV is the angular cone of visibility of the sensor. The IFOV is dependent on the altitude of the sensor above the ground level and the viewing angle of the sensor. For all types of scanners, the full cross-track angular coverage is called the Field Of View (FOV) and the corresponding ground coverage is called the Ground-projected Field Of View (GFOV) also called the swath width, or sometimes, the footprint of the sensor.
Spectral Resolution: Represents the spectral band width of the filter and the sensitiveness of the detector. A spectral band is a limited range of values the sensor is set to detect along a spectrum. The electromagnetic spectrum (**Figure 2.1**) is the range of frequencies (the spectrum) of electromagnetic radiation and their respective wavelengths and photon energies. It can be divided into wavelength regions:

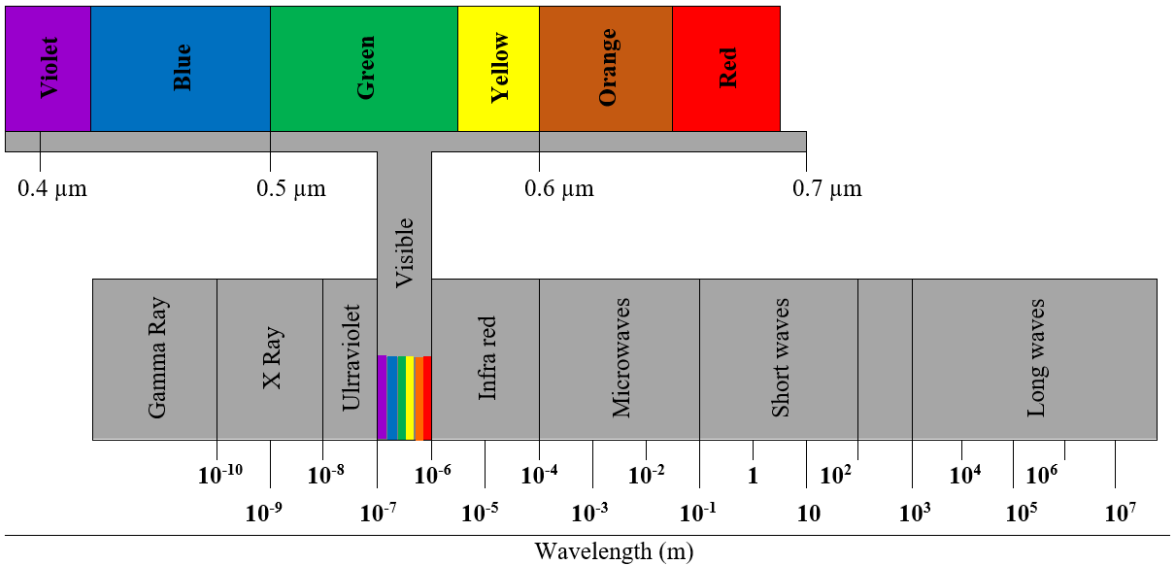


Figure 2.1. Electromagnetic spectrum.

Multispectral remote sensing is defined as the collection of reflected, emitted, or backscattered energy from an object or area of interest in multiple bands of electromagnetic spectrum, thus the spectral resolution of a sensor system is defined by its number of bands and the bandwidth of each band. While hyperspectral remote sensing involves data collection in hundreds of bands, with very small bandwidths.

Temporal Resolution: describes the ability to collect data of the same area of the Earth's surface at a given time period. The revisit period refers to the length of time it takes for a satellite to complete one entire orbit cycle. The interval between revisits depends solely on the particulars of the satellite orbit for sensors that have a fixed view direction (nadir).

Radiometric Resolution: describes the ability to discriminate very slight differences in energy. The finer the radiometric resolution of a sensor, the more sensitive it is to detecting small differences in reflected or emitted energy.

Optical Radiation Models: Passive remote sensing in the optical regime (visible through thermal) depends on two sources of radiation.

- In the visible to shortwave infrared, the radiation collected by a remote sensing system originates with the sun. Part of the radiation received by a sensor has been reflected at the earth’s surface and part has been scattered by the atmosphere, without ever reaching the earth surface.

- In the thermal infrared, thermal radiation is emitted directly by materials on the earth and combines with self-emitted thermal radiation in the atmosphere as it propagates upward.

2.2 Radiative transfer

2.2.1 Radiative transfer (processes)

Radiative transfer describes the interaction between radiation and matter (gases, aerosols, cloud droplets). The three key processes to be considered are: emission, absorption and scattering (Sportisse, 2010):

Emissions: the radiative energy emitted by a "body" at equilibrium depends on its temperature. The maximum of radiative energy that can be emitted per unit area of surface and per time unit defines the so-called *blackbody emission* (blackbody, defined as a substance that absorbs all incident radiation).

The plank radiation law quantifies the Energy, E, emitted by an ideal black body at an absolute temperature of T across a spectrum of wavelengths, λ:

$$E(\lambda, T) = \frac{2hc^2}{\lambda^5} \frac{1}{\exp\left(\frac{ch}{\lambda k_B T}\right) - 1} \qquad \textbf{Equation 2.1}$$

where $k_B = 1.38 \times 10^{-23}$ JK^{-1} is the Boltzmann's constant, *c is* the speed of light in the medium and $h = 6.626 \times 10^{-34}$ Js is the Plank's constant. Emission is expressed typically in terms of W m^{-2} nm^{-1}.

Absorption: corresponds to a decrease of the radiative energy in the incident direction. A fraction of the incident radiation is absorbed along the path of propagation in a medium (here the atmosphere). The Beer-Lambert law (also referred to as the Beer-Lambert-Bouguer law) governs the reduction in the radiation intensity I_λ at wavelength λ (Dimov, 2017).

Scattering: corresponds to a dispersion redistribution of the radiative energy in all directions. Scattering is dependent upon the particle size in relation to the wavelength. Four scattering regimes are usually distinguished: i. the Rayleigh scattering, also referred to as molecular scattering (typically for gases, where the molecule size is smaller than wavelength); ii. the scattering represented by the optical geometry's laws (typically for liquid water drops, where the particle size is larger than the wavelength.); iii. the Mie scattering (for aerosols), which

occurs when the radiation wavelength is of similar size of the scattering particles; and iv. nonselective scattering, that occurs when the particle size is much larger than the wavelength.

For atmospheric aerosols the scattering calculation is given by the *Mie theory*: the intensity of the scattered radiation (I) in the direction with a scattering angle (θ) to the incident direction, at a distance (r) between the particle and the observer, is

$$I(\theta, r) = I_0 \frac{\lambda^2 (i_1 + i_2)}{4\pi^2 r^2} \qquad \textbf{Equation 2.2}$$

where I_0 is the light intensity before the interaction with the particle. i_1 and i_2 are the intensity Mie parameters.

2.2.2 Radiative transfer equation - visible radiation

Beer's Law describes the fundamental one-dimensional (1D) radiative transfer in the visible part of the spectrum:

$$I(\lambda) = I_0(\lambda) e^{-\alpha(\lambda) Z} \qquad \textbf{Equation 2.3}$$

where α is the extinction coefficient of the atmosphere (m^{-1}), λ is the wavelength, and z is the physical path in meters. I_0 is the source radiance and I is the measured radiance after passing through the path (z). If the path is vertical, z is the altitude of the source. In atmospheric radiative transfer, the path is rarely vertical (Hoff and Christopher, 2009).

The extinction is a sum of the contribution of electronic, vibrational, and rotational transitions in the molecule interacting with radiation. The extinction properties of a particle are determined by its *extinction efficiency* (Q_λ^{ext}), defined as the ratio of the scattering cross section to the interception surface (Sportisse, 2010)**:**

$$Q_\lambda^{ext} = Q_\lambda^a + Q_\lambda^s = \frac{\sigma_\lambda^a + \sigma_\lambda^s}{\pi (D/2)^2} \qquad \textbf{Equation 2.4}$$

where Q_λ^a and Q_λ^s are the absorption and scattering efficiencies, respectively. $D = 2r$ is the particle diameter, r is the radius of the sphere. σ_λ^a is the absorption cross-section and σ_λ^s is the scattering cross-section.

For aerosols, scattering and absorption (which sum to extinction) have spectral dependence governed by the particle size, particle complex index of refraction, and have an angular

dependence between the source and scattered radiation. Then the extinction, absorption and scattering efficiencies, are functions of the *size parameter* (α_λ):

$$\alpha_\lambda = \frac{\pi D}{\lambda} \qquad \textbf{Equation 2.5}$$

and the properties of the medium defined by the particle (due to its chemical composition), described by its *complex refractive index*:

$$m_\lambda = n_\lambda + ik_\lambda, \;\; i^2 = -1 \qquad \textbf{Equation 2.6}$$

The real part n_λ is related to scattering while the imaginary part k_λ is related to absorption (Fu, 2015).

2.3 Aerosol properties

2.3.1 Aerosol Optical Depth

The wavelength dependent aerosol optical depth (AOD, or τ_a) is the integral of the aerosol extinction due to scattering and absorption in a vertical column. It is the total optical depth of the atmosphere corrected for absorption and scattering of gases in the atmosphere (e.g., Rayleigh scattering for O_3, nitrogen dioxide [NO_2], and water absorption) (Hoff and Christopher, 2009). AOD is a measure of the extinction of the solar beam by aerosols (NOAA, 2019). It is the degree to which aerosols prevent the transmission of light by absorption or scattering of light. AOD is defined as the integrated extinction coefficient over a vertical column of unit cross section (Barry and Hall-McKim, 2014).

Because the path through the entire atmosphere has a pressure-dependence (α), the optical depth (τ) of the atmosphere down to height z is defined as

$$\tau_\lambda = \int_Z^{TOA} \alpha(\lambda, z')dz' \qquad \textbf{Equation 2.7}$$

At the surface, $z = 0$ and τ is represented by τ^*. At the top of the atmosphere (TOA) $\tau = 0$.

2.3.2 Single Scattering Albedo

The single scattering albedo (SSA) is the fraction of light that is scattered compared to the total extinction (scattering plus absorption) optical depth (Hsu, n.d.). The SSA, that is, the ratio of scattering and extinction cross-sections (Moosmüller and Sorensen, 2018) can be written in terms of efficiencies as:

$$SSA = \frac{Q_{sca}}{Q_{ext}} = \frac{Q_{sca}}{Q_{sca} + Q_{abs}}$$ **Equation 2.8**

A value of 1 indicates that an aerosol only scatters light, and values closer to 0 indicate that an aerosol mostly absorbs light (Hsu, n.d.)..

2.3.3 Ångström Exponent

The Ångström Exponent (AE) is a measure of how the AOD changes relative to the various wavelengths of light (known as 'spectral dependence'.) The AE, which is the slope of the logarithm of aerosol optical depth (τ_a) versus the logarithm of wavelength, is commonly used to characterize the wavelength dependence of τ_a and to provide some basic information on the aerosol size distribution (Eck et al., 1999):

$$AE = \frac{d\ln\tau_a}{d\ln\lambda} = -\frac{\ln\left(\frac{\tau_{a2}}{\tau_{a1}}\right)}{\ln\left(\frac{\lambda_2}{\lambda_1}\right)}$$ **Equation 2.9**

where τ_{a1} is the AOD at wavelength λ_1 and τ_{a2} is the AOD at wavelength λ_2.

2.4 AOD - $PM_{2.5}$ relationship

The AOD-$PM_{2.5}$ relationship can be defined by the following equation (Koelemeijer et al., 2006; Hoff and Christopher, 2009):

$$AOD = PM_{2.5}\, H\, f(RH)\, \frac{3Q_{\text{ext,dry}}}{4\rho r_{\text{eff}}} = PM_{2.5} H\, S$$ **Equation 2.10**

where *f*(*RH*) is the ratio of ambient and dry extinction coefficients, ρ is the aerosol mass density (g m^{-3}), $Q_{ext,dry}$ is the Mie extinction efficiency, and r_{eff} is the particle effective radius (the ratio of the third to second moments of the size distribution). *S* is the specific extinction efficiency (m^2 g^{-1}) of the aerosol at ambient relative humidity (RH).

The assumptions made for correlating AOD with ground-based $PM_{2.5}$ mass are: cloud-free skies, well-mixed boundary layer of height (*H*) with no overlying aerosols, and aerosols that have similar optical properties (Hoff and Christopher, 2009).

2.5 MODIS aerosol modeling

The Moderate Resolution Imaging Spectroradiometer (MODIS) Atmospheres data product suite includes three algorithms for retrieving AOD: i) Dark Target (DT) algorithms over land, ii) DT over-ocean (water) algorithm, and iii) Deep Blue (DB) algorithms (over land). All

three have been refined in the "Collection 6" (C6) MODIS reprocessing. Algorithms over land are of particular interest in this research.

Fundamentals of four AOD MODIS aerosol datasets used in this research are described in this section, including: AOD retrieval procedure of both 10k and 3k product from the DT algorithm, AOD retrieval procedure from the DB algorithm, and the procedure to estimate AOD used in the "merged" data set released in C6.

2.5.1 MODIS - Dark Target Algorithm (10 km)

2.5.1.1 Theoretical background

The upward reflectance (normalized solar radiance), at the top of the atmosphere (TOA), is a function of successive orders of radiation interactions, within the coupled surface-atmosphere system. The TOA angular (θ_0, θ and Φ = solar zenith, view zenith and relative azimuth angles) spectral reflectance ($\rho_\lambda(\theta_0, \theta, \Phi)$) at a wavelength results from (ATBD, 2013):

- The scattering of radiation within the atmosphere without interaction with the surface (known as the 'atmospheric path reflectance'),
- The reflection of radiation off the surface that is directly transmitted to the TOA (the 'surface function'), and
- The reflection of radiation from outside the sensor's field of view (the 'environment function').

The environment function is neglected so that to a good approximation can be stated as:

$$\rho_\lambda^*(\theta_0, \theta, \phi) = \rho_\lambda^a(\theta_0, \theta, \phi) + \frac{T_\lambda(\theta_0)T_\lambda(\theta)\rho_\lambda^s(\theta_0,\theta,\phi)}{1-s_\lambda\rho_\lambda^s(\theta_0,\theta,\phi)} \qquad \textbf{Equation 2.11}$$

where ρ_λ^a is the atmospheric 'path reflectance', $T_\lambda(\theta_0)$ and $T_\lambda(\theta)$ is the downward and upward atmospheric transmissions (reciprocity implied), s_λ is the atmospheric backscattering ratios, and $\rho^s{}_\lambda$ is the angular spectral surface reflectance. Except for the surface reflectance, each term on the right-hand side of Equation 2.11 is a function of the aerosol type and loading (τ) (ATBD, 2013).

Simultaneously inverting the aerosol and surface information in three channels (0.47 μm, 0.65 μm and 2.11 μm) yields three parameters: AOD τ ($\tau_{0.55}$), the weighted sum η ($\eta_{0.55}$) of

the spectral reflectance from a combination of fine and coarse dominated aerosol models and the surface reflectance ($\rho^{s}_{2.11}$). It is assumed that the 2.11 μm channel contains information about coarse mode aerosol as well as the surface reflectance.

In the Dark Target algorithm (DT), Equation 2.11 is rewritten, the calculated spectral total reflectance ρ^{*}_{λ} at the top of the atmosphere is the weighted sum (η) of the spectral reflectance from a combination of fine and coarse dominated aerosol models, i.e.

$$\rho_{\lambda}^{*} = \eta\rho_{\lambda}^{*f} + (1-\eta)\rho_{\lambda}^{*c} \qquad \textbf{Equation 2.12}$$

where ρ^{*f}_{λ} and ρ^{*c}_{λ} are each composites of surface reflectance ρ^{s}_{λ} and atmospheric path reflectance of the separate aerosol models. That is:

$$\rho_{\lambda}^{*f} = \rho_{\lambda}^{af} + F_{d\lambda}^{f}T_{\lambda}^{f}\rho_{\lambda}^{s}/\left(1 - s_{\lambda}^{f}\rho_{\lambda}^{s}\right) \qquad \textbf{Equation 2.13}$$

and

$$\rho_{\lambda}^{*c} = \rho_{\lambda}^{ac} + F_{d\lambda}^{c}T_{\lambda}^{c}\rho_{\lambda}^{s}/(1 - s_{\lambda}^{c}\rho_{\lambda}^{s}) \qquad \textbf{Equation 2.14}$$

where ρ_{λ}^{af} and ρ_{λ}^{ac} are the fine and coarse model atmospheric path reflectance, $F_{d\lambda}^{f}$ and $F_{d\lambda}^{c}$ are normalized downward fluxes for zero surface reflectance, T_{λ}^{f} and T_{λ}^{c} represent upward total transmission into the satellite field of view, and s_{λ}^{f} and s_{λ}^{c} are atmospheric backscattering ratios. Note the angular and τ dependence of some of the terms: $\rho^{a} = \rho^{a}(\tau, \theta_0, \theta, \Phi)$, $F = F(\tau, \theta_0)$, $T = T(\tau, \theta)$, $s = s(\tau)$ and $\rho^{s} = \rho^{s}(\theta_0, \theta, \Phi)$. Whereas the other terms are a function of the aerosol model and are contained within the lookup tables (which will be described later in this document).

DT should retrieve AOD within the Expected Error ($EE_{DT} = 0.05 \pm 0.15\tau$), which was determined on global average (Levy et al., 2010, 2013).

2.5.1.2 Masking procedures

Because the MODIS DT over-land retrieval algorithm employs only three channels (and suffers from surface and other contaminations), several masking procedures are performed in order to minimize residual contamination over most situations (ATBD, 2013).

First, the measured reflectance (from L1B) is organized into nominal (at nadir) 10 km by 10 km boxes (corresponding to 20 by 20, or 40 by 40 pixels, depending on the channel). The

400 pixels in the box are evaluated pixel by pixel to identify whether the pixel is suitable for aerosol retrieval (Levy et al., 2013). The masking procedure includes to filter out: i) shallow water pixels contaminated by underwater sediment (Li et al., 2003), ii) ice/snow pixels (Li et al., 2005), iii) bright land scenes, glint over water, etc., and iv) cloudy pixels (Martins et al., 2002). Once these truly unsuitable pixels are removed, the brightest 50% and darkest 20% over land are discarded. Furthermore, the retrieval performs corrections for absorption by atmospheric gases (water vapor, ozone, etc.). Finally, the pixels that remain, after all de-selection and gas corrections are applied, are averaged, yielding a final set of mean spectral reflectance (set of "observed" top-of-atmosphere (TOA) spectral reflectance) (Levy et al., 2013).

2.5.1.3 Surface Scheme

The surface scheme in the DT algorithm in Collection 6 (C6) for dark target pixels is based on the SWIR–VIS surface relationships, considering the band correlation based on the normalized difference vegetation index $NDVI_{SWIR}$ and the scattering angle (Θ). Once the pixels are discarded, and the remaining 30%, or at least 12 pixels inside the box are used for $NDVI_{SWIR}$ calculation:

$$NDVI_{SWIR} = \frac{(\rho_{1.24\mu m} - \rho_{2.12\mu m})}{(\rho_{1.24\mu m} + \rho_{2.12\mu m})} \qquad \textbf{Equation 2.15}$$

Where $\rho^m_{1.24}$ and $\rho^m_{2.12}$ are the measured reflectance by MODIS at wavelengths 1.24 and 2.12 μm, respectively. Following, pixels are classified into three categories based on the $NDVI_{SWIR}$ (Levy et al., 2013):

$$NDVI_{SWIR} < 0.25$$

$$0.25 < NDVI_{SWIR} < 0.75$$

$$NDVI_{SWIR} > 0.75$$

and the f_1 and f_2 linear equations are applied (**Equation 2.16 and 2.17, respectively**) to those three categories. The 0.65 μm surface reflectance ($\rho^s_{0.65}$) is calculated from that in 2.11 μm ($\rho^s_{2.12}$), followed by calculating 0.47 μm ($\rho^s_{0.47}$) from the 0.65 μm, i.e. (ATBD, 2013):

$$\rho^s_{0.65} = f_1(\rho^s_{2.12}) = \rho^s_{2.12} * slope_{0.65/2.12} + yint_{0.65/2.12} \qquad \textbf{Equation 2.16}$$

$$\rho^{s}_{0.47} = g(\rho^{s}_{0.65}) = \rho^{s}_{0.65} * slope_{0.47/0.65} + yint_{0.47/0.65} \quad \textbf{Equation 2.17}$$

where

$$slope_{0.65/2.12} = slope^{NDVI_{SWIR}}_{0.65/2.12} + 0.002\Theta - 27, \quad \textbf{Equation 2.18}$$

$$yint_{0.65/2.12} = -0.00025\Theta + 0.033, \quad \textbf{Equation 2.19}$$

$$slope_{0.47/0.65} = 0.49, \quad \textbf{Equation 2.20}$$

and

$$yint_{0.47/0.65} = 0.005 \quad \textbf{Equation 2.21}$$

where in turn

$$slope^{NDVI_{SWIR}}_{0.65/2.12} = 0.58; NDVI_{SWIR} < 0.25,$$

$$slope^{NDVI_{SWIR}}_{0.65/2.12} = 0.48; NDVI_{SWIR} > 0.75$$

$$slope^{NDVI_{SWIR}}_{0.65/2.12} = 0.58 - 0.2(NDVI_{SWIR} - 0.25); 0.25 \leq NDVI_{SWIR} \leq 0.75$$

Over bright surfaces, only one band at the 0.47 μm wavelength is used and continental model is assigned for aerosol retrieval during the lookup table (LUT) calculation, therefore the quality of AOD over bright surfaces is deemed poor. The retrieval process performance is then evaluated as "No confidence" and the worst quality assurance flag (QA) is assigned (QA = 0).

2.5.1.4 Aerosol optical models

It is assumed that a small set of aerosol types (five) and loadings can describe the range of global aerosol. The five aerosol types are comprised of two or more log-normal modes. These may be "fine-dominated" or "coarse-dominated" depending on which mode dominates the size distribution.

Since the land algorithm tries to deal with larger surface uncertainty with only three spectral bands, both fine and coarse model aerosol types must be prescribed as a function of season and location. The land algorithm must assign the fine aerosol model before the retrieval. (Levy et al., 2013).

Comparisons between fine modes and coarse modes, in addition to the criteria for performing the cluster analysis for identifying three fine aerosol types are presented in **Table 2.2.**

2.5.1.5 Lookup tables

The DT algorithm takes the set of "observed" top-of-atmosphere (TOA) spectral reflectance, and tries to match to values within lookup tables (LUTs). The lookup table is used to determine the conditions that best mimic the MODIS-observed spectral reflectance ρ^{m}_{λ}, and retrieve the associated aerosol properties (including τ and η) (Levy et al., 2013).

LUT contains pre-computed optical properties of the five types of aerosol (**Table 2.3**) at four discrete wavelengths (0.466, 0.553, 0.644 and 2.119 μm, representing MODIS channels 3, 4, 1 and 7, respectively) for several values of aerosol total loadings, and for a variety of geometries.

For discrete optical depths (described by the τ or τ at 0.55 μm), the spectral scattering and extinction coefficients are calculated using a Mie code (Wiscombe, 1980) for assumed spherical particles or using the Dubovik T-matrix code (Dubovik and King, 2000; Dubovik et al., 2002b) for assumed spheroid model (Dust). The Radiative Transfer (RT) code (Evans and Stephens, 1991) is used to simulate TOA reflectance and total fluxes (ATBD, 2013).

2.5.1.6 Quality Assurance Confidence

At selected stages, the algorithm evaluates the retrieval process performance (Hubanks, 2012). Quality assurance flags in DT C6 have four different levels: ideal performance is given the highest QA "confidence" value (QA = 3), with good, marginal and no confidence retrievals given QA values of 2, 1 and 0, respectively. Many QA flags are included in C6 related to product quality, retrieval processing, processing path and input data quality. A detailed list of the product quality and retrieval processing QA flags over land is included in Table C1 in Levy et al. (2013).

Finally, **Figure 2.2** illustrates the main steps of the Dark Target over land Collection 6.

Table 2.2. Aerosol modes, characteristics and differences

	"fine" modes (FM)	**"coarse" modes (CM)**
Effective radius (r_{eff})	FM have $r_{eff} < 0.25$ μm	CM have $r_{eff} > 1.0$ μm
Geometric shapes	FM are assumed to be spherical particles	CM are assumed to be non-spherical particles (spheroids)
Aerosol models	'non-absorbing' aerosol model 'absorbing' aerosol model 'neutral' aerosol model	This model is based on dust which is extremely absorbing.
Aerosol types	1.Continental[1] 2. Moderately absorbing (default) 3. Strongly absorbing smoke 4. Weakly absorbing (Urban/Industrial)	5. Spheroid (Dust)
Cluster analysis	Moderately absorbing (default) - This cluster represents the middle combination (SSA and ASYM) values ('neutral' aerosol model). - Neutral SSA~0.90 model (presumably background, forest smoke and developing world aerosol) - The Neutral aerosol model was set as the default, and would be overwritten only if clear dominance of one of the other two aerosol types was observed. Strongly absorbing (smoke) - This cluster represents the lowest SSA and lowest ASYM ('absorbing' aerosol model). - Highly absorbing SSA~0.85 model (presumably savanna/grassland smoke aerosol) Weakly absorbing (Urban/Industrial) - This cluster represents the combination of highest SSA and highest ASYM ('non-absorbing' aerosol model), - The non-absorbing SSA~0.95 model (presumably urban/industrial aerosol)	Spheroid (Dust) DT assumed that the spheroid model represents coarse ('dust') aerosol. This assumption is based on Dubovik et al. (2006), where they observed that the sites which contributing to spheroid data were primarily those known to be in dust regions.

Based on Levy et al. (2013) and ATBD (2013). [1] Continental model is not derived from a cluster procedure, and it is only derived from "Procedure B: Alternative Retrieval for Brighter Surfaces". [2]Two optical parameters (SSA at 0.67 μm and the asymmetry parameter at 0.44 μm) are used for clustering three fine aerosol types: Presumably the SSA would differentiate between non-absorbing aerosols (such as urban/industrial pollution – (Remer et al., 1998; Dubovik et al., 2002a) and much more absorbing aerosols (such as savanna burning smoke – (Ichoku et al., 2003; Dubovik et al., 2002a), and the asymmetry parameter (ASYM) at 0.44 μm would help differentiate between the phase functions of different (mainly fine mode – size similar to wavelength) aerosols. (ATBD, 2013)

Table 2.3. Lookup table used in DT algorithm (ATBD, 2013)

Index	Model	Mode	r_v (μm)	σ	V_0 (μm³/μm²)	Refractive Index: k		ω₀/ g (0.47/0.55/0.65/2.1 μm) For $\tau_{0.55}$ = 0.5
1	Continental	Soluble	0.176	1.090	3.050	1.53-0.005i; 1.53-0.006i; 1.53-0.006i; 1.42-0.010i;	0.47 μm 0.55 μm 0.65 μm 2.11 μm	0.90/0.89/0.88/0.67 0.64/0.63/0.63/0.79
		Dust	17.600	1.090	7.364	1.53-0.008i; 1.53-0.008i; 1.53-0.008i; 1.22-0.009i;	0.47 μm 0.55 μm 0.65 μm 2.11 μm	
		Soot	0.050	0.693	0.105	1.75-0.045i; 1.75-0.044i; 1.75-0.043i; 1.81-0.050i;	0.47 μm 0.55 μm 0.65 μm 2.11 μm	
2	Moderately Absorbing (Smoke)	Accum	$0.0203\tau + 0.145$	$0.1365\tau + 0.3738$	$0.1642\ \tau^{0.7747}$	$1.43 - (+0.002\tau + 0.008)i$;		0.93/0.92/0.91/0.87 0.68/0.65/0.61/0.68
		Coarse	$0.3364\tau + 3.101$	$0.098\tau + 0.7292$	$0.1482\tau^{0.6846}$	$1.43 - (+0.002\tau + 0.008)i$;		
3	Strongly Absorbing (Smoke)	Accum	$0.0096\tau + 0.1335$	$0.0794\tau + 0.3834$	$0.1748\ \tau^{0.8914}$	$1.51 - 0.02i$		0.88/0.87/0.85/0.70 0.64/0.60/0.56/0.64
		Coarse	$0.9489\tau + 3.4479$	$0.0409\tau + 0.7433$	$0.1043\tau^{0.6824}$	$1.51 - 0.02i$		
4	Weakly-absorbing (Urban/Industrial)	Accum	$0.0434\tau + 0.1604$	$0.1529\tau + 0.3642$	$0.1718\ \tau^{0.8213}$	$1.42 - (0.0015\tau + 0.007)i$;		0.95/0.95/0.94/0.90 0.71/0.68/0.65/0.64
		Coarse	$0.1411\tau + 3.3252$	$0.1638\tau + 0.7595$	$0.09434\tau^{0.6394}$	$1.42 - (0.0015\tau + 0.007)i$;		
5	Spheroid/ Dust	Accum	$0.1416\ \tau^{-0.0519}$	$0.7561\ \tau^{0.148}$	$0.0871\tau^{1.026}$	$1.48\tau^{-0.021} - (0.0025\tau^{0.132})i$; $1.48\tau^{-0.021} - 0.002i$; $1.48\tau^{-0.021} - (0.0018\tau^{0.08})i$; $1.46\tau^{-0.040} - (0.0018\tau^{0.30})i$;	0.47 μm 0.55 μm 0.65 μm 2.11 μm	0.94/0.95/0.96/0.98 0.71/0.70/0.69/0.71
		Coarse	2.2	$0.554\tau^{-0.0519}$	$0.6786\tau^{1.0569}$	$1.48\tau^{-0.021} - (0.0025\tau^{0.132})i$; $1.48\tau^{-0.021} - 0.002i$; $1.48\tau^{-0.021} - (0.0018\tau^{0.08})i$; $1.46\tau^{-0.040} - (0.0018\tau^{0.30})i$;	0.47 μm 0.55 μm 0.65 μm 2.11 μm	

Listed for each model are the individual lognormal modes, and the final **ω₀**, at different wavelengths. Listed for each mode are the mean radius $\mathbf{r_v}$, standard deviation σ of the volume distribution, and total volume of the mode, V_0. The complex refractive index is assumed for all wavelengths (0.47, 055. 0.65 and 2.1 μm), unless otherwise noted. The Absorbing and Moderately absorbing model parameters (r_v, σ and k) are defined for $\tau \leq 2.0$; for $\tau > 2.0$, $\tau = 2.0$ is assumed. Likewise, the Non-absorbing and Spheroid model parameters are defined for $\tau \leq 1.0$. V_0 (for all models) is define for all τ.

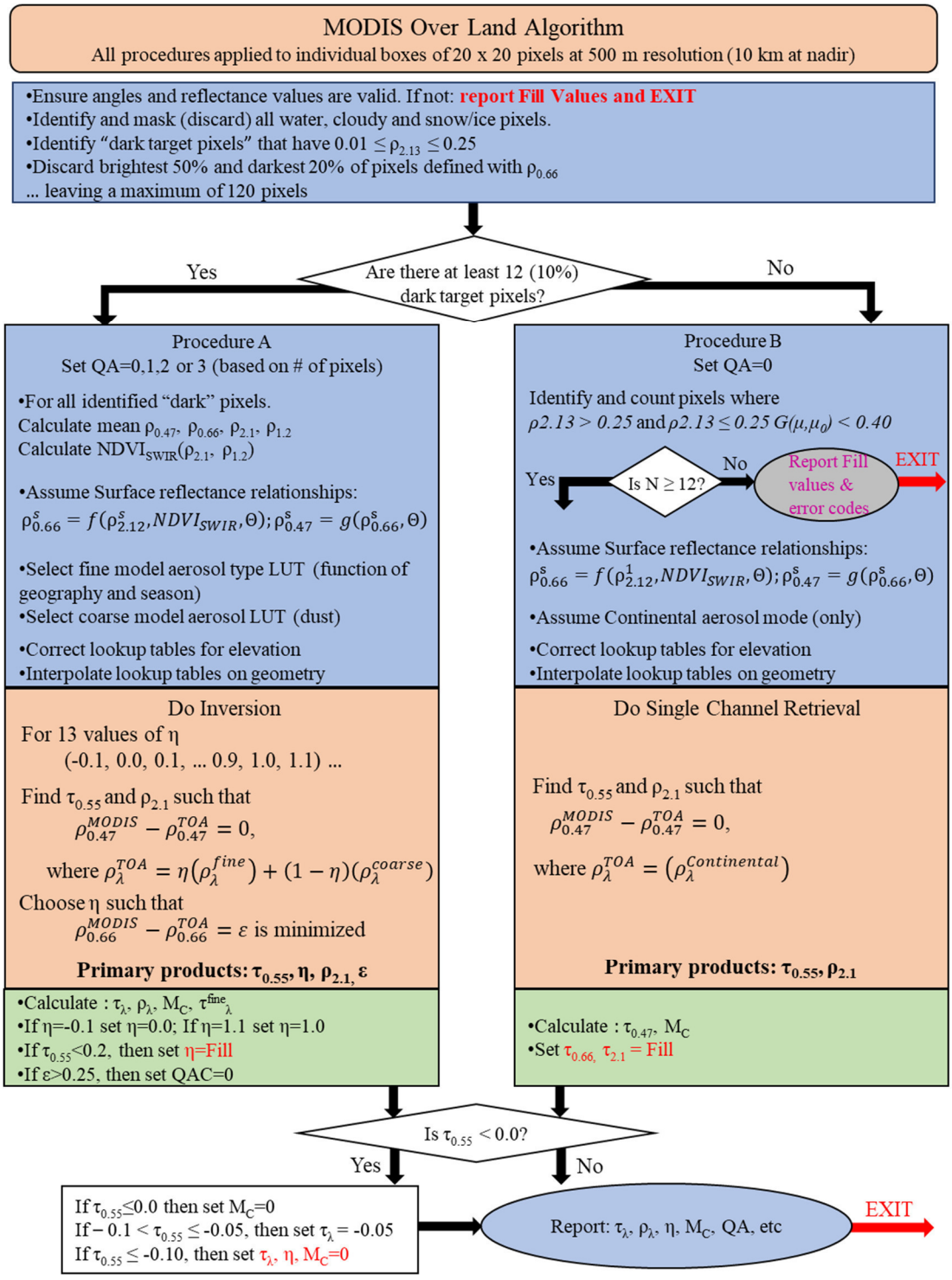


Figure 2.2. Flowchart of the Dark Target algorithm over land (modified from Levy et al. [2013]).

2.5.2 MODIS - Dark Target Algorithm (3 km product)

2.5.2.1 Retrieval procedure

The 10 km resolution is not fine enough to resolve local variability, especially near and within cities where most of the human population lives. Therefore, in recent years, the air quality community has been advocating for higher resolution aerosol retrieval data to monitor and model air pollution. The 3 km product was designed to provide insight into the aerosol situation on a focused local basis (Levy et al., 2013).

For C6, data product at 3 km resolution is archived as a separate Level 2 aerosol product "MxD04_3K". Compared to the standard 10 km algorithm, the inputs at 3 km algorithm are identical, as are the masking procedures. The same aerosols and surface assumptions are used, the same methodology and structure, the same lookup tables, the same numerical inversion and the same criteria to determine a good fit (Remer et al., 2013).

In the 3 km retrieval, the 0.5 km input reflectances pixels are organized in retrieval boxes of 6 × 6 arrays of 36 pixels, versus 20 × 20 pixels for the 10 km algorithm (Remer et al., 2013). Therefore, the sorting and discarding processes are slightly different: pixels that might be discarded during the sorting and discarding procedure at 10 km might be kept at 3 km. This has the potential to make the 3 km product noisier than at 10 km (Levy et al., 2013).

2.5.2.2 Quality Assurance Confidence

The data set includes an integer Quality Flag (Land-Ocean Quality Flag) that designates each 3 km retrieval as "3", "2", "1" or "0". The same meaning of these quality flags meaning apply to the 3 km product as to the original 10 km product (with the highest confidence (QA = 3), with good (QA = 2), marginal (QA = 1), and no confidence retrievals (QA = 0)) (Remer et al., 2013).

There is a possibility that the 3 km product introduces additional noise, especially over land where pixels representing inhomogeneous surfaces would have been eliminated in the deselection process of the larger box, but are included in the 3 km retrieval. However, this tendency towards noise may be mitigated by the slightly more stringent requirements in deselection and the minimum number of pixels (5 pixels) needed to represent the box (Remer et al., 2013).

2.5.3 MODIS – Deep Blue Algorithm (10 km)

2.5.3.1 Theoretical background

The aerosol properties over land are determined according to the model used by Dave (1972):

$$R(\mu, \mu_0, \emptyset) = \frac{\pi I(\mu,\mu_0,\emptyset)}{\mu_0 F_0} \qquad \textbf{Equation 2.22}$$

where R is the normalized radiance (or apparent reflectance), F_0 is the extraterrestrial solar flux, I is the radiance at the top of the atmosphere, μ is the cosine of the view zenith angle, μ_0 is the cosine of the solar zenith angle, and $\emptyset$ is the relative azimuth angle between the direction of propagation of scattered radiation and the incident solar direction.

Inverting surface information and aerosol information from TOA reflectances at 0.412, 0.47, and 0.65 μm yields three parameters: Aerosol Optical Depth (also called Aerosol Optical Thickness), Single Scattering Albedo and the Ångström exponent.

DB should retrieve AOD within the Expected Error ($EE_{DT} = \pm 0.03 + 0.20\tau$), which was determined on global average (Sayer et al., 2014)

2.5.3.2 Masking procedures

Similar to DT, only clear pixels are included for performing the deep blue algorithm retrieval. The masking procedure in DB includes to filter out: i) clouds, cloud edges and thick aerosol plumes; and ii) seasonal/transient inland water. Once this truly unsuitable pixels are removed, a correction for the Rayleigh scattering is made and the values are averaged into a daily mean for the given grid. Atmospheric correction is necessary to better determine the spectral surface reflectance relationship in the blue wavelength range of 0.412 to 0.490 μm, which is vital for DB retrieval.

2.5.3.3 Surface Scheme

To determine the surface reflectance for the 0.412, 0.47, and 0.65 μm channels there are three different methods that can be applied, and the decision of which one to use will depend on the TOA reflectance at shortwave-infrared (SWIR) or near-infrared (NIR) wavelengths and the normalized difference vegetation index (NDVI). These methods are:

i) by a dynamic surface reflectance approach, used over arid and semiarid regions. A surface database was compiled based upon the minimum reflectivity method at the resolution of 0.1°

latitude × 0.1° longitude for each season using MODIS TOA reflectances at 0.412, 0.47 and 0.65 μm (see Hsu et al. [2004] for details). The derived surface reflectance database depends upon the scattering angle, NDVI, and season.

ii) based upon its geolocation using a pre-calculated surface reflectance database created from the MODIS measurements, used over general vegetation.

Over vegetated land surfaces, aerosol properties are retrieved based on the SWIR–VIS surface relationships (defined by Equation 2.15) described in the DT surface scheme.

The spectral surface reflectance relationships can vary depending upon the land cover type (i.e., naturally vegetated area and cropland) and season. For MODIS data, the surface reflectances is estimated at visible channels based upon the TOA reflectances at 2.1 μm ($\rho_{2.12}$) and land cover type using the following expressions:

$$ESR_{0.65} = \mathrm{a} + \mathrm{b} * \rho_{2.12} + \mathrm{c} * (\rho_{2.12})^2 \qquad \textbf{Equation 2.23}$$

$$ESR_{0.47} = \mathrm{d} + \mathrm{e} * ESR_{0.65} \qquad \textbf{Equation 2.24}$$

where *ESR* is the estimated surface reflectance and is calculated using the coefficients shown in **Table 2.4.**

Table 2.4. Surface Reflectance Coefficients[1] for MODIS 0.47 and 0.65 μm (Hsu et al., 2013)

	$NDVI_{SWIR} < 0.35$				
	$\rho_{0.65}/\rho_{2.12}$			$\rho_{0.47}/\rho_{0.65}$	
Months[2]	a,	b,	c,	d,	e
DJF/MAM	6.2828	0.1658	0.0	2.6884	0.2751
JJA	5.2935	0.2077	0.0	0.2451	0.5442
SON	-2.2642	0.6781	0.0	1.2493	0.3576

	$NDVI_{SWIR} < 0.35$				
	$\rho_{0.65}/\rho_{2.12}$			$\rho_{0.47}/\rho_{0.65}$	
Months	a,	b,	c,	d,	e
DJF/MAM	-0.9766	0.6213	0.0	0.9126	0.3982
JJA	-0.1187	0.5036	0.0	-0.0736	0.5345
SON	-1.2799	0.6161	0.0	1.2724	0.2039

[1] Coefficients (a, b, c, d, and e) were determined by a least squares fitting to the derived benchmark surface reflectance data over the AERONET sites. The Aerosol Robotic Network (AERONET [Holben et al., 1998]) project is an optical ground-based remote sensing aerosol monitoring network and data archive, established by NASA and expanded by federation with many non-NASA institutions. The network hardware consists of identical automatic Sun-sky scanning spectral radiometers owned by national agencies and universities (Duvobick et al., 2000). The network provides globally distributed observations of spectral AOD, aerosol size distributions, aerosol radiative properties etc., for aerosol research including validation of satellite retrievals (Kaufman et al., 2002).

[2] DJF/MAM corresponds to December, January, February / March, April, May. JJA corresponds to June, July and August. SON corresponds to September, October and November.

iii) a combination of the first two approaches, used over urban/built-up and transitional regions.

In urban/built-up and transitional zones, the surface reflectances at visible wavelengths do not exhibit simple and well-behaved relationships with NDVI as those for the densely vegetated regions and thus a hybrid approach was developed for these types of land cover by combining the Deep Blue surface database with the angular shapes of a surface bidirectional reflectance distribution function (BRDF) derived using AERONET measurements.

The selection of which method is used depends on the TOA reflectance at SWIR or near-infrared NIR wavelengths (i.e., 2.1 μm for MODIS) and the normalized difference vegetation index (NDVI), which is defined as

$$NDVI = \frac{(\rho_{0.86} - \rho_{0.65})}{(\rho_{0.86} + \rho_{0.65})} \qquad \textbf{Equation 2.25}$$

where $\rho_{0.65}$ and $\rho_{0.86}$ correspond to the TOA reflectance measured at 0.65 and 0.86 μm, respectively.

Also, to account for the effect of vegetation changes, the resulting surface reflectances were divided into three different groups according to their NDVI values: NDVI ≤ 0.19, 0.19 < NDVI ≤ 0.24, and NDVI > 0.24. Regression lines were estimated as a function of scattering angle using a second-order polynomial fit for each NDVI group and each season to obtain the shapes of surface BRDFs (Hsu et al., 2004). Finally, these derived angular shapes are combined with surface reflectance values from the Deep Blue surface database at 135° scattering angle to compute surface reflectance for aerosol retrievals over urban/built-up and transitional zones.

2.5.3.4 Aerosol optical models

The procedures to retrieve aerosols using DB algorithm are summarized as following.

i) Retrieval for Dust Aerosols (Hsu et al., 2004):

For heavy dust loading, the algorithm executes a three-channel technique (0.412, 0.490 and 0.670 μm), in which the relationship of apparent reflectances between two channels is a function of τ and ω_0.

For thick dust layers, the DB algorithm uses the maximum-likelihood method with lookup tables to solve for the values of τ and single-scattering albedo at 0.412 and 0.490 μm simultaneously, which give the best fit to the satellite measured radiances at these three channels.

ii) Retrieval for Mixture of Dust/Smoke Aerosols:

To retrieve the aerosol properties under the condition of mixed aerosol types (dust mixed with smoke), the radiance (R_λ) received by the satellite is described by:

$$R_\lambda(\mu, \mu_0, \phi) = aR_\lambda^D(\mu, \mu_0, \phi) + (1 - a)R_\lambda^S(\mu, \mu_0, \phi) \qquad \textbf{Equation 2.26}$$

where R_λ^D and R_λ^S are the radiances of the dust and smoke modes, respectively, and α is the ratio of the contributions from the dust and smoke modes to the total radiance. μ is the cosine of the view zenith angle, μ_0 is the cosine of the solar zenith angle, and ∅ is the relative azimuth

angle between the direction of propagation of scattered radiation and the incident solar direction.

iii) Retrieval for strongly absorbing dust

For detecting strongly absorbing dust such as the silicates (e.g., quartz, clays, etc.), DB uses the brightness temperature difference (BTD) between 8.6 and 11 μm (BTD8–11).

Mineral dust that has strong absorption of visible light, for blue wavelengths, are no identified in the Deep Blue retrieval algorithm, this lead to the underestimation of AOD. A heavy dust flag was added in the C6 algorithm, which is based upon the *D** value developed by Hansell et al. (2007). *D** is defined as:

$$D^* = \exp\left\{\frac{[(\mathrm{BTD11-12})-A]}{[(\mathrm{BTD8-11})-B]}\right\} \qquad \textbf{Equation 2.27}$$

where parameters *A* and *B* are the thermal offsets for BTD11– 12 and BTD8–11, respectively. In C6, the values of –0.05 and 10.0 are used for *A* and *B* for the *D** calculation. When the condition *D** > 1.1 is detected, the retrieval algorithm will go directly to the three-wavelength (0.412, 0.47, and 0.65 μm) approach and bypass the two-wavelength (0.412 and 0.47 μm) method (Hansell et al., 2007).

2.5.3.5 Lookup tables

A set of lookup tables using a polarized radiative transfer model are used to simulate the radiances for a range of solar and viewing geometries at the top of the atmosphere that encompass measurements from a satellite. In these calculations, it is assumed that the radiance was dominated by Rayleigh scattering and bounded by a Lambertian surface. The Lambert equivalent reflectivity (LER) is then calculated for each pixel using these lookup tables by matching the surface albedo to the measured satellite radiance (Hsu et al., 2004). The radiative transfer code includes full multiple scattering and considers polarization.

The lookup table contains information regarding solar, satellite (viewing) zenith, azimuth angles, surface reflectance, aerosol optical thickness, and single-scattering albedo. A maximum-likelihood method is used to match the appropriate values of aerosol optical thickness and single-scattering albedo to the measured reflectances.

2.5.3.6 Quality Assurance Confidence

As DT, the quality flags in DB also have four different levels (i.e., QA= 0, 1, 2, 3, with 0 for no retrieval, 1 as the worst quality retrieval and 3 for the best data quality). Pixels with QA=1 for the Deep Blue AOT product could potentially still have cloud contamination issues. The Scientific Dataset (SDS) named "Deep_Blue_Aerosol_Optical_Depth_550_Land_Best-_Estimate" was created in C6 to report good quality pixels with QA = 2 or 3.

Finally, **Figure 2.3** illustrates the main steps of the Deep Blue in Collection 6.

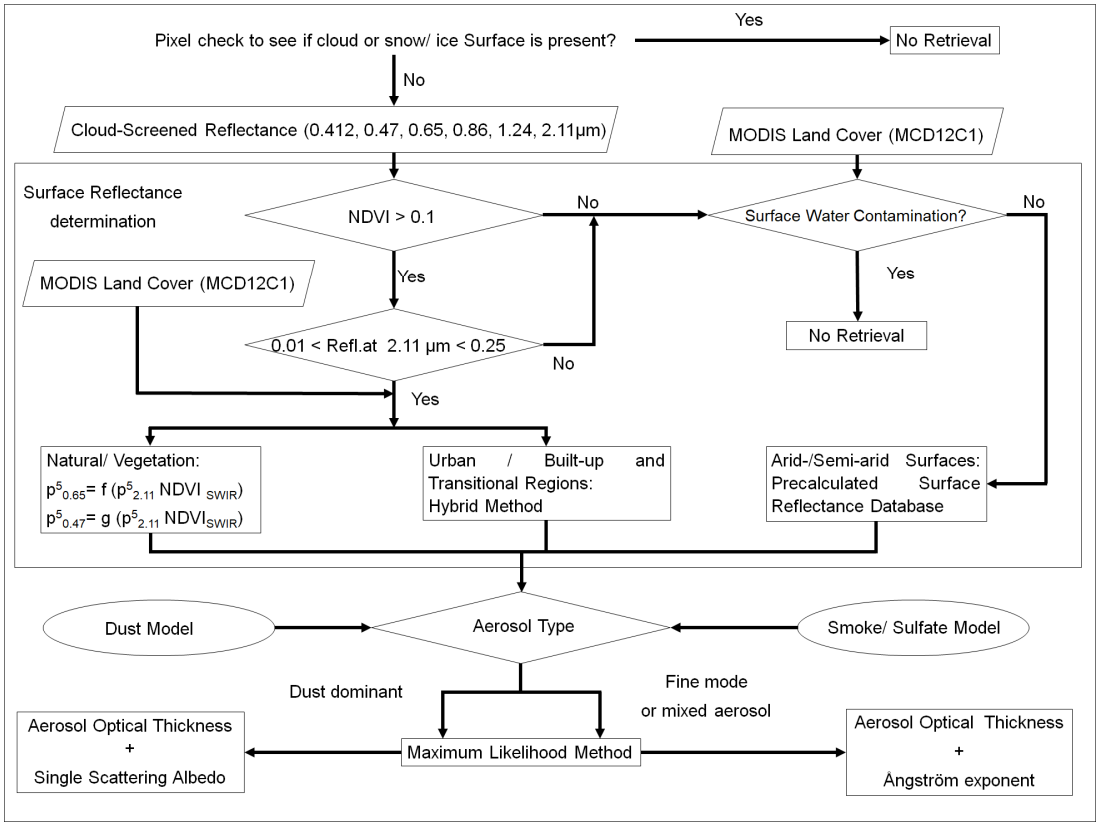


Figure 2.3. Flowchart of the Deep Blue algorithm over land (Hsu et al., 2013).

2.5.4 MODIS – Combined product (Merged DT and DB, 10 km)

A merged SDS is included within the C6 aimed to provide a more gap-filled data set than is available from the individual algorithms alone, both algorithms report AOD at 0.55 μm with high QA confidence (Sayer et al., 2014). There are land areas that may be retrieved by both DB and DT algorithms, as only DB data are available for bright arid regions, there is no choice to be made in this case. Conversely, in the areas with densest vegetation, the DT algorithm is more mature than the DB algorithm, and performs well; thus, a sensible choice

is to use DT in these areas. This leaves several transition regions which have comparatively low vegetation cover but are sufficiently dark for the DT algorithm to be applied (Levy et al., 2013).

For C6, the solution was to simply merge the products from the two algorithms in these transition regions, thus creating a "best-of" AOD product that combines DB, DT-land and DT-ocean. This is reported as the SDS named "Dark_Target_Deep_Blue_Optical_Depth-_550_Combined", it was designed for providing a single data set combining the best of DB and DT into a data product with fewer gaps (Levy et al., 2013)

Three classifications based on NDVI are determined over land: when NDVI ≤ 0.2, DB data are used to populate the merged AOD SDS. When NDVI ≥ 0.3, DT data are used. For the transition areas 0.2 ≤ NDVI ≤ 0.3 (often transition zones between arid and vegetated land), the algorithm whose retrieval returns the higher QA flag is used: $QA_{DT} = 3$ or $Q_{DB} \geq 2$. If both are high confidence (QA=3), then the mean AOD value is used (Sayer et al., 2014):

$$\tau_{Comb} = \frac{(\tau_{DT}+\tau_{DB})}{2} \qquad \textbf{Equation 2.28}$$

where τ_{Comb} represents the AOD estimated to populate the merged SDS, τ_{DT} represents AOD from Dark Target Algorith and τ_{DB} represents the AOD from Deep Blue Algorithm.

2.6 MERRA-2 reanalysis model.

The Modern-Era Retrospective analysis for Research and Applications, Version 2 (MERRA-2) is a NASA atmospheric reanalysis for the modern satellite era (1980 - onward). The reanalysis process is based on forecasting where model fields and observations distributed irregularly in space and time are combined into a spatially complete gridded dataset, using an unchanging data assimilation system and spanning an extended segment of historical data record (Rienecker et al., 2011). The resulting gridded datasets are composed by a broad range of variables, including ones that are sparsely or not directly observed (Gelaro et al., 2017).

The MERRA-2 is performed at 0.5° × 0.625° latitude by longitude with 72 vertical levels, from the surface to 0.01 hPa (about 80 km), resolving both the troposphere and the stratosphere. The MERRA-2 data is available in netCDF format, each parameter within a MERRA-2 file is referred to as a "Scientific DataSet" (SDS) and each netCDF file provides metadata and SDS.

The MERRA-2 is mainly based on the Goddard Earth Observing System Model, Version 5 (GEOS-5), a weather and climate capable model, with its atmospheric data assimilation system (ADAS), version 5.12.4. The GEOS-5 system is composed of atmospheric circulation and composition, oceanic and land components (Randles et al., 2017). MERRA-2 is not only a meteorological reanalysis but it includes also the assimilation of aerosol observations based on a version of the GEOS-5 model that is radiatively coupled to the Goddard Chemistry, Aerosol, Radiation, and Transport (GOCART) and includes assimilation of satellite AOD (Buchard et al., 2015; Provencal et al., 2017). MERRA-2 is the first multi-decadal reanalysis within which meteorological and aerosol observations are jointly assimilated into a global assimilation system (Randles et al., 2017).

2.6.1 GEOS-5 model

The GEOS-5 atmospheric model is a weather-and-climate capable model being used for atmospheric analyses, weather forecasts, uncoupled and coupled climate simulations and predictions, and for coupled chemistry-climate simulations. The GEOS-5 atmospheric general circulation model (AGCM) used for MERRA-2 is based on finite-volume dynamics (Lin, 2004).

The atmospheric data assimilation subsystem of GEOS-5 uses a three-dimensional variational data assimilation (3DVAR) analysis algorithm based on the Grid-point Statistical Interpolation scheme (GSI; Wu et al., 2002; Derber et al., 2003; Purser et al., 2003a,b) with a 6-h update cycle. The GSI algorithm is able to assimilate all the in-situ and remotely- sensed atmospheric data (Buchard et al., 2015).

Four major groups of physical processes are included in the GEOS-5 physics package: moist processes, radiation, turbulent mixing, and surface processes which are parameterized as described by Rienecher et al., 2008. The prognostic variables are discretized on a staggered D-grid. The system has an effective time step of 30 minutes, which is the time step of its physics parameterization. The GEOS-5 terrain-following Lagrangian control volume (lcv) coordinates are similar to an eta coordinate system. The primary variables are: wind components (u, v); scaled virtual potential temperature (θ^*); pressure thickness (of the Lagrangian control volume, p); and specific humidity (q). The model outputs are lcv products. The pressures reported are on a hybrid-sigma coordinate. Indexing in the vertical

starts at the top, i.e., lcv layer 1 is the layer at the top of the atmosphere; lcv layer 72 is adjacent to the earth's surface (Rienecher et al., 2008).

2.6.2 GOCART aerosol module

The GOCART model (Chin et al., 2002; Colarco et al., 2010) is coupled with the GEOS atmospheric model to simulate the life cycles of five externally mixed aerosol species, including dust, sea salt, black carbon, organic carbon, and sulfate (Gelaro et al., 2017). GOCART aerosol module provide gridded aerosol data set where the aerosol species are assumed to be external mixtures that do not interact with each other (Buchard et al., 2015).

Emissions of both dust and sea salt are surface wind speed dependent and are estimated for five size bins, following the parameterization proposed by Marticorena and Bergametti (1995) and Gong (2003), respectively.

Sulfate and carbonaceous aerosol emissions are derived from both natural and anthropogenic sources. Primary sulfate and carbonaceous species are emitted mainly from fossil fuel combustion, biomass burning, and biofuel consumption, with additional biogenic sources of organic carbon. Carbonaceous aerosol and sulfate emissions are prescribed in GOCART module from standard inventories (Randles et al., 2017). Chemical oxidation of sulfur dioxide gas (SO_2) and dimethyl sulfide (DMS) are included to represent secondary sources of sulfate. A database of volcanic SO_2 emissions and injection heights are also included in the aerosol module. Natural emissions of SO_2 from volcanoes are derived from the AeroCom Phase II project (Diehl et al. 2012; http://aerocom.met.no/). The Emission Database for Global Atmospheric Research 4.2 (EDGARv4.2; European Commission 2011) inventory is used to prescribe emissions of SO_2 from anthropogenic sources (energy and non-energy sectors).

Emissions of biomass burning OC, BC, and SO_2 are derived using satellite observations and are based on the Reanalysis of the Tropospheric Chemical Composition, version 2 (RETRO-2; Schultz et al., 2008), the Global Fire Emissions Database, version 3.1 (GFED-3.1; van derWerf et al., 2006), and the NASA Quick Fire Emissions Dataset (QFED) version 2.4-r6 (Darmenov and da Silva, 2015). MODIS level-2 fire and geolocation products are used to locate fires. Biomass burning emissions are distributed uniformly throughout the

planetary boundary layer (PBL) in the grid box where the fire emission occurs (Randles et al., 2017).

The AOD at 550nm from MERRA-2 is a column- and species-integrated optical quantity, which is calculated as the summed product of each species mass and its extinction coefficient (Randles et al., 2017). Mass extinction efficiency, single scattering albedo, and asymmetry parameter, associated with the five aerosol species are derived from the Optical Properties of Aerosols and Clouds (OPAC) data set (Hess et al., 1998).

2.7 Artificial Neural Network

2.7.1 Multi-layer perceptron artificial neural networks

An artificial neural network (ANN) is one type of computing model applied to problem-solving and machine learning, ANNs are considered efficient tools for modeling complex and non-linear processes (Poznyak et al., 2019). ANNs serve to create approximate models relating the input and output information in several applications including classification, clustering, pattern recognition and prediction in many disciplines (Abiodun et al., 2019).

ANN consists of many interconnected processing units called neurons, each producing a sequence of real-valued activations. Multilayer perceptron (MLP) is a forward-structured ANN, where the input vectors are mapped to output vectors (Schmidhuber, 2015). MLP is formed by layers (several nodes arranged in groups). A back-propagation neural network (BPN) is a multilayered network that uses forward nonlinear differentiable functions for weight training. A BPN is composed of input layer, output, and single or multiple hidden layers, each of which are connected to all the cells in the adjacent layer, but without allowing connection between cells of the same layer.

The process is initiated when the network is presented with a pair of learning samples, which triggers the propagation of the neuron's activation value from the input to the output layer. The neurons of the output layer receive an input response of the network, and connection weights are modified using the following sequence (in a layer by layer basis): output layer through middle layer, and back to the input layer (Zhang et al., 2019).

The nodes are linked via weighted connections w and v. The values of these connections are adaptively modified during the process of training. Each node performs a weighted sum of

its inputs and filters it through the so-called activation function (Rowiński et al., 2015). The sigmoidal function is widely used as an activation function in the hidden layer (Piotrowski and Napiorkowski, 2011):

$$f(Z_j) = \frac{1}{1+e^{-Z_j}} \qquad \textbf{Equation 2.29}$$

Where Z_j is the weighted sum of I input variables (x_i)

$$Z_j = f\left(w_{j0} + \sum_{i=1}^{I} w_{ji} x_i\right) \qquad \textbf{Equation 2.30}$$

Then the weighted signals Z_j are multiplied by proper weights v_j and then transferred to the neuron of the third (output) layer, where the new weighted sum is computed, giving finally:

$$y^p = v_0 + \sum_{j=1}^{J} v_j f\left(w_{j0} + \sum_{i=1}^{I} w_{ji} x_i\right) \qquad \textbf{Equation 2.31}$$

Where v_j and w_j are neural networks weights, J is the number of hidden nodes and y^p is the estimated value of dependent variable y. Then, the estimated output value is obtained ($PM_{2.5}$ in this study).

2.7.2 Levenberg-Marquardt algorithm, generalization and overfitting

Levenberg-Marquard (LM) algorithm is a backpropagation algorithm and a gradient-based method that requires a differentiable objective function. The goal of MLP training can be formulated as minimization of a cost function (Lawrence and Giles, 2000):

$$E_{true} = \int_{\mathrm{x,d}} e(f(\mathbf{x}, \mathbf{w}), \mathbf{d}) p(\mathbf{x}, \mathbf{d}) d\mathbf{x} d\mathbf{d} \qquad \textbf{Equation 2.32}$$

E_{true} is the generalization error, i.e. the expected performance of the MLP on new patterns randomly chosen from $p(\mathbf{x}, \mathbf{d})$. e is a local cost function, f is the function implemented by the MLP, **x** is the input to the model, **d** is the desired output of the model, **w** corresponds to the weights in the networks, and p represents the probability distribution (Bengio, 1996).

The objective of training is to optimize the parameters **w** such that E_{true} is minimized:

$$\widehat{\boldsymbol{w}} = argmin_{\boldsymbol{w}} \int_{x,d} e(f(\boldsymbol{x}, \boldsymbol{w}), \boldsymbol{d}) p(\boldsymbol{x}, \boldsymbol{d}) d\boldsymbol{x} d\boldsymbol{d}. \qquad \textbf{Equation 2.33}$$

In practice $p(\mathbf{x}, \mathbf{d})$ is not known. Instead, a training set $T$is given,

$$T = \{\mathbf{x}_p, \mathbf{d}_p\}_1^{N_p} \qquad \textbf{Equation 2.34}$$

where N_p is the number of patterns, and an approximation of E_{true} is minimized which is called the *empirical error* or training error:

$$E = \sum_{p=1}^{N_p} e(\mathbf{x}_p, \mathbf{d}_p). \qquad \textbf{Equation 2.35}$$

The most important characteristic of a neural network model (NN) is its ability to perform well under new situations that were not used during the training process. NNs and other machine learning techniques are susceptible to "overfitting" (poor generalization), i.e. generating spuriously small error on the training data but poor results in real applications. The Levenberg–Marquardt algorithm computes derivations according to training data and it is possible that the minimization for the training data results in a worsening of objective function for the validation data. Hence, when low E (performance on the training set) does not necessarily result in low E_{true} (expected performance on new patterns), this may be a sign that the trained model is overfitting.

2.7.3 Artificial Neural Network to estimate $PM_{2.5}$

There is a growing interest in recent years on the use of ANNs in predicting and forecasting air pollution. The use of ANNs does not require an in-depth understanding of the dynamics between air pollution concentration levels and other explanatory variables. On the other hand, other methods as the common deterministic approaches are sensitive to several factors, including the scale and quality of the parameters involved, computationally expensive, and dependent to large databases of several input parameters (Cabaneros et al., 2019). Similarly, traditional regression models are restricted by some assumptions, like the independence of observations and data distribution of $PM_{2.5}$ (Hu et al., 2017).

ANNs can be very effective to estimate $PM_{2.5}$ if appropriately trained. However, one of the main disadvantages of using NNs for estimating $PM_{2.5}$ is that they do not provide information about the physical mechanisms leading to specific concentrations. Also, the quality of the results obtained with NNs will depend on the quality of the training data, which can also mean that unusual conditions, not represented in the training data, will not be adequately estimated by the NNs. Therefore, for $PM_{2.5}$ estimations using satellite data, NNs cannot replace completely the ground-based measurements (Gupta et al., 2016a).

Chapter 3 METHODOLOGY

This study was performed in three stages: i. model development and validation for establishing the relationship between the AOD and ground-based data to estimate $PM_{2.5}$ within the MMA; ii. developing of a local application: proposal of an alternative methodology to calculate air quality indexes (AQI); and iii. developing of a regional application: analysis of spatial and temporal distribution of $PM_{2.5}$ pollution over Northeastern Mexico (NEM) using MERRA-2 reanalysis datasets.

Overall, the first stage was performed by steps. The first step involved data acquisition, spatial and temporal collocation and testing of twenty-four datasets (four AOD Scientific DataSets [AOD – SDS] in both three domains and in two temporal schemes) using stepwise regression in order to identify the meteorological variables that influence the AOD - $PM_{2.5}$ relationship in the MMA. A second step followed, in which twenty-four refined datasets (which include only the influential factors in the AOD - $PM_{2.5}$ relationship) were analyzed with the aid of multiple linear regression model (MLR) in order to identify which AOD-SDS, domain and temporal schemes showed the best correlation in MMA. The resulting dataset was named as the "*best dataset*". In a third step, an ensemble model based on MLR was developed for validating the *best dataset* robustness. Then, neural networks were trained and validated in order to estimate $PM_{2.5}$.

The second and third stage (local and regional applications, respectively) were based on the results obtained from the first stage. In the local application, we categorized the $PM_{2.5}$ estimations (from the first stage) in air quality indexes. Finally, in the regional application, we trained and validated neural networks using the influential factors (found in the first stage) with additional data provided by MERRA-2 (a global reanalysis model) aimed to analyze spatial and temporal distribution of $PM_{2.5}$ pollution in NEM.

Complete description about methods used in the three stages are detailed later in the current chapter. **Figure 3.1** provides an overview of the methodology adopted in this research.

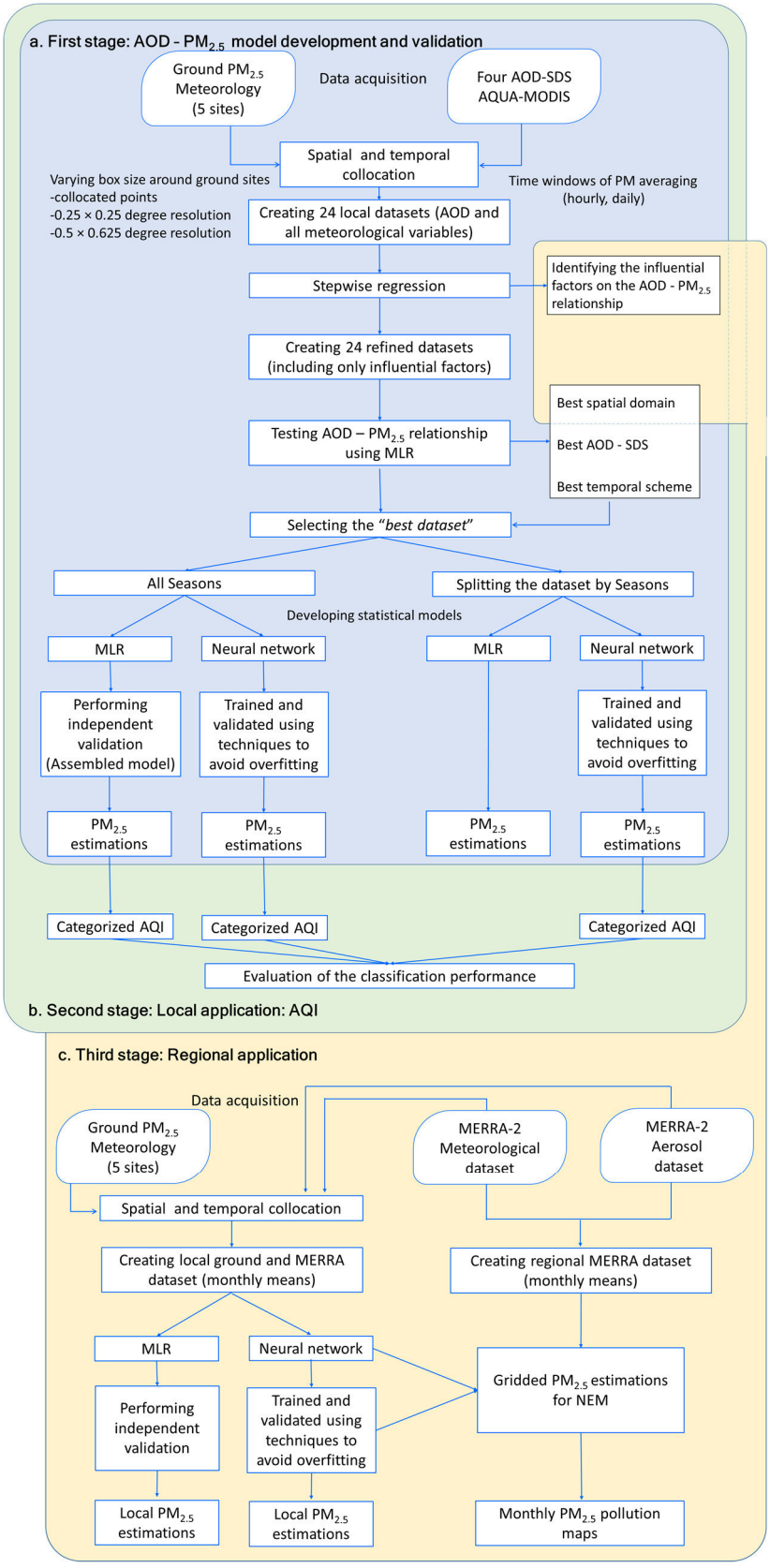


Figure 3.1. Research methodology framework.

3.1 AOD and ground-based data relationship

3.1.1 Model development and validation

$PM_{2.5}$ was estimated by developing relational models between AOD and ground-based data within the MMA. The model development and validation for estimating $PM_{2.5}$ was performed in a series of steps. The first step involved data acquisition, spatial and temporal collocation. In a second step, twenty-four datasets (four AOD Scientific DataSets [AOD – SDS] in three spatial domains and in two temporal schemes, detailed later) were tested using stepwise regression in order to identify the meteorological variables that influence the AOD - $PM_{2.5}$ relationship in the MMA. In a third step, twenty-four "refined" datasets (which include only the influential factors in the AOD - $PM_{2.5}$ relationship) were analyzed with the aid of multiple linear regression model (MLR) in order to identify which AOD-SDS, spatial domain and temporal schemes showed the best correlation performance in MMA. The resulting dataset was named as the "*best dataset*". In a fourth step, an ensemble model based on MLR was developed for validating the robustness of the *best dataset*.

Finally, neural network models were trained and validated to estimate $PM_{2.5}$. This task was performed because regression models tend to under-predict the higher concentrations and over-predict the lower ones (EPA, 2003). A Neural Network (NN) approach, if well-trained, should be precise and accurate for this type of problem.

Figure 3.2 provides an overview of the methodology adopted to estimate $PM_{2.5}$ concentrations using MODIS data over MMA. The full methodology is explained in detail in the following sections.

3.1.2 Ground- based $PM_{2.5}$ and meteorological data

For this study, $PM_{2.5}$ and meteorological data were retrieved for the period covering from 2010 to 2017, at five urban air-monitoring sites from the SIMA network (southeast, northeast, downtown, northwest and southwest; **Figure 1.2**). Because they were the first monitoring sites installed in the MMA and have been operating since 1992, they provide the long-term data availability required for this research. The monitoring sites operate continuously and automatically. Specifically, $PM_{2.5}$ instruments are based on the Beta-ray Attenuation method

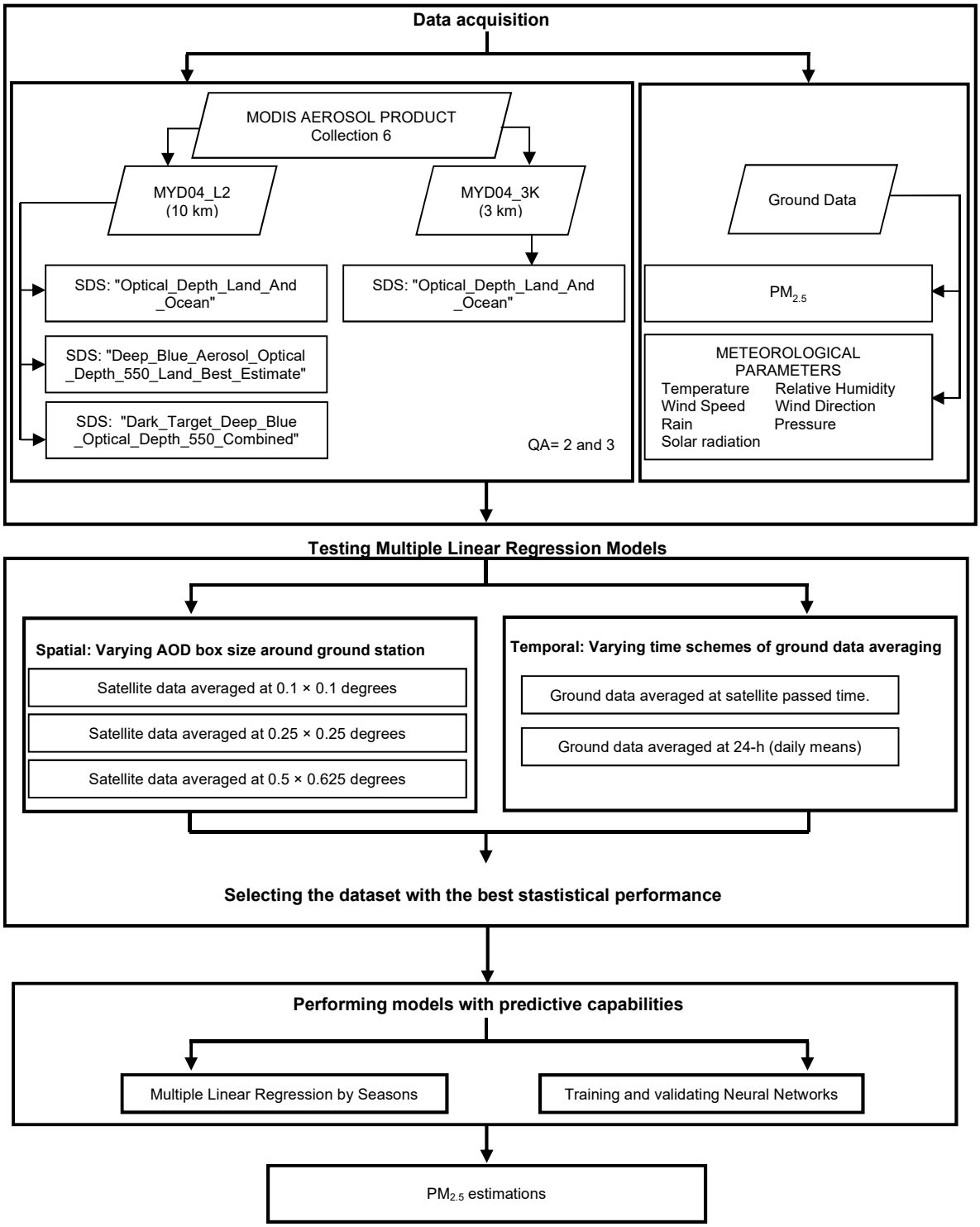


Figure 3.2. Methodology framework for establishing the relationship between the AOD and ground-based data to estimate $PM_{2.5}$ within the MMA

(BAM) and continuous weighing with microbalance that provides automatic, continuous measurement and concentration recording. The BAM instruments are designated as EPA Class III Federal Equivalency Method (EQPM-1013-209; EPA, 2016). The measurement range is from –15 μg m^{-3} to 10,000 μg m^{-3}, the data resolution is 0.1 μg m^{-3}, the lower detection limit is < 2.4 μg m^{-3} (hourly, 2σ) and < 0.5 μg m^{-3} (24 hour, 2σ) and the accuracy meets the US-EPA Requirements for Class III $PM_{2.5}$ FEM. The instruments provide continuous air sampling with hourly tape advance at a sample flow rate of 16.7 liters per minute (EPA, 2018).

SIMA instrumentation records $PM_{2.5}$ data every minute, which is then validated and archived as 1-h averages. After collecting the hourly data of $PM_{2.5}$ and meteorological parameters, daily averages are calculated per sampling site. In this research, a threshold of 75% data capture (18 records) was defined to consider data valid and representative on daily basis as established by the Mexican standard NOM-025-SSA1-2014. Missing values were excluded from our analysis.

3.1.3 MODIS AOD data sets

MODIS (Moderate Resolution Imaging Spectroradiometer) is a passive instrument that flies aboard both the Terra and Aqua satellites. Terra has a north to south trajectory (descending node) across the equator in the morning (at approximately 10:30 AM local time), while Aqua has a south to north trajectory (ascending node) across the equator in the afternoon (at approximately 1:30 PM local time). Terra and Aqua satellites cover the globe every 1-2 days (MODIS - LAADS DAAC, n.d.).

As a passive instrument, MODIS detects solar electromagnetic radiation that is being absorbed and reemitted, reflected, and scattered by the earth and atmosphere. The incoming radiation passes through a spectrometer, a device that measures energy intensity as a function of wavelength, to create a spectrum of wavelengths that are then detected (Duncan et al., 2014). MODIS provides data in moderate spatial resolution (between 250 m and 1 km at nadir –depending on the band) with 2,330 km in swath and 36 spectral bands covering from 0.405 μm to 14.385 μm. Among the many physical parameters derived from MODIS, spectral radiances are processed to a suite of products characterizing aerosol particles, for example

AOD, which is a measure of the integrated extinction by aerosols of light passing through the entire atmospheric column from the surface to the top of the atmosphere (TOA).

In this study, AOD data derived from Aqua-MODIS instrument -particularly, the collection 6 MYD04_3K and the MYD04_L2 - MODIS aerosol product released at a spatial resolution of 3 and 10 km at nadir- were used. The data is distributed freely by the Atmosphere Archive and Distribution System (LAADS), and was downloaded from http://ladsweb.nascom.nasa.gov/ in Hierarchical Data Format (HDF). Each parameter within a MODIS file is referred to as a "Scientific DataSet" (SDS) and each HDF file provides metadata and SDS. AOD retrieval procedures in collection 6 (C6) are described in the fundamentals section in this document (**Chapter 2, Section 2.5**) and a detailed description of the retrieval algorithms and aerosol products is provided by Levy et al. (2013), Remer et al. (2013), Hsu et al. (2013) and Sayer et al. (2014).

AOD data extracted from four SDS in C6 were evaluated to identify which dataset best described the relationship between AOD and ground-based data to estimate $PM_{2.5}$ at the MMA. **Table 3.1** provides the names for the relevant AOD SDS used in this research. Quality Assurance flags (QA) for the SDSs derived from Dark Target algorithm (3k and DT, in **Table 3.1**) are provided in a separate SDS, as mentioned in the fundamental section in this document (**Chapter 2, Section 2.5.1** and **2.5.2**).

Table 3.1. AOD-SDS from MODIS aerosol products.

Product	SDS	Resolution
3k	"Optical_Depth_Land_And_Ocean" "Land_Ocean_Quality_Flag" QA= 2 and 3	3 km
DT	"Optical_Depth_Land_And_Ocean" "Land_Ocean_Quality_Flag" QA= 2 and 3	10 km
DB	"Deep_Blue_Aerosol_Optical_Depth_550_Land_Best_Estimate" *	10 km
Cb	"AOD_550_Dark_Target_Deep_Blue_Combined" **	10 km

*This SDS contains deep blue AOD at 0.55 micron for land with higher quality data (QA= 2 and 3)
** "AOD_550_Dark_Target_Deep_Blue_Combined_Algorithm_Flag" QA= 2 and 3

Comparing the four AOD SDS is a non-trivial task because of the different assumptions made in their retrieval procedures, bestowing strengths and limitations that must be considered

when satellite data is used for local $PM_{2.5}$ estimations. Detailed description of the products was provided in **Chapter 2, Section 2.5**. However, some fundamentals are summarized below, in order to illustrate the potentialities and constraints for using these products in this particular research:

- Both the 10k_DT AOD-SDS from MYD04_L2 product (10 km at nadir) and the 3k_DT AOD-SDS from MYD04_3k product (3 km at nadir) are retrieved using the DT algorithm. The surface scheme used in the DT algorithm identifies dark target pixels (mainly vegetated areas); this does not completely represent the highly heterogeneous surfaces in the MMA. In the DT land algorithm, fine-dominated and coarse-dominated aerosol *models* (each composed of multiple modes) are combined to match with the observed spectral reflectance. Four distinct fine-dominated models are identified in the DT and one of them represents the urban/industrial aerosol type, which is of interest in MMA.
- Global products at 3 km were released in response to the air quality community's need for high resolution aerosol retrieval for example applications focused on a local basis. 3 km algorithm differ to the 10 km algorithm in the way that the pixels are organized and the number of pixels required to retrieve data after all masking and deselection are accomplished (Remer et al., 2013). The input reflectance (from L1B) are organized into groups of 6 × 6 pixels for the 3 km algorithm, instead 20 × 20 pixels as the case of the 10 km algorithm. Therefore, pixels representing bright or inhomogeneous surfaces (as is the case of the MMA surface) that might be discarded during the sorting and discarding procedure at 10 km might be kept at 3 km retrieval. This has the potential limitation to make the 3 km product noisier than at 10 km, but offers the advantage of retaining useful information discarded by the 10 km procedure.
- The DB AOD-SDS from MYD04_L2 product is retrieved using DB algorithm at 10 km at nadir. Three surface categories are derived from MODIS DB C6 including: (1) arid and semiarid regions, (2) general vegetation, and (3) urban/built-up and transitional regions. MMA surfaces presumably fall in at least one of these categories. However, the DB algorithm performs aerosol retrieval for non-spherical and absorbing aerosols including: (1) dust aerosols, (2) for mixture of dust/smoke aerosols and (3) for strongly absorbing dust. These types of aerosols do not represent completely the mixture of

aerosols (as $PM_{2.5}$) present in the MMA, where dust represents only 10% of its mass, with organic carbon (50%), elemental carbon (5%) and soluble inorganic salts such as nitrates, sulfates and ammonium (35%) accounting for the balance (Mancilla et al., 2016).

- Finally, the merged AOD-SDS from MYD04_L2 product is reported at 10 km. This dataset was designed to provide a single dataset that combines the best of DB and DT into a data product with fewer gaps (Sayer et al., 2014). The merge procedure is based on NDVI criteria (NDVI > 0.30 = DT; NDVI < 0.20 = DB; 0.2 < NDVI < 0.30 = merge). The NDVI criteria for the merged product describe the transition areas. Transition zones between vegetated and non-vegetated areas are typically present in MMA. The "Combined" products are not validated in C6. The primary intended purpose of this merged SDS at C6 was visualization or for applications where coverage is more important than quantitative accuracy. In an ideal case, users could perform analyses using both DT and DB data where available.

3.1.4 $PM_{2.5}$ concentration and MODIS AOD spatio-temporal collocation

AOD values from the four SDS's (**Section 3.1.3**) were tested using three spatial domains (**Section 3.1.4.1**) and two temporal schemes (**Section 3.1.4.2**) to obtain the most representative sample of AOD data for developing statistical models that would establish the relationship among $PM_{2.5,}$ AOD and meteorological data.

The MODIS AOD and the ground-based data were geo-referenced to geographic latitude/longitude (WGS84) coordinates. All the variables were collocated in space and time.

3.1.4.1 Spatial collocation

For exploring spatial differences in the AOD selection, the effect of box size (number of pixels in spatial collocation) around the ground stations on the mean AOD were evaluated. In this study, three different box sizes (**Figure 3.3**) centered on MMA were evaluated.

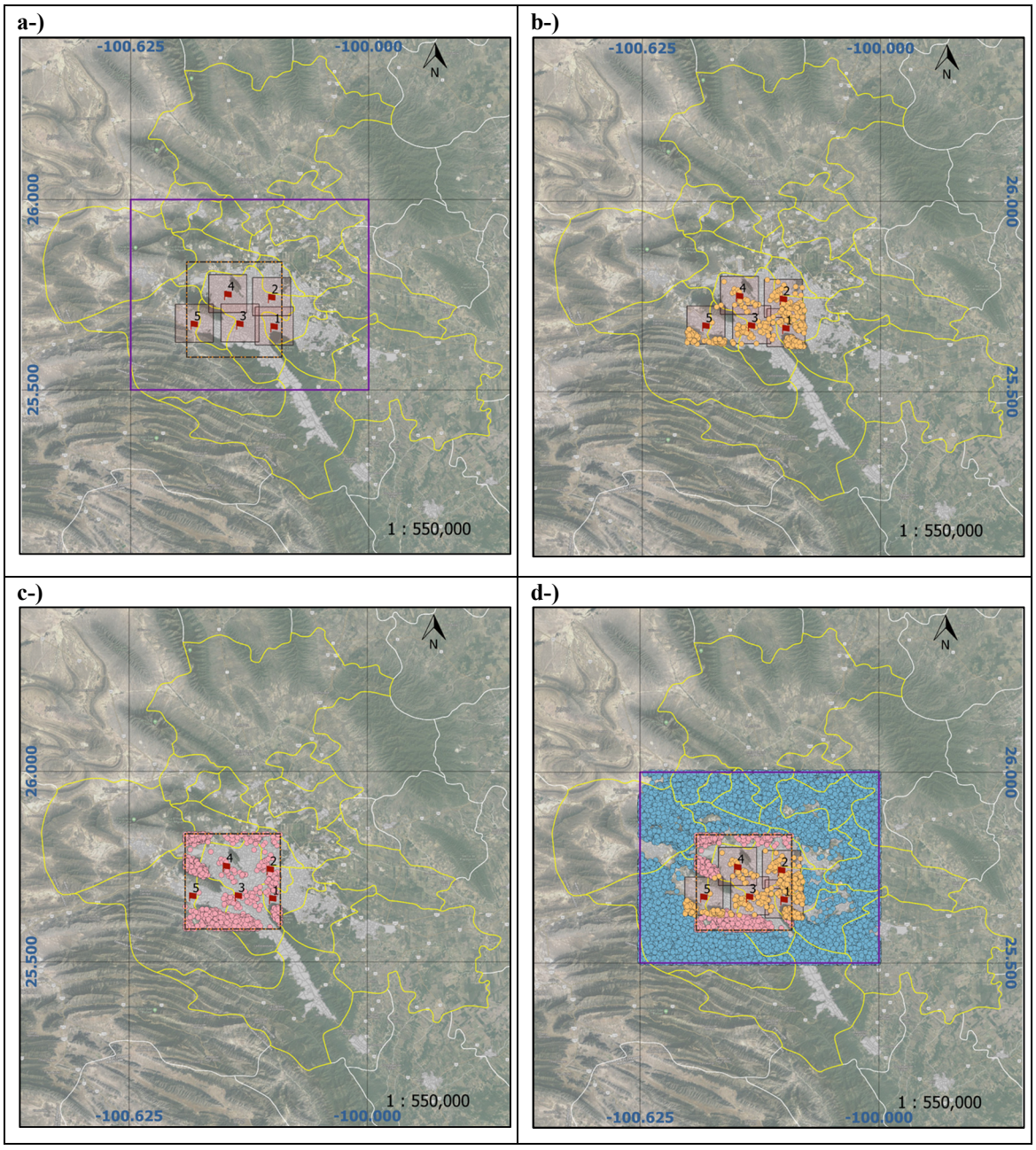


Figure 3.3. Ground-based data and MODIS AOD spatial collocation.
Figure 3.3a, Three domains. Boxes of 0.10 × 0.10 deg size correspond to the first domain (D1); collocated AOD values are contained in this domain. Box size of 0.25 × 0.25 deg corresponds to the second domain (D2). Box size of 0.500 × 0.625 deg corresponds to the third domain (D3).
Figure 3.3b, Ground data from each site vs AOD averaged at 0.10 × 0.10 º
Figure 3.3c, Ground data (averaged from the 5 sites) vs AOD averaged at 0.25 × 0.25 º
Figure 3.3d, Ground data (averaged from the 5 sites) vs AOD averaged at 0.500 × 0.625 º

3.1.4.2 Temporal collocation

The first temporal scheme corresponds to the satellite overpass time (hourly scheme), since Aqua crosses the MMA at 1:30 p.m. local time (± 55 min). The average AOD at the satellite overpass time was used to represent both hourly and daily AOD mean value. Accordingly, hourly $PM_{2.5}$ and meteorological values were processed by averaging $PM_{2.5}$ and meteorological observations measured from 12:00 p.m. to 3:00 p.m. local time. This hourly scheme is used to represent a similar air mass as observed by MODIS.

The second temporal scheme corresponds to daily scheme (24 h). Data were processed by averaging $PM_{2.5}$ and meteorological observations measured from 12:00 a.m. to 12:00 p.m. local time, considering the threshold record criteria (18 records) for data validation and representativeness. This daily scheme is used to represent the $PM_{2.5}$ data according to the 24 h Mexican standard NOM-025-SSA1-2014.

3.1.5 Model development and validation

3.1.5.1 Multiple Linear Regression

Twenty-four different datasets (AOD from four MODIS SDS, three spatial domains and two temporal schemes) were analyzed to identify the "*best dataset*", namely, the dataset that best described the relationship among $PM_{2.5,}$ AOD and meteorological data at MMA. The impact of various spatio-temporal average schemes on this relationship was evaluated by performing Multiple Linear Regression (MLR) techniques where both the average of AOD satellite data at the three domains and the average of $PM_{2.5}$ and meteorological parameters were tested using hourly and daily schemes. **Figure 3.4** provides an overview of the methodology adopted to select and validate the *best dataset.*

To sequentially identify the best subset of independent variables to be incorporated into the regression equation, all the meteorological variables recorded by SIMA (Temperature, Relative Humidity, Wind Speed, Wind Direction, Rain, Pressure and Solar Radiation) along with the AOD data were included in the twenty-four datasets. The final datasets (refined datasets) were selected through a stepwise method to achieve the highest level of estimation accuracy, and only those meteorological variables that were statistically significant and most significantly correlated with the dependent variable ($PM_{2.5}$) were retained. Forward and

backward stepwise procedures were applied using Minitab® 16 software (State College, PA, USA), and a significance level of 0.05 was used for the inclusion and exclusion of variables.

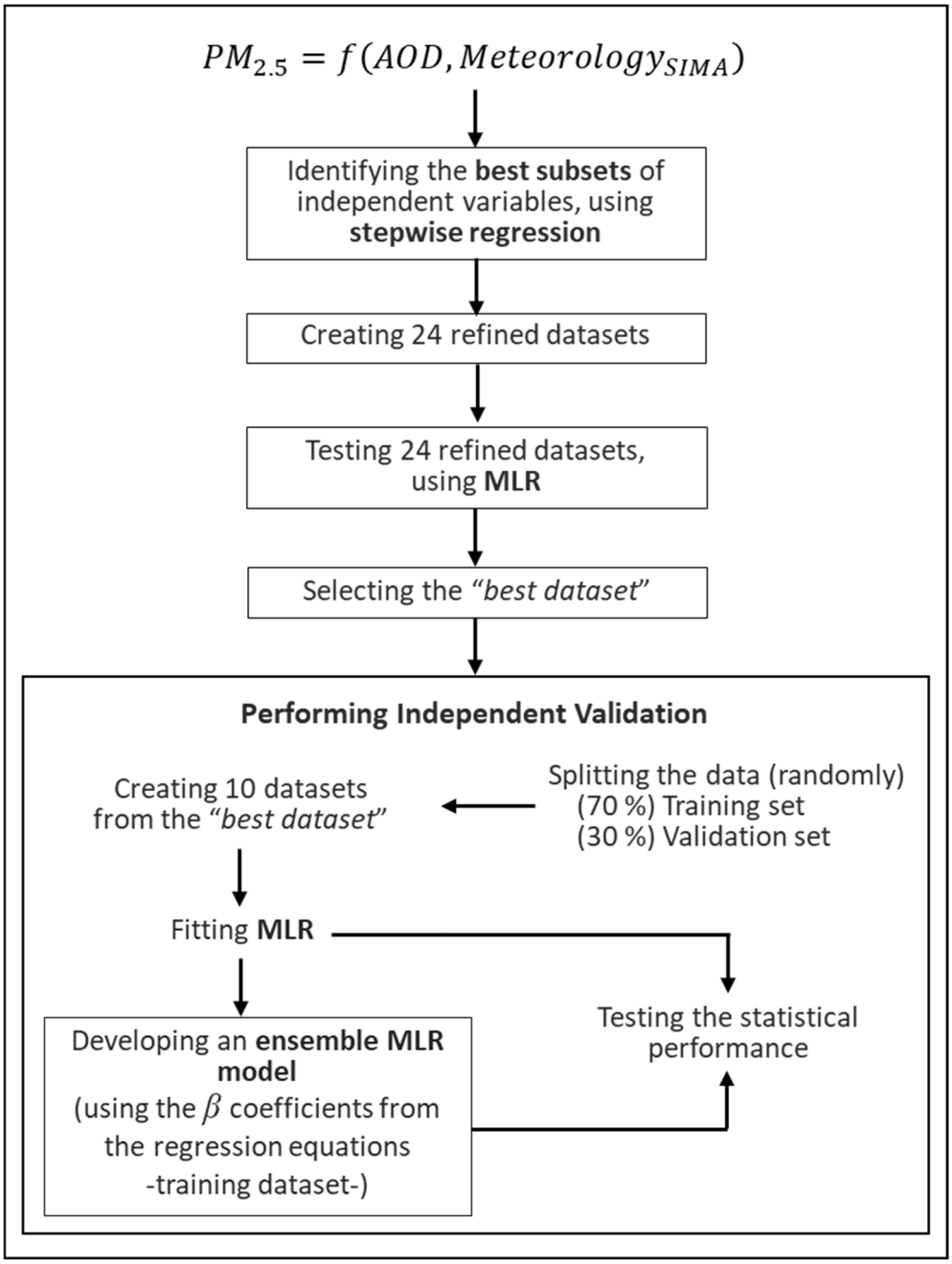


Figure 3.4. Selection and validation of the "*best dataset*".

The refined datasets were tested to identify the "*best dataset*" with the aid of MLR. The "*best dataset*" was selected according to the following criteria: first, dataset should have data representativeness (number of data, *N*). Second, once the MLR is performed, the mean and

variability (standard deviation) for the estimated model should be similar to the mean and variability from the observed dataset. Finally, the dataset should have a higher value in the correlation coefficient (R) and lower values in the error estimators, i.e. Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), Mean Bias Error (MBE) and Mean Percentage Error (MPE). The calculation of these error estimators is conducted as indicated in **Section 3.1.6**.

Because model performance can be sensitive to the dataset, the robustness of the MLR was tested by performing an independent validation. Once the *best dataset* was selected, an ensemble model was developed following the next steps: first, ten iterations were performed, creating datasets randomly from the *best dataset* and splitting the data in 70% and 30% to create training and validation datasets, respectively. Then, a MLR was performed independently for each training dataset. The regression coefficients (β) from the training dataset were saved at each iteration. Finally, an ensemble MLR model was built by averaging each β coefficients from the regression equations at the training datasets. It was expected that the ensemble model has similar statistical performance than that obtained in the MLR using the original *best dataset*.

3.1.5.2 Artificial Neural Network

To overcome overfitting, in this research, the NNs were trained by combining two strategies. First, the NN architecture was designed under the "simpler-structure principle" (Jiang et al., 2004). Second, the early stopping method was applied as the regularization technique (Sarle, 1995). The "simpler-structure principle" assumes that the optimal structure of a neural network should be fairly simple, with several nodes in one hidden layer. Under this principle, the training starts with the simplest structure (only one node in the hidden layer). The aim is to detect a local optimum on the validation error curve. This procedure is performed again with a new node, and the resulting value is compared to the previous minimum value. This will be repeated until a better performance with a simpler structure is obtained., i.e. a structure having fewer nodes in the hidden layer (Mao et al., 2017). The stopping criteria were determined by a predefined threshold number of training iterations, cross-validation approach (Prechelt, 1998) and the reduction of root mean squared error (RMSE) for a validation data set (Shao et al., 2011).

The general scheme used in this research for estimating $PM_{2.5}$ based on NN is shown in **Figure 3.5.** For this study, a MATLAB® (version R2014a) code for the NN trained using the "nntool" was extracted and modified to include the strategies to overcome overfitting. A multilayer perceptron (MLP) type back-propagation NN was chosen. The number of nodes in the hidden layer was determined by the training process. The input layer has an identity activation function. A sigmoid function was used as the transfer function in the hidden layer (**Equation 3.1**) and the Levenberg-Marquardt (LM) algorithm (Shepherd, 1997) was employed for training. Finally, independent validation method was applied to test the predictive power from the trained NN models.

$$y\,(x) = \frac{1}{1+ e^{-x}} \qquad \textbf{Equation 3.1}$$

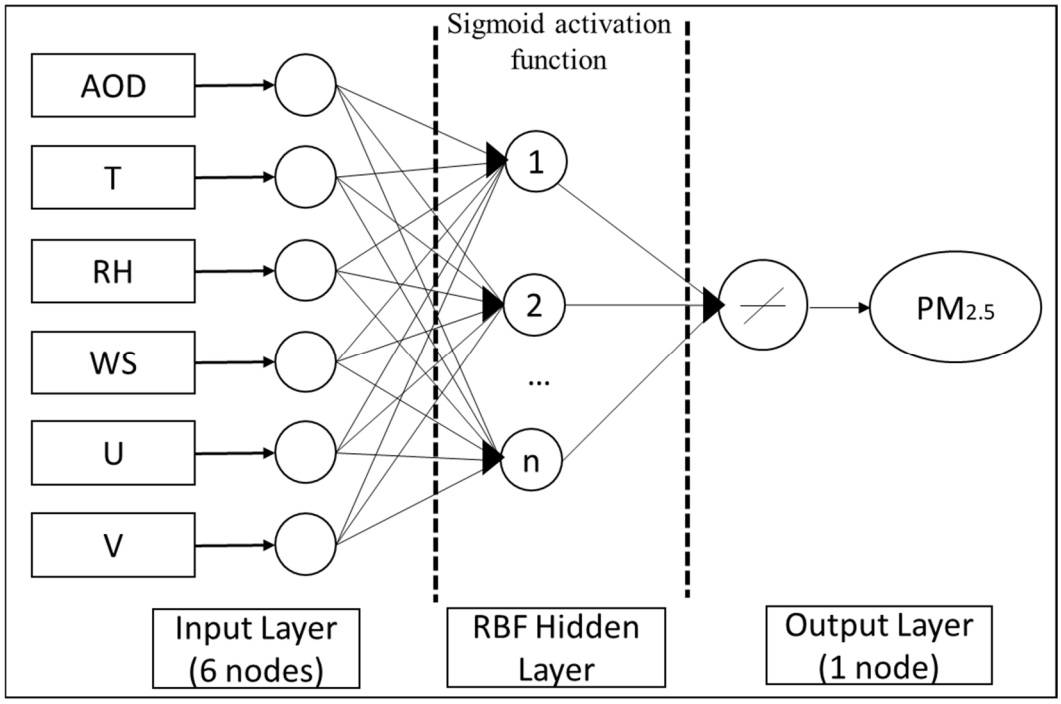


Figure 3.5. A schematic of the NN model used in this study for estimation of $PM_{2.5}$.

3.1.5.3 Seasonal and Temporal Trends

Aimed to explore the seasonal relationship and to further understand the effects of temporal variability on model performance improvement, conventional MLR and NN models were developed by seasons. The *"best dataset"* was divided into four subsets corresponding to

four seasons: winter (December, January, and February), spring (March, April, and May), summer (June, July, and August) and fall (September, October, and November).

3.1.6 Statistical performance

With the purpose of measuring model accuracy, validating results and analyzing uncertainty, the following statistical parameters were calculated **(Equation 3.2 - Equation 3.5**): Pearson's correlation coefficient (R), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), Mean Bias Error (MBE) and Mean Percentage Error (MPE). The correlation coefficient is a statistic used to evaluate the fit of the model. It measures the strength and the direction of the relationship between two variables (observations and the estimations) and is interpreted as the degree of variation in the response that is explained by the model. RMSE is a quadratic scoring rule that measures the average magnitude of the errors in a set of predictions. MAE also measures the average magnitude of the error but does not consider its direction. If the absolute value is not considered –that is, the signs of the errors are not removed-, the average error becomes the Mean Bias Error (MBE), which is a metric commonly used to measure average model bias, but cancellation of errors could appear (positive and negative errors will cancel out). In order to avoid this situation, the mean percentage error (MPE) was computed. The MPE represents the calculated average of percentage errors by which the estimated $PM_{2.5}$ ($PM_{2.5\ (Est)}$) of a model differ from the ($PM_{2.5\ (Obs)}$) of the quantity estimated. Both MAE and RMSE are in μg m^{-3}. Both metrics can range from 0 to ∞ and are indifferent to the direction of errors. RMSE assigns a relatively high weight to large errors. The RMSE result will always be larger or equal to the MAE. If all of the errors have the same magnitude, then RMSE is equal to MAE.

$$R = r_{xy} = \frac{N\sum x_i y_i - (\sum x_i)(\sum y_i)}{\sqrt{N(\sum {x_i}^2) - (\sum x_i)^2}\ \sqrt{N(\sum {y_i}^2) - (\sum y_i)^2}} \qquad \textbf{Equation 3.2}$$

Where N is the number of data. x is the observed $PM_{2.5}$ ($PM_{2.5\ (Obs)}$) and y is the estimated $PM_{2.5}$ ($PM_{2.5\ (Est)}$)

$$RMSE\ (\mu g\ m^{-3}) = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(PM_{2.5\ (Obs)i} - PM_{2.5\ (Est)i}\right)^2} \qquad \textbf{Equation 3.3}$$

$$MAE\ (\mu g\ m^{-3}) = \frac{1}{N}\sum_{i=1}^{N}\left|PM_{2.5\ (Obs)i} - PM_{2.5\ (Est)i}\right| \quad \textbf{Equation 3.4}$$

$$MBE\ (\mu g\ m^{-3}) = \frac{1}{N}\sum_{i=1}^{N}\left(PM_{2.5\ (Obs)i} - PM_{2.5\ (Est)i}\right) \quad \textbf{Equation 3.4}$$

$$MPE\ (\%) = \frac{100\%}{N}\sum_{i=1}^{N}\left(\frac{PM_{2.5\ (Obs)i} - PM_{2.5\ (Est)i}}{PM_{2.5\ (Obs)i}}\right) \quad \textbf{Equation 3.5}$$

3.2 Developing a local application: Categorization of metropolitan air quality index based on satellite-derived data and meteorological ground-based measurements.

3.2.1 Metropolitan Air Quality Index (MX-AQI)

The Metropolitan Air Quality Index (IMECA), MX-AQI in this document, is a numerical index based on measured concentrations of selected ambient air pollutants and its compliance with regulations (Norma Ambiental para el Distrito Federal NADF-009-AMBT-2017). The MX-AQI uses data collected from air quality monitoring sites and reports the severity of air pollution to the public, adverse effects on human health, and preventive measures. The MX-AQI magnitude is proportional to pollutants concentration in the ambient air, therefore greater MX-AQI value indicates more serious health implications (Norma Ambiental para el Distrito Federal NADF-009-AMBT-2017).

The MX-AQI is obtained from the calculation of individual AQI values for each of the six criteria pollutants in ambient air: carbon monoxide (CO), nitrogen dioxide (NO_2), ozone (O_3), sulfur dioxide (SO_2), coarse particulate matter (PM_{10}) and fine particulate matter ($PM_{2.5}$). Of these, the highest individual AQI value is considered to be the most critical health concern and is chosen to represent the daily AQI category (Trejo, 2006).

3.2.2 Data and methods

In this study, daily $PM_{2.5}$ concentrations, as estimated from the developed ensemble MLR and NN models (both NN and NNbS) –described in **Section 3.1**-, were used to determine the air quality categories of the $PM_{2.5}$ MX-AQI. **Figure 3.6.** provides an overview of the methodology adopted to calculate $PM_{2.5}$ AQI using ground-based and MODIS data over MMA.

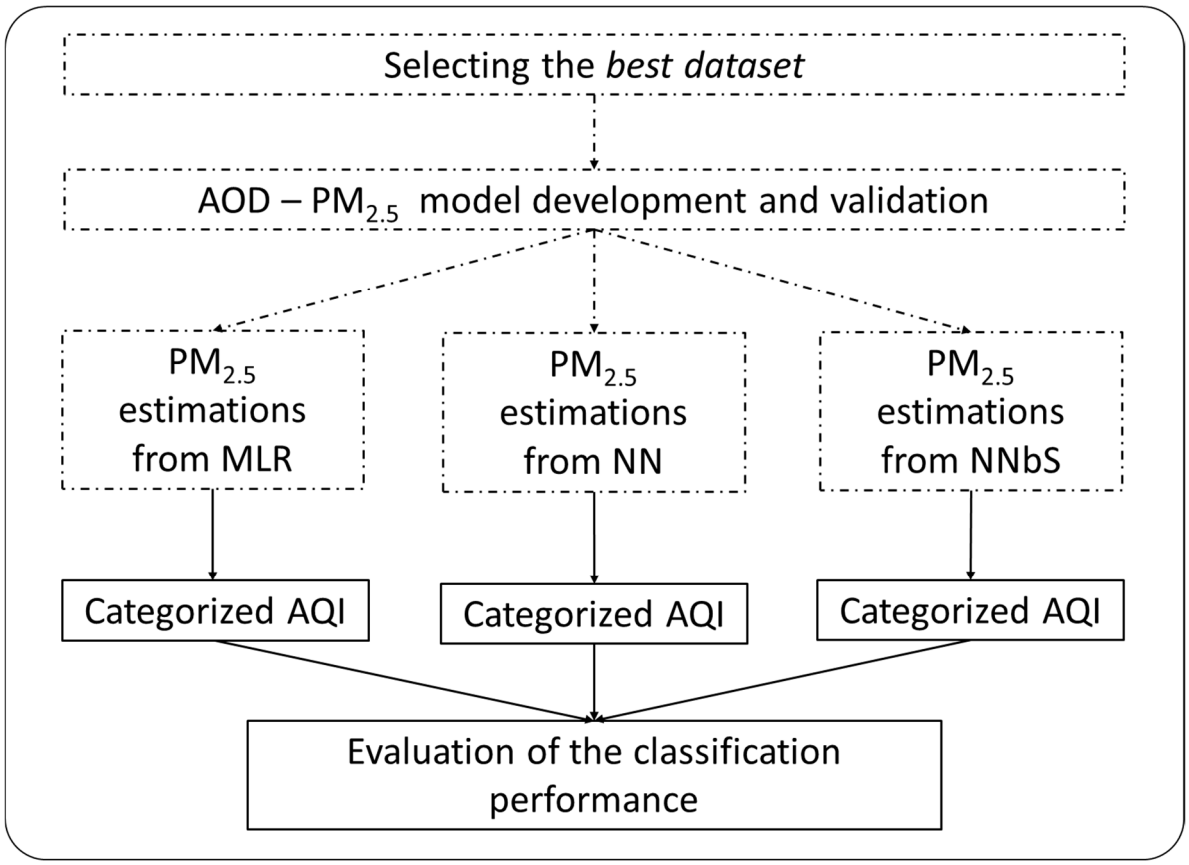


Figure 3.6. Methodology framework for categorizing daily $PM_{2.5}$ AQI (the dash line represents methods from **Section 3.1**.)

3.2.3 Classification strategy

Daily $PM_{2.5}$ concentrations were classified according to the air quality categories of the $PM_{2.5}$ MX-AQI, as shown in **Table 3.2.**, where relevant health effects are also included for each color-coded MX-AQI (EPA, 2009; EPA, 2014; EPA, 2015; NADF-009-AMBT-2017)

3.2.4 Performance evaluation

Classification performance was evaluated using a confusion matrix which describe the performance of a classification model on a set of test data for which the true values are known. The standard evaluation metrics contained in the confusion matrix, the equations used to calculate them and two examples of the meaning of the standard evaluation metrics are presented in **Table 3.3.**

Table 3.2. Categories based on $PM_{2.5}$ MX-AQI and relevant information (EPA, 2009; EPA, 2014; EPA, 2015; NADF-009-AMBT-2017.

Color Code/ Range / Categories	Breakpoints $PM_{2.5}$ ($\mu g\ m^{-3}$)	Effects on human health	Suggested measures
0–50 Good	0 – 15.4	There are no health concerns.	All kinds of people are allowed to take normal activities.
51–100 Moderate	15.5 – 40.4	There might be a weak effect on health of some highly sensitive people	Such sensitive people should reduce outdoor activities.
101–150 Bad	40.5 – 65.4	Aggravation of symptoms of susceptible people who may be influenced on the heart and respiratory system.	Children, the elderly, and patients with heart disease and respiratory disease should reduce long-time and high-strength outdoor exercises.
151–200 Very bad	65.5 – 150.4	Dramatic aggravation of symptoms of vulnerable groups who may be influenced on the heart and respiratory systems.	Children, the elderly, and patients with heart disease and respiratory disease should avoid long-time and high-strength outdoor exercises. Common people should reduce outdoor exercises.
> 200 Extremely bad	> 150.4	Aggravation of symptoms of people with heart attack and pulmonary disease who may have their exercise tolerance degraded; common healthy people may suffer from certain symptoms.	Children, the elderly, and patients should stay indoors and avoid physical consumption. Common people avoid outdoor activities.

matrix, the equations used to calculate them and two examples of the meaning of the standard evaluation metrics are presented in **Table 3.3.**

Table 3.3. Standard evaluation metrics contained in the confusion matrix

Evaluation metrics	Equations	Evaluating the "Good – Moderate" Level	Evaluating "Moderate – Bad" Level
True positives (TP)		Predicted "Good", and actually "Good".	Predicted "Moderate", and actually "Moderate".
True negatives (TN)		Predicted "Moderate", and actually "Moderate".	Predicted "Bad", and actually "Bad".
False positives (FP)		Predicted "Good", actually "Moderate". ("Type I error")	Predicted "Moderate", actually "Bad". ("Type I error")
False negatives (FN)		Predicted "Moderate", actually "Good". ("Type II error.")	Predicted "Bad", actually "Moderate". ("Type II error.")
False Positive Rate (FPR)	$FPR = \frac{FP}{(FP + TN)}$	When it's actually "Moderate", how often does it predict "Good"?	When it's actually "Bad", how often does it predict "Moderate"?
False Negative Rate (FNR)	$FNR = \frac{FN}{(FN + TN)}$	When it's actually "Good", how often does it predict "Moderate"?	When it's actually "Moderate", how often does it predict "Bad"?
True positive Rate (TPR)	$TPR = \frac{TP}{(TP + FN)}$	When it's actually "Good", how often does it predict "Good"?	When it's actually "Moderate", how often does it predict "Moderate"?
True negative Rate (TNR)	$TNR = \frac{TN}{(TN + FP)}$	When it's actually "Moderate", how often does it predict "Moderate"?	When it's actually "Bad", how often does it predict "Bad"?
Accuracy (AC)	$AC = \frac{(TP + TN)}{(TP + TN + FP + FN)}$	Overall, how often is the classifier correct?	Overall, how often is the classifier correct?
Error rate (ER	$ER = \frac{(FP + FN)}{N}$	Overall, how often is the classifier wrong?	Overall, how often is the classifier wrong?

The scheme of the confusion matrix used in this research is presented in **Figure 3.7.**

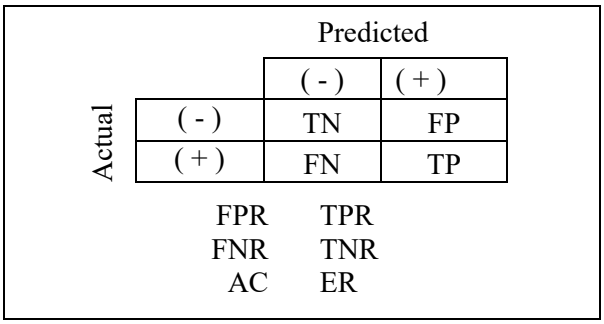


Figure 3.7. Scheme of the confusion matrix (definition of evaluation metrics are presented in **Table 3.3**).

3.3 Developing a regional application: Analysis of spatial and temporal distribution of $PM_{2.5}$ pollution over Northeastern Mexico using MERRA-2 Reanalysis Datasets.

3.3.1 Data and methods

We trained and validated neural networks using the influential factors (found in the first stage -**Section 3.1**-) with additional data provided by MERRA-2 (a global reanalysis model described in **Section 2.6**) aimed to analyze spatial and temporal distribution of $PM_{2.5}$ pollution. First, a method that uses neural networks (NN) was developed to combine five years of $PM_{2.5}$ data recorded at local monitoring sites with MERRA-2 reanalysis data – including Aerosol Optical Depth (AOD), meteorological fields and $PM_{2.5}$ components. Based on these data, monthly $PM_{2.5}$ pollution maps were generated to analyze the spatial and temporal patterns of $PM_{2.5}$ concentration fields.

Figure 3.8. provides an overview of the methodology adopted to develop the regional application: analyze the spatial and temporal distribution of $PM_{2.5}$ pollution.

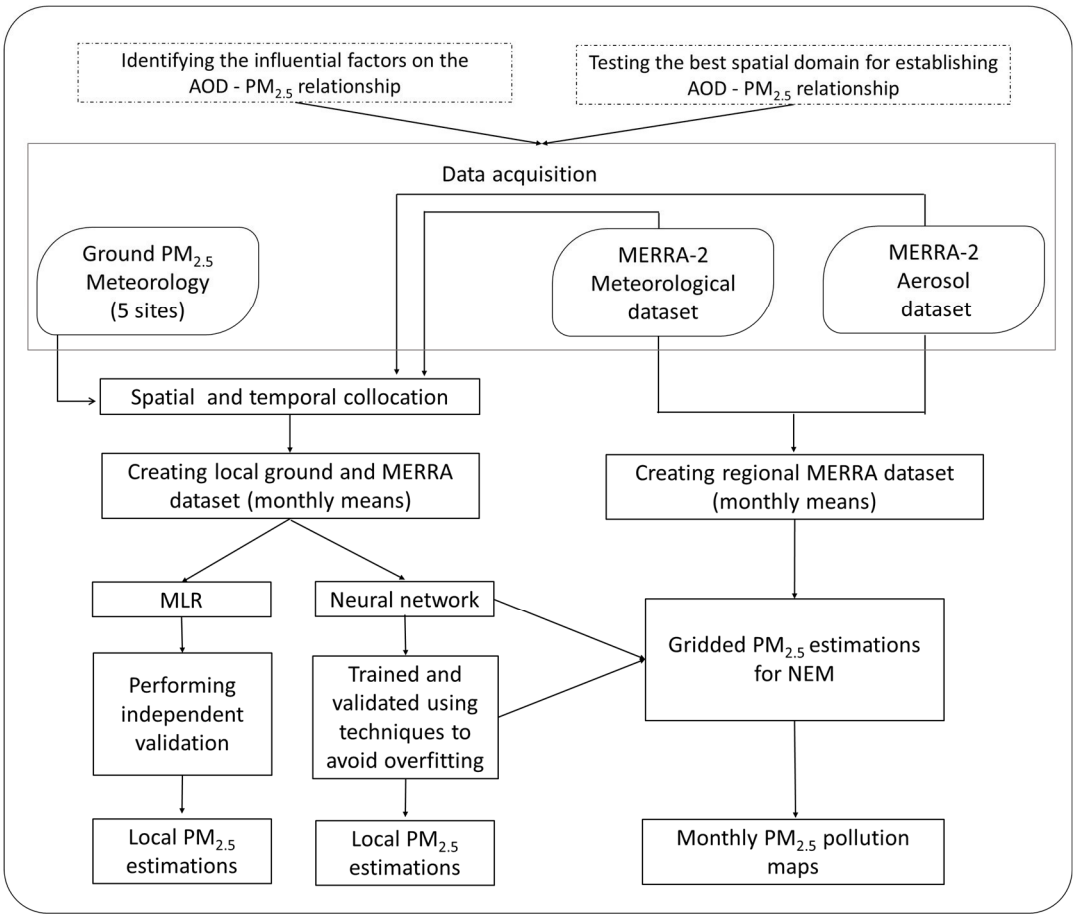


Figure 3.8. Methodology framework for analyzing the spatial - temporal distribution of $PM_{2.5}$

3.3.2 MERRA-2 datasets

3.3.2.1 Meteorological datasets:

The meteorological fields were extracted from the M2TMNXSLV -MERRA-2 tavgM_2d_slv_Nx: 2d, Monthly mean, Time-Averaged, Single-Level, Assimilation, Single-Level Diagnostics V5.12.4 files, at 0.5 ° x 0.625 °spatial resolution and monthly temporal resolution (GMAO, 2015a). The meteorological SDS used for developing the regional application in this research are presented in **Table 3.4.**

Table 3.4. Meteorological SDS from MERRA-2

Name	Long Name
T	"air_temperature"
RH	"relative_humidity_after_moist"
U	"eastward_wind"
V	"northward_wind"

3.3.2.2 Aerosol datasets

The aerosol components were extracted from the M2TMNXAER - MERRA-2 tavgM_2d_aer_Nx: 2d, Monthly mean, Time-averaged, Single-Level, Assimilation, Aerosol Diagnostics V5.12.4 files, at 0.5 ° x 0.625 °spatial resolution and monthly temporal resolution (GMAO, 2015b). The aerosol SDS used for developing the regional application in this research are presented in **Table 3.5.**

Table 3.5. Aerosol SDS from MERRA-2

Name	Long Name
BCSMASS	"Black_Carbon_Surface_Mass_Concentration"
DUSMASS25	"Dust_Surface_Mass_Concentration_PM_2.5"
OCSMASS	"Organic_Carbon_Surface_Mass_Concentration_ENSEMBLE"
SO4SMASS	"SO4_Surface_Mass_Concentration_ENSEMBLE"
SSSMASS25	"Sea_Salt_Surface_Mass_Concentration_PM_2.5"
TOTEXTTAU	"Total_Aerosol_Extinction_AOT_[550_nm]"

3.3.3 Model development

Multiple Linear Regression and Artificial Neural network were developed over MMA for estimating monthly $PM_{2.5}$ using both local meteorological data from SIMA records and meteorological and aerosol components from MERRA-2.

Monthly $PM_{2.5}$ pollution maps were generated using only MERRA-2 data as input in the NN to analyze the spatial and temporal patterns of $PM_{2.5}$ concentration fields in order to identify the regional $PM_{2.5}$ distributions over Northeastern Mexico.

Chapter 4 RESULTS

4.1 AOD and ground-based data relationship

4.1.1 Predictor variables and model selection

1,156 meteorological records were considered as valid data once the threshold criteria were applied and missing values and untrustworthy data were excluded from our analysis. Four out of seven meteorological variables recorded by SIMA were selected from the stepwise regression within the best subset of independent variables that can best describe the AOD - $PM_{2.5}$ relationship at MMA. The stepwise regression showed that the subset of meteorological parameters, along with AOD data, that best estimates $PM_{2.5}$ concentrations over MMA includes Temperature, Relative Humidity, Wind Speed and Wind Direction (*p-value* < 0.05 and *Mallows Cp* = 7.00):

$$PM_{2.5} = f(AOD_{MODIS}, T, RH, WS, WD)$$

The twenty-four refined datasets (four AOD – SDS, three spatial domains and two temporal schemes) are shown in **Table 4.1**.

Table 4.1. Refined datasets developed to evaluate the relationship among AOD, meteorological variables and $PM_{2.5}$ over the MMA

Model	Spatial domain	Temporal Scheme
$PM_{2.5} = f(AOD_{3K}, T, RH, WS, WD)$	D1, D2, D3	Hourly / Daily
$PM_{2.5} = f(AOD_{DT}, T, RH, WS, WD)$	D1, D2, D3	Hourly / Daily
$PM_{2.5} = f(AOD_{DB}, T, RH, WS, WD)$	D1, D2, D3	Hourly / Daily
$PM_{2.5} = f(AOD_{Cb}, T, RH, WS, WD)$	D1, D2, D3	Hourly / Daily

The selected meteorological predictors have clear physical meanings that should be considered when establishing relationships between ground-level $PM_{2.5}$ and satellite AOD, for example (Jiang et al., 2016; EPA, 2003):

Temperature: Warm temperatures are associated with increased evaporative, biogenic, and power plant emissions, which act to increase $PM_{2.5}$. High air temperature accelerates the generation of secondary particles or enhance vertical mixing of particles. $PM_{2.5}$ emissions may increase in cold temperatures for example due to home heating on winter nights. Although warm surface temperatures are generally associated with poor air quality conditions, very warm temperatures can decrease ground $PM_{2.5}$ levels by increasing the

vertical mixing and dispersion of pollutants. Heated aerosol plumes will be lofted to higher elevations, so it will affect the aerosol vertical distribution but not necessarily the AOD magnitude. This effect is also related to solar radiation.

Relative Humidity: Moisture acts to increase the production of secondary $PM_{2.5}$ including sulfates and nitrates. Relative humidity affects particle sizes and light extinction efficiency, and the neglect of it might bring overestimated or underestimated $PM_{2.5}$ concentrations from AOD. Some particles absorb water (hydrophilic) but certain aerosol particles, such as desert dust, are hydrophobic (do not absorb water). The size of hydrophilic aerosols depends on RH, because in the presence of water vapor, they tend to be swollen. Because the additional particle mass, AOD will increase.

Wind components: Strong surface winds tend to disperse $PM_{2.5}$ regardless of the season. However, strong winds can resuspend dust which can increase $PM_{2.5}$ concentrations. Wind components affect the mixing and dilution of pollutants. Winds are responsible for the long-range transport of aerosol plumes. As the aerosol plumes spread over larger regions, their concentration goes down and, therefore, the AOD decreases too.

Daily temperature (T) and relative humidity (RH) frequency distributions are shown in **Figure 4.1**. Wind roses from daily wind data are shown in **Figure 4.2**. The meteorological variables included in these figures were the only ones found to be statistically significant to our model.

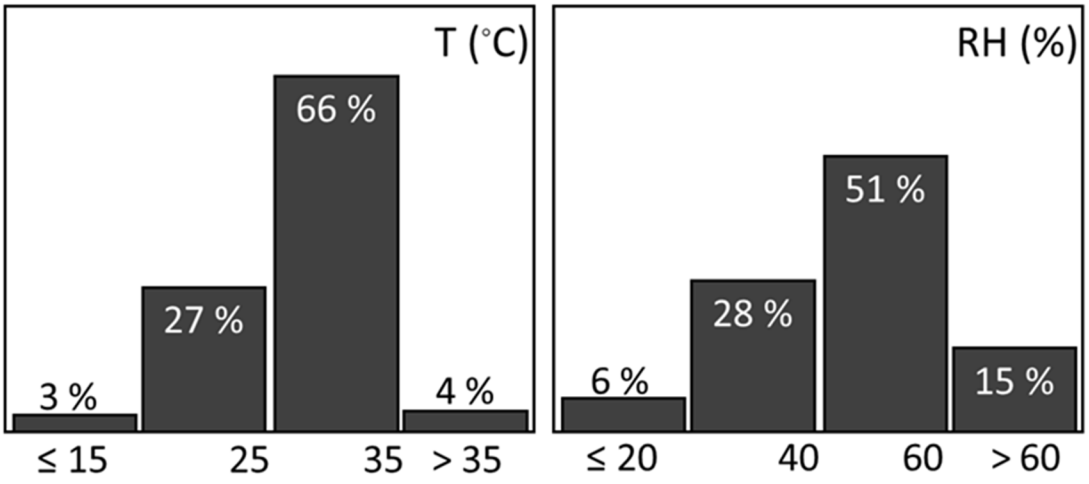


Figure 4.1. Histograms of temperature and relative humidity data included in the daily scheme over domain 3 (N = 694).

Based on SIMA data, the temperature at the MMA increases consistently from April to August. July and August are the hottest months with mean daily maximum and minimum temperatures at about 33 ◦C and 25 ◦C, respectively. After mid-October, temperature becomes progressively cooler. At the end of November temperatures drop rapidly till January. December and January are the coldest months with the mean daily maximum (22 ºC) and minimum (3 ºC) temperature. Monthly RH mean showed peak values of about 75% and above in the month of July. Monthly RH mean value decreased after February and continued to do so until April. In May, it showed a small increment and afterwards it followed, more or less, an increasing trend in July. Daily data showed that RH reached 96% and above on a few days in the month of December.

The east-to-west direction was the predominant wind pattern in the MMA for the period 2010-2017 (**Figure 4.2)**, which is typical for this region (Hernández Paniagua et al., 2017; Mancilla et al., 2019). During spring and summer, wind speed reaches maximum values, above 10 km h^{-1}, blowing primarily from the East (E) to West (W). Minimum wind speeds, with values of about 3 - 4 km h^{-1} occurred in winter and fall seasons, blowing mainly from the Northeast (NE) and only 10% of the days blowing from the Northwest (NW).

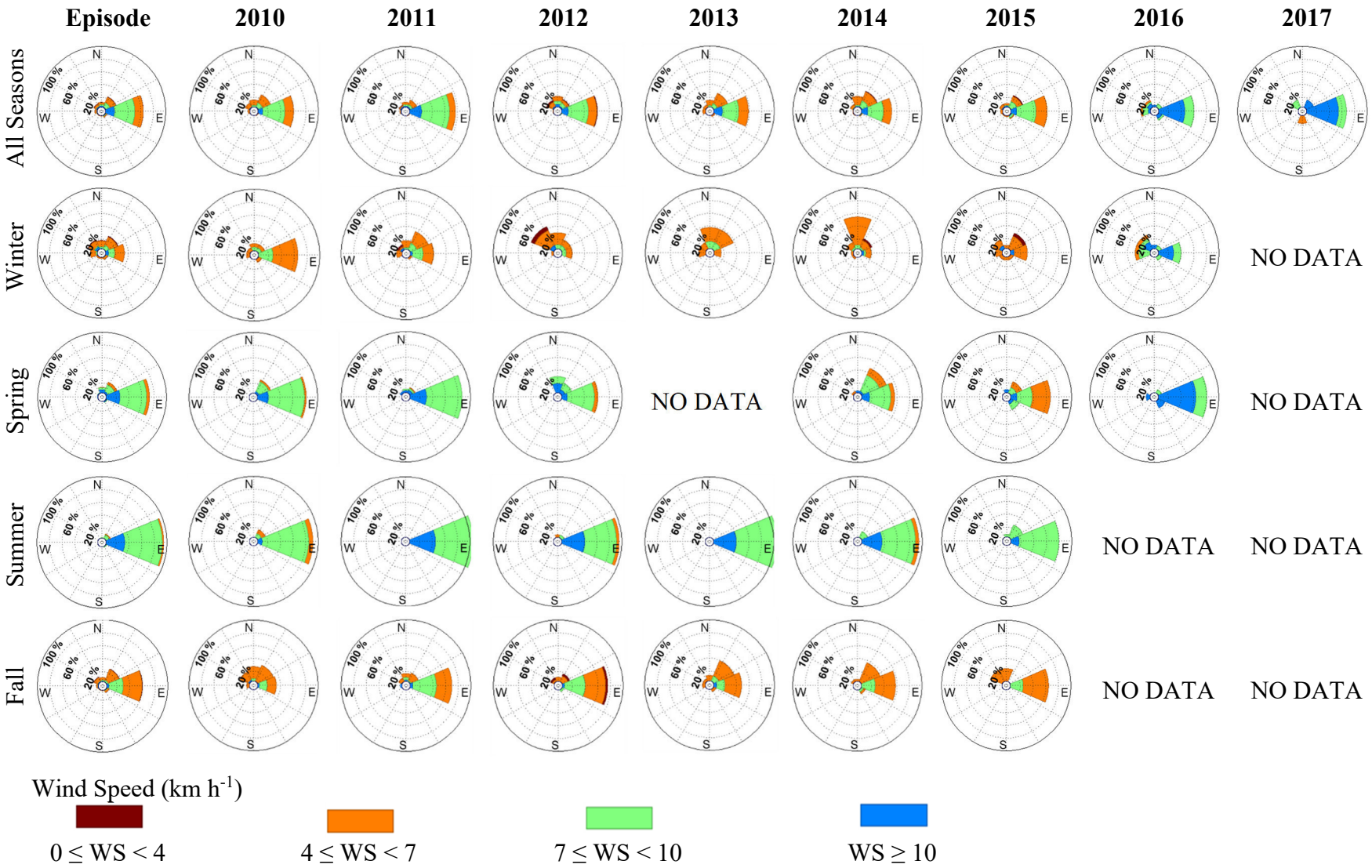


Figure 4.2. Wind rose diagrams for the entire study episode. (N = 694).

4.1.2 The spatio-temporal variability of the correlation among AOD, meteorological variables and $PM_{2.5}$

4.1.2.1 Spatial differences in the AOD selection

The descriptive statistics of the four AOD - SDS averaged in the three tested domains is presented in **Table 4.2**, including: the number of days (*N*) that AOD was available in each domain, percentage of days (% days) with complete data (AOD data availability matches SIMA meteorological valid data [1,156 records]), and the minimum, maximum, mean, and standard deviation values. It was noted that a smaller box size yields a higher AOD mean.

Table 4.2. The descriptive statistics of the four AOD - SDS averaged in the three tested domains.

	Days (*N*)	(%) days	Mean	Std. Dev	Min.	Max.
AOD_3K						
D1= 0.10 × 0.10 º	215	18.58	0.489	0.452	0.007	1.989
D2= 0.25 × 0.25 º	476	41.21	0.257	0.263	0.005	2.012
D3= 0.50 × 0.625 º	694	60.00	0.185	0.139	0.008	0.888
AOD_DT						
D1= 0.10 × 0.10 º	4	0.35	- -	- -	- -	- -
D2= 0.25 × 0.25 º	77	6.58	0.162	0.150	0.008	0.682
D3= 0.50 × 0.625 º	640	55.50	0.155	0.132	0.002	0.812
AOD_DB						
D1= 0.10 × 0.10 º	137	11.86	0.237	0.125	0.031	0.651
D2= 0.25 × 0.25 º	189	16.36	0.202	0.126	0.020	0.651
D3= 0.50 × 0.625 º	575	49.78	0.177	0.113	0.017	0.656
AOD_Cb						
D1= 0.10 × 0.10 º	4	0.35	- -	- -	- -	- -
D2= 0.25 × 0.25 º	77	6.67	0.162	0.150	0.008	0.682
D3= 0.50 × 0.625 º	653	56.54	0.154	0.131	0.009	0.812

Almost no pixels are available over D1 using the 10k_DT product. The bright surface from the urban area leads to pixels being identified as not appropriate for dark-target aerosol retrieval (Remer et al., 2013). There were no absolute differences between AOD values from 10k_DT SDS and Cb SDS at the smallest domains, where the number of available data was

not statistically significant. Some possibilities that explain this behavior is that probably in the smallest domain the NDVI > 0.30, therefore AOD is retrieved using dark target algorithm (NDVI > 0.30 = DT; NDVI < 0.20 = DB; 0.2 < NDVI < 0.30 = Cb –merged-), or it could indicate that in the smallest domain, 10k_DT has better confidence than DB, because the AOD in the Cb SDS is assigned to the algorithm (DT or DB) that has the higher confidence (Levy et al., 2013). Although the AOD_Cb data availability was slightly higher than the AOD_DT at the D3, descriptive statistics do not change up to the second digit of significance.

AOD_3K was the SDS with the highest number of available data, highest variability (standard deviation) and wider difference between minimum and maximum AOD at all domains. On the other hand, AOD_DB was the SDS with the least variability among the three domains. The resulting AOD_DB means were similar among the three domains and the difference between minimum and maximum AOD was bounded. The observed weak ability to report high (maximum) and low (minimum) AOD values from DB SDS (compared to AOD_3k) may increase the possibility of under- or overestimation of $PM_{2.5}$ using this SDS.

A significant increase in the availability of data in all SDSs is observed for domain 3, where AOD means were apparently similar from the four compared SDSs. A One-way ANOVA was performed and it revealed significant statistical differences among the four tested AOD SDS (*p-value* < 0.05). A boxplot chart was constructed to further understand the differences and similarities in the AOD distribution from the four SDS in domain 3 (**Figure 4.3**). The observed similarities were: asymmetric AOD distribution and positive skewness represented by the outlier at the upper end and longer upper whiskers. The observed differences were: interquartile range box indicates 50% of the AOD data are from 0.07 to 0.20 for both AOD_3k and AOD_DB, and for both AOD_DT and AOD_Cb 50% of the AOD data are from 0.05 to 0.15. The median (vertical bar) and interquartile range (box) were different among the SDSs. These statistics are better measures of central tendency and spread for highly skewed data because they are not affected by extreme observations as is the case of mean and standard deviation.

In conclusion, the AOD values from the four tested AOD – SDS were different. Spatial differences in the AOD selection were identified, showing that there is an effect of box size

on the mean AOD: smaller box size yields a higher AOD mean but bigger box size yields higher number of AOD available data.

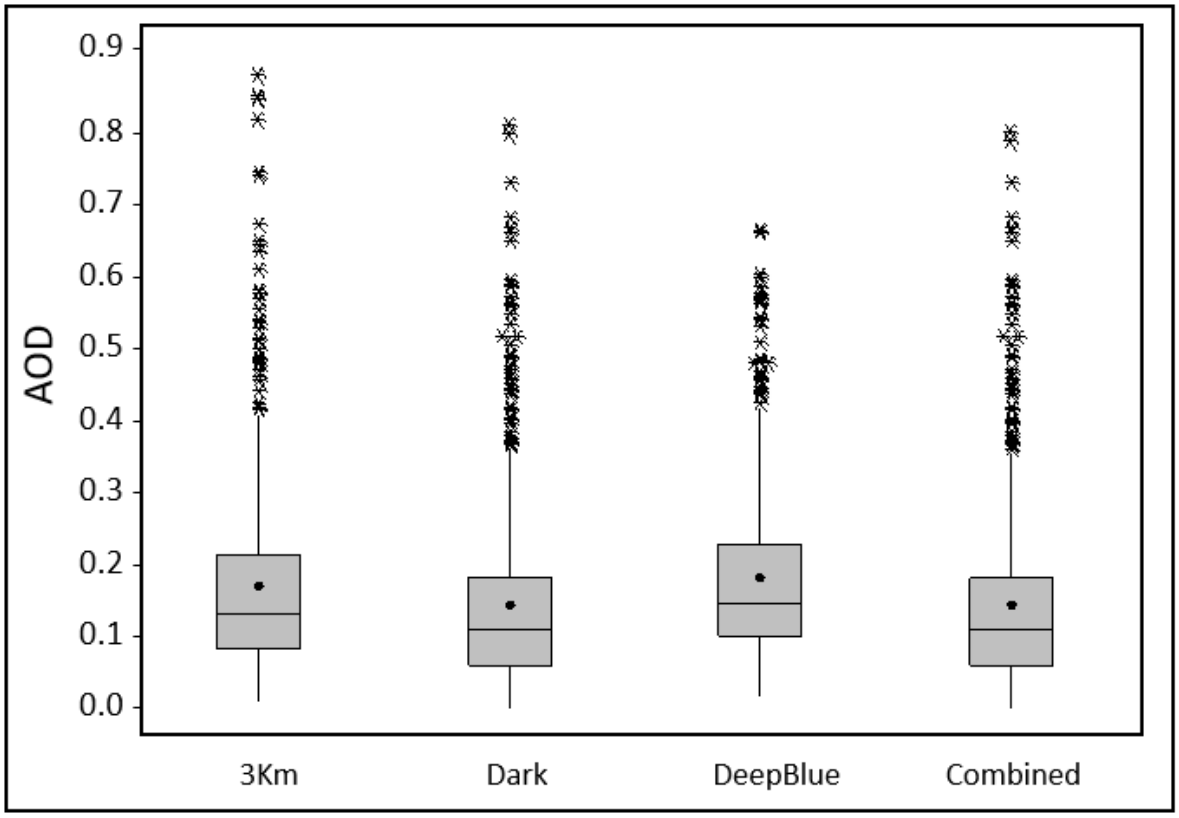


Figure 4.3. Similarities and differences in the AOD distribution from the four SDS in domain 3.
The median is represented by a vertical bar in the center of the box.
The mean is represented by filled dots
The outliers are represented by asterisk symbol

4.1.2.2 The spatio-temporal variability of the correlation among AOD, meteorological variables and $PM_{2.5}$

The statistical performance obtained using the MLR from each twenty-four refined datasets, which represent various spatio-temporal average schemes on the AOD and ground-based data relationship, is presented in **Table 4.3**. MPE was calculated instead of MBE (commonly used to measure average model bias) because error cancellation was observed. The negative sign observed in MPE indicates that in general the MLR models are overestimating the observed $PM_{2.5}$. RMSE assigns a relatively greater weight to large errors than MAE, as influenced especially by episodes of high $PM_{2.5}$ load in MMA that were underestimated by the MLR model.

Table 4.3. Impact of various spatio-temporal average schemes on the correlation among AOD, meteorological variables and $PM_{2.5}$.

	Hourly scheme						
	N	Mean $PM_{2.5}$* ($\mu g\ m^{-3}$)	Std. Dev*	*R*	MPE (%)	RMSE ($\mu g\ m^{-3}$)	MAE ($\mu g\ m^{-3}$)
$PM_{2.5\ Observed}$	694	29.39	12.32				
MLR_1 D1	215	28.85	13.24	0.35	-20.14	15.06	9.29
MLR_1 D2	476	29.54	10.39	0.53	-11.43	10.31	7.58
MLR_1 D3	694	29.39	10.80	0.50	-12.66	10.75	7.90
MLR_2 D1	4	- -	- -	- -	- -	- -	- -
MLR_2 D2	77	28.85	8.90	0.67	-9.86	8.47	7.14
MLR_2 D3	640	29.53	10.70	0.58	-10.67	10.64	7.84
MLR_3 D1	137	29.70	13.87	* *	-24.11	13.64	7.87
MLR_3 D2	189	31.29	10.60	0.59	-10.56	10.40	7.79
MLR_3 D3	575	29.71	10.73	0.52	-12.18	10.65	7.71
MLR_4 D1	4	- -	- -	- -	- -	- -	- -
MLR_4 D2	77	28.85	8.90	0.67	-9.86	8.47	7.14
MLR_4 D3	653	29.40	10.65	0.49	-12.45	10.59	7.79
	Daily scheme						
$PM_{2.5\ Observed}$	694	27.88	8.52				
MLR_1 D1	215	28.07	9.08	0.48	-10.57	8.95	6.14
MLR_1 D2	476	28.35	7.24	0.55	-5.72	7.19	5.44
MLR_1 D3	694	27.88	7.03	0.57	-5.77	7.00	5.29
MLR_2 D1	4	- -	- -	- -	- -	- -	- -
MLR_2 D2	77	28.16	5.87	0.71	-3.43	5.59	4.55
MLR_2 D3	640	28.10	6.95	0.57	-5.49	6.91	5.34
MLR_3 D1	137	29.92	9.08	* *	-10.92	9.57	5.91
MLR_3 D2	189	30.26	7.40	0.57	-5.13	7.26	5.84
MLR_3 D3	575	28.34	6.88	0.60	-5.18	6.84	5.06
MLR_4 D1	4	- -	- -	- -	- -	- -	- -
MLR_4 D2	77	28.16	5.87	0.71	-3.43	5.59	4.55
MLR_4 D3	653	28.01	6.92	0.59	-5.45	6.88	5.20

* For the MLR, mean $PM_{2.5}$ and standard deviation are estimated from the MLR models

* * AOD is not statistically significant in the MLR

MLR_1 Multiple Linear Regression based on $PM_{2.5} = f(AOD_3K, T, RH, WS, WD)$

MLR_2 Multiple Linear Regression based on $PM_{2.5} = f(AOD_DT, T, RH, WS, WD)$

MLR_3 Multiple Linear Regression based on $PM_{2.5} = f(AOD_DB, T, RH, WS, WD)$

MLR_4 Multiple Linear Regression based on $PM_{2.5} = f(AOD_Cb, T, RH, WS, WD)$

When MLR is estimated using AOD_3k, it is observed that domain 1 shows the greatest variability, the lowest correlation coefficient and the highest errors, independently of the temporal averaging scheme. Domain 2 and domain 3, have similar performance among them. Using the hourly scheme, performance drops from domain 2 to domain 3, unlike using the daily scheme, in which performance increases from domain 2 to domain 3. Domain 3 has greater data availability among the three compared domains. The best performance using AOD from 3k_DT is obtained using the daily scheme over domain 3.

When MLR is estimated using AOD values from AOD_DT, it is observed that over domain 1, almost no pixels are available and the statistical parameters could not be calculated. In domain 2, using the hourly and daily schemes, both AOD_DT and AOD_Cb show the best performance compared to the four SDS and the three evaluated domains. However, data availability is still low to build a good predictor model. In domain 3, the statistical performance declines when compared to domain 2 in spite of a substantial increase in data availability from domain 2 to domain 3. In comparison to the other evaluated SDSs, with the hourly scheme, the inclusion of the AOD_DT shows the best performance in the domain 3.

When MLR is estimated using AOD values from DB, it is observed that, using the hourly and daily schemes, AOD_DB is not an explanatory variable in the MLR at domain 1, where this model shows the worst performance for both the four SDSs over the three evaluated domains. Using the hourly scheme, AOD_DB in domain 2 shows the best performance, but there is low data availability. In domain 3, the statistical performance declines compared to domain 2, but with a substantial increase in data availability, in contrast to the daily scheme, where performance increases from domain 2 to domain 3. Among the compared SDSs, DB is the one that has the best performance over the domain 3 using the daily scheme.

AOD from 10k_DT and AOD_Cb SDSs presents the same values in domain 1 and domain 2, therefore the same statistical performance occurs in these domains. In domain 3, possibly due to the heterogeneity of surface conditions (urban, rural, natural), the combined SDS reports few extra AOD pixels than the 10k_DT. However, using the hourly scheme, these extra AOD pixels do not enhance the model performance causing a decline in the statistical metrics compared to AOD_DT over domain 2 and domain 3. Using the daily scheme, AOD

values from AOD_Cb slightly improves the $PM_{2.5}$ estimations with respect to AOD_DT over the domain 3.

Using the hourly scheme, domain 1 represents the spatial collocation to the satellite data on the ground-based monitoring sites, coinciding with the overpass time of the AQUA satellite. The AOD data availability in domain 1 is scarce, where AOD_DT and AOD_Cb have almost no pixels. AOD_DB is not an explanatory variable in the MLR using the hourly and daily schemes. While data availability from AOD_3k is the highest in domain 1, a poor performance is observed using the hourly scheme, improving the performance using the daily scheme. Over domain 2, using the hourly scheme, the four evaluated SDSs yield the best performance in the MLR. Using the daily scheme, the best performance is yielded when AOD values from 10k_DT or Cb are used. Although the AOD data availability over domain 2 is better than over domain 1, the availability of data is still limiting to develop a good predictive model. Domain 3 is the spatial scheme with higher availability of AOD data from the four SDSs. In domain 3, the models that include AOD_3k and AOD_DB get the best performance using the daily scheme.

The MLR using AOD MODIS from collection 6 is limited to represent the urban $PM_{2.5}$ pollution at the finest domain. Recently, the MODIS Collection 6.1 was released. This version incorporates the surface reflectance scheme proposed by Gupta et al. (2016b), for retrieving DT AOD over urban surfaces. It is expected that the use of DT AOD from this collection will greatly reduce the high bias observed in the DT AOD over cities, becoming the product more useful for urban air quality applications than previous collections.

Overall, data availability increases when the size of domain increases, so domain 3 provides the best AOD data representativeness. Therefore, data availability depends on the spatial scheme rather than on the temporal scheme. Correlation coefficient and error estimators are dependent of both spatial and temporal schemes: it is found that correlation coefficients changed significantly as a function of box size. Errors are larger in the smallest domain and larger using the hourly scheme than using the daily scheme. In conclusion, the correlation between $PM_{2.5}$ and AOD, to some extent, depends on spatial and temporal average scheme of MODIS AOD (Guo et al., 2017).

4.1.2.3 Differences between temporal schemes

The mean and standard deviation from the MLR performed at the twenty-four refined datasets are in the same range of the mean and variability from the observed dataset at both hourly and daily schemes. The daily scheme is the temporal scheme that best related AOD and ground-based data. In contrast, Guo et al. (2017) found that the best model performance is achieved when the hourly scheme is used.

A one-way ANOVA revealed that there is statistical difference (*p-value* < 0.05) between observed hourly-averaged $PM_{2.5}$ values (29.39 ± 12.32 µg m^{-3}) and observed daily-averaged $PM_{2.5}$ values (27.88 ± 8.52 µg m^{-3}). As shown in **Figure 4.4**, the observed hourly $PM_{2.5}$ and daily $PM_{2.5}$, have similar means and medians, but different interquartile range box. It is clear that the 24-h average used in the daily scheme softened the extreme $PM_{2.5}$ values observed in the hourly scheme.

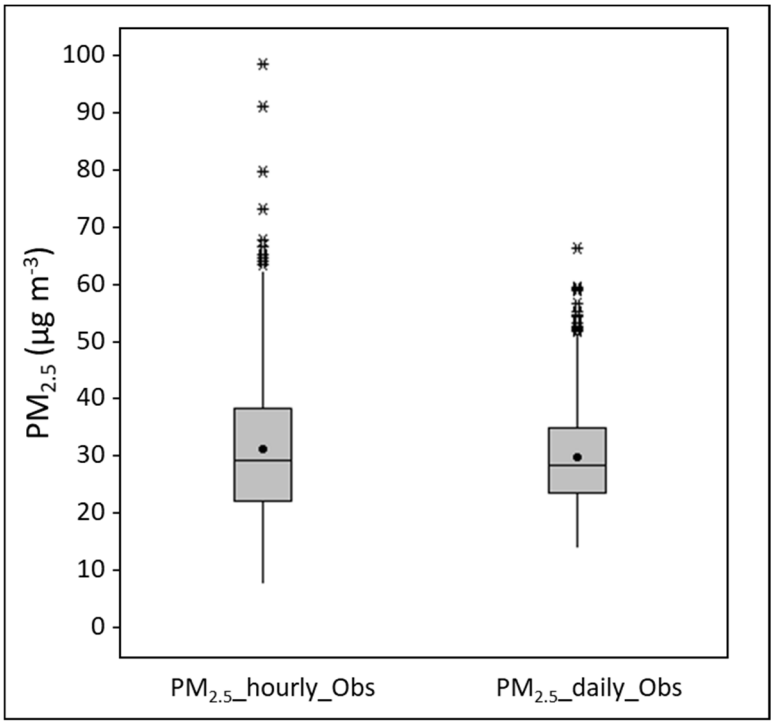


Figure 4.4. Similarities and differences in the data distribution of the observed hourly $PM_{2.5}$ means and the observed daily $PM_{2.5}$ means.

Diurnal fluctuation in $PM_{2.5}$ concentrations is observed (**Figure 4.5**). The $PM_{2.5}$ concentration displays a multimodal pattern, with peaks between 12:00 and 14:00 hrs (probably due to emissions and secondary formation), between 18:00 and 20:00 hrs (probably

due to emissions) and between 23:00 and 01:00 (possibly attributed to a decrease in mixing layer height). The minimum concentration generally occurs between 8:00 and 10:00 hrs. Between 12:00 and 15:00 hrs, which is the satellite over pass time, the observed $PM_{2.5}$ averaged for the entire period falls within the ranges of observed hourly mean and observed daily mean. However, the latter is more of a coincidence, and not always the case for all regions (e.g. the hourly and daily means do not necessarily fall between the hourly behavior confidence intervals in other places of the world). This might explain why, despite the fact that AOD is reported only at the satellite over pass time, good statistical performance in the daily scheme is achieved.

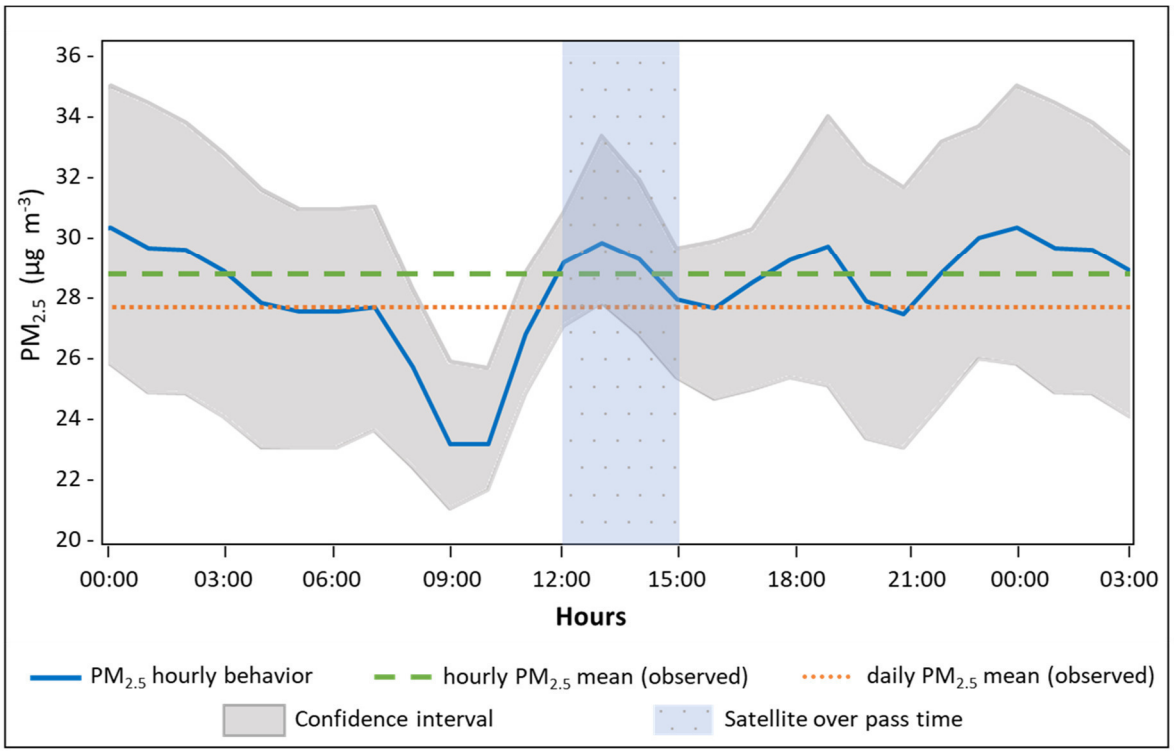


Figure 4.5. Diurnal $PM_{2.5}$ variability.

4.1.2.4 Selection of the "best dataset"

The statistical performance from the MLR models is similar in terms of correlation coefficient and error analysis when using any SDS (no matter what SDS is chosen) in domain 3 with daily scheme, but differs in the data availability, as shown in **Section 4.1.2.2**. The refined dataset that include AOD from 3k_DT, over domain 3 and daily scheme provides the greatest number of available data. Hence, the refined dataset "$PM_{2.5} = f(AOD_{3K}, T, RH, WS, WD)$, over

domain 3 and using the daily scheme" is selected in this study as the "best dataset" for its statistical performance and data availability.

4.1.3 Model development and validation

4.1.3.1 MLR model validation

A one-way ANOVA revealed that there is no difference between the $PM_{2.5}$ estimated from the conventional MLR and the $PM_{2.5}$ estimated from the ensemble model (*p-value* < 0.05). Therefore, we conclude that performing MLR using the *best dataset* is robust and the model performance was not influenced by the dataset. Accordingly, the relation $PM_{2.5}$-AOD from the *best dataset* could be tested for improving it using NN models. The regression coefficients (β) obtained from each iteration, the ensemble MLR model and the MLR using the original *best dataset* (AOD_3K averaged over domain 3 in combination with meteorological *variables at the* daily scheme) are shown in **Table 4.4**.

Table 4.4. Ensemble model development and comparison with the MLR using the original *best dataset*

			AOD_3K		T		RH		WS		U		V
	β_0		β_1		β_2		β_3		β_4		β_5		β_6
Iteration 1	18.2	+	24.2	+	0.316	+	0.156	−	0.837	−	11.5	−	2.53
Iteration 2	20.7	+	22.9	+	0.21	+	0.157	−	0.826	−	11.4	−	2.98
Iteration 3	21.1	+	24.6	+	0.242	+	0.135	−	0.912	−	9.91	−	2.53
Iteration 4	19.7	+	24.7	+	0.28	+	0.152	−	0.827	−	12.4	−	3.3
Iteration 5	19.8	+	20.9	+	0.3	+	0.157	−	0.952	−	10.5	−	2.48
Iteration 6	17.5	+	23.0	+	0.276	+	0.174	−	0.745	−	11.3	−	2.9
Iteration 7	17.9	+	25.2	+	0.232	+	0.195	−	0.718	−	12.8	−	4.63
Iteration 8	21.2	+	26.8	+	0.217	+	0.148	−	0.963	−	11	−	2.63
Iteration 9	18.8	+	28.1	+	0.231	+	0.149	−	0.669	−	11.4	−	3.35
Iteration 10	19.2	+	22.2	+	0.276	+	0.166	−	0.801	−	12.3	−	4.5
Ensemble	19.41	+	24.26	+	0.26	+	0.16	−	0.83	−	11.45	−	3.18
Non Ensemble	19.40	+	23.80	+	0.26	+	0.16	−	0.80	−	11.60	−	3.51

4.1.3.2 NN model validation

The same number of neurons was used in the hidden layer as in the input layer as this produced generally good results. The best validation performance was yielded with a RMSE = 5.91 μg m^{-3} in the validation dataset, as shown in **Figure 4.6**

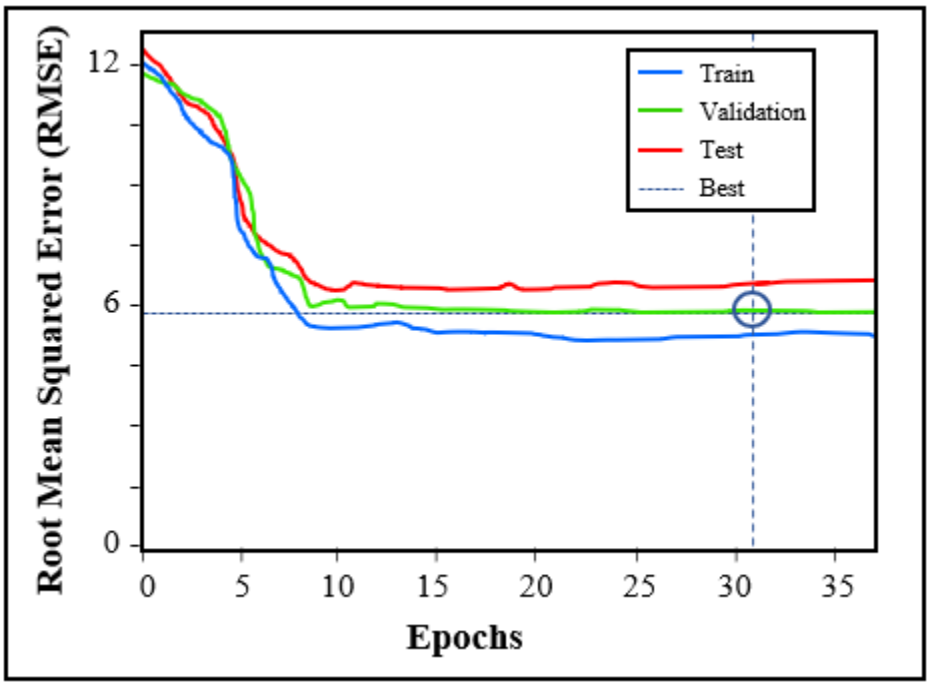


Figure 4.6. NN model validation. Train, validation and test lines should lie on or near the best (dotted) line, then it can be confirmed that training has been done successfully. In other words, if any of the 3 (training, validation and test) lines meet or pass near the best line, it means convergence has been obtained. One Epoch is when the dataset is passed forward and backward through the neural network only once.

4.1.3.3 Performance of statistical models: MLR and NN (entire episode)

The predictive abilities of the models are shown in **Figure 4.7**. A scatterplot from the ensemble MLR model is shown in **Figure 4.7a**. The correlation between observed and estimated $PM_{2.5}$ via MLR model exhibits a moderate but statistically significant correlation at a 95% confidence level (R = 0.56, MPE = − 5.77%, RMSE = 7.00 μg m^{-3}, and MAE= 5.28 μg m^{-3}). As in Guo et al. (2017), the MLR model tended to overestimate lower observed $PM_{2.5}$ levels and underestimate higher observed $PM_{2.5}$ levels

Scatterplot from NN model is shown in **Figure 4.7b**, the NN model improves the $PM_{2.5}$ estimations compared to MLR, with the following performance: R = 0.73, MPE = − 2.59%, RMSE = 5.83 μg m^{-3}, MAE= 4.24 μg m^{-3} and *p-value* < 0.05. The underestimation and

overestimation in $PM_{2.5}$ that occurs in MLR is lower using NN. Similarly, to our findings, Gupta and Christopher (2009a and 2009b) and Kamarul et al. (2017) found that NN model improves the PM estimates compared to linear models.

Model fitting for both MLR and NN exhibited a large degree of seasonal variability. AOD and $PM_{2.5}$ showed seasonal differences. For example, $PM_{2.5}$ values are higher in winter (64.05 µg m^{-3}) whereas AOD is higher in spring (0.89). According to Guo et al. (2017) the different observed seasonal patterns found between $PM_{2.5}$ and MODIS AOD arise from their complex relationship, which can be influenced by various factors such as aerosol types, aerosol chemical compositions, aerosol size distributions, aerosol vertical profiles, and so on (Ma et al., 2014; Song et al., 2014; You et al., 2015).

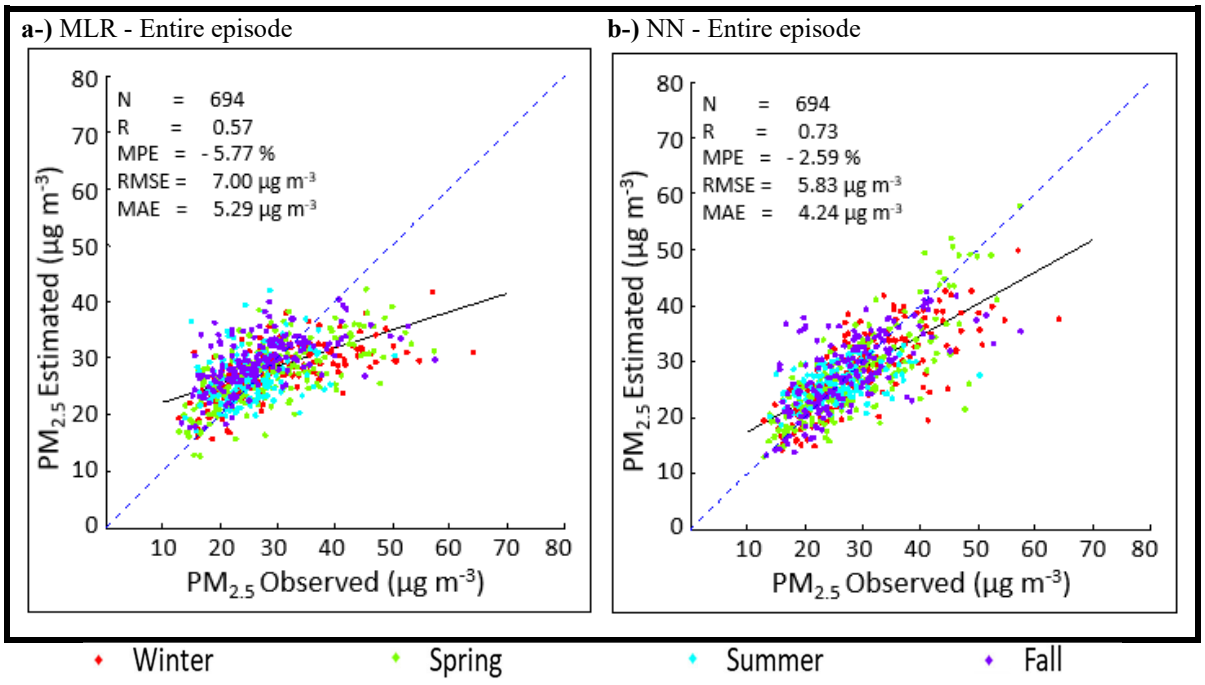


Figure 4.7. Statistical performance obtained using MLR and NN models to estimate daily $PM_{2.5}$ means, over the MMA. (all episode)
Figure 4.7a, Statistical performance obtained using MLR (non-ensemble).
Figure 4.7b, Statistical performance obtained using NN model.
The dash line is the 1:1 line as a reference.

4.1.3.4 Performance of statistical models: MLR and NN (by seasons)

A scatterplot from MLR performed for each season (non-ensemble) is shown in **Figure 4.8a**. When compared to MLR using the entire dataset, overall, the correlation coefficients increased (except in summer), and the error parameters decreased (except in winter). The

predicted - observed plot showed weak correlation between summer's predictions and actual observations. MLR per season underpredicts high $PM_{2.5}$ levels, and overpredicts low $PM_{2.5}$ levels (as it was observed using the entire dataset). For example, in fall, high $PM_{2.5}$ is predicted using the MODIS AOD while lower concentrations are measured at the ground-based monitoring sites. According to Tesche et al. (2016), two possible explanations for this kind of discrepancy are the influence of clouds in the AOD retrieval and/or the presence of elevated aerosol layers above the boundary layer.

A scatterplot from NN trained per season is shown in **Figure 4.8b**. Correlation coefficient improved almost in 20% when compared to MLR performed by seasons. Despite the good performance obtained in wintertime, extreme underestimated outliers are observed.

When MLR and NN models were run by season, the best model performance was obtained for the spring period. In spring, the lower relative humidity -when compared to other seasons- (Carrillo et al., 2017) and the strong atmospheric dispersion (Mancilla et al., 2015) may enhance the correlation between AOD and $PM_{2.5}$ (Kong et al., 2017).

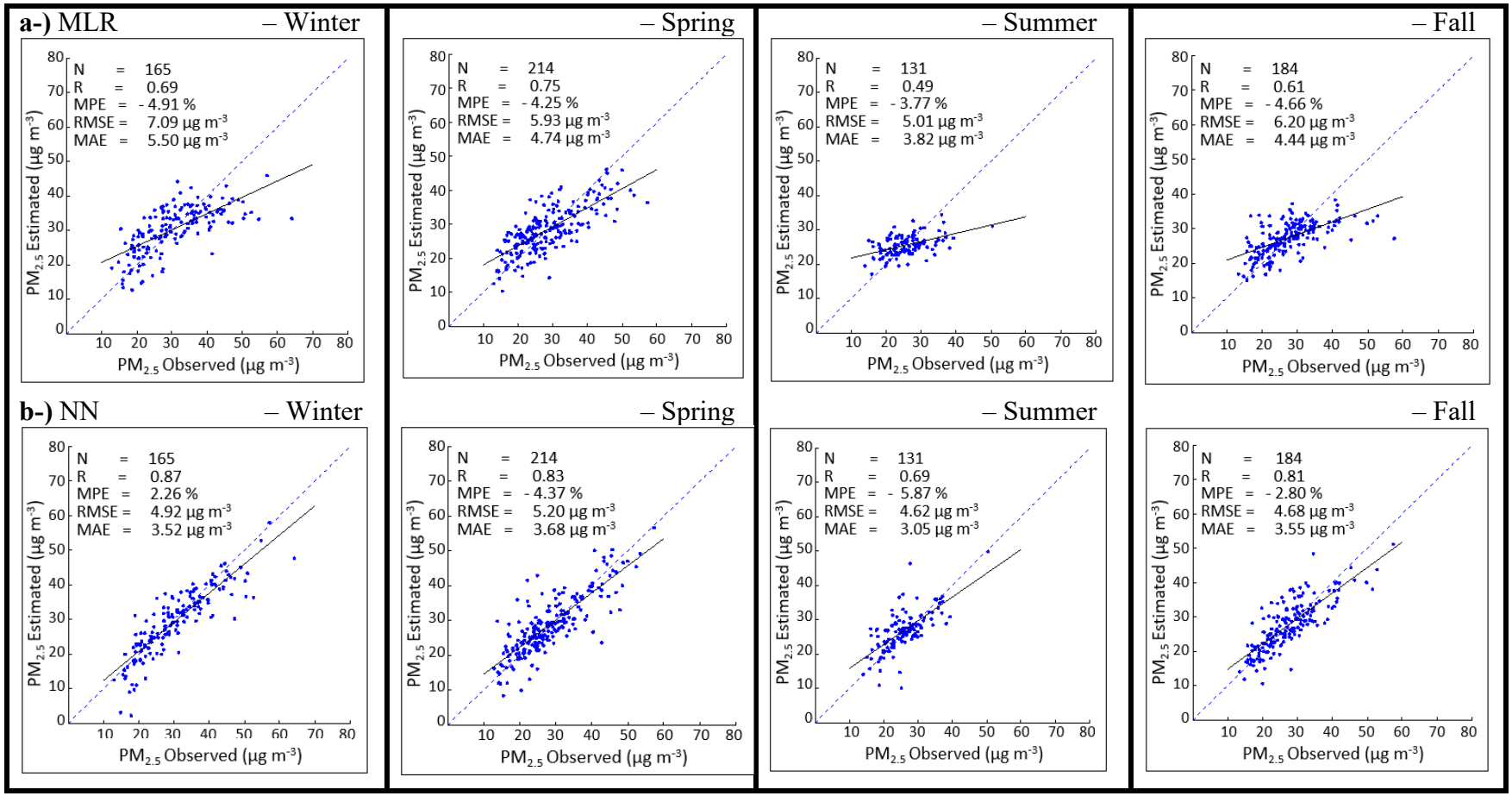


Figure 4.8. Statistical performance obtained by each season using MLR and NN models to estimate daily $PM_{2.5}$ means, over the MMA.

Figure 4.8a, Statistical performance obtained by each season using MLR.
Figure 4.8b, Statistical performance obtained by each season using NN.

Finally, a scatterplot from MLR merged by seasons and a scatterplot from NN trained by seasons (NNbS) are shown in **Figure 4.9a** and **Figure 4.9b**, respectively. The fitted results in NNbS were able to explain 86% of the variability in the corresponding $PM_{2.5}$ mass concentrations, which offered improvement on model accuracy from approximately 0.57 (MLR all data) to 0.83.

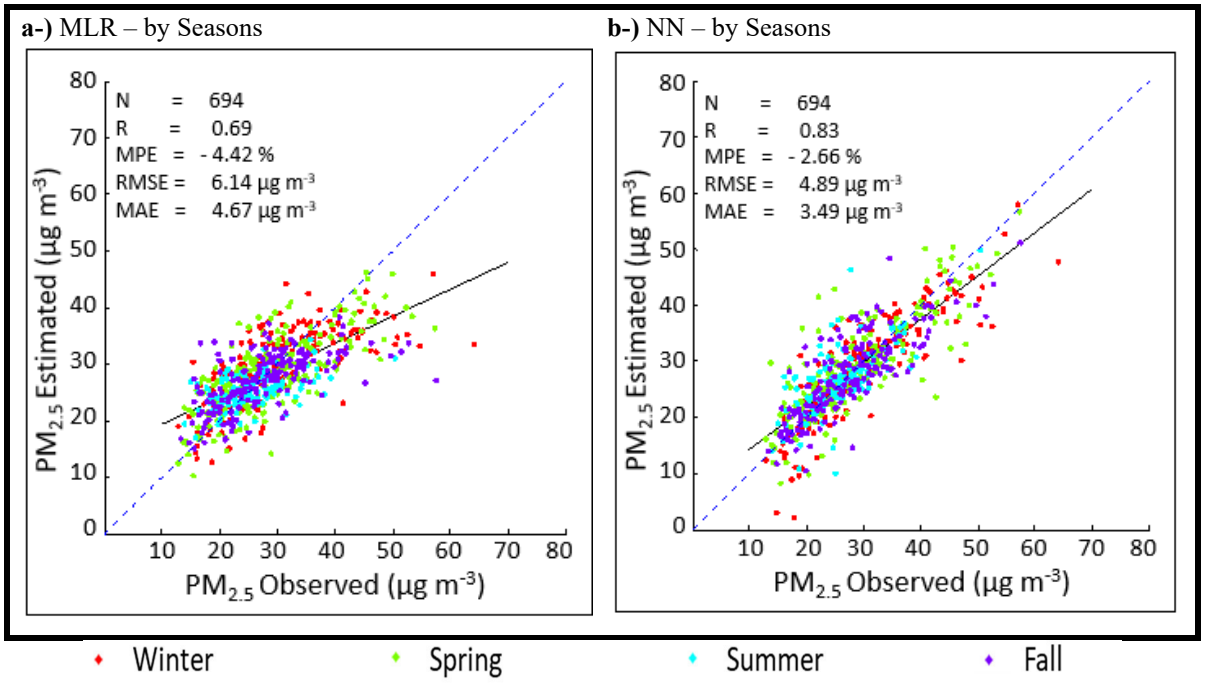


Figure 4.9. Statistical performance obtained using the daily $PM_{2.5}$ means estimated by the MLRs performed by seasons and the daily $PM_{2.5}$ means estimated by the NNs trained by seasons, over the MMA. (The fitted models results were merged by seasons).
Figure 4.9a, Statistical performance obtained using MLR, merged by seasons
Figure 4.9b, Statistical performance obtained using NN, merged by seasons

4.1.3.5 Comparison of developed models.

The statistical performance obtained from the models developed in this research are summarized in **Table 4.5**. The $PM_{2.5}$ estimations agree generally well with the observations exhibiting the fluctuations observed in the time series plot (**Figure 4.10.**).

These results suggest that this approach is useful for estimating large-scale ground-level $PM_{2.5}$ distributions especially in urban areas where none or without extensive ground-based $PM_{2.5}$ monitoring networks

Table 4.5. Summary of the descriptive statistics of the daily $PM_{2.5}$ (observed and estimated) and statistical performance obtained using the models developed in this study.

			W	Sp	Su	F						
$PM_{2.5}$ observed												
Mean	27.88		30.27	28.18	25.45	27.12						
Std Dev	8.52		9.77	8.99	5.78	7.83						
Min	12.80		12.80	12.82	13.85	13.09						
Max	64.05		64.05	57.15	50.35	57.42						
			MLR				NN				MLR merged	NN trained
	MLR	NN	W	Sp	Su	F	W	Sp	Su	F	by Seasons	by Seasons
$PM_{2.5}$ estimated												
Mean	27.88	27.59	30.27	28.18	25.45	27.12	29.37	28.71	26.49	27.37	27.88	28.09
Std Dev	4.84	6.65	6.70	6.74	2.85	4.76	9.41	8.45	5.76	7.17	5.89	7.99
Min	12.76	13.05	12.80	10.28	17.08	15.25	2.24	8.42	10.17	10.50	10.28	2.24
Max	42.31	57.78	46.01	46.26	34.53	38.44	57.94	56.78	50.04	51.20	46.26	57.94
N	694	694	165	214	131	184	165	214	131	184	694	694
R^2	0.32	0.53	0.47	0.56	0.24	0.37	0.76	0.68	0.48	0.65	0.48	0.68
R	0.57	0.73	0.69	0.75	0.49	0.61	0.87	0.83	0.69	0.81	0.69	0.83
MPE (%)	-5.77	-2.59	-4.91	-4.25	-3.77	-4.66	**2.26**	-4.37	-5.87	-2.80	-4.42	-2.66
RMSE ($\mu g\ m^{-3}$)	7.00	5.83	7.09	5.93	5.01	6.20	4.92	5.20	4.62	4.68	6.14	4.89
MAE ($\mu g\ m^{-3}$)	5.29	4.24	5.50	4.74	3.82	4.44	3.52	3.68	3.05	3.55	4.67	3.49
MBE ($\mu g\ m^{-3}$)	0.00	0.29	0.00	0.00	0.00	0.00	0.89	-0.53	-1.04	-0.25	0.00	-0.22

W: Winter. Sp: Spring. Su: Summer. F: Fall

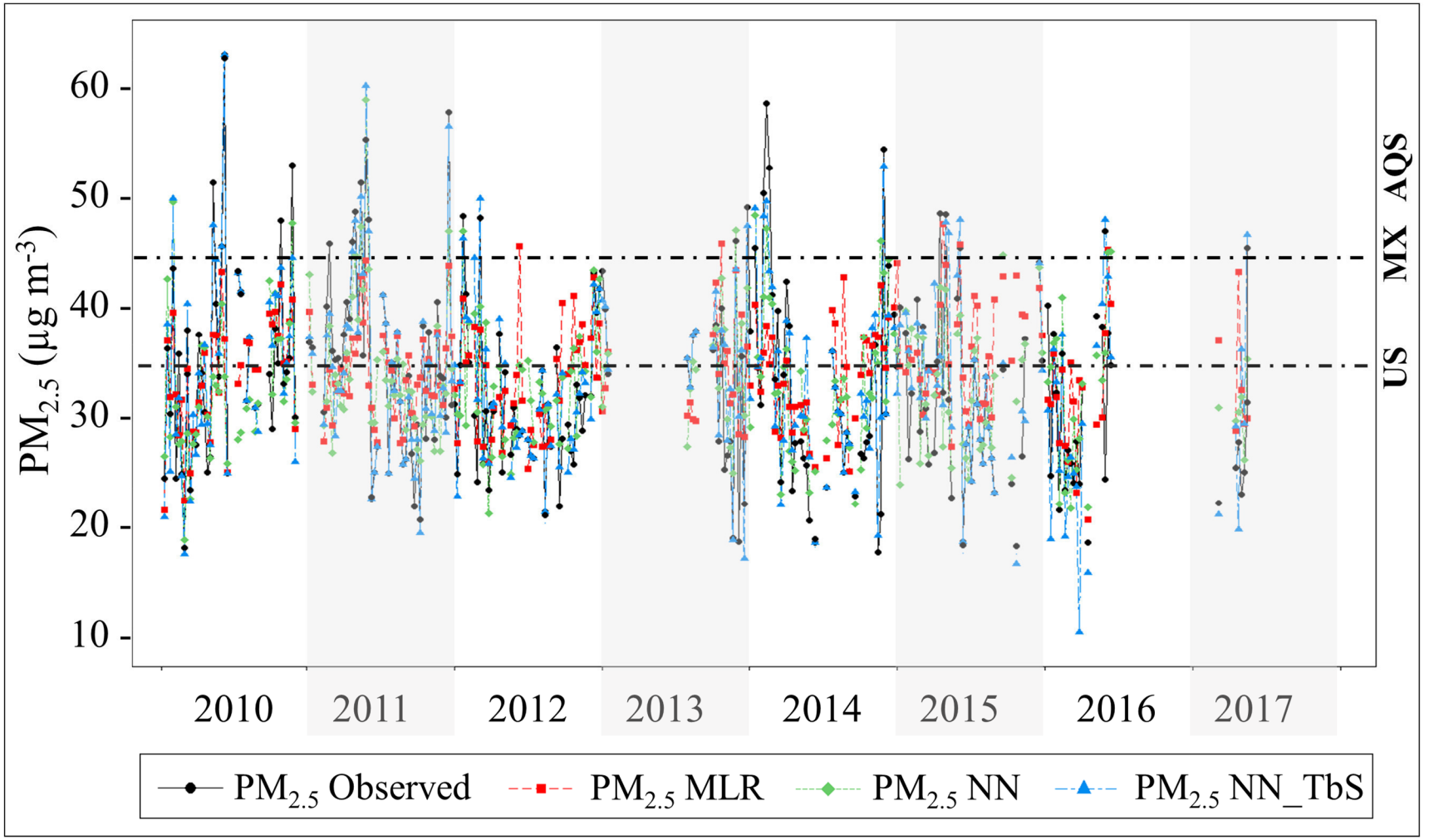


Figure 4.10. Comparison of daily $PM_{2.5}$ means estimated by the statistical models developed in this study and the daily $PM_{2.5}$ means recorded by SIMA ($PM_{2.5}$ observed).

4.2 Developing a local application: Categorization of metropolitan air quality index based on satellite-derived data and meteorological ground-based measurements.

4.2.1 Air quality index classification

The $PM_{2.5}$ MX-AQI classifications derived from the application of the MLR and NN models using the *best dataset* are presented through frequencies plots in **Figure 4.11.**

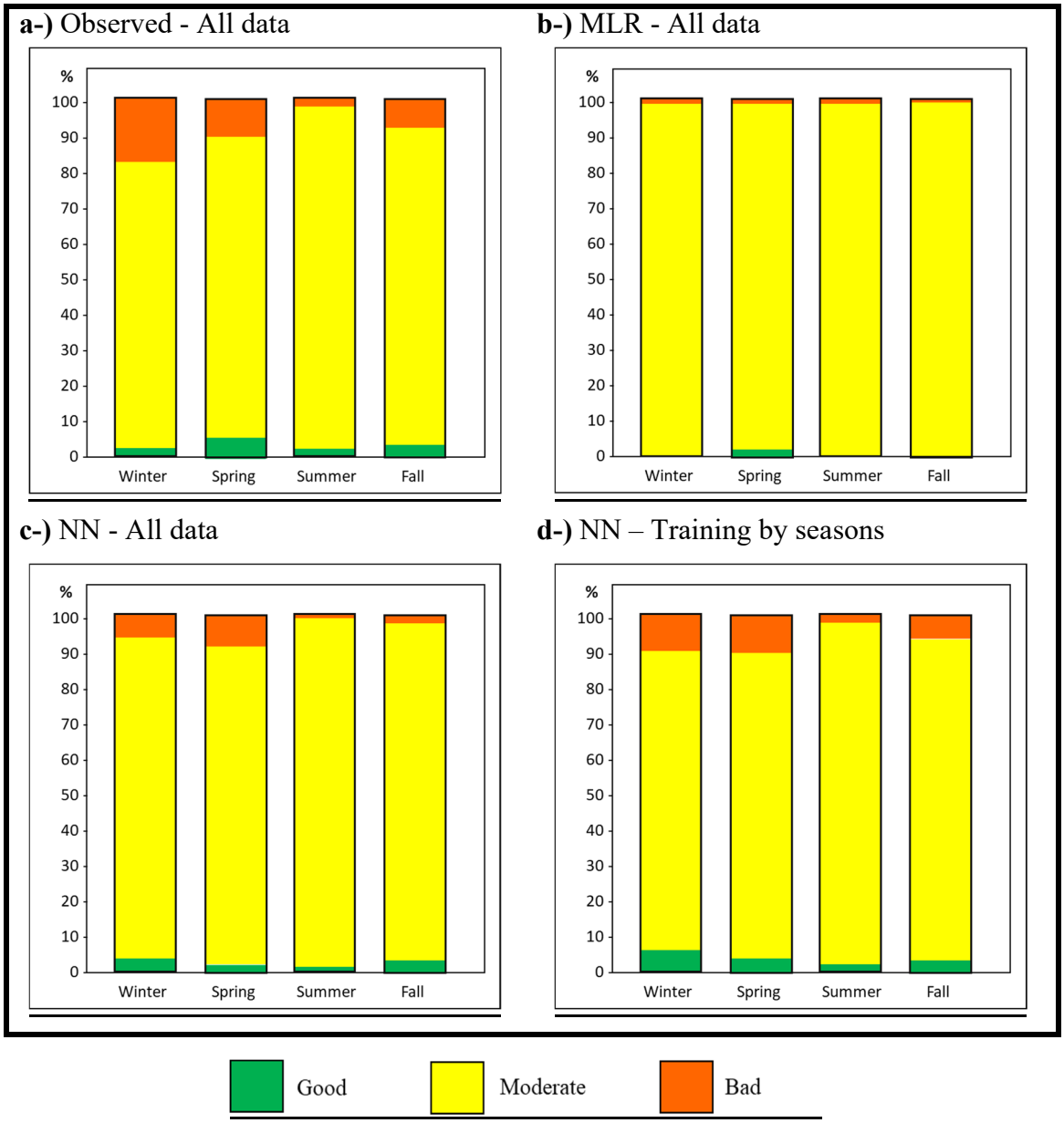


Figure 4.11. Frequencies of MX-AQI categories for $PM_{2.5}$.
a) MX-AQI frequencies plot for observed $PM_{2.5}$
b) MX-AQI frequencies plot for estimated $PM_{2.5}$ using MLR
c) MX-AQI frequencies plot for estimated $PM_{2.5}$ using NN
d) MX-AQI frequencies plot for estimated $PM_{2.5}$ using NN training by seasons

$PM_{2.5}$ MX-AQI categories for the observed $PM_{2.5}$ were found to be good, moderate and bad. The observed $PM_{2.5}$ MX-AQI reveals a distinct seasonal variation: a decreasing trend is observed from spring to summer, but then it gradually increases from fall to winter, indicating that urban air quality across the region is much worse in winter, but relatively better in summer. Meteorological conditions in winter can be unfavorable for pollutant dispersion/removal, including the frequent occurrence of stagnant weather associated with high pressures, temperature inversion and less precipitation. In addition, the bad air quality in winter is generally attributed to increased air pollutant emissions due to increased use in fossil fuels for heating purposes (Wang et al., 2017).

The $PM_{2.5}$ MX-AQI classification using the MLR showed a stable seasonal trend, where the moderate level classification was predominant. On the other hand, the $PM_{2.5}$ MX-AQI classification using both NN and NNbS presented similar seasonal variation to that obtained with observed $PM_{2.5}$-AQI.

4.2.2 Classification behavior

The overall $PM_{2.5}$ MX-AQI classification behavior using the tested statistical models is presented in **Figure 4.12.**

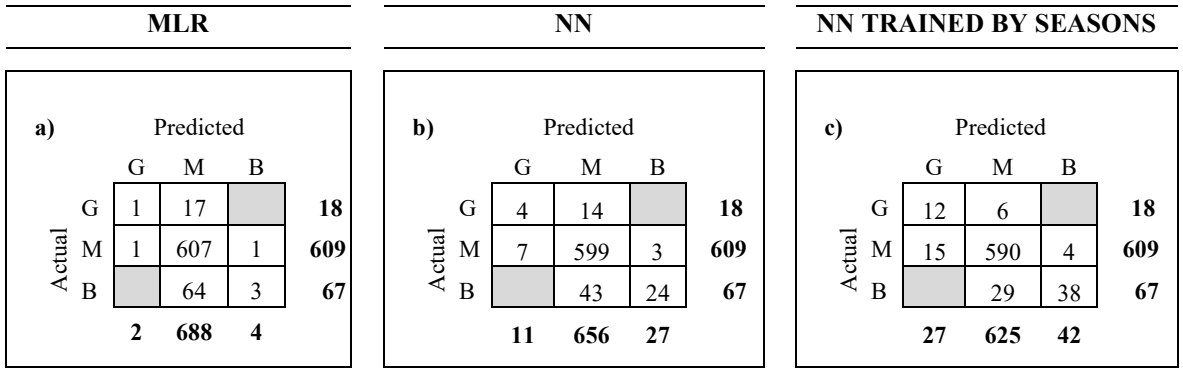


Figure 4.12. Overall classification of MX-AQI categories for $PM_{2.5}$.

4.2.2.1 Evaluation of the overall classification of individual MX-AQI categories for $PM_{2.5}$

The overall statistical performance of the $PM_{2.5}$ MX-AQI classification using the developed models is presented in **Table 4.6**. The use of NNbS yielded the best performance, while MLR produced the worst.

Table 4.6. Overall performance of the $PM_{2.5}$ MX-AQI classification obtained using the models developed in this study.

	MLR	NN	NNbS
Accurate	88.04%	90.35%	92.22%
Overestimation	2.59%	2.45%	1.44%
Underestimation	9.37%	7.20%	6.34%

The evaluation of overall performance by AQI categories using the tested statistical models is presented in the frequency plots shown in **Figure 4.13.** The NNbS is more effective in estimating extreme categories, while similar performance at the "moderate" category is observed using any of the tested statistical models

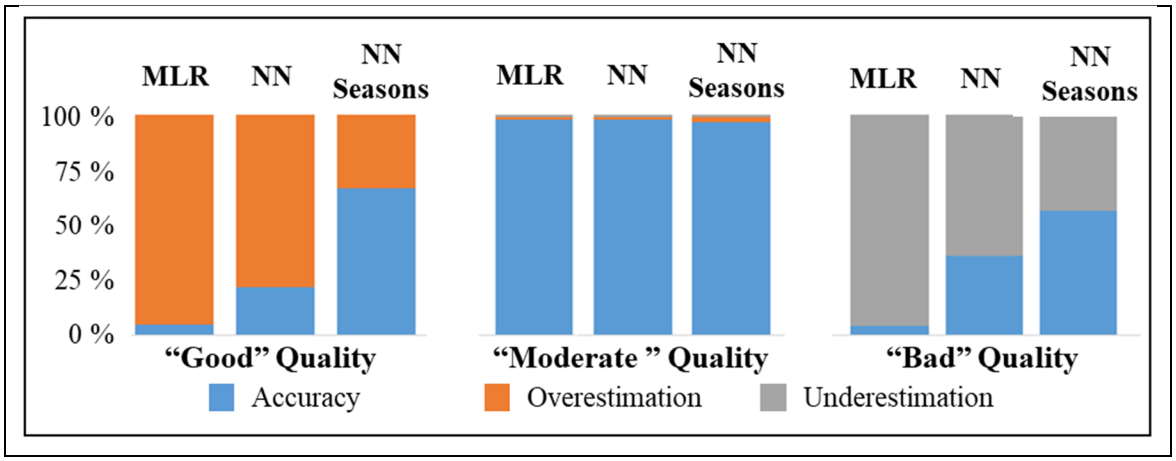


Figure 4.13. Overall performance by MX-AQI categories

4.2.2.2 Evaluation of the classification performance by levels

Evaluating the "Good – Moderate" level of the $PM_{2.5}$ MX-AQI classification performance using the tested statistical models is presented through confusion matrixes in **Figure 4.14.** In this level, the accuracy using MLR is slightly better than the NN models. However, false negative rate is higher using MLR than the NN models (predicted "Moderate", actually "Good"). False positive rate (predicted "Good", actually "Moderate") remains almost stable using any of the tested statistical models. The highest true positive rate (predicted "Good", actually "Good") is yielded using the NNbS, this means that NNbS is more effective in estimating the extreme lower category.

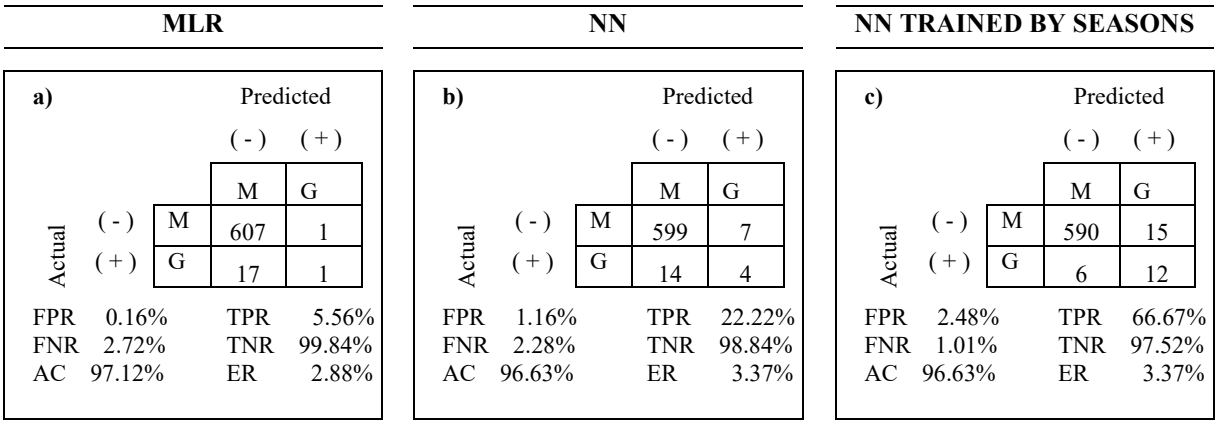


Figure 4.14. Confusion Matrix, Evaluating "Good" – "Moderate" Level
G= Good, M= Moderate

Evaluating Moderate – Bad level of the $PM_{2.5}$ MX-AQI classification performance using the tested statistical models is presented through confusion matrices in **Figure 4.15.** In this level, the accuracy using NNbS is better than MLR and NN. The highest True Negative Rate (predicted "Bad", actually "Bad") is yielded using the NNbS. This means that NNbS is more effective in estimating the "bad air" category. False Positive Rate and False Negative Rate are more common when evaluating "Moderate - Bad" level than "Good-Moderate" level. In other words, the tested models show better performance when estimating lower $PM_{2.5}$ AQ categories. The True Positive Rate (predicted "Moderate", actually "Moderate") shows no difference among the tested models (~99.50%). In the "Moderate - Bad" level, MLR showed the worst performance in most metrics (error rate, false positive rate, and false negative rate). This is important in terms of its public health implications, e.g. the possibility of issuing erroneous alerts to the public regarding the health risk of $PM_{2.5}$ pollution. It is preferable to alert and protect the population from higher $PM_{2.5}$ levels (false negative) than from lower ones.

MLR	NN	NN TRAINED BY SEASONS

a)

Actual \ Predicted		(-)	(+)
		B	M
(-)	B	3	64
(+)	M	1	607

FPR	95.52%	TPR	99.84%
FNR	25.00%	TNR	4.48%
AC	90.37%	ER	9.63%

b)

Actual \ Predicted		(-)	(+)
		B	M
(-)	B	24	43
(+)	M	3	599

FPR	64.18%	TPR	99.50%
FNR	11.11%	TNR	35.82%
AC	93.12%	ER	6.88%

c)

Actual \ Predicted		(-)	(+)
		B	M
(-)	B	38	29
(+)	M	4	590

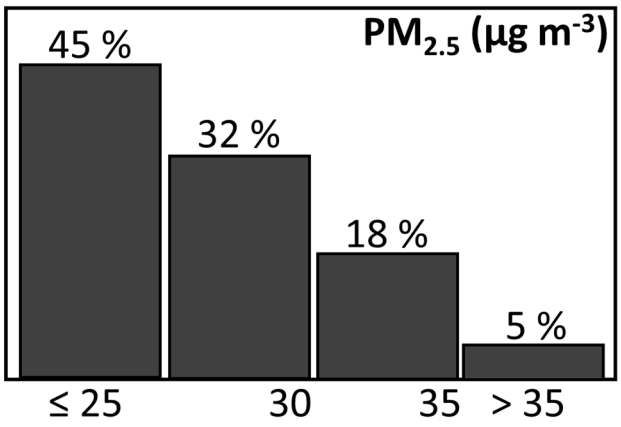
FPR	43.28%	TPR	99.33%
FNR	9.52%	TNR	56.72%
AC	95.01%	ER	4.99%

Figure 4.15. Confusion Matrix, Evaluating "Moderate" – "Bad" Level
M= Moderate, B= Bad

4.3 Developing a regional application: Analysis of spatial and temporal distribution of $PM_{2.5}$ pollution over Northeastern Mexico using MERRA-2 Reanalysis Datasets.

4.3.1 Distribution of data: $PM_{2.5}$ from SIMA and AOD from MERRA-2

Analyzing the distribution of data recorded by SIMA (**Figure 4.16**), it was observed that the monthly $PM_{2.5}$ means concentrations, only 45% of the available data corresponded to concentrations below 25 $\mu g\ m^{-3}$. 38% of the recorded temperatures (monthly means) were in the range of 20 to 25 ^{o} C, while 80% of the relative humidity data (monthly means) corresponded to values between 40% and 60%.



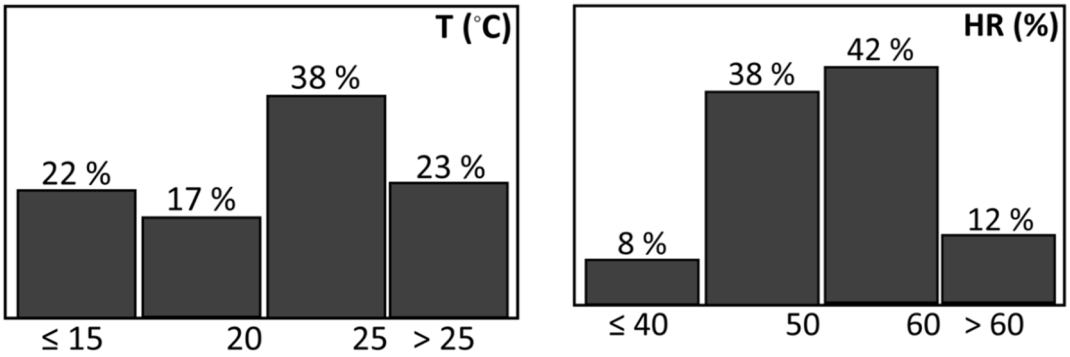
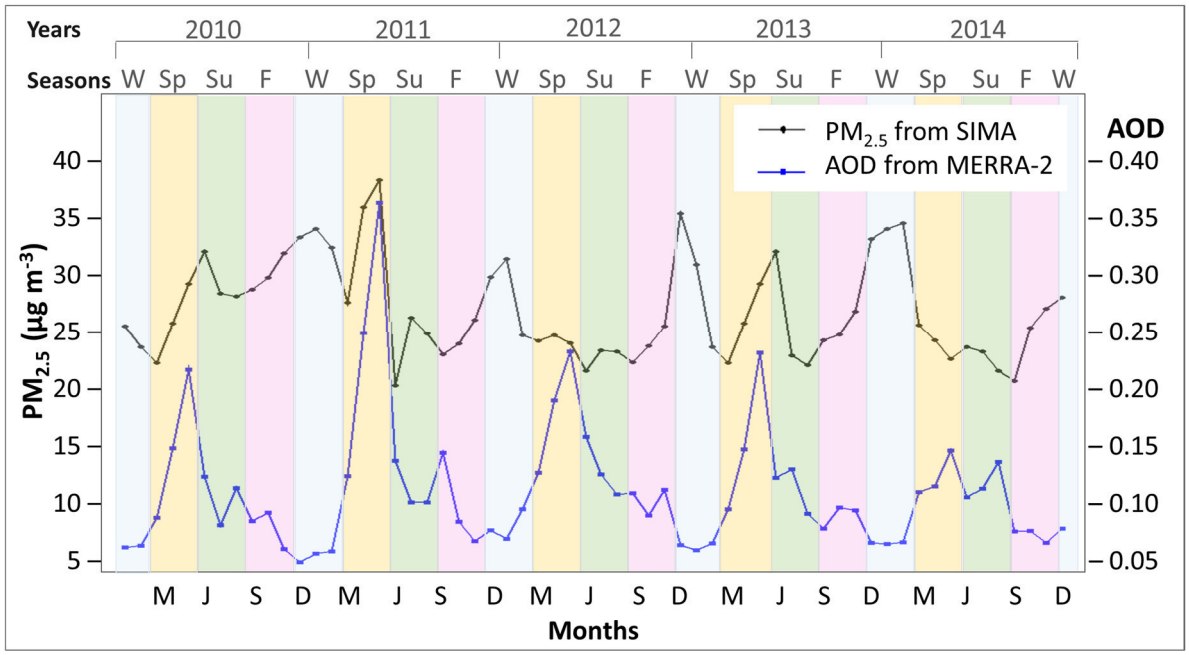


Figure 4.16. Distribution of meteorological variables recorded by SIMA that influenced the AOD-$PM_{2.5}$ relationship in the MMA
(upper panel, monthly $PM_{2.5}$ means concentration. Lower left panel, monthly temperature means. Lower right panel, monthly relative humidity means.

Temporal behavior of AOD obtained from MERRA-2 and monthly $PM_{2.5}$ means recorded by SIMA, for the period covering from January 2010 to December 2014 is shown in **Figure 4.17.** It was observed that - during the study period - the average AOD varies in the range of 0.05 to 0.25, although it registers a maximum value of 0.36 in May 2011. AOD represents amount of aerosol in the vertical column of the atmosphere, is a dimensionless number and values lower than 0.01 correspond to an extremely clean atmosphere, and values greater than 0.4 corresponds to a high pollution condition (NOAA, 2019). $PM_{2.5}$ also registered a maximum value of 38 $\mu g\ m^{-3}$ in May 2011. It should be noted that less absolute difference between $PM_{2.5}$ and AOD is usually observed in spring.

AOD and $PM_{2.5}$ showed seasonal differences: $PM_{2.5}$ values were higher in winter ($\geq 30\ \mu g\ m^{-3}$) while AOD was highest in spring (≥ 0.20). As mentioned in **Section 4.1.3** these differences are due to the complex relationship between $PM_{2.5}$ and AOD, which can be influenced by aerosol types, chemical composition and particle size, vertical concentration profiles, etc. (Guo et al; 2017; Liu et al., 2007; Ma et al., 2014; Song et al., 2014; You et al., 2015).



Seasons: W: Winter, Sp: Spring, Su: Summer, F: Fall. Months: M: March; J: June, S: September, D: December.

Figure 4.17. Comparison of temporal behavior of monthly AOD means obtained from MERRA-2 and monthly $PM_{2.5}$ means recorded by SIMA, over MMA for the period covering from January 2010 to December 2014.

4.3.2 $PM_{2.5}$ components from MERRA-2

MERRA-2 reports five $PM_{2.5}$ components: Dust, Sea Salt (SS), Organic Carbon (OC), Black Carbon (BC) and sulfates (SO_4^{2-}), which represent almost the 70% of the mass contained in $PM_{2.5}$ in the MMA (Martinez et al., 2012; Blanco et al., 2015; Mancilla et al., 2019).

Annual profiles for $PM_{2.5}$ components obtained from MERRA-2 are shown in **Figure 4.18.** It was observed that during almost the entire episode, sulfates and OC were the principal components of $PM_{2.5}$ as reported by Martinez et al. (2012), Blanco et al. (2015) and Mancilla et al. (2019).

SO_4^{2-} is the component with the lowest temporal variability. Monthly SO_4^{2-} means remains in the range between 2 and 3 µg m^{-3} during the period of study, except for summer months when the lowest values (> 2 µg m^{-3}) and variability were observed. During the summer high percentages of sulfate ions were detected by Gonzalez et al. (2018) over the MMA, however, these decreased significantly in winter due to the low conversion of SO_2 to SO_4^{2-} (possibly caused by the low solar radiation and relative humidity).

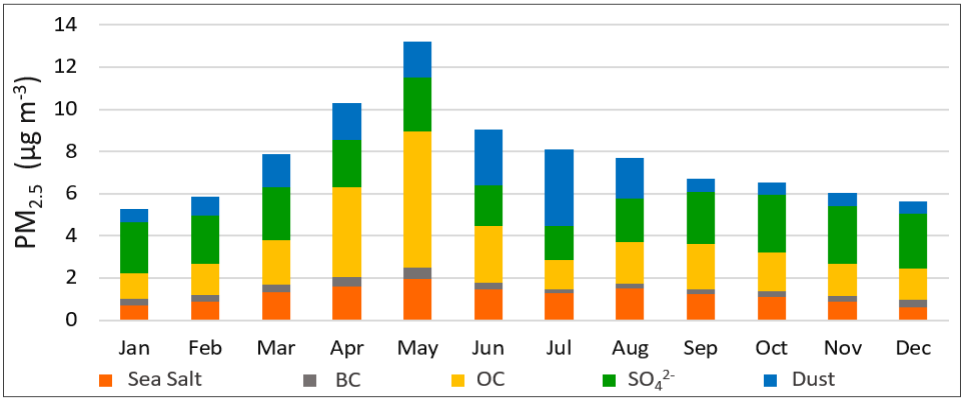
OC is the component with the highest temporal variability. High monthly OC means (> 5 μg m^{-3}) were observed, particularly from April to June. During these months the photochemical activity and transport is predominant over MMA as reported by Mancilla et al. (2015), who carried out a chemical speciation of the OC for spring 2011 and confirmed the aging of primary emissions due to photochemical activity.

BC is the component with the lowest monthly means (0.19 - 0.56 μg m^{-3}). The monthly BC means fluctuated in similar way than monthly OC means. Peaks of high monthly BC means were observed in April over the MMA as reported by INECC (2016).

The highest monthly dust means (1.94 - 3.62 μg m^{-3}) were observed in summer. Dust is a wind dependent component and it has been observed that the highest wind velocities occur in summer over the MMA (Hernández Paniagua et al., 2017). Since the geological material is the principal component (45%) of PMc (PMc = $PM_{10} – PM_{2.5}$), highest PM_{10} concentrations (even exceeding the 24-h standard) are observed in summer over the MMA (Mancilla et al., 2019).

The highest monthly sea salt means were observed in spring (1.96 μg m^{-3}) followed by summer (1.96 μg m^{-3}). As the MMA is not a local source of sea salt, the sea salt concentrations over MMA can be attributed mainly from long-range transport. This means that SS over MMA is also a wind dependent component. As mentioned in **Section 4.1.1** during spring and summer, wind speed reaches maximum values, above 10 km h^{-1}, blowing primarily from the East (E) -Gulf of Mexico region- to West (W).

a)



b)

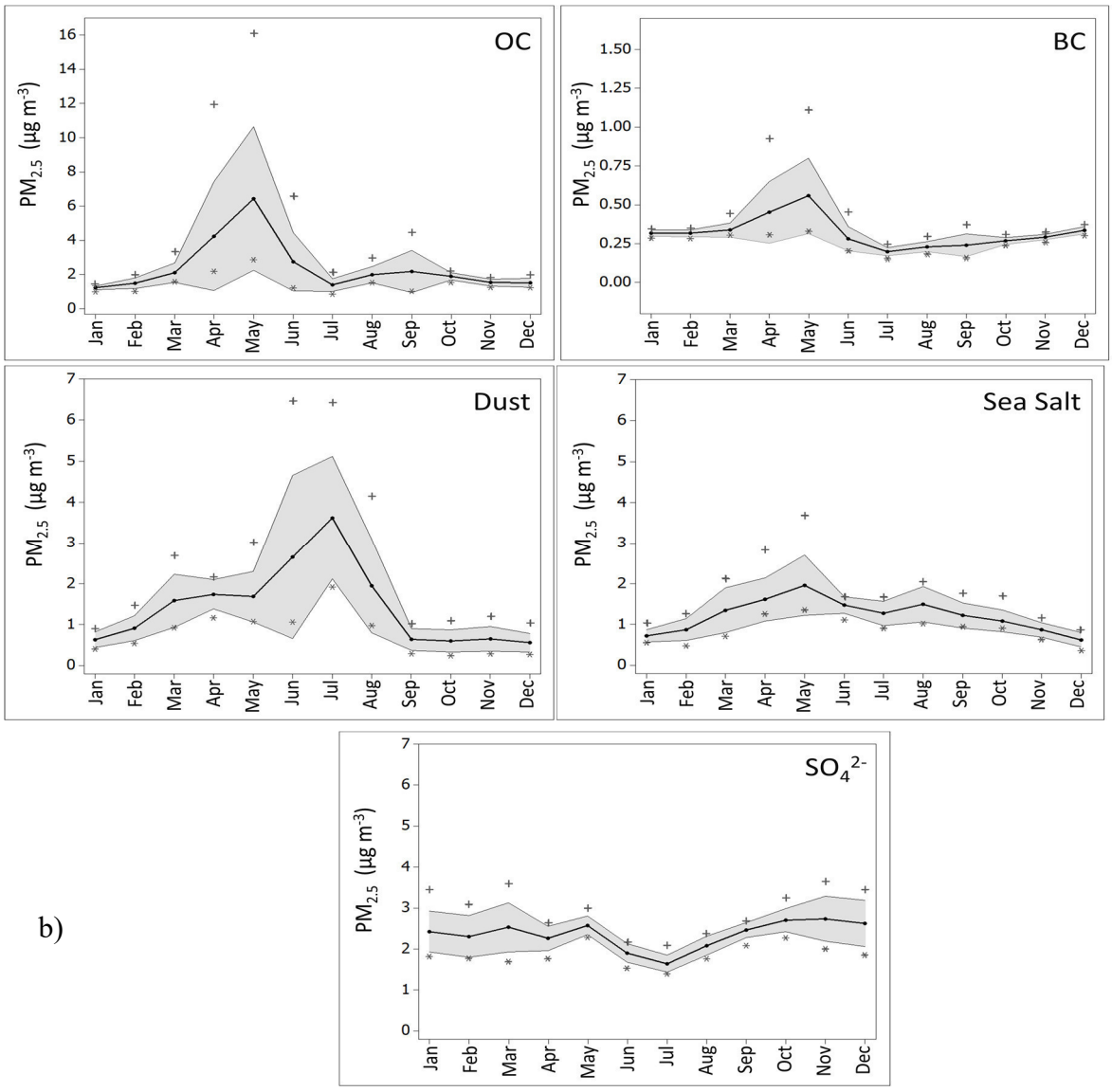


Figure 4.18. Annual profile $PM_{2.5}$ components from MERRA-2, over MMA for the period covering from January 2010 to December 2014.

Figure 4.18 a) Annual profile (accumulated)

Figure 4.18 b) Annual profile (by components)

A comparison of concentrations (μg m^{-3}) from the $PM_{2.5}$ components reported by MERRA-2 with other studies conducted in the MMA (Martinez et al., 2012; Blanco et al., 2015 and Mancilla et al., 2019) is shown in **Table 4.7**. The aerosol data from MERRA-2 were obtained for the same periods of study reported by the local studies.

The compared studies were carried out throughout monitoring campaigns which represents the air quality observed over the study monitoring sites. Components from these studies are reported in this study as daily mean concentrations. Conversely, aerosol data from MERRA-2 represent the air quality observed over a box size of 0.500 × 0.625 deg spatial resolution. Components from MERRA-2 are reported in this study as monthly means concentrations.

Despite the temporal and spatial resolution difference, some similarities were observed. SO_4^{2-} derived by MERRA-2 was consistently ~2 times lower than the SO_4^{2-} reported by the local studies. This suggests that the SO_4^{2-} is well distributed (both spatial and temporal) over the MMA. Additionally, the OC-to-BC ratio for $PM_{2.5}$ derived by MERRA-2 ranged from 3.94 to 6.65, whereas those reported by the local studies ranged from 2.38 to 5.40. This means that MERRA-2 produces similar temporal OC/BC variations as observed in MMA, i.e., seasonal variations of OC/BC are captured by MERRA-2. OC/BC ratios exceeding 2 which might reflect the combined contributions from charcoal combustion, motor-vehicle exhaust, and biomass burning sources, also can be an indicator of SOA formation (Cao et al., 2005). However, they may reflect contributions from other primary sources with a high OC/EC emission ratio (Cabada et al., 2004).

Table 4.7. Comparison of $PM_{2.5}$ concentrations (μg m^{-3}) from components reported by MERRA-2 (M-2) with those reported by other MMA studies.

Period	Nov – Dec 2007			Dec 2014 – Mar 2015			Jun – Jul 2015		
Component	(1)	M-2	Ratio	(2)	M-2	Ratio	(3)	M-2	Ratio
BC	2.74	0.27	10.15	2.07	0.33	6.27	0.35	0.20	1.75
OC	13.6	1.18	11.53	4.92	1.30	3.78	1.89	1.33	1.42
Dust	4.67	0.55	8.49	7.67	0.52	14.75	0.53	6.43	0.08
SO_4^{2-}	**6.4**	**2.92**	**2.19**	**6.79**	**3.23**	**2.10**	**3.39**	**1.47**	**2.31**
Sea Salt	1.4	0.94	1.49	0.59	0.60	0.98	0.09	1.67	0.05
OC/BC	**4.96**	**4.37**	**1.13**	**2.38**	**3.94**	**0.60**	**5.40**	**6.65**	**0.80**

(1) Martínez et al. (2012), (2) Blanco et al. (2015), (3) Mancilla et al. (2019).

A comparison of concentrations (µg m^{-3}) from BC reported by MERRA-2 with a local study conducted in the MMA during 2015 is shown in **Figure 4.19**. The local study was carried out throughout monitoring campaigns which represents the air quality observed over the study monitoring sites. Both BC from the local study and BC derived by MERRA-2 are reported in this study as monthly means concentrations.

Despite the spatial resolution difference, from January to April and from August to November BC derived by MERRA-2 was approximated 7.5 times lower than the BC reported by INECC (2016). Some discrepancies are observed in summer months, when the study conducted by INECC (2016) reported the highest BC concentrations (> 3 µg m^{-3}) and MERRA-2 reported the lowest BC concentrations (~0.2 µg m^{-3}). In these months BC derived by MERRA-2 was approximated 13 times lower than the BC reported by INECC (2016).

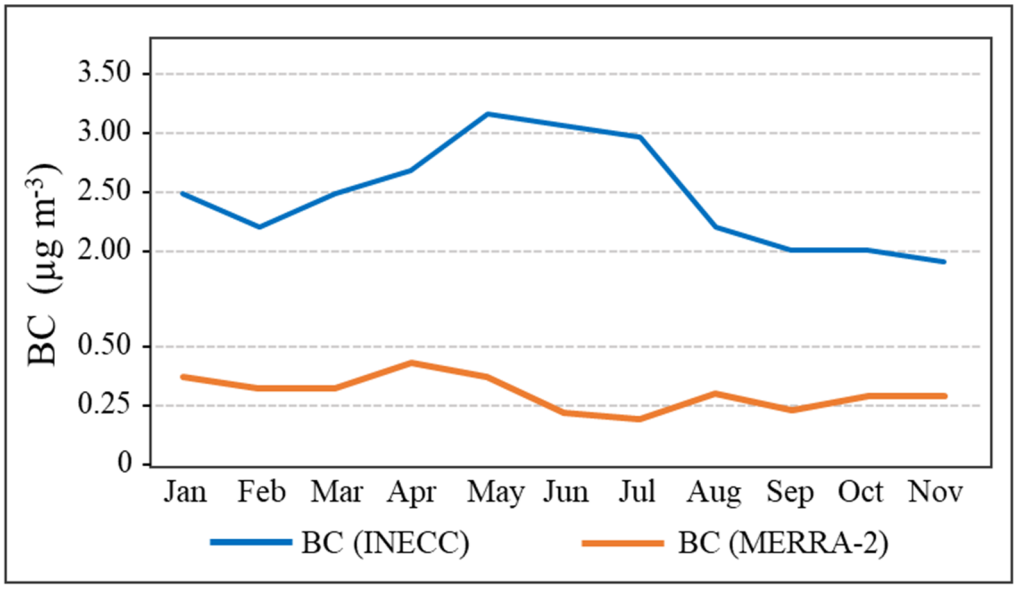


Figure 4.19. Comparison of monthly BC mean concentration estimated by INECC (2016) with monthly BC mean concentration estimated by MERRA-2 during 2015.

4.3.3 Model performance

The model incorporated in the neural network for estimating the monthly $PM_{2.5}$ means using the MERRA-2 data was:

$$PM_{2.5} = f(AOD, T, RH, Dust_{2.5}, SS_{2.5}, BC, OC, {SO_4}^{2-})$$

Wind components from MERRA-2 were not able to represent the predominant wind pattern (east-to-west direction) and the wind speeds observed in the MMA for the period 2010-2014. Thus, wind components from MERRA-2 were not incorporated to the NN model.

The model performance over MMA using the NN is presented through scatterplots in **Figure 4.20.** The NN estimated a monthly $PM_{2.5}$ mean of 29.37 µg m^{-3}, with the following metrics: $R \sim 0.88$, MPE = -2.70%, RMSE = 3.41 µg m^{-3} and MAE = 1.44 µg m^{-3}, which are comparable or even better than those obtained in other studies carried out in other regions of the world (e.g. Li et al. 2017; Shen et al., 2018; He et al, 2019).

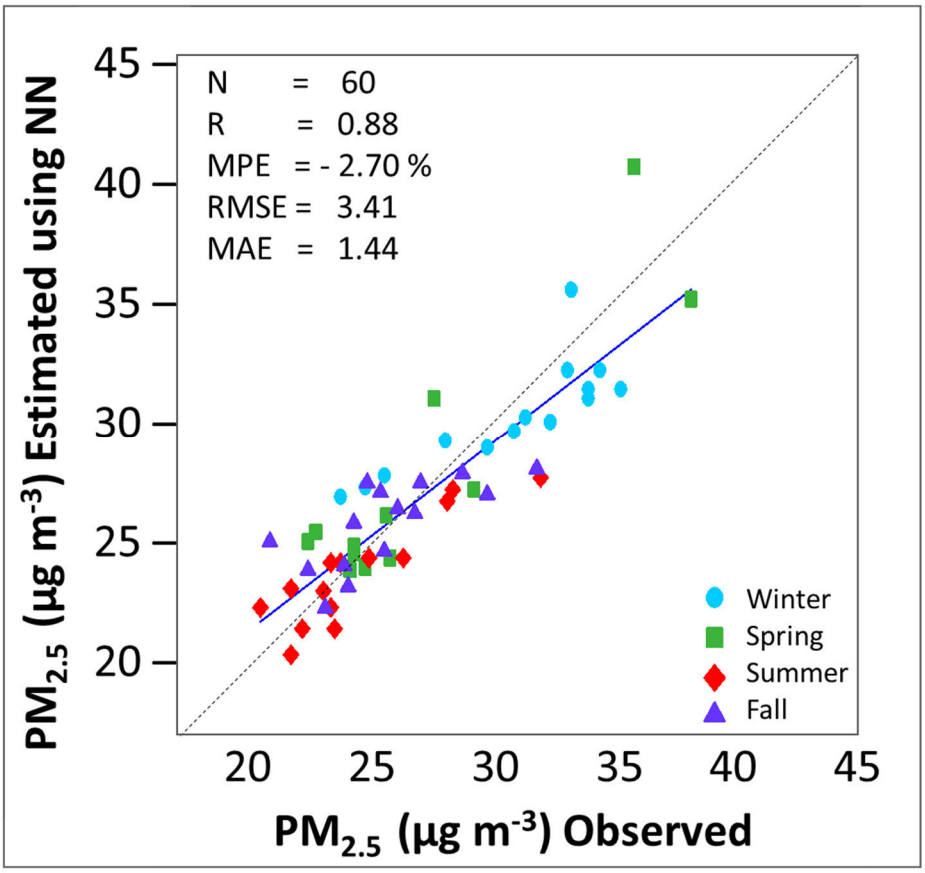


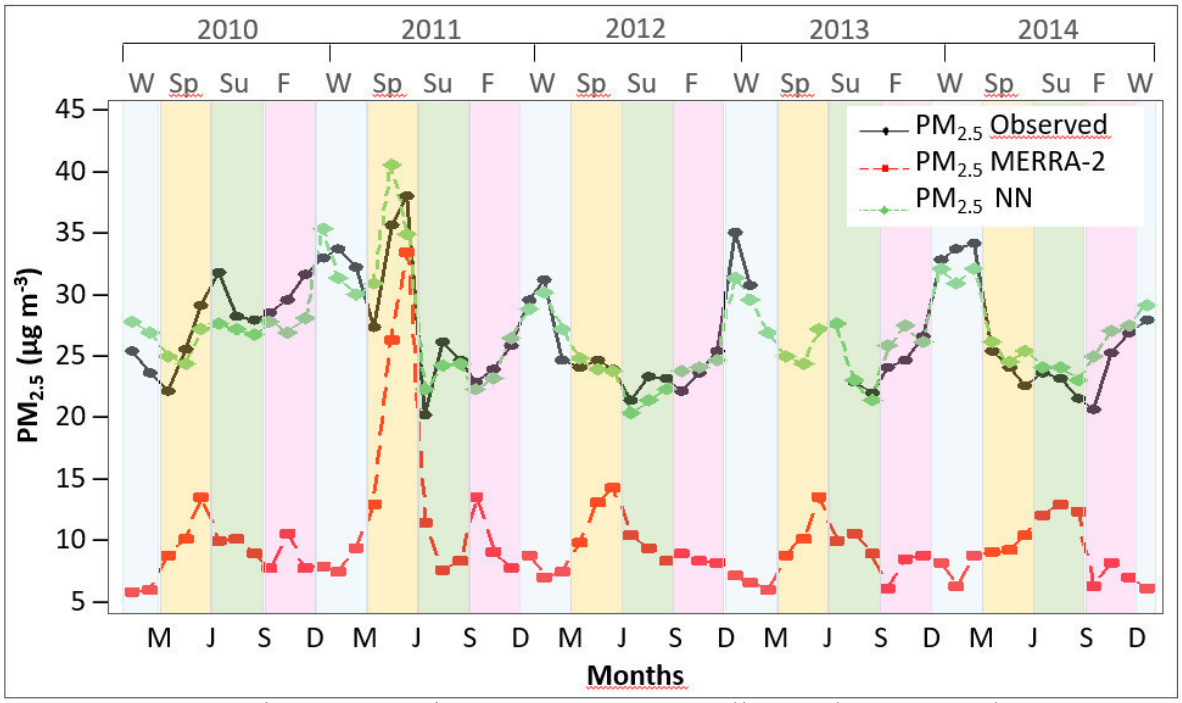
Figure 4.20. Statistical performance obtained using NN to estimate monthly $PM_{2.5}$ means, over the MMA for the period covering from January 2010 to December 2014.

The results showed wide temporal variability. In the MMA the concentrations of $PM_{2.5}$ present seasonal variations, where most of the highest concentrations are observed in winter, peaks of high concentrations occur in spring while those of low concentration are observed in summer and early autumn (Blanco et al., 2015; Mancilla et al., 2019).

A comparison of monthly $PM_{2.5}$ means (µg m-3) estimated by the NN developed in this study with the monthly $PM_{2.5}$ means from MERRA-2 derived data and monthly $PM_{2.5}$ means recorded by SIMA is shown in **Figure 4.21**. $PM_{2.5}$ from MERRA-2 was calculated as suggested by Buchard et al. (2015):

$$PM_{2.5} = [Dust_{2.5}] + [SS_{2.5}] + [BC] + 1.4 \times [OC] + 1.375 \times [SO_4] \quad \textbf{Equation 4.1}$$

The MERRA-2 estimations using **Equation 4.1**, which include only $PM_{2.5}$ components, underestimate the $PM_{2.5}$ observed by those recorded by SIMA. However, MERRA-2 estimations were able to represent the peak of high concentrations observed in spring 2011. On the other hand, $PM_{2.5}$ estimated by the NN developed in this study, which include AOD, meteorological variables and $PM_{2.5}$ components, follow the same pattern and have the same temporal behavior than those observed by SIMA. These findings reinforce the usefulness of the methodology developed here.



Seasons: W: Winter, Sp: Spring, Su: Summer, F: Fall. Months: M: March; J: June, S: September, D: December.

Figure 4.21. Comparison among the monthly $PM_{2.5}$ means (µg m-3) estimated by the NN developed in this study, MERRA-2 derived data and SIMA data.

4.3.4 Regional $PM_{2.5}$ distribution

The regional distribution of $PM_{2.5}$ was estimated using only MERRA-2 data as the input to the NN. Monthly $PM_{2.5}$ means maps were generated using QGIS 3.2, and the same grid mesh to which MERRA-2 data was reported. Specifically, this study made use of data contained within the lower and upper bounds of (23 ° Lat, -101.25 ° Lon) and (28 ° Lat, -98.125 ° Lon), respectively.

Monthly $PM_{2.5}$ means are presented in **Figure 4.22.** It is observed that, in general, $PM_{2.5}$ levels are higher in winter and spring. Significant spatial clustering of high $PM_{2.5}$ was also observed, with the highest $PM_{2.5}$ levels located in the MMA, which is the main source of air pollution throughout this entire area. Estimated data indicate that $PM_{2.5}$ are not evenly distributed throughout the region, varying spatially as well as temporally, since the mixed mountainous-and-valley topography complicates the dispersion of local emissions.

The regional distribution of $PM_{2.5}$ also presents seasonal variations: decreasing trends are observed from spring to summer, then increase from fall to winter, concluding that the air quality in the area is worse in winter and slightly better in summer. Specifically, the month of May usually has the highest concentrations (particularly the month of May 2011 and the month of May 2013) followed by the winter months. In winter, in addition to the increase in concentrations associated with the increase in emissions from burning fossil fuels, for example, for heating (Wang et al., 2017), transportation, or cooking of food; meteorological conditions, especially those related to high pressure systems, thermal inversions and low precipitation, also play an important role in making it difficult to disperse and remove contaminants.

The results obtained here may allow the identification of needs and locations of new monitoring sites. In addition, they may allow the identification of regions to carry out campaigns for training citizen science groups and deploying low-cost air quality monitor sensors (Gupta et al., 2018).

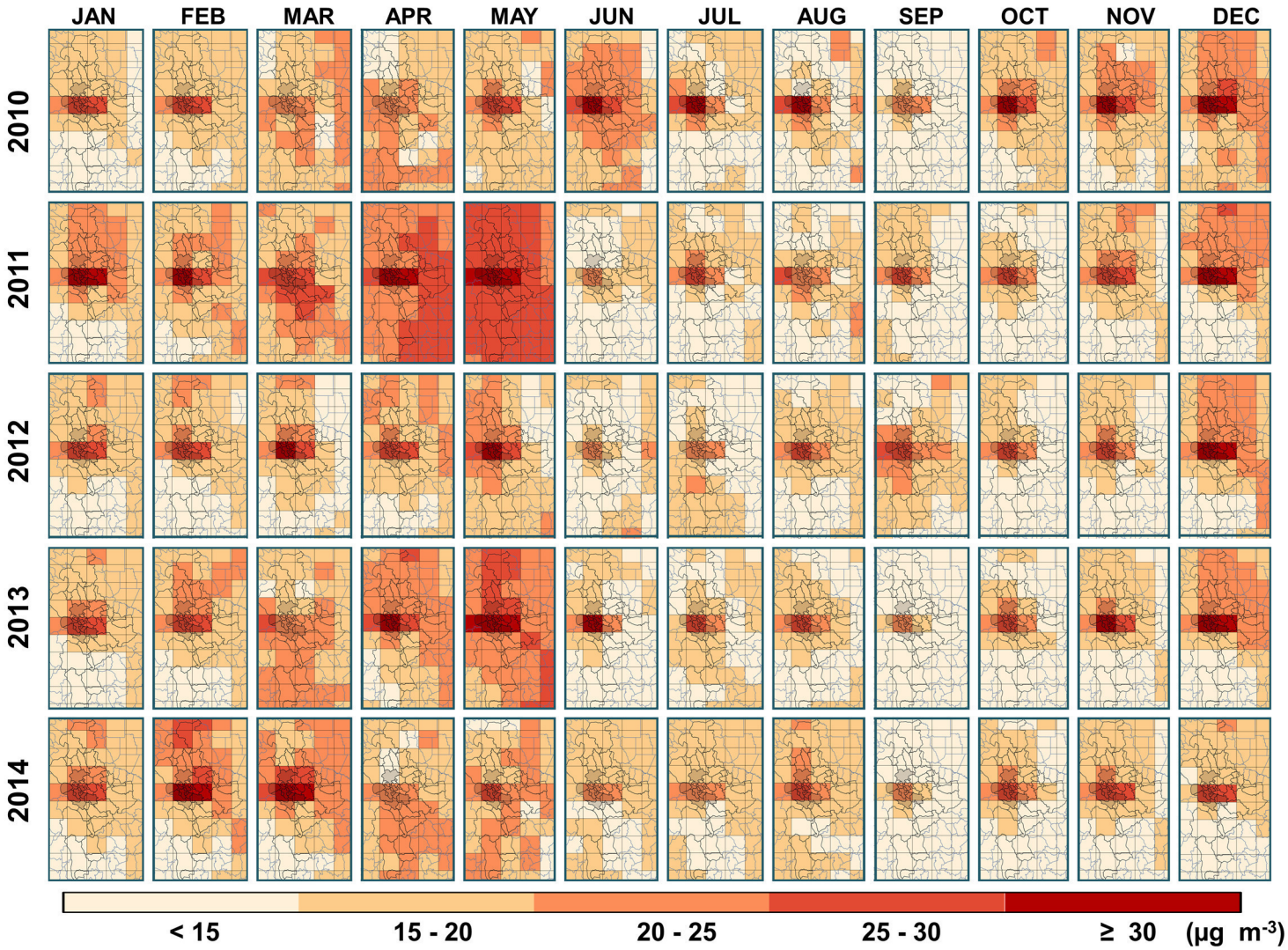


Figure 4.22. Regional distribution of monthly $PM_{2.5}$ means in NEM

Chapter 5 CONCLUSIONS, SUMMARY AND FUTURE WORK

5.1 Conclusions

The effect of different spatio-temporal average schemes on the AOD - $PM_{2.5}$ relationship was evaluated by comparing four MODIS-AOD Scientific DataSets (SDS) for studying the AOD - $PM_{2.5}$ relationships at the Monterrey Metropolitan Area, in Northeastern Mexico. To our best knowledge, this is study is one of the first researches of this type (evaluation of spatio-temporal average schemes and comparison of AOD data extracted from four SDS) that is conducted for a Latin-American urban area. Temperature, relative humidity, wind speed and wind direction were identified as the meteorological variables that influence the AOD - $PM_{2.5}$ relationship in the MMA. Additionally, AOD_{3K}, over domain 3 and using the daily scheme were identified as the AOD-SDS, spatial domain and temporal scheme, respectively, that best describe the $PM_{2.5}$ in MMA using MODIS data. These findings served as the base line to design air quality applications.

Robust models for predicting $PM_{2.5}$ using AOD_{3K} and meteorological data were developed and validated. MLR showed the following metrics: correlation coefficient R = 0.57; MPE = -5.77%; RMSE = 7.00 μg m^{-3}, and MAE = 5.29 μg m^{-3}. NN showed better statistical performance than MLR: correlation coefficient of R = 0.73; MPE = -2.59%; RMSE = 5.83 μg m^{-3} and MAE = 4.24 μg m^{-3}. Neural network training by season (NNbS) showed the best statistical performance among the tested statistical models: correlation coefficient R = 0.83; MPE = -2.66%; RMSE = 4.89 μg m^{-3} and MAE = 3.49 μg m^{-3}. The developed models underpredicted the lowest $PM_{2.5}$ concentrations, while overpredicted the highest $PM_{2.5}$ concentrations, this was most evident using MLR and less evident using NNbS.

An alternative methodology to categorize daily $PM_{2.5}$ air quality indexes based on satellite-derived data and meteorological ground-based measurements were proposed. The validated models (MLR, NN, and NNbS) were able to accurately predict the AQI categories for $PM_{2.5}$ in a local application over MMA. Overall classification accuracy resulted to be 88.04%; 90.35%; and 92.22% for MLR, NN and NNbs, respectively. The methodology proposed here, which takes advantage of both the wide spatial coverage provided by satellite data and the AOD – $PM_{2.5}$ relationship, suggests that models trained using satellite data could be used instead of (or complement) ground-based measurements for calculating AQI, especially in regions where none or without extensive ground-based $PM_{2.5}$ monitoring networks but

meteorological ground-based measurements (or modeled) are available. This methodology can serve as a complementary technique to identify "hot-spot" regions where people is exposed to $PM_{2.5}$ pollution. The developed daily $PM_{2.5}$ air quality index could be used as an input for exposure models and can yield a great impact in studying health effects associations of $PM_{2.5}$ exposure. Mapping historical $PM_{2.5}$ AQI could facilitate studies of long-term $PM_{2.5}$ exposure and to identify changes over time.

A method that uses NN were developed for predicting local monthly $PM_{2.5}$ using MERRA-2 datasets. Then, the developed NN was used in conjunction with MERRA-2 data to analyze spatial and temporal distribution of $PM_{2.5}$ pollution over Northeastern Mexico. This study is one of the first researches using reanalysis data from MERRA-2 for air quality purposes in Latin America, specifically for estimating $PM_{2.5}$ concentrations. It was found that the highest levels of $PM_{2.5}$ were located over the MMA, where the NN estimated monthly $PM_{2.5}$ mean of 29.37 $\mu g\ m^{-3}$. High concentration levels were observed in winter and spring, moderate in fall and low concentration in summer. These results are consistent with the trends observed in previous studies carried out in the region, specifically for the MMA. The global coverage from the reanalysis datasets used in this study along with the developed methodology here could be easily replicated, and of great contribution to asses local and regional air quality applications specially in regions where none or few $PM_{2.5}$ ground-based observations are placed. These contributions are expected to help improve current control strategies aimed to reduce $PM_{2.5}$ pollution in NEM, as well as to provide a better understanding of the regional $PM_{2.5}$ distribution.

The methodologies and robust models developed here could be used to identify new monitoring sites needs and can help to face information gaps. These findings can also help epidemiologists to better understand the exposure-response relationships associated to $PM_{2.5}$ and serve as a stepping stone towards designing effective environmental regulation.

It was concluded that the use of satellite data, along with meteorological data has tremendous potential for assessing local and regional distributions of $PM_{2.5}$ in NEM.

5.2 Summary

This study demonstrates the feasibility of using satellite data in conjunction with surface data and the use of neural networks for the study of $PM_{2.5}$ pollution in the northeast region of Mexico (NEM). This study constitutes the first application of satellite data and local data to estimate $PM_{2.5}$ in the region. Also, it constitutes one of the first applications of the MERRA-2 re-analysis model for estimating $PM_{2.5}$ in Latin America. The use of satellite data along with meteorological data showed tremendous potential for assessing local and regional distributions of $PM_{2.5}$ in NEM.

Forward and backward stepwise procedures were applied to develop a model that relating AOD and ground-based data to estimate $PM_{2.5}$ with the highest level of accuracy. Only those meteorological variables that were statistically significant to the AOD-$PM_{2.5}$ relationship were retained. The best subset for predicting $PM_{2.5}$ included AOD, T, RH, WS and WD.

Twenty-four datasets (four AOD–SDSs, three box sizes around the ground sites and two temporal schemes) were tested using MLR to select the *best dataset* to estimate local $PM_{2.5}$. The refined dataset "$PM_{2.5} = f(AOD_{3K}, T, RH, WS, WD)$, over domain 3 and using the daily scheme" was selected as the "best dataset". Independent model validation was performed by developing an ensemble model. It was concluded that the use of MLR with the *best dataset* is robust and that model performance is not sensitive to the dataset.

The statistical performance from the *best dataset* using MLR showed the following metrics: correlation coefficient R = 0.57; MPE = -5.77%; RMSE = 7.00 μg m^{-3}, and MAE = 5.29 μg m^{-3}. The model underpredicted the lowest $PM_{2.5}$ concentrations, while overpredicted the highest $PM_{2.5}$ concentrations.

Also, a neural network (NN) was tested using the same variables as in MLR (*best dataset*). NN showed better statistical performance than MLR: correlation coefficient of R = 0.73; MPE = -2.59%; RMSE = 5.83 μg m^{-3} and MAE = 4.24 μg m^{-3}. The under-estimation and over-estimation observed in NN was less evident than in MLR.

In order to identify seasonal patterns, the *best dataset* was divided in four sub-datasets (one for each season). Neural network training by season (NNbS) showed the best statistical performance among the tested statistical models: correlation coefficient R = 0.83;

MPE = -2.66%; RMSE = 4.89 μg m^{-3} and MAE = 3.49 μg m^{-3}. The under-estimation and over-estimation observed in NNbS was less evident than in MLR and NN.

The developed models (MLR, NN, and NNbS) were used to estimate daily $PM_{2.5}$ means and determine local air quality categories according to the Mexican $PM_{2.5}$ AQI. Overall accuracy resulted to be 88.04%; 90.35%; and 92.22% for MLR, NN and NNbs, respectively.

A regional application was also carried out. A method that uses NN was developed to study the regional $PM_{2.5}$ distribution. Five years of $PM_{2.5}$ data recorded at local monitoring sites were combined with MERRA-2 reanalysis data –including Aerosol Optical Depth (AOD), meteorological fields and $PM_{2.5}$ components-. It was found that the highest levels of $PM_{2.5}$ were located over the MMA, where the NN estimated monthly $PM_{2.5}$ mean of 29.37 μg m^{-3}. High concentration levels were observed in winter and spring, moderate in fall and low concentration in summer. These results are consistent with the trends observed in previous studies carried out in the region, specifically for the MMA.

The magnitude of $PM_{2.5}$ contamination varied in space and time, showing great seasonality and heterogeneous distribution throughout the region, which reinforces the usefulness of the methodology developed here, by generating information on places where no surface measurements are available. The results obtained here may allow the identification of needs and locations of new monitoring sites and campaigns, for example.

5.3 Future Work

This research demonstrated the AOD – $PM_{2.5}$ relationship in the MMA. However, spatial and temporal limitations were faced to stablish this relationship. Since new algorithms (e.g. MAIAC, 1 km resolution), collections (e.g. MODIS Collection 6.1) and technology (e.g. Geostationary Satellites as GOES -) are now available; the inclusion of these new approaches may enhance the useful of satellite data for estimating $PM_{2.5}$ in the region.

New missions focused in air pollution, e.g. MAIA and Tempo, are going to be released soon. It is also expected that using these data in the region could serve to improve our understanding of air pollution process and could break the temporal and spatial barriers observed using data from previous missions.

The models developed here were able to represent the AOD-$PM_{2.5}$ relationship. It would be interesting to apply optimization techniques to develop the neural networks or exploring new machine learning techniques for improving the $PM_{2.5}$ estimations

Currently, there is a growing interest in using ground-based low-cost sensor for estimating $PM_{2.5}$ in the region. A suggested future step is using this data to establish new relationships with satellite data, or improve and validate the developed models.

We used AOD satellite derived products for estimating $PM_{2.5}$. It is suggested to use other satellite products for tracking extreme events (i.e., wildfires or dust storms). For example, in this research a high episode of $PM_{2.5}$ pollution was observed in the entire region in the spring 2011, it would be interesting to analyze the source and periodicity of this kind of events.

The validated models were able to accurately predict the AQI categories for $PM_{2.5}$ in a local application over MMA. It is suggested to run the models in a regional application, for example, for mapping the historical $PM_{2.5}$ AQI in the entire region and analyzing the exposure-response relationships associated to $PM_{2.5}$.

It is suggested that, at another stage of the research, new monitoring campaigns over the MMA and other places in the NEM be carry out. This in order to make validations and comparisons with satellite data (i.e., MODIS) as well meteorological data and aerosol components reported by MERRA-2. It is suggested to segregate the regional database by seasons in order to improve the analysis of seasonal behavior at regional scale.

REFERENCES

Abiodun O. I., Jantan A., Omolara A. E., Dada K. D., Mohamed N. A., Arshad H. (2018). Review: State-of-the-art in artificial neural network applications: A survey. J. Heliyon. https://doi.org/10.1016/j.heliyon.2018.e00938

Agrawal A., Sharma A. R., Tayal S. (2014). Environmental monitoring and assessment. Assessment of regional climatic changes in the Eastern Himalayan region: a study using multi-satellite remote sensing data sets. Volume 186 Isuue 10: 6521-6536 DOI: 10.1007/s10661-014-3871-x

Aguilar-Barajas I., Sisto N., Ramírez-Orozco. A.I. (2015). Agua para Monterrey Logros, retos y oportunidades para Nuevo León y México. México.

Alvesa C., Vicentea A., Pioa C., Kissb G., Hofferb A., Decesaric S., Prevôtd S.H., Minguillóne M.C., Querole X., Hillamof R., Spindlerg G., Swietlickih E. (2012). Organic compounds in aerosols from selected European sites – Biogenic versus anthropogenic sources. Atmospheric Environment. Volume 59, Pages 243–255. DOI: 10.1016/j.atmosenv.2012.06.013

ATBD. (2013). Algorithm Theoretical Basis Document: The Collection 6 (C6) version of the "Dark-Target" (DT) algorithms for retrieving aerosol properties over global land and ocean, as applied to spectral reflectance observed by Moderate resolution Imaging Spectroradiometer (MODIS) aboard NASA's Terra and Aqua satellites.

Atkinson R.W., Mills I.C., Walton H.A., Anderson H.R. (2015). Fine particle components and health —a systematic review and meta-analysis of epidemiological time series studies of daily mortality and hospital admissions. Journal of Exposure Science and Environmental Epidemiology (2015) vol 25, pp. 208-214. DOI: 10.1038/jes.2014.63

Balmes J.R., Cisternas M., Quinlan P.J., Trupin L., Lurmann F.W., Katza P.P., Blanca P.D. (2014). Annual average ambient particulate matter exposure estimates, measured home particulate matter, and hair nicotine are associated with respiratory outcomes in adults with asthma. Environmental Research. Volume 129, February 2014, Pages 1–10. DOI: 10.1016/j.envres.2013.12.007

Barry R. G., Hall-McKim E. A. (2014). Essentials of the Earth's Climate System. 1st Edición. Cambridge University Press. ISBN: 978-1-107-62049-0

Bengio, Y. (1996). Neural Networks for Speech and Sequence Recognition. London: Van Nostrand Reinhold.

Blanco-Jiménez S., Altúzar F., Jiménez B., Aguilar G., Pablo M., and Benítez M. A. (2015). Evaluación de partículas suspendidas $PM_{2.5}$ en el Área Metropolitana de Monterrey, 34. México: Instituto Nacional de Ecología y Cambio Climático (INECC).

Buchard V., da Silva A.M., Colarco P.R., Darmenov A., Randles C.A., Govindaraju R., Torres O., Campbell J., Spurr R. (2015). Using the OMI aerosol index and absorption aerosol optical depth to evaluate the NASA MERRA Aerosol Reanalysis. At-mos. Chem. Phys. 15 (10), 5743–5760. http://www.atmos-chem-phys.net/ 15/5743/2015/.

Cabada J. C., Spyros N. P., Subramanian R., Robinson A. L., Polidori A., and Turpin B. (2004). Estimating the secondary organic aerosol contribution to $PM_{2.5}$ using the EC tracer method. Aerosol. Sci. Technol. 38:140–55. doi:10.1080/ 02786820390229084

Cabaneros S. M., Calautit J. K., and Hughes B. R. (2019). A review of artificial neural network models for ambient air pollution prediction. J. Environmental Modelling and Software 119 (2019) 285–304. https://doi.org/10.1016/j.envsoft.2019.06.014

Cao J. J., Wu F., Chow J. C., Lee S. C., Li Y., Chen S. W., An Z.S., Fung K., Watson J. G., Zhu C. S., and Liu S. X. (2005). Characterization and source apportionment of atmospheric organic and elemental carbon during fall and winter of 2003 in Xi'an, China, Atmos. Chem. Phys., 5, 3127–3137. http://www.atmos-chem-phys.net/5/3127/2005

Carmona J. M., Mendoza A., Lozano D. F. (2014). Particulate Matter levels estimations from MODIS data in the Northeastern Area of Mexico. Paper presented at the XVI International Symposium SELPER 2014. Medellin, Colombia.

Carrillo E, Hernández I. Y, and Mendoza A. (2017). Use of Combined Observational- and Model-Derived Photochemical Indicators to Assess the O3-NOx-VOC System Sensitivity in Urban Areas. Atmosphere 2017, 8, 22. doi:10.3390/atmos8020022.

Chang P., Yang J., Den W., Wu C. (2014). Characterizing and locating air pollution sources in a complex industrial district using optical remote sensing technology and multivariate statistical modeling. Environmental science and pollution research international. Vol 21 Issue: 18:10852-66. DOI:10.1007/s11356-014-2962-0

Chen C., Zhu P., Lan L., Zhou L., Liu R., Sun Q., Ban J., Wang W., Xu D., Li T. (2018). Short-term exposures to $PM_{2.5}$ and cause-specific mortality of cardiovascular health in China. Environmental Research (2018) vol 161, pp. 188–194. doi.org/10.1016/j.envres.2017.10.046

Chin M., Ginoux P., Kinne S., Torres o., Holben B. N., Duncan B. N, Martin R. V., Logan J. A., Higurashi A., Nakajima T. (2002). Tropospheric aerosol optical thickness from the GOCART model and comparisons with satellite and sun photometer measurements. J. Atmos. Sci., 59, 461–483, doi:10.1175/1520-0469(2002)059,0461:TAOTFT.2.0.CO;2.

Chu Y., Liu Y., Li X., Liu Z., Lu H., Lu Y., Mao Z., Chen X., Li N., Ren M., Liu F., Tian L., Zhu Z., and Xiang H. (2016). A Review on Predicting Ground $PM_{2.5}$ Concentration Using Satellite Aerosol Optical Depth. Atmosphere 2016, 7, 129; doi:10.3390/atmos7100129

Colarco P., da Silva A., Chin M., and Diehl T. (2010). Online simulations of global aerosol distributions in theNASAGEOS-4 model and comparisons to satellite and ground-based aerosol optical depth. J. Geophys. Res., 115, D14207, doi:10.1029/2009JD012820.

Collaud C.M., Andrews E., Asmi A., Baltensperger U., Bukowiecki N., Day D., Fiebig M., et al. (2013). Aerosol Decadal Trends – Part 1: In-Situ Optical Measurements at GAW and IMPROVE Stations. Atmospheric Chemistry and Physics. https://doi.org/10.5194/acp-13-869-2013.

CONAPO, SEDATU e INEGI. (2018). Delimitación de las zonas metropolitanas de México 2015.

Conway E.D. (1997). An Introduction to Satellite Image Interpretation. Mariland Space Grant Consortium. The Johns Hopkins University Press. ISBN 0-8018-5576-4

Crumeyrolle S., Chen G., Ziemba L., Beyersdorf A., Thornhill L., Winstead E., Moore R.H., Shook M.A., Hudgins C., Anderson B.E. (2014). Factors that influence surface $PM_{2.5}$ values inferred from satellite observations: perspective gained for the US Baltimore - Washington metropolitan area during DISCOVER-AQ. Atmos. Chem. Phys. 14, 2139-2153. http://dx.doi.org/10.5194/acp-14-2139-2014.

Darmenov A., and da Silva A. (2015). The Quick Fire Emissions Dataset (QFED): Documentation of versions 2.1, 2.2 and 2.4. Technical Report Series on Global Modeling and Data Assimilation, Vol. 32, NASA Tech. Rep. NASA/TM–2015–104606, 201 pp. [Available online at https://gmao.gsfc.nasa.gov/pubs/docs/Darmenov796.pdf.]

Dave J. V. (1972). Development of Programs for Computing Characteristics of Ultraviolet Radiation, Tech. Rep., Vector Case, IBM Corp, Fed. Syst. Div, Gaithersburg, MD.

Derber J. C., Purser R. J., Wu W.-S., Treadon R., Pondeca M., Parrish D., and Kleist D. (2003). Flow dependent Jb in a global grid-point 3D-Var. Proc. ECMWF annual seminar on recent developments in data assimilation for atmosphere and ocean. Reading, UK, 8-12 Sept. 2003.

Diehl T., Heil A., Chin M., Pan X., Streets D., Schultz M., and Kinne S. (2012). Anthropogenic, biomass burning, and volcanic emissions of black carbon, organic carbon, and SO_2 from 1980 to 2010 for hindcast model experiments. Atmos. Chem. Phys. Discuss., 12, 24 895–24 954, doi:10.5194/acpd-12-24895-2012.

Du Y., Xu X., Chu M., Guo Y., Wang J. (2016). Review: Air particulate matter and cardiovascular disease: the epidemiological, biomedical and clinical evidence. Journal of Thoracic Disease 8 (1): E8-E19. doi: 10.3978/j.issn.2072-1439.2015.11.37

Dubovik O., and King M. D. (2000). A flexible inversion algorithm for retrieval of aerosol optical properties from Sun and sky radiance measurements, J. Geophys. Res., 105, 20673–20696, 2000.

Dubovik O., Smirnov A., Holbe B. N., King M.D., Kaufman Y. J., Eck T.F., Slutsker I. (2000). Accuracy assessments of aerosol optical properties retrieved from AERONET sun and sky radiance measurements. J. Geophys. Res., 105, 9791– 9806.

Dubovik O., Holben B. N., Eck T. F., Smirnov A., Kaufman Y. J., King M. D, Tanré D., and Slutsker I. (2002a). Variability of absorption and optical properties of key aerosol types observed in worldwide locations J. Atmos. Sci.,59,590–608

Dubovik O., Holben B. N., Lapyonok T., Sinyk A., Mishchenko M. I., Yang P., and Slutsker I. (2002b). Non-spherical aerosol retrieval method employing light scattering by spheroids. Geophys.Res. Lett.,29,1415, doi:10.1029/2001GL014506.

Dubovik O., Sinyuk A., Lapyonok T., Holben B. N., Mishchenko M., Yang P., Eck T. F., Volten H., Munoz O., Veihelmann B., van der Zande W. J., Leon J.-F., Sorokin M., and Slutsker I. (2006). Application of spheroid models to account for aerosol particle

nonsphericity in remote sensing of desert dust. J. Geophys. Res. 111, 11208. doi:10.1029/2005JD006619 2006

Duncan B., Prados A., Lamsal L., Yang L., Streets D., Gupta P., Hilsenrath E., Kahn R., Nielsen J.E., Beyersdorf A., Burton S., Fiore A., Fishman J., Henze D., Hostetler C., Krotkov N., Lee P., Meiyun L., Pawson S., Pfister G., Pickering K. E., Bradley Pierce R., Yoshida Y., Ziemba L. D. (2014). Review: Satellite data of atmospheric pollution for U.S. air quality applications: Examples of applications, summary of data end-user resources, answers to FAQs, and common mistakes to avoid. Atmospheric Environment. Sep 2014, Vol. 94, p647-662. 16p. http://dx.doi.org/10.1016/j.atmosenv.2014.05.061

Eck T. F., Holben,B. N., Reid J. S., Dubovik O., Smirnov A., O'Neill, N. T. (1999). Wavelength dependence of the optical Depth of biomass burning, urban, and desert dust aerosols, J. Geophys.Res.-Atmos., 104(D24), 31333–31349, 1999.

Eguiluz De A.S., Aranda G.M., Marrett R. (2000). Tectónica de la Sierra Madre Oriental, México. Boletín de la Sociedad Geológica Mexicana, v. LIII, pp. 1-26.

Engel-Cox J., Holloman C., Coutant B., Hoff R. (2004). Qualitative and quantitative evaluation of MODIS satellite sensor data for regional and urban scale air quality. Atmospheric Environment 38, 2495–2509.

EPA - US Environmental Protection Agency. (1997). Guidance for Network Design and Optimun Site Exposure for $PM_{2.5}$ and PM_{10}. EPA-454/R99-002, December 1997.

EPA - US Environmental Protection Agency. (2003). Guidelines for Developing an Air Quality (Ozone and $PM_{2.5}$) Forecasting Program. EPA-456/R-03-002.Research Triangle Park, North Carolina.

EPA - US Environmental Protection Agency. (2004). Air Quality Criteria for Particulate Matter, vols. 1 of 2. Research Triangle Park, North Carolina.

EPA - US Environmental Protection Agency. (2009). Technical assistance document for reporting of daily air quality—the air quality index (AQI). EPA-454/B-09-001. North Carolina.

EPA - US Environmental Protection Agency. (2014). AQI Air Quality Index: A guide to air quality and your health.

EPA - US Environmental Protection Agency. (2015). Air Quality Guide for Particle Pollution. Office of Air Quality and Radiation (6301A). EPA-456/F-15-005.

EPA - US Environmental Protection Agency. (2016). List of designated reference and equivalent methods. National exposure research laboratory. Research Triangle Park, North Carolina.

EPA - US Environmental Protection Agency. (2018). List of designated reference and equivalent methods. National exposure research laboratory. Exposure methods and measurement division (MD-D205-03). Research Triangle Park, NC.

European Commission. (2011). European Commission/Joint Research Centre (JRC)/Netherlands Environmental Assessment agency (PBL): Emission Database for Global

Atmospheric Research (EDGAR), release version 4.2. [Available online at http://edgar.jrc.ec.europa.eu.]

Evans K. F., and Stephens G. L. (1991). A new polarized atmospheric radiative-transfer model. J. Quant. Spectrosc. Ra., 46(5), 413–423, 1991.

Feng S., Gao D., Liao F., Zhou F., Wang X. (2016). Review: The health effects of ambient $PM_{2.5}$ and potential mechanisms. Ecotoxicology and Environmental Safety Volume 128, June 2016, Pages 67-74. https://doi.org/10.1016/j.ecoenv.2016.01.030

Fishman J., Bowman K.W., Burrow J. P., Richter A., Chance K. V., Edwards D. P., Martin R. V., Morris G. A., Pierce R. B., Ziemke J. R., Al-Saadi J. A., Creilson J. K., Schaack T. K., Thompson A.M. (2008). Remote sensing of tropospheric pollution from space. Bull. Am. Meteorol. Soc., 805e821 http://dx.doi.org/10.1175/2008BAMS2526.1.

Fu P., Guo X., Cheung F. M. O, Yung K. K. L. (2019). The association between $PM_{2.5}$ exposure and neurological disorders: A systematic review and meta-analysis. Science of the Total Environment 655 pp. 1240–1248. https://doi.org/10.1016/j.scitotenv.2018.11.218.

Fu, Q. (2015) RADIATION TRANSFER IN THE ATMOSPHERE | Cloud-Radiative Processes». Encyclopedia of Atmospheric Sciences (Second Edition), editado por Gerald R. North, John Pyle, y Fuqing Zhang, 13-15. Oxford: Academic Press, 2015. https://doi.org/10.1016/B978-0-12-382225-3.00338-8.

Gelaro R., McCarty W., Suárez M. J., Todling R., Molod A.; Takacs L., Randles C., Darmenov A., Bosilovich M.G, Reichle R., Wargan K., Coy L., Cullather R., Draper C., Akella S., Buchard V., Conaty A., Da Silva A., Gu W., Zhao B. et al. (2017). The Modern-Era Retrospective Analysis for Research and Applications, Version 2 (MERRA-2). Journal of Climate. 30. 10.1175/JCLI-D-16-0758.1.

GMAO - Global Modeling and Assimilation Office. (2015a). M2TMNXSLV -MERRA-2 tavgM_2d_slv_Nx: 2d, Monthly mean, Time-Averaged, Single-Level, Assimilation, Single-Level Diagnostics V5.12.4 files. GES DISC.DOI:10.5067/AP1B0BA5PD2K

GMAO - Global Modeling and Assimilation Office. (2015b).: M2TMNXAER - MERRA-2 tavgM_2d_aer_Nx: 2d, Monthly mean, Time-averaged, Single-Level, Assimilation, Aerosol Diagnostics V5.12.4 files. GES DISC. DOI:10.5067/FH9A0MLJPC7N

Gong S. L. (2003). A parameterization of sea-salt aerosol source function for sub- and super-micron particles. Global Biogeochem. Cycles, 17, 1097, doi:10.1029/2003GB002079.

Gonzalez L. T, Longoria F. E, Sanchez M., Leyva C., Acuña K., Kharissov B., Arizpe A., Alfaro J. M. (2018). Seasonal variation and chemical composition of particulate matter: A study by XPS, ICP-AES and sequential microanalysis using Raman with SEM/EDS. Journal of environmental sciences 74 (2018) 32 -49. https://doi.org/10.1016/j.jes.2018.02.002

Guo Y., Tang Q., Gong D-Y., Zhang Z. (2017). Estimating ground-level $PM_{2.5}$ concentrations in Beijing using a satellite-based geographically and temporally weighted regression model. Remote Sensing of Environment – Article.

Gupta P., Christopher S. A, Wang J., Gehrig R., Lee Yc., Kumar N. (2006). Satellite remote sensing of particulate matter and air quality assessment over global cities. Atmospheric Environment. doi:10.1016/j.atmosenv.2006.03.016

Gupta P., and Christopher S. A. (2008). Seven year particulate matter air quality assessment from surface and satellite measurements, Atmos. Chem. Phys., 8, 3311–3324. http://www.atmos-chem-phys.net/8/3311/2008/.

Gupta P., and Christopher S. A. (2009a). Particulate matter air quality assessment using integrated surface, satellite, and meteorological products: Multiple regression approach, J. Geophys. Res., 114, D14205, doi:10.1029/2008JD011496, 2009.

Gupta P., and Christopher S. A. (2009b). Particulate matter air quality assessment using integrated surface, satellite, and meteorological products: 2 A neural network approach, J. Geophys. Res., 114, D14205, doi:10.1029/2008JD011497, 2009.

Gupta P., Joiner J., Vasilkov A., and Bhartia P. (2016a). Top-of-the-atmosphere shortwave flux estimation from satellite observations: an empirical neural network approach applied with data from the A-train constellation. Atmos. Meas. Tech., 9, 2813–2826, 2016. doi:10.5194/amt-9-2813-2016

Gupta P., Levy R. C., Mattoo S., Remer L. A., and Munchak L. A. (2016b) A surface reflectance scheme for retrieving aerosol optical depth over urban surfaces in MODIS Dark Target retrieval algorithm, Atmos. Meas. Tech., 9, 3293-3308, https://doi.org/10.5194/amt-9-3293-2016, 2016.

Gupta P., Doraiswamy P., Levy R., Pikelnaya O., Maibach J., Feenstra B., Polidori A., Kiros F., and Mills K. C. (2018). Impact of California Fires on Local and Regional Air Quality: The Role of a Low-Cost Sensor Network and Satellite Observations. GeoHealth, 2. https://doi.org/10.1029/2018GH000136

Hadjimitsis D.G., Clayton C, Toulios L. (2010). Retrieving visibility values using satellite remote sensing data. J. Physics and Chemistry of the Earth. Volume 35, Issues 1–2: 121–124. DOI: 10.1016/j.pce.2010.03.002

Hansell R. A.,. Ou S. C, Liou K. N., Roskovensky J. K., Tsay S. C., Hsu C.,and Ji Q. (2007). Simultaneous detection/separation of mineral dust and cirrus clouds using MODIS thermal infrared window data. Geophysical Research Letters. Volume 34, Issue 11. doi:10.1029/2007GL029388

Harrison R.M., Shi J.P., Xi S., Khan A., Mark D., Kinnersley R. and Yin J. (2000). Measurement of number, mass and size distribution of particles in the atmosphere. Philosophical Transactions of the royal society a mathematical, physical and engineering sciences, volume 358, issue 1775, pp. 2567-2580. Online ISSN:1471-2962. https://doi.org/10.1098/rsta.2000.0669

He L., Lin A., Chen X., Zhou H., Zhou Z., He P. (2019). Assessment of MERRA-2 Surface $PM_{2.5}$ over the Yangtze River Basin: Ground-based Verification, Spatiotemporal Distribution and Meteorological Dependence. Remote Sensing, vol. 11, p. 460, 2019.

He Q., Li C., Mao J., Lau A.K.-H., Chu D.A. (2008). Analysis of aerosol vertical distribution and variability in Hong Kong. Journal of geophysical research, vol. 113. http:// dx.doi.org/10.1029/2008JD009778

Hess M., Koepke P., Schult I. (1998). Optical properties of aerosols and clouds: the software package OPAC. Bull. Am. Meteorol. Soc. 79, 831–844.

Hernández Paniagua I. Y., Clemitshaw K. C., and Mendoza A. (2017). Observed trends in ground-level O3 in Monterrey, Mexico during 1993-2014: Comparison with Mexico City and Guadalajara. Atmos. Chem. Phys. 17:9163–85. doi:10.5194/acp-17-9163-2017.

Hoff R., Christopher S. (2009). Remote sensing of particulate pollution from space: have we reached the promised land?. J. Air Waste Manag. Assoc. 59 (6), 645-675.

Holben B. N., Eck T. F., Slutsker I., Tanre D., Buis J. P., Setzer A., et al. (1998). AERONET – A federated instrument network and data archive for aerosol characterization, Remote Sens. Environ.,66(1), 1–16.

Hsu, N.C. (2020) How Aerosols Are Measured: The Science of Deep Blue | Deep Blue. Susannah Pearce and Nate Perrin. Accedido 10 de junio de 2020. /science.

Hsu N. C., Tsay S.-C., King M. D., and Herman J. R. (2004). Aerosol properties over bright-reflecting source regions, IEEE Trans. Geosci. Remote Sens., 42(3), 557–569, doi:10.1109/TGRS.2004.824067.

Hsu N. C., Jeong M. J., Bettenhausen C., Sayer A. M., Hansell R., Seftor C. S., et al. (2013). Enhanced Deep Blue aerosol retrieval algorithm: the second generation. J Geophys Res 118:9296–9315

Hu X., Belle J.H., Meng X., Wildani A., Waller L., Strickland M., and Liu Y. (2017). Estimating $PM_{2.5}$ concentrations in the conterminous United States using the random forest approach. Environ. Sci. Technol. 51, 6936–6944. https://doi.org/10.1021/acs.est.7b01210.

Hubanks P. A. (2012). MODIS Atmosphere QA Plan for Collection 005. Greenbelt, MD USA, NASA Goddard Space Flight Center.

Ichoku C., Remer L. A., Kaufman Y. J., Levy R., Chu D. A., Tanré D., and Holben B. N. (2003). MODIS observation of aerosols and estimation of aerosol radiative forcing over southern Africa during SAFARI 2000. Journal of Geophysical Research Atmospheres. Volume 108, Issue D13. doi:10.1029/2002JD002366

INECC - Instituto Nacional de Ecología y Cambio Climático. (2016). Monitoreo Atmosférico Nacional de Carbono Negro 2015-2016.

IPCC - Intergovernmental Panel on Climate Change. (2014). IPCC Fifht Assessment Report: Climate Change 2013. The Physical Science Basis

Jiang D., Zhang Y., Hu X., Zeng Y., Tan J., Shao D. (2004). Progress in developing an ANN model for air pollution index forecast. Atmospheric Environment. Volume 38, Issue 40. Pages 7055-7064. https://doi.org/10.1016/j.atmosenv.2003.10.066

Jiang M., Sun W. (2016). Investigating metrological and geographical effect in remote sensing retrieval of $PM_{2.5}$ concentration in Yangtze River Delta. *2016 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, 4108-4111.

Just A.C., Wright R.O., Schwartz J., Coull B.A., Baccarelli A.A., Tellez-Rojo M.M., Moody E., Wang Y., Lyapustin A., Kloog I. (2015). Using high-resolution satellite aerosol optical depth to estimate daily $PM_{2.5}$ geographical distribution in Mexico City. Environ. Sci. Technol. 2015, 49, 8576–8584. doi:10.1021/acs.est.5b00859.

Kahna R. A., Garay M.J., Nelson D.L., Levy R.C., Bull M.A., Diner D.J., Martonchik J.V., Hansen E.G., Remer L.A., Tanré D. (2011). Response to "Toward unified satellite climatology of aerosol properties. 3. MODIS versus MISR versus AERONET". Journal of Quantitative Spectroscopy and Radiative Transfer. Volume 112, Issue 5, March 2011, Pages 901–909. DOI: 10.1016/j.jqsrt.2010.11.001

Kahnert M., Nousiainen T., and Lindqvist H. (2014). Review: Model particles in atmospheric optics. Journal of Quantitative Spectroscopy and Radiative Transfer 146 (2014) 41–58

Kamarul Z., NAF., Kanniah K. D., Kaskaoutis D.G. (2017). Estimating Particulate Matter using satellite based aerosol optical depth and meteorological variables in Malaysia Atmospheric Research 2017 Sep 1;193:142-162. Available from, DOI: 10.1016/j.atmosres.2017.04.019

Karagulian F., Belis C.A., Francisco C., Dora C., PrüssUstün A.M., Bonjour S. (2015). Contributions to cities' ambient particulate matter (PM): A systematic review of local source contributions at global level. Atmospheric Environment, 120 (November) (2015), pp. 475-483, 10.1016/j.atmosenv.2015.08.087. ISSN 13522310

Kaufman Y. J., Gobron N., Pinty B., Widlowski J., and Verstraete M. (2002). Relationship between surface reflectance in the visible andmid-IR used in MODIS aerosol algorithm – theory, Geophys.Res. Lett., 29(23), 2116, doi:10.1029/2001GL014492, 2002.

Kloog I., Chudnovsky A. A., Just A. C., Nordio F., Koutrakis P., Coull B.A., Lyapustin A., Wang Y., Schwartz J. (2014). A new hybrid spatio-temporal model for estimating daily multi-year $PM_{2.5}$ concentrations across northeastern USA using high resolution aerosol optical depth data. Atmospheric Environment Volume 95, October 2014, Pages 581–590. doi:10.1016/j.atmosenv.2014.07.014

Kloog I., Koutrakis P., Coull B. A., Lee H .J., Schwartz J. (2011). Assessing temporally and spatially resolved $PM_{2.5}$ exposures for epidemiological studies using satellite aerosol optical depth measurements. Atmospheric Environment 45, 6267-6275.

Kokhanovsky A. (2008). Aerosol Optics: Light Absorption and Scattering by Particles in the Atmosphere. Springer Berlin Heidelberg New York. Praxis Publishing Ltd, Chichester, UK. ISBN 978-3-540-23734-1

Koelemeijer R.B.A., Homan C.D., Matthijsen J. (2006). Comparison of Spatial and Temporal Variations of Aerosol Optical Thickness and Particulate Matter over Europe. Atmospheric Environment. 40, 5304-5315. doi:10.1016/j.atmosenv.2006.04.044

Kong L., Xin J., Liu Z., Zhang K., Tang G., Zhang W., Wang Y. (2017). The $PM_{2.5}$ threshold for aerosol extinction in the Beijing megacity. Atmospheric Environment. 167. 10.1016/j.atmosenv.2017.08.047.

Lawrence S., and Giles C. L. (2000) Overfitting and Neural Networks: Conjugate Gradient and Backpropagation, International Joint Conference on Neural Networks, Como, Italy, July 24–27, IEEE Computer Society, Los Alamitos, CA, pp. 114–119, 2000.

Lee H. J., Kang C. M., Coull B .A., Bell M .L., Koutrakis P. (2014). Assessment of primary and secondary ambient particle trends using satellite aerosol optical depth and ground speciation data in the New England region, United States. Environmental Research [Environ Res]. Vol. 133, pp. 103-10.

Lee H. J., Chatfield R. B., Strawa A. W. (2016). Enhancing the applicability of satellite remote sensing for $PM_{2.5}$ estimation using MODIS Deep Blue AOD and land use regression in California, United States. Environmental Science & Technology 50 pp 6546–6555. https://doi. org/10.1021/acs.est.6b01438

Levy R. C., Remer L. A., Kleidman R. G, Mattoo S., Ichoku C., Kahn R., and Eck T. F. (2010). Global evaluation of the Collection 5 MODIS dark-target aerosol products over land, Atmos. Chem. Phys., 10 , 103,999–10,420, doi:10.5194/acp-10-10399-2010.

Levy R. C., Mattoo S., Munchak L.A., Remer L.A, Sayer A.M., Patadia F., and Hsu N.C. (2013). The Collection 6 MODIS aerosol products over land and ocean. Atmospheric Measurment Techniques (6) 2989–3034. doi:10.5194/amt-6-2989-2013

Li R.-R., Kaufman Y. J., Gao B.-C., and Davis C. O. (2003). Remote sensing of suspended sediments and shallow coastal waters. IEEE Trans. Geosci. Remote Sens. ,41,559–566.

Li R., Remer L., Kaufman,Y., Mattoo S., Gao B., and Vermote E. (2005). Snow and ice mask for the MODIS aerosol products. IeeeGeosci. Remote S., 2(3), 306–310, 2005.

Li J., Carlson B., Lacis A.A. (2015). How well do satellite AOD observations represent the spatial and temporal variability of $PM_{2.5}$ concentration for the United States?. J. Atmospheric Environment 102 (2015) 260 - 273. http://dx.doi.org/10.1016/j.atmosenv.2014.12.010

Li T., Shen H., Zeng C., Yuan Q., Zhang L. (2017). Point-surface fusion of station measurements and satellite observations for mapping $PM_{2.5}$ distribution in China: Methods and assessment", Atmospheric environment, vol.152, p. 477-489, 2017.

Liang C.S., Duan F.K., He K.B., Ma Y.L. (2016). Review on recent progress in observations, source identifications and countermeasures of $PM_{2.5}$. Environment International, 86 (January) (2016), pp. 150-170, 10.1016/j.envint.2015.10.016. ISSN 01604120

Lin S.-J. (2004). A''vertically Lagrangian'' finite-volume dynamical core for global models. Mon. Wea. Rev., 132, 2293–2307, doi:10.1175/1520-0493(2004)132,2293: AVLFDC.2.0.CO;2.

Liu Y., Franklin M., Kahn R., Koutrakis P. (2007). Using aerosol optical thickness to predict ground-level $PM_{2.5}$ concentrations in the St. Louis area: A comparison between MISR and MODIS. J of Remote Sensing of Environment. Volume 107, Issues 1–2, 15 March 2007, Pages 33–44. doi:10.1016/j.rse.2006.05.022

López-Ramos E. (2008). Geología general de México. Editorial Trillas S.A de CV. ISBN 978-968-24-1176-2. México.

Ma X., Yu F. (2014). Seasonal variability of aerosol vertical profiles over east US and west Europe: GEOS-Chem/APM simulation and comparison with CALIPSO observations. Atmos. Res. 140-141, 28–37.

Mancilla Y., Mendoza A. (2012). A tunnel study to characterize $PM_{2.5}$ emissions from gasoline-powered vehicles in Monterrey, Mexico. Atmospheric Environment 59 (2012) 449–460. doi:10.1016/j.atmosenv.2012.05.025.

Mancilla Y., Herckes P., Fraser M. P., and Mendoza A. (2015). Secondary organic aerosol contributions to $PM_{2.5}$ in Monterrey, Mexico: Temporal and seasonal variation. Atmos. Res. 153:348–59. doi:10.1016/j.atmosres.2014.09.009.

Mancilla Y., Mendoza A., Fraser M. P., and Herckes P.. (2016). Organic composition and source apportionment of fine aerosol at Monterrey, Mexico, based on organic markers. Atmos. Chem. Phys. 16:953–70. doi:10.5194/acp-16-953-2016.

Mancilla Y., Hernandez Paniagua I. Y., and Mendoza A. (2019). Spatial differences in ambient coarse and fine particles in the Monterrey metropolitan area, Mexico: Implications for source contribution. Journal of the Air & Waste Management Association. DOI: 10.1080/10962247.2018.1549121

Mao X., Shen T., Feng X. (2017). Prediction of hourly ground-level $PM_{2.5}$ concentrations 3 days in advance using neural networks with satellite data in eastern China. Atmospheric Pollution Research. 8. 10.1016/j.apr.2017.04.002.

Marticorena B., and Bergametti G. (1995). Modeling the atmospheric dust cycle: 1. Design of a soil-derived dust emission scheme. J. Geophys. Res., 100, 16 415–16 430, doi:10.1029/95JD00690.

Martínez M. A., Caballero P., Carrillo O., Mendoza A., and Mejía G. M. (2012). Chemical characterization and factor analysis of $PM_{2.5}$ in two sites of Monterrey, Mexico. J. Air Waste Mange. Assoc. 62:817–27. doi:10.1080/10962247.2012.681421.

Martins J., Tanre D., Remer,L., Kaufman Y., Mattoo S.,and Levy R. (2002). MODIS Cloud screening for remote sensing ofaerosols over oceans using spatial variability, Geophys. Res.Lett., 29(12), 1619, doi:10.1029/2001GL013252.

Martins N.R., Carrilho da Graça G. (2018). Impact of $PM_{2.5}$ in indoor urban environments: A review. Sustainable Cities and Society 42 (2018) pp. 259–275. https://doi.org/10.1016/j.scs.2018.07.011

MATLAB® (version R2014a) and Neural Network Toolbox. The MathWorks, Inc., Natick, Massachusetts, United States.

Medina Gaitán G. E. (2015). Desarrollo de perfiles de emisión de la fracción orgánica del $PM_{2.5}$ en el Área Metropolitana de Monterrey. Master Thesis, Tecnologico de Monterrey, December.

Mhawish A., Kumar M., Mishra A.K., Srivastava P.K., Banerjee T. (2018). Chapter 3 - Remote Sensing of Aerosols From Space: Retrieval of Properties and Applications. Remote Sensing of Aerosols, Clouds, and Precipitation. Pages 45-83. https://doi.org/10.1016/B978-0-12-810437-8.00003-7

Minitab® 16 statistical software (State College, PA, USA)

MODIS - LAADS (2020). https://ladsweb.modaps.eosdis.nasa.gov/missions-and-measurements/modis/. Accedido 10 de junio de 2020.

Moosmüller H., and Sorensen C.M. (2018). Small and large particle limits of single scattering albedo for homogeneous, spherical particles. Journal of Quantitative Spectroscopy and Radiative Transfer. Volume 204. Pages 250-255. https://doi.org/10.1016/j.jqsrt.2017.09.029

NOAA - National Oceanic and Atmospheric Administration. (2019). Aerosol Optical Depth. Retrieved from http://www.esrl.noaa.gov/gmd/grad/surfrad/aod/

Péré J.-C., Pont V., Mallet M., Bessagnet B. (2009). Review: Mapping of PM_{10} surface concentrations derived from satellite observations of aerosol optical thickness over South-Eastern France. Atmospheric Research 91 (2009) 1–8. doi:10.1016/j.atmosres.2008.05.001

Piotrowski A.P., Napiorkowski J. J. (2011). Optimizing neural networks for river flow forecasting – Evolutionary Computation methods versus the Levenberg–Marquardt approach. Journal of Hydrology 407 (2011) 12–27. doi:10.1016/j.jhydrol.2011.06.019

Poznyak A., Chairez I., Poznyak T. (2019). Review article: A survey on artificial neural networks application for identification and control in environmental engineering: Biological and chemical systems with uncertain models. Annual Reviews in Control. https://doi.org/10.1016/j.arcontrol.2019.07.003

Prechelt L. (1998). Contributed article: Automatic early stopping using cross validation: quantifying the criteria. Neural Networks 11 (1998) 761–767.

Provençal S., Kishcha P., Silva A.M., Elhacham E., Alpert P. (2017). AOD distributions and trends of major aerosol species over a selection of the world's most populated cities based on the 1st Version of NASA's MERRA Aerosol Reanalysis. Urban climate, 20, 168-191.

Prud'homme G., Dobbin N., Sun L., Burnett R., Martin R. V, Davidson A., Cakmak S., Villeneuve P., Lamsal L., Van Donkelaar A., Peters P. A., Johnson M. (2013). Comparison of remote sensing and fixed-site monitoring approaches for examining air pollution and health in a national study population. Atmospheric Environment. Volume 80, Pages 161–171. DOI: 10.1016/j.atmosenv.2013.07.020

Purser R.J., Wu W.-S., Parrish D.F., and Roberts N.M. (2003a). Numerical aspects of the application of recursive filters to variational statistical analysis. Part I: Spatially homogeneous and isotropic Gaussian covariances. Mon. Wea. Rev., 131, 1524-1535.

Purser R.J., Wu W.-S., Parrish D.F., and Roberts N.M. (2003b). Numerical aspects of the application of recursive filters to variational statistical analysis. Part II: Spatially inhomogeneous and anisotropic general covariances. Mon. Wea. Rev., 131, pp. 1536-1548.

QGIS Development Team. (2019). QGIS Geographic Information System. Version QGIS 3.2. Open Source Geospatial Foundation Project. http://qgis.osgeo.org

Quispe-Medina F.M. (2005). Análisis de susceptibilidad a la ocurrencia de deslizamientos en la Sierra de las Mitras, utilizando Sistemas de Información Geográfica (SIG), Tesis de Maestría, Instituto Tecnológico y de Estudios Superiores de Monterrey.

Randles C.A., da Silva A.M., Buchard V., Colarco P.R., Darmenov A., Govindaraju R., Smirnov A., Holben B., Ferrare R., Hair J., Shinozuka Y., Flynn C.J. (2017). The MERRA-2 aerosol reanalysis, 1980 onward. part I: system description and data assimilation evaluation. J. Clim. 30, 6823–6850. http://dx.doi.org/10.1175/JCLI-D-16-0609.s1.

Remer L. A., and Y. J. Kaufman. (1998). Dynamical aerosol model:Urban/industrial aerosol. J. Geophys. Res.,103,13 859–13 871.

Remer L. A., Mattoo S., Levy R. C., and Munchak L. A. (2013). MODIS 3 km aerosol product: Algorithm and global perspective. Atmos. Meas. Tech., 6, 1829–1844, doi:10.5194/amt-6-1829-2013.

Rienecker M.M., Suarez M.J., Todling R., Bacmeister J., Takacs L., Liu H.-C., Gu W., Sienkiewicz M., Koster R.D., Gelaro R., Stajner I., Nielsen J. E. (2008). The GEOS-5 Data Assimilation System—Documentation of versions 5.0.1, 5.1.0, and 5.2.0. NASA/TM-2014-104606, Vol 27, NASA Global Modeling and Assimilation Office, 101 pp. [Available online at https://gmao.gsfc.nasa.gov/pubs/docs/Rienecker369.pdf.]

Rienecker M.M., Suarez M.J., Gelaro R., Todling R., Bacmeister J., Liu E., Bosilovich M.G., Schubert S.D., Takacs L., Kim G.K., Bloom S., Chen J., Collins D., Conaty A., da Silva A., Gu W., Joiner J., Koster R.D., Lucchesi R., Molod A., Owens T., Pawson S., Pegion P., Redder C.R., Reichle R., Robertson F.R., Ruddick A.G., Sienkiewicz M., Woollen J. (2011). MERRA: NASA's Modern-Era retrospective analysis for research and application. J. Clim. 24, 3624–3648. DOI: 10.1175/JCLI-D-11-00015.1

Riojas H., Soares A., Texcalac J.L., Moreno G.L. (2016). Air pollution management and control in Latin America and the Caribbean: implications for climate change. Rev Panam Salud Publica. 2016;40(3):150–59.

Robertson S., Miller M.R. (2018). Ambient air pollution and thrombosis. J. Particle and Fibre Toxicology. DOI 10.1186/s12989-017-0237-x

Rowiński P.M., Piotrowski A., and Napiórkowski J.J. (2005). Are Artificial Neural Network Techniques Relevant for the Estimation of Longitudinal Dispersion Coefficient in Rivers? / Les Techniques de Réseaux de Neurones Artificiels Sont-Elles Pertinentes Pour Estimer Le Coefficient de Dispersion Longitudinale En Rivières?. Hydrological Sciences Journal 50, n.o 1 (febrero de 2005): 6. https://doi.org/10.1623/hysj.50.1.175.56339.

Sarle W. S. (1995). Stopped training and other remedies for overfitting. In: Proceedings of the 27th Symposium on the Interface of Computer Science and Statistics, pp. 352e360.

Sayer A. M., Munchak L.A., Hsu N. C., Levy R. C., Bettenhausen C., and Jeong M.-J. (2014). MODIS Collection 6 aerosol products: Comparison between Aqua's e-Deep Blue, Dark

Target, and “merged” data sets, and usage recommendations. J. Geophys. Res. Atmos., 119, 13,965–13,989, doi:10.1002/2014JD022453.

Schaap M., Apituley A., Timmermans R. M. A., Koelemeijer R. B. A., de Leeuw G. (2009). Exploring the relation between aerosol optical depth and $PM_{2.5}$ at Cabauw, the Netherlands. Atmos. Chem. Phys., 9, 909–925, 2009. www.atmos-chem-phys.net/9/909/2009/

Schafer K., Harbusch A., Emeis S., Koepke P., Wiegner M. (2008). Correlation of aerosol mass near the ground with aerosol optical depth during two seasons in Munich. Atmospheric Environment Vol. 42 Issue 18 pp 4036-4046. https://doi.org/10.1016/j.atmosenv.2008.01.060

Schmidhuber J. (2015). Review, Deep learning in neural networks: An overview. Neural Networks 61 (2015) 85–117. http://dx.doi.org/10.1016/j.neunet.2014.09.003

Schowengerdt, R.A. (2007). Remote Sensing: Models and Methods for Image Processing. Third Edition. Academic Press of Elsevier. ISBN 13: 978-0-12-369407-2.

Schultz M. G., Heil A., Hoelzemann J.J., Spessa A., Thonicke K., Goldammer J. G., Held A. C., Pereira J. M. C., van het Bolscher M. (2008): Global wildland fire emissions from 1960 to 2000. Global Biogeochem. Cycles, 22, GB2002. doi:10.1029/2007GB003031.

Secretaría de Medio Ambiente y Recursos Naturales. Norma Oficial Mexicana NOM-036-SEMARNAT-1993. Métodos de medición para determinar la concentración de ozono en el aire ambiente y los procedimientos para la calibración de los equipos de medición. D.O.F., 18 de octubre de 1993.

Secretaría de Medio Ambiente y Recursos Naturales. Norma Oficial Mexicana NOM-156-SEMARNAT-2012, Establecimiento y operación de sistemas de monitoreo de la calidad del aire. D.O.F., 16 de julio de 2012.

Secretaría del Medio Ambiente. Norma Ambiental para el Distrito Federal NADF-009-AMBT-2017. Requisitos para Elaborar el Índice de Calidad del Aire en la Ciudad de México. Gaceta oficial de la Ciudad de México. 14 de noviembre de 2008.

Secretaría de Salud. Norma Oficial Mexicana NOM-025-SSA1-2014. Salud ambiental. Valores límite permisibles para la concentración de partículas suspendidas PM_{10} y $PM_{2.5}$ en el aire ambiente y criterios para su evaluación. D.O.F., 20 de agosto de 2014.

Shao Y., Taff G.N., Walsh S. J. (2011). Comparison of Early Stopping Criteria for Neural-Network-Based Subpixel Classification. IEEE Geoscience and remote sensing letters. Vol. 8, No. 1, January 2011.

Shen H., Li T., Yuan Q., Zhang L. (2018). Estimating regional ground-level $PM_{2.5}$ directly from satellite top-of-atmosphere reflectance using deep belief networks. Journal of Geophysical Research: Atmospheres, vol. 123, 13, p. 875–13, 886, 2018.

Shepherd A. J. (1997). Second-order Methods for Neural Networks: fast and reliable training methods for Multi-Layer perceptrons. ISBN-13:978-3-540-76100-6. Springer-Verlag Berlin Heidelberg New York.

Song W., Jia H., Huang J., Zhang, Y. (2014). A satellite-based geographically weighted regression model for regional $PM_{2.5}$ estimation over the Pearl River Delta region in China. Remote Sens. Environ. 2014, 154, 1–7.

Soni M., Payra S., Verma S. (2018). Particulate matter estimation over a semi-arid region Jaipur, India using satellite AOD and meteorological parameters. Atmospheric Pollution Research Volume 9, Issue 5, Pages 949-958. https://doi.org/10.1016/j.apr.2018.03.001

Sportisse B. (2010). Fundamentals in Air Pollution: From Processes to Modelling. Springer Dordrecht Heidelberg London New York. ISBN 978-90-481-2970-6.

Stirnberg R., Cermak J., Andersen H. (2018). An Analysis of Factors Influencing the Relationship between Satellite-Derived AOD and Ground-Level PM_{10}. Remote Sensing (10), 1353. doi:10.3390/rs10091353

Strawa A.W., Chatfield R. B., Legg M., Scarnato B., Esswein R. (2013). Improving Retrievals of Regional $PM_{2.5}$ Concentrations from MODIS and OMI Multi-Satellite Observations, Journal of the Air & Waste Management Association, DOI:10.1080/10962247.2013.822838

Streets D.G., Canty T., Carmichael G.R., de Foy B., Dickerson R.R., Duncan B.N., Edwards D.P., Haynes J.A., Henze D.K., Houyoux M.R., Jacob D.J., Krotkov N.A., Lamsal L.N., Liu Y., Lu Z,. Martin R.V., Pfister G. G., Pinder R.W., Salawitch R. J., and Wechtj K. J. (2013). Review: Emissions estimation from satellite retrievals: A review of current capability. Atmospheric Environment 77 (2013) 1011-1042. http://dx.doi.org/10.1016/j.atmosenv.2013.05.051

Taiwo A., Harrison R., Shi Z. (2014). A review of receptor modeling of industrially emitted particulate matter. Atmospheric Environment 97 (2014) 109-120 DOI: 10.1016/j.atmosenv.2014.07.051

Tesche M., Glantz P., Johansson C. (2016). Spaceborne observations of low surface aerosol concentrations in the Stockholm region, Tellus B: Chemical and Physical Meteorology, 68:1, DOI: 10.3402/tellusb.v68.28951.

Trejo V. (2006). El IMECA: Indicador del Grado de Contaminación de la Atmósfera, Nota de Divulgación. Conciencia Tecnológica No. 31. ISSN: 1405-5597

van der Werf G. R., Randerson J. T., Giglio L., Collatz G. J., Kasibhatla P. S., and Arellano Jr. A. F. (2006). Interannual variability in global biomass burning emissions from 1997 to 2004. Atmos. Chem. Phys., 6, 3423–3441, doi:10.5194/acp-6-3423-2006.

Van Donkelaar A., Martin R.V., Brauer M., Kahn R., Levy R., Verduzco C., Villeneuve P. J. (2010). Global estimates of ambient fine particulate matter concentrations from satellite-based aerosol optical depth: development and application. Environmental Health Perspectives 118, 847-855.

Van Donkelaar A., Martin R. V., Brauer M., and Boys B. L. (2015). Use of Satellite Observations for Long-Term Exposure Assessment of Global Concentrations of Fine Particulate Matter. Environ. Health Perspect. 2015, 123 (2), 135−143

Van Donkelaar A., Martin R. V., Brauer M., Hsu C., Kahn R. A., Levy R.C., Lyapustin A., Sayer A. M., and Winker D. (2016). Global Estimates of Fine Particulate Matter using a Combined Geophysical-Statistical Method with Information from Satellites, Models, and Monitors. Environ. Sci. Technol. 2016, 50, 3762−3772. DOI: 10.1021/acs.est.5b05833

Vianaa M., Peya J., Querola X., Alastueya A., de Leeuwb F., Lükewillec A. (2014). Natural sources of atmospheric aerosols influencing air quality across Europe. Science of the Total Environment. Volume 472, Pages 825–833

Vu T., Delgado-Saborit J.M., and Harrison R.M. (2015). Review: Particle number size distributions from seven major sources and implications for source apportionment studies. Atmospheric Environment (122) pp. 114-132. doi.org/10.1016/j.atmosenv.2015.09.027

Wakamatsu S., Kanda I., Okazaki Y., Saito M., Yamamoto M., Watanabe T., Maeda T., and Mizohata A. (2017). A Comparative Study of Urban Air Quality in Megacities in Mexico and Japan: Based on Japan-Mexico Joint Research Project on Formation Mechanism of Ozone, VOCs and $PM_{2.5}$, and Proposal of Countermeasure Scenario. JICA-RI Working Paper.

Wang E., Eliot M.N., Wellenius G.A. (2014). Short-term Changes in Ambient Particulate Matter and Risk of Stroke: A Systematic Review and Meta-analysis. Journal of the American Heart Association. doi: 10.1161/JAHA.114.000983

Wang S., Zhou C., Wang Z., Feng K., Hubacek K. (2017). The characteristics and drivers of fine particulate matter ($PM_{2.5}$) distribution in China. J. Clean. Prod.142, 1800e1809.

WHO - World Health Organization. (2005). WHO Air quality guidelines for particulate matter, ozone, nitrogen dioxide and sulfur dioxide. Global update 2005. Summary of risk assessment.

Wiscombe W. (1980). Improved mie scattering algorithms, Appl. Optics,19(9), 1505–1509, 1980.

Wu W.-S., Purser R., Parrish D. (2002). Three-dimensional variational analysis with spatially inhomogeneous covariances. Mon. Wea. Rev., 130, 2905–2916, doi:10.1175/1520-0493(2002)130,2905:TDVAWS.2.0.CO;2.

Xing Y. F., Xu Y. H., Shi M. H., Lian Y. X. (2016). Review: The impact of $PM_{2.5}$ on the human respiratory system. Journal of Thoracic Disease. doi: 10.3978/j.issn.2072-1439.2016.01.19

You W., Zang Z., Zhang L., Li Y., Wang W. National-Scale Estimates of Ground-Level $PM_{2.5}$ Concentration in China Using Geographically. (2015). Institute of Meteorology and Oceanography, PLA University of Science and Technology, Nanjing 211101, China.

Zhang B., Zhang M., Kang J., Hong D., Xu J., and Zhu X. (2019). Estimation of PMx Concentrations from Landsat 8 OLI Images Based on a Multilayer Perceptron Neural Network. Remote Sensing 11 (2019). doi:10.3390/rs11060646

Zheng C., Zhao C., Zhu Y., Wang Y., Shi X., Wu X., Chen T., Wu F., Qiu Y. (2017). Analysis of influential factors for the relationship between $PM_{2.5}$ and AOD in Beijing. Atmos. Chem. Phys., 17, 13473–13489, 2017. https://doi.org/10.5194/acp-17-13473-2017.

PUBLISHED PAPERS

Publications

(2020) **Carmona J.M**., Gupta P., Lozano-García D.F., Vanoye A.Y., Yépez F.D., and Mendoza A. "Spatial and Temporal Distribution of $PM_{2.5}$ Pollution over Northeastern Mexico: Application of MERRA-2 Reanalysis Datasets". *Remote Sensing* (under review)

(2020) **Carmona J.M**., Gupta P., Lozano-García D.F., Vanoye A.Y., Yépez F.D., Hernández-Paniagua I., and Mendoza A. "Satellite Data and Ground-Based Measurements: Model developmen for $PM_{2.5}$ estimations in Northeast Mexico". *Atmospheric Pollution Research* (unpublished; manuscript in preparation)

(2020) Huertas M.E., Madrid O., **Carmona J.M**., Mendoza A., and González A. "Historical trends and spatio-temporal analysis of hydrometeorological variables in the Colombian Caribbean Region". *International Journal of Climatology* (unpublished; manuscript in preparation)

(2016) Yepez F.D., Huerta R., Ramirez N., Lozano D. F., **Carmona J.M**., and Lasso F. "Evaluación de tres productos de sensores remotos para la actualización de inventarios de árboles urbanos [Not available in English]". *Estudios territoriales en México: Percepción remota y Sistemas de Información Espacial*. ISBN 978-607-520-215-0

(2015) **Carmona J.M**., Vanoye A.Y., Lozano-García D.F., and Mendoza A. "Dust Emission Modeling for the Western Border Region of Mexico and the United States". J. Environ. Earth Sciences. DOI 10.1007/s12665-015-4173-5

Conference Proceedings

(2019) **Carmona J.M**, Lozano D. F, Vanoye A. Y, Yépez F. D, Mendoza A. "Mapeo de Concentraciones y Determinación de Variables Influyentes en la Distribución Regional de $PM_{2.5}$ utilizando Redes Neuronales y el Modelo de Re-análisis MERRA-2 [Not available in English]," 2019 *Congreso Colombiano y Conferencia Internacional de Calidad de Aire y Salud Pública* (CASP), Barranquilla, Colombia, 2019, pp. 1-7, doi: 10.1109/CASAP.2019.8916672.

(2014) **Carmona J.**, Mendoza A. and Lozano F. "Estimación de los Niveles de Partículas Atmosféricas por Medio de Datos MODIS en el Noreste De México [Not available in English]," Presented at the *XVI International Simposium SELPER 2014*. Medellín, Colombia. https://selper.org.co/papers-XVI-Simposio/Sensores-y-plataforma-de-PR/SP24-Integrado-JMCG-Selper2014.pdf

Abstract and Talks

(2019) **Carmona J.**, Gupta P., Lozano D. F and Mendoza A. "Categorization of Mexican Air Quality Index Based On Satellite-Derived Data and Meteorological Ground-Based Measurements". Abstract [GH21B-1199] presented at *2019 AGU Fall Meeting*, San Francisco, CA, 9 -13 Dec.

(2018) **Carmona J.**, Gupta P., Buchard V.; Vanoye A., Yepez F. and Mendoza A. "Analysis of Spatial and Temporal Distribution of $PM_{2.5}$ Pollution Over Northeastern Mexico Using

MERRA-2 Reanalysis Datasets". Abstract [A13M-2648] presented at *2018 AGU Fall Meeting*, Washington, D.C, 10-14 Dec.

(2018) **Carmona J.**, Lozano F., Gupta P., Rios J., Mejia G and Mendoza A. "Modeling local and regional distributions of $PM_{2.5}$ based on satellite-derived data and ground-based measurements". *48th Congress of research and development at Tecnologico de Monterrey. Monterrey*, Mexico.

(2017) **Carmona J.**, Lozano F., Gupta P., Rios J., Mejia G., Hernandez I. and Mendoza A. "Comparison of Satellite Data with Ground-Based Measurements for Assessing Local Distributions of $PM_{2.5}$ in Northeast Mexico". Abstract [A53A-2192] presented at *2017 Fall Meeting, AGU*, New Orleans, LA, 11-15 Dec.

(2017) **Carmona J.**, Lozano F. and Mendoza A. "Predicting Air Quality Indexes for $PM_{2.5}$ combining MODIS data, surface measurements and neural network modeling in a local application over Monterrey Metropolitan Area". *Seventh national congress of research on climate change*. Mexico City, México

CURRICULUM VITAE

Johana Margarita Carmona García

She was born in Pereira, Risaralda, Colombia, where she completed all her basic education and professional studies. In 2006, she obtained her professional degree Environmental Management at Universidad Tecnológica de Pereira. Subsequently, she worked in environmental, safety and food auditing and certifications for the private sector in Colombia and Mexico.

In 2010, she obtained a scholarship from Instituto Tecnológico y de Estudios Superiores de Monterrey, Monterrey campus, to pursue a Master of Science with a specialty in Environmental Systems and being part of the Air Pollution Research Group.

In 2012, she was granted the Excellence Scholarship to pursue a Doctorate in Engineering Sciences, where she was member of the Energy and Climate Change Research Group. In addition, she obtained a fellowship through a funding proposal presented to COLCIENCIAS Colombia. In 2015, she made her research stay through the Universities Space Research Association at NASA- Goddard Space Flight Center working on the project "Exploring the relationship between Aerosol Optical Depth and $PM_{2.5}$ over Monterrey, Mexico".

During her PhD studies, she has participated in complementary projects through the CONACYT funding proposal "Aerosols direct climate forcing in the Metropolitan Area of Monterrey: evaluation and sources attribution". Moreover, she has served as a mentor for environmental engineering students in Colombia, Mexico, and Peru.

Her research interest includes air quality modeling with particular emphasis on the use of satellite data along surface measurements to estimate atmospheric pollutants concentrations; identifying and quantifying temporal and spatial distribution of environmental pollution; use of emissions, dispersion, and reanalysis models to derive air quality data; and development of statistical models and data analysis to model a variety of environmental issues.

Her permanent contact is through email johanamargarita@gmail.com