PREDICTIVE ANALYSIS OF FLIGHT TRAFFIC AT NEW YORK JFK AIRPORT ON AIR POLLUTION USING MACHINE LEARNING

by

Batuhan Burç Türkoğlu

B.Sc. in Chemical Engineering, İstanbul Technical University, 2010

Submitted to the Institute of Environmental Sciences

in partial fulfillment of the requirements for the degree of

Master of Science

in

Environmental Technology

Boğaziçi University

2019

PREDICTIVE ANALYSIS OF FLIGHT TRAFFIC AT NEW YORK JFK AIRPORT ON AIR POLLUTION USING MACHINE LEARNING

APPROVED BY:

Prof. Dr. Nadim Copty . . . . . .
Thesis Advisor

Prof. Dr. Mete Tayanç . . . . . .

Prof. Dr. Orhan Yenigün . . . . . .

DATE OF APPROVAL: 03/12/2019

ACKNOWLEDGEMENTS

Firstly, I would like to express my sincere thanks to my advisor Prof. Dr. Nadim Copty for his continuous support of my study and research, for his patience, motivation, and tremendous knowledge. His guidance helped me in all the time of my research and writing of this thesis. Without him, this thesis would never be completed.

I would also like to express my gratitude to my jury members Prof. Dr. Orhan Yenigün and Prof. Dr. Mete Tayanç for their valuable comments and suggestions.

I would like to thank my friends from the Institute of Environmental Sciences Oğuz Sarıgül and F. Öykü Sefiloğlu for their friendship and support throughout this process.

Finally, I would like to express my very profound gratitude to my family and my wife Özgücan Eken Türkoğlu for providing me with constant support and encouragement in the process of writing this thesis.

ABSTRACT

PREDICTIVE ANALYSIS OF FLIGHT TRAFFIC AT NEW YORK JFK AIRPORT ON AIR POLLUTION USING MACHINE LEARNING

Anthropogenic activities like transportation result in the emission of numerous pollutants like NO_2, CO, SO_2, $PM_{2.5-10}$, O_3, and Pb, which are identified as major air pollutants by environmental and health agencies. Emissions of those major pollutants result in negative health impacts like cardiovascular and respiratory diseases. Even though emissions from on-road traffic have decreased in recent years due to stricter regulations and technological advancements, less strict regulations on aircraft have resulted in an increase in emissions with the increasing air traffic. This study aims to estimate NO_2 emissions from commercial flights at John F. Kennedy Airport (JFK) in New York and their impacts on air quality. The study combines numerical modeling using the AERMOD air dispersion model along with a machine learning model to predict NO_2 concentration distributions as a function of space and time. To achieve this goal, departure and arrival flight data of John F. Kennedy Airport (JFK) in New York for the year 2018 is used. After the data is cleaned and prepared for the analysis, AERMOD is used to simulate atmospheric pollutant dispersion. The results of this study indicate that aircraft emissions can lead to significant NO_2 concentrations in the vicinity of the airport. The simulated concentrations are then used in the training of a machine learning model. Decision tree-based extreme gradient boosting (XGBoost) is used as a machine learning model. It is shown that training in the emission prediction model has resulted in a well-generalized and well-performing model. Overall, this study demonstrates that machine learning modeling can be an effective tool for estimating pollutant dispersion.

ÖZET

NEW YORK JFK HAVALİMANINDAKİ UÇUŞLARIN HAVA KİRLİLİĞİNE ETKİSİNİ MAKİNE ÖĞRENMESİ YARDIMIYLA TAHMİN ANALİZİ

Ulaşım gibi antropojenik aktiviteler, çevre ve sağlık kuruluşları tarafından başlıca hava kirletici maddeler olarak tanımlanan, NO_2, CO, SO_2, $PM_{2.5-10}$, O_3 ve Pb gibi kimyasalların emisyonuna yol açmaktadır. Bu kirleticilerin emisyonları, kardiyovasküler ve solunum yolu hastalıkları gibi olumsuz sağlık etkilerine sebep olmaktadır. Karayolu taşıtlarındaki standartlar ve yönetmelikler yardımıyla yıllar boyunca emisyonların azaltılmasına rağmen, uçaklarla ilgili daha az katı düzenlemeler, artan hava trafiğiyle emisyonların artmasına neden olmaktadır. Bu çalışma, New York'taki John F. Kennedy (JFK) Havalimanı'nda gerçekleşen ticari uçuşların neden olduğu NO_2 emisyonlarını tahmin etmeyi amaçlamaktadır ve belirli bir lokasyonda NO_2 emisyonlarını tahmin etmek amacıyla AERMOD hava dispersiyon modeli ile bir makine öğrenme modelini kullanarak sayısal modellemeyi birleştirmektedir. Bu çalışmayı gerçekleştirmek için, 2018 yılında New York'taki John F. Kennedy (JFK) Havalimanı'ndaki uçakların kalkış ve varış uçuş verileri kullanılmıştır. Veriler temizlendikten ve analiz için hazırlandıktan sonra, AERMOD, kirletici dağılımını simüle etmek ve bir makine öğrenim modelinin eğitiminde kullanılmak üzere emisyon dağılım verileri üretmek için kullanılmıştır. Karar ağacı esaslı ileri seviye gradyan artırma (XGBoost), bu çalışmada makine öğrenme modeli olarak kullanılmıştır. Bu çalışmanın sonucunda, emisyon tahmini modelinin eğitimi iyi genelleştirilmiş ve iyi performans gösteren bir modelle sonuçlanmıştır. Genel olarak, bu çalışma makine öğrenim modelinin kirletici dağılımını tahmin etmede etkili bir araç olabileceğini göstermektedir.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF SYMBOLS/ABBREVIATIONS

Symbol	Explanation
NO_2	Nitrogen Dioxide
SO_2	Sulfur Dioxide
CO	Carbon Monoxide
$PM_{2.5}$	Particulate Matter 2.5µm
PM_{10}	Particulate Matter 10µm
O_3	Ozone
Pb	Lead
NOx	Oxides of Nitrogen

Abbreviation	Explanation
µm	Micrometer
WHO	World Health Organization
EPA	Environmental Protection Agency
EEA	European Environmental Agency
VOC	Volatile Organic Compounds
LTO	Landing and Take-off
ICAO	International Civil Aviation Organization
JFK	John F. Kennedy Airport
AI	Artificial Intelligence
AQI	Air Quality Index
EI	Emission Index
XGBoost	Extreme Gradient Boosting

1. INTRODUCTION

Air pollution can be defined as, the release of chemicals, which are both toxic to human health and to the environment, into the atmosphere as a result of natural and anthropogenic activities. Volcanic eruptions and biological activities are examples of natural sources of air pollutions, while fossil fuel combustion and industrial processes are examples of anthropogenic activities. It is possible to partition anthropogenic sources into stationary or point and areal sources, and mobile or linear sources. Stacks, flues, and chimneys are among point sources, whereas clusters of typically small sources dispersed across the area like industrial complexes such as steelworks or emissions from urban household heating and landfills are areal sources. Vehicles like cars, trains, ships, and airplanes form mobile sources. Routes of vehicles are the main linear sources.

Environmental agencies and health organizations around the world like Environmental Protection Agency (EPA), European Environmental Agency (EEA), and World Health Organization (WHO) state the followings as major air pollutants,

- Nitrogen dioxide (NO_2),
- Sulfur dioxide (SO_2),
- Carbon monoxide (CO),
- Particulate matter (mainly $PM_{2.5}$ and PM_{10}),
- Ground-level ozone (O_3),
- Lead (Pb)

Pollutants can also be classified into two categories based on their release into the atmosphere. The first is *primary pollutants*, which are directly emitted to the atmosphere, such as SO_2 from industrial plants and power stations, and CO as a result of incomplete combustions. The other category is *secondary pollutants* which are formed in the atmosphere as a result of chemical reactions between atmospheric chemicals and pollutants, for instance, O_3 produced by photochemical reactions in the atmosphere (Tiwary et al., 2018).

Once released into the atmosphere, discharged air pollutants are continuously distributed and diluted. Meteorological conditions such as atmospheric stability, turbulence and the direction and magnitude of wind speed affect the distribution and dilution of the air pollutants particularly. Since they are affected by shortwave radiation, air temperature, and humidity, chemical reactions are also dependent on the ambient weather conditions (Mayer, 1999).

Urban air pollution started to gain importance after the 1952 London Smog when sulfurous smoky fog released from coal-burning stoves and local factories settled in London and caused the death of over 4000 people (Wilkins, 1954). With the establishment of the Clean Air Act in 1956 in the US, smoke and SO_2 concentrations have decreased over the years as a result of the use of clean fuels such as natural gas, the decline of some heavy industry, the use of more environmentally friendly process designs in power plants, and the relocation of power plants outside of cities (Giussani, 1994). Even though gasoline and natural gas are considered as a cleaner alternative compared to other fossil fuels such as coal and lignite, emissions from various sectors such as transportation have continued to increase with time due to the significant increase in land, air and marine transport. Nitrogen dioxide and particulate matter (PM) are the primary pollutants of most concern that are emitted into the atmosphere within exhaust gases from the fuel. Photochemical reactions in the atmosphere also result in an increase in ground-level ozone concentrations (Brimblecombe, 2006).

Inhalation and ingestion are the main routes of exposure to air pollutants for humans. Dermal contact has relatively less importance. Air pollution leads to a range of food and water contamination, making ingestion the main route of the intake of contaminants in several instances (Thron, 1996). Air pollution is a major environmental risk to health by having negative impacts on cardiovascular and respiratory health. In recent studies, it is found that a 10 $\mu g/m^3$ increase in NO_2 concentration causes a 0.41% increase in cardiovascular disease mortality, 0.34% increase in respiratory disease mortality, and a 0.30% increase in total mortality (T. M. Chen et al., 2007). Moreover, lung functions have been negatively associated with PM_{10} (particulate matter small than 10 μm), nitrogen dioxide, and sulfur dioxide in a series of studies from different communities around the world, with various symptoms such as bronchitis (Brunekreef et al., 2002). Acute human NO_2 exposure at concentrations above 150 ppm (282 mg/m^3) can cause death, either quickly due to pulmonary edema or after a few weeks due to severe fibrosis bronchiolitis obliterans (Last et al., 1994). Another study shows that exposure to NO_2 increased the risk of dementia indirectly as cardiovascular diseases are linked with cognitive decline and dementia risk (Paul et al., 2019).

Since the establishment of the Clean Air Act in the US, important developments in relation to air pollution have been taking place. For example, in 1974, Global Environment Monitoring System (GEMS) was founded by the United Nations Environment Program (UNEP) and WHO, to monitor air quality in over 50 cities in 35 countries worldwide (Mage et al., 1996). Throughout the years, with the mission of reducing the concentrations of pollutants, national or regional environmental agencies had been established, guidelines had been published, regulations and standards had been formed in different countries according to their diverse characteristics (Baldasano et al., 2003). On the other

hand, developing countries with high concentrations of air pollutants, show a tendency to increase in concentrations while they continue to develop.

Following the Clean Air Act, the EPA developed the National Ambient Air Quality Standard (NAAQS) to indicate standards for major air pollutants and formed the Air Quality Index (AQI) to inform the public with the air quality data, both standard and index are used today in the USA. Table 1.1 presents standards for various chemicals corresponding to different exposure periods. Table 1.2 shows air quality index levels and their color codes as well as the meaning of every level.

Table 1.1. National Ambient Air Quality Standard (NAAQS).

Pollutant	**Standard Type**	**Standards**	**Averaging Times**
Carbon Monoxide (CO)	primary	9 ppm	8 hours
		35 ppm	1 hour
Lead (Pb)	primary and secondary	0.15 μg/m^3	Quarterly average
Nitrogen Dioxide (NO_2)	primary	100 ppb (188 μg/m^3)	1 hour
	primary and secondary	53 ppb (100 μg/m^3)	Annual Mean
Particulate matter (PM_{10})	primary and secondary	150 μg/m^3	24 hours
Particulate matter ($PM_{2.5}$)	primary	12 μg/m^3	Annual Mean
	secondary	15 μg/m^3	Annual Mean
	primary and secondary	35 μg/m^3	24 hours
Ozone (O_3)	primary and secondary	0.07 ppm	8 hours
Sulfur Dioxide (SO_2)	primary	75 ppb	1 hour
	secondary	0.5 ppm	3 hours

In Turkey revised air quality regulations were announced in 2008. These regulations call for the gradual decrease of major pollutant standards starting in 2014 until they are in line with the EU standards. The Turkish, EU and World Health Organization (WHO, 2006) standards for key emissions related to fossil fuel burning: SO_2, NO_2, $PM_{10,}$ and CO are presented in Table 1.3. The standards are defined for different exposure times ranging from 1 hour to annual for the different pollutants. The Turkish Regulations allow for some exceedances of the standards as indicated in Table 1.3. Moreover, the Turkish Air Quality regulations allow some time lag before the new standards come in effect. The lag time varies with parameter and exposure time.

Table 1.2. Air Quality Index (AQI).

Category	AQI Value and Color	Meaning
Good	0 to 50 (Green)	Air quality is considered satisfactory, and air pollution poses little or no risk.
Moderate	51 to 100 (Yellow)	Air quality is acceptable; however, for some pollutants, there may be a moderate health concern for a very small number of people who are unusually sensitive to air pollution.
Unhealthy for Sensitive Groups	101 to 150 (Orange)	Members of sensitive groups may experience health effects. The general public is not likely to be affected.
Unhealthy	151 to 200 (Red)	Everyone may begin to experience health effects; members of sensitive groups may experience more serious health effects.
Very Unhealthy	201 to 300 (Purple)	Health alert: everyone may experience more serious health effects.
Hazardous	301 to 500 (Maroon)	Health warnings of emergency conditions. The entire population is more likely to be affected.

Table 1.3. Turkish, EU and WHO regulations for SO_2, NO_2 and PM_{10}.

		Concentration (µg/m³)		
Pollutant	***Exposure Time***	***Turkish Regulation[5]***	***EU Regulation 2008/50/EC***	***WHO (2006)***
Sulfur Dioxide (SO_2)	1-hour[1]	350 by 01-2019 410 in 2017	350	
	24-hour[2]	125	125	20
	Annual	20		
Nitrogen Dioxide (NO_2)	1-hour[3]	200 by 01-2024 270 in 2017	200	200
	Annual	40	40	40

Table 1.3. Turkish, EU and WHO regulations for SO_2, NO_2 and PM_{10} (continued).

		Concentration (μg/m³)		
Pollutant	***Exposure Time***	***Turkish Regulation[5]***	***EU Regulation 2008/50/EC***	***WHO (2006)***
Particulate Matter (PM_{10})	24-hour[4]	50 by 01-2019 70 in 2017	50	50
Carbon Monoxide (CO)	Annual	40	40	20

[1]can be exceeded up to 24 times per year
[2] can be exceeded up to 3 times per year
[3]can be exceeded up to 18 times per year
[4] can be exceeded up to 35 times per year
[5] from the 2008 Turkish air quality regulations.

The major sources of outdoor air pollution in cities include transportation like road vehicles, aviation, and naval traffic. In the USA, besides PM production, it is estimated that mobile sources contribute up to 45% of NO_2 emissions, 2% of SO_2 emissions, 81% of CO emissions, and 37% of Volatile Organic Compounds (VOC) (Peden, 2008). Pollution resulting from transportation also causes respiratory diseases. In a study, 1759 children from 12 cities in Southern California, were observed from age 10 until they are 18. At the end of the study, it was reported that pollutants from vehicular fuel use like NO_2 and $PM_{2.5}$ are correlated to a decrease in lung functions (Gauderman et al., 2004).

Over the years, with enhanced regulations and improved technology, pollution levels of six common pollutants have decreased. However, regulations on airplanes and ships are not as strict as road vehicle regulations (Harrison et al., 2015). Aviation fuels and automotive fuels have similar volatility range; however, the sulfur content of automotive fuels has decreased by regulations to lower than 10 ppm whereas, the limit for aviation fuel has stayed at 3000 ppm and the real concentrations are reported to be in the range of 300 to 1100 ppm (Harrison, 2015). Therefore, an increase in air traffic remains a major problem in terms of air pollution and its health effects. The increase in air traffic and passenger counts in both worldwide and the USA is shown in Figure 1.1. It is clear that air traffic has increased rapidly worldwide and will continue to increase in the future.

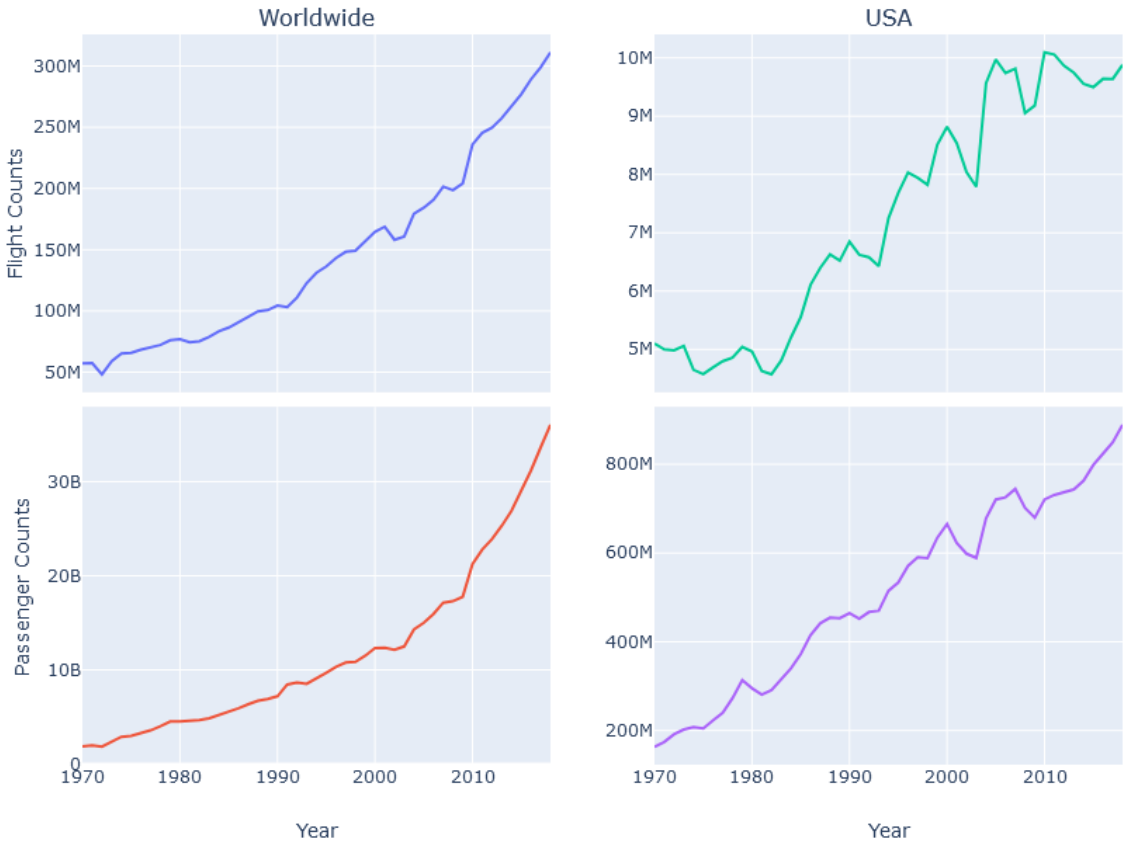

Figure 1.1. Registered carrier departure counts and passenger counts worldwide and in the USA since 1970 (data retrieved from International Civil Aviation Organization (ICAO)).

Ideal combustion in aircraft engines is not possible; hence, it results in unwanted chemical emissions like CO, SO_2, PM (soot), VOC and oxides of nitrogen (NO_x) as depicted in Figure 1.2. NO_x is a collective term used to refer to nitrogen monoxide (nitric oxide or NO) and NO_2, which are formed as a result of the oxidization of atmospheric nitrogen (N_2) during combustion. Improved combustion techniques can reduce the emissions of NO_x, CO, and VOC. However, over the past decades, there is an increase in the emission of NO_x as a result of burning kerosene at high temperatures, which is the major trace gas emission during a flight (Ruijgrok et al., 2005).

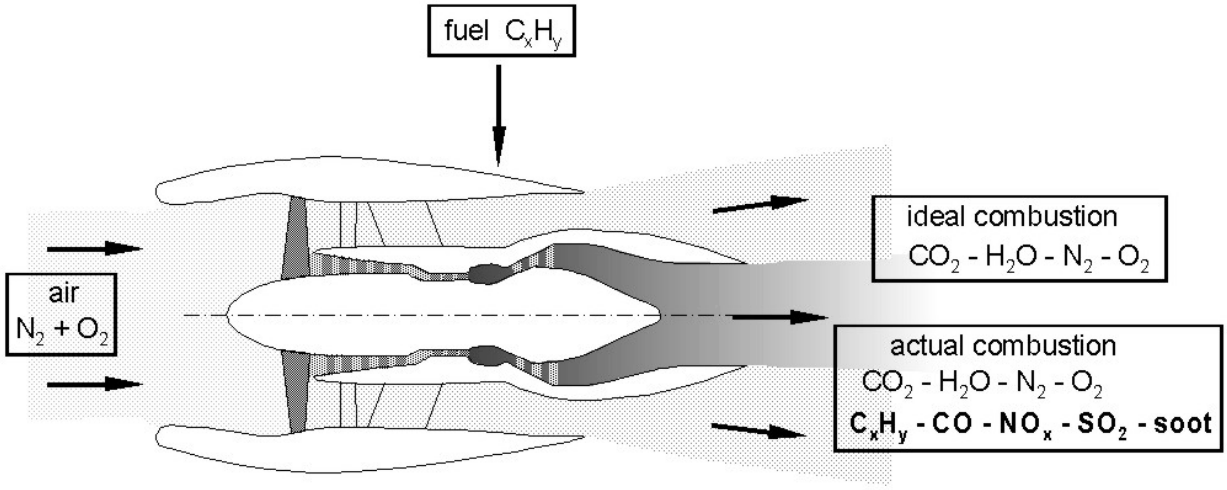

Figure 1.2. The combustion process in an aircraft engine (Ruijgrok, 2005).

There has been an unambiguity on the effect of low-altitude aircraft emissions on air quality. Aircraft emissions show changes with the different engine thrust settings during the landing and take-off (LTO) phase. While CO and unburnt hydrocarbons (HC) have predominant emissions in lower thrust settings, NO_x and $PM_{2.5}$ emissions increase with the thrust settings. Table 1.4 shows fuel consumption and emission factors at different thrust settings for Airbus A306 aircraft.

Table 1.4. Fuel consumption and emission factors of Airbus A306 type aircraft for different engine thrust settings (retrieved from EEA, 2019).

	Engine thrust setting (% of maximum thrust)			
	Taxi in/out	**Approach**	**Climb out**	**Take off**
	7	30	85	100
Rate of fuel burn (kg/s/engine)	0.2110	0.6820	2.004	2.481
Rate of emission of CO (kg/s/engine)	0.004429	0.001282	0.001082	0.0009924
Rate of emission of HC (kg/s/engine)	0.0003756	0.00009548	0.00004008	0.0002233
Rate of emission of NO_X (kg/s/engine)	0.001013	0.008048	0.04749	0.07493
Rate of emission of $PM_{2.5}$ (kg/s/engine)	0.000016	0.000048	0.000166	0.000229

As mentioned before, NO_x describes the sum of NO and NO_2. The ratio of NO/NO_2 in NO_x changes with different thrust settings; NO_x is dominated by NO at higher thrust levels, at low powers, more than 80% of the total NO_x can consist of NO_2 (Wormhoudt et al., 2007).It is estimated that the

toxicity of NO_2 based on pulmonary reactions as a result of acute exposures is about thirty times that of NO (Last, 1994). Moreover, NO_2 is a reactive chemical in the air; it causes O_3 formation in the troposphere by reacting with hydrocarbons in the presence of sunlight. It also causes acid rain by reacting with O_2 in the air.

The aim of this study is to estimate NO_2 emissions caused by commercial flights John F. Kennedy Airport (JFK) in New York. JFK Airport is selected because it is one of the most crowded airports in the world and because of the existence of abundance of data to estimate NO_2 emissions and to perform both air dispersion model and machine learning model. Specifically, the study will combine numerical modeling using the AERMOD air dispersion model along with a machine learning model to predict NO_2 spatial and temporal distributions in the vicinity of the JFK airport. To conduct this study, departure and arrival flight data of John F. Kennedy Airport (JFK) in New York for the year 2018 are used. After the data is cleaned and prepared for the analysis, AERMOD was used to simulate pollutant dispersion and to generate emission dispersion data for use in the training of a machine learning model. The performance of XGBoost, a decision tree based gradient boosting model, to reproduce the complex air pollution data is evaluated.

2. LITERATURE REVIEW

2.1. Study Area

New York, John F. Kennedy Airport, which is the subject of this study, is the fifth busiest airport in the U.S. and twenty-first in the world with 455,542 flights and 61.6 million passengers in the year 2018. As New York City's main airport, JFK sees arrivals and departures from nearly every international airline in the world. It is located in the southeast part of New York City as shown in Figure 2.1, which is the most populated city in the U.S. with 8.6 million people.

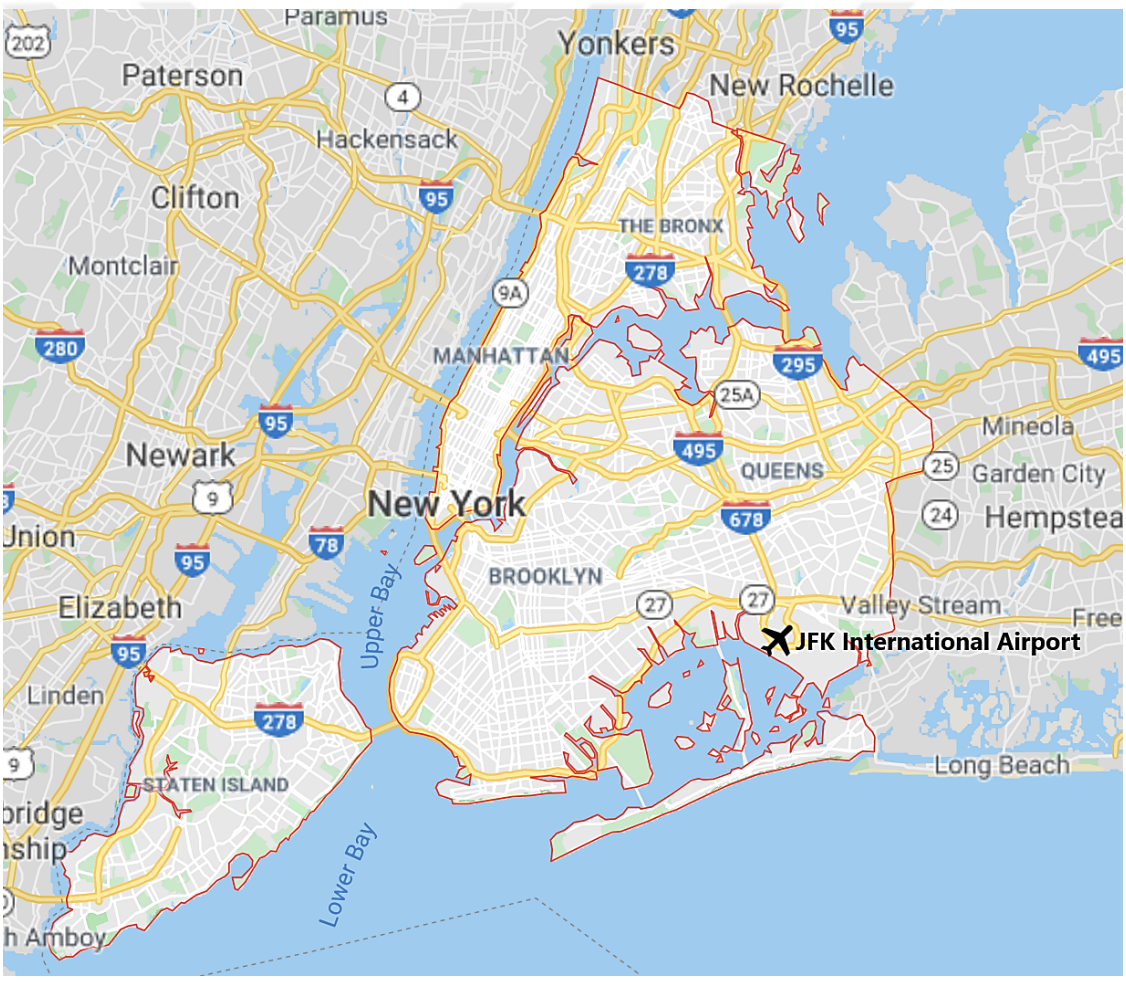

Figure 2.1. JFK International Airport, New York City on the map.

JFK has six operational terminals numbered 1-8, but Terminal 3 and Terminal 6 were demolished after Terminal 5 was expanded in 2011 and 2013. Also, the JFK runway system consists of two pairs of parallel runways aligned with right angles: 4L-22R, 4R-22L, 13L-31R and 13R-31L. The total

runway length is nearly fifteen kilometers. The terminals and runways of JFK Airport are shown in Figure 2.2.

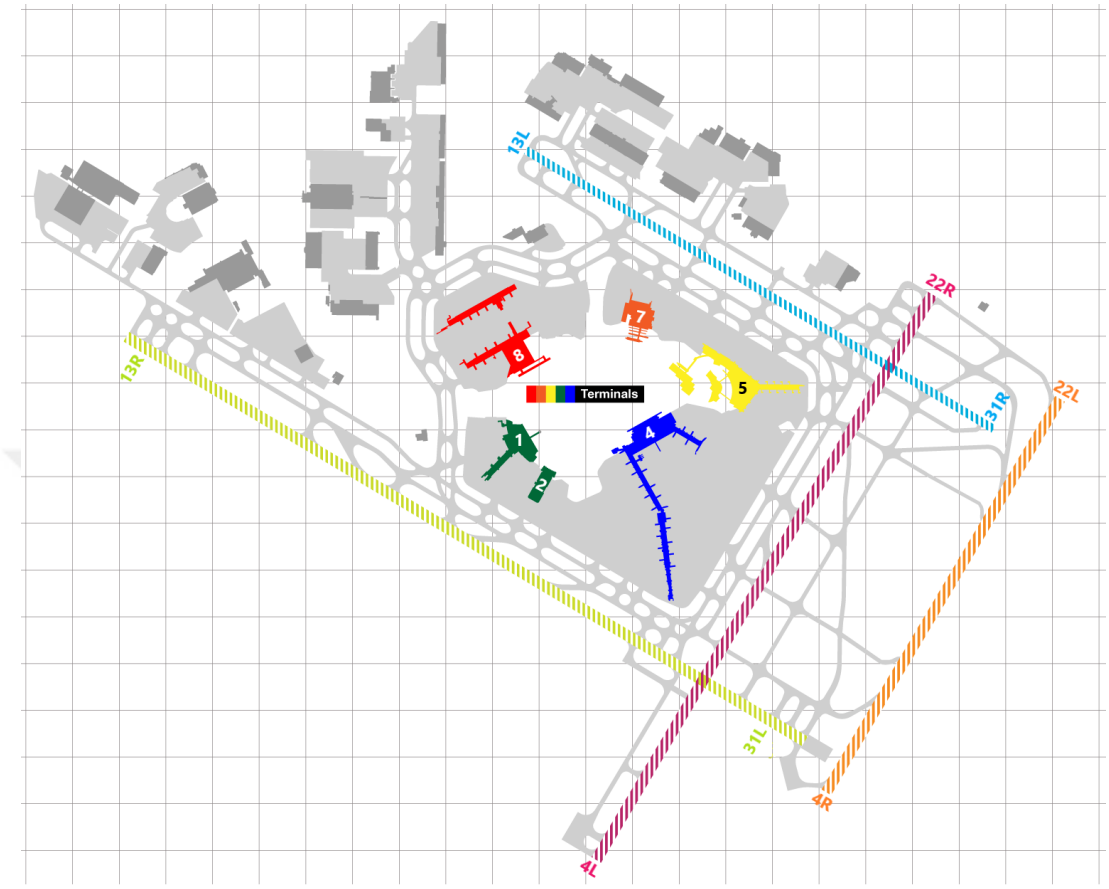

Figure 2.2. Terminals and runways of JFK Airport, New York City.

Over the years, air quality has improved in New York, but within a year there are still days of unhealthy status in the AQI index. Daily AQI values in 2018 for both total and NO_2 in New York City can be seen in Figure 2.3 and Figure 2.4.

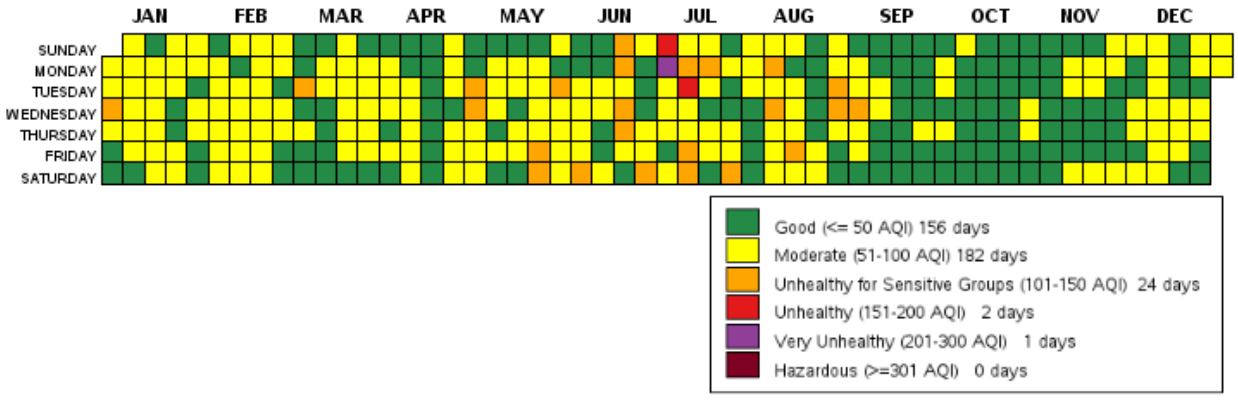

Figure 2.3. Daily AQI values of New York City in 2018 for all pollutants (retrieved from U.S. EPA AirData).

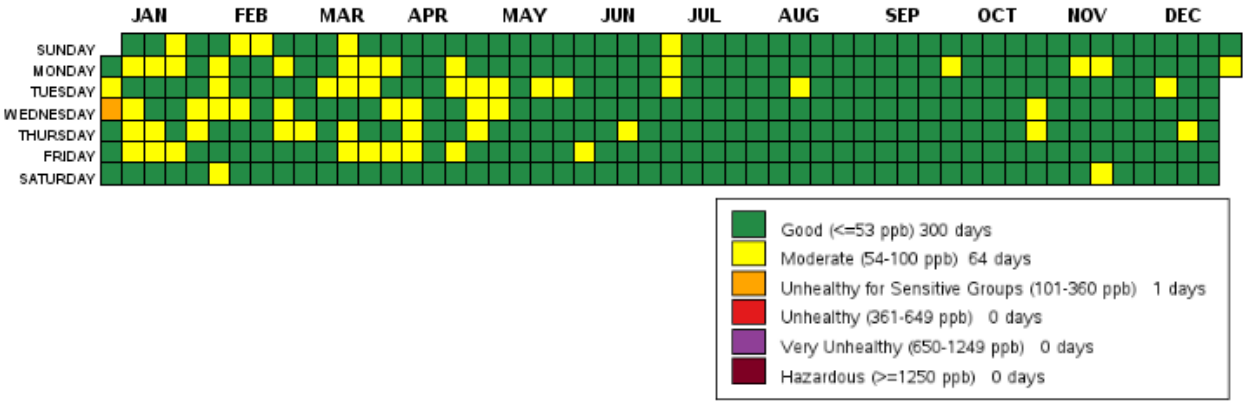

Figure 2.4. Daily AQI values of New York City in 2018 for NO_2 (retrieved from U.S. EPA AirData).

2.2. Aircraft Emissions and Landing/Take-off (LTO) Cycle

Air pollutants emitted from aircraft primarily originate from the burning of jet fuel and aviation gasoline used for aircraft fuel. The main pollutants emitted from aircraft are:

- CO_2
- NO_x
- H_2O vapor
- Methane (CH_4)
- CO
- Sulfur oxides (SO_x)
- VOCs
- PMs

Figure 2.5 shows the combustion process in aircraft, what it results, and also shows the composition of input and output gas.

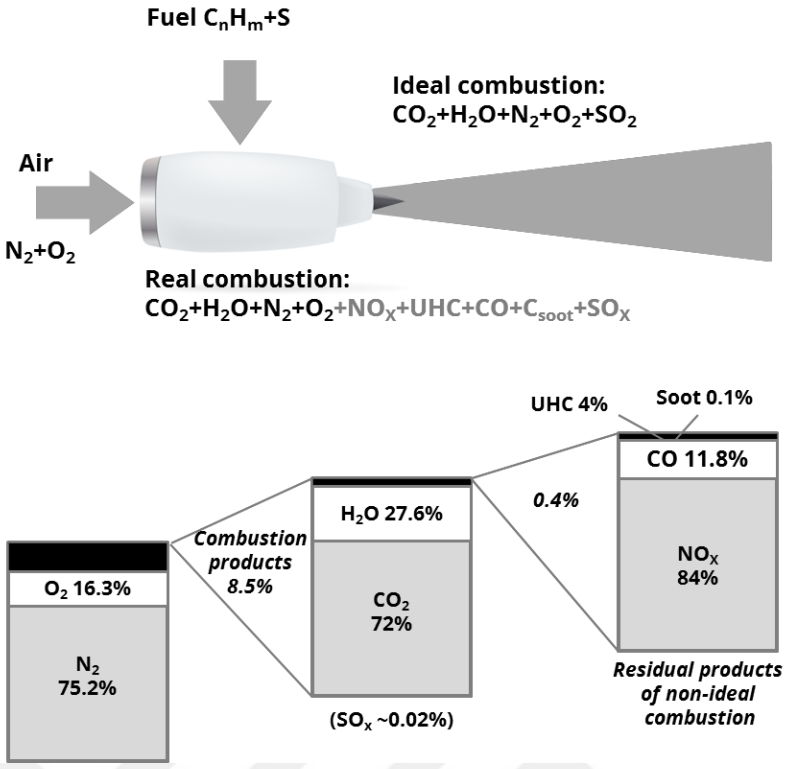

Figure 2.5. Aircraft engine combustion process and composition of its input and output gas (retrieved from EEA, 2019).

Aircraft emissions can change with respect to engine model, each engine has its own emission factors set by environmental agencies like EPA and EEA. Those emission factors also change with the different parts of the aircraft movement phases. LTO is the phase when an aircraft is moving near to the airport under 3000 ft (914 m). On the other hand, when an aircraft is moving above 3000 ft, it is called the Cruise phase as illustrated in Figure 2.6.

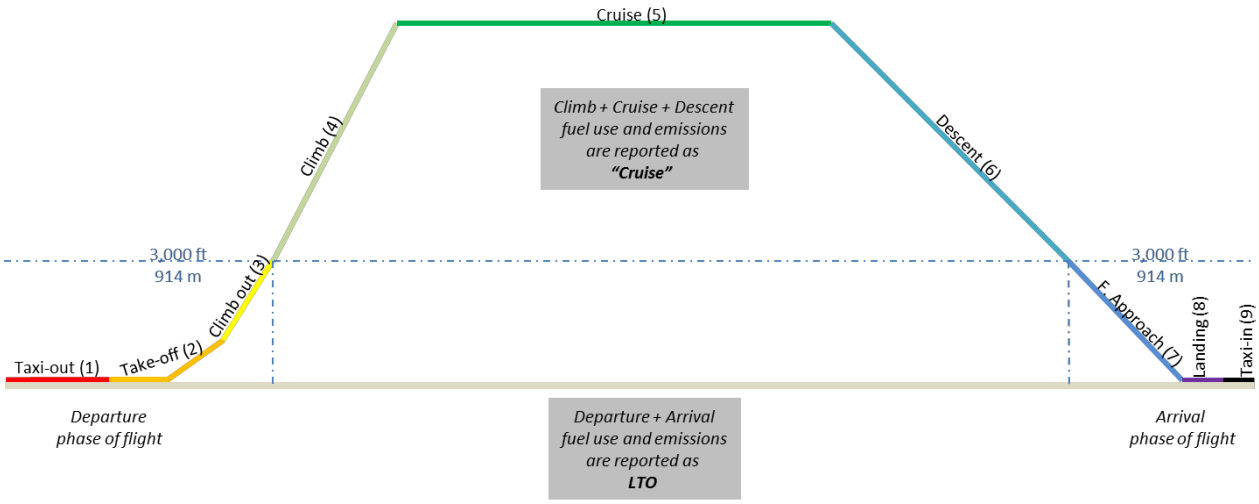

Figure 2.6. Movement phases of an aircraft (from EEA, 2019).

Duration of take-off, approach and climb out phases have standard durations set by ICAO, however, taxi-in and taxi-out times differ by airports. In Table 2.1, duration of take-off, approach and climb out, set by ICAO, and taxi-in, taxi-out set by FAA for JFK Airport is given.

Table 2.1 Duration of LTO cycle phases (from FAA, 2018; ICAO, 2019).

Phase	*Duration (minutes)*
Taxi-out	27.5
Take-off	0.7
Climb out	2.2
Approach	4
Taxi-in	9

Emissions of PM and NO_x in some Airbus and Boeing airframes are observed to be higher in landing and take-off phases compared to the idle phase (Mazaheri et al., 2009). In the world's second busiest airport, Beijing Capital International Airport (ZBAA), aircraft are the main source of emissions around the airport. The emission rates may vary with the seasons, but on average, NO_x, CO, $PM_{2.5}$, SO_2 are accounted for 86.3%, 78.7%, 48.0%, and 95.6% of total emissions in ZBAA, respectively (Yang et al., 2018). Moreover, a study held in London Heathrow Airport shows that NO_x emissions caused by aircraft in the airport area can be easily detected from 2.6 km away. While airport accounts for 27% of NO_x emissions around the airport area, emissions are diluted by a factor of 5 in 1-1.5 km distance from the airport, and it accounts for 12-14% of NO_x emissions at that area (Carslaw et al., 2006).

2.3. Atmospheric Dispersion Modelling of Aircraft Emissions and Predictive Analysis

After entering the environment, air pollutants can spread through the air, water, soil, living organisms, and food. The distribution paths differ considerably depending on both the source of emissions and the relevant pollutants. Distribution rates and patterns are also largely dependent on environmental conditions. Dispersion of the air pollutants is affected by several factors like meteorology, the height of emission, geographical features and source type (Peden, 2008).

Pollutants are subjected to a broad variety of transformations and transportations during dispersion. Dilution takes place due to incorporation with air. Depending on the physical properties of the pollutant, deposition, separation or accumulation of pollutants take place. Chemical reactions can lead to the fragmentation of pollutants or turn them into new compounds. Besides, some contaminants can be eliminated from the transport mechanism through sedimentation, for instance, by gravity, precipitation or retention by plants and other obstacles (Arya, 1999).

In environments like cities and towns, many pollutants show very complicated forms of distribution, because these environments have a variety of emission sources and a wide range of environmental conditions. This complex spatial pattern signifies that it is usually challenging to model or determine pollution patterns and trends, and so to estimate the grade of human exposure (Mayer, 1999).

Over the years, complex mathematical equations were developed to analyze and understand the transfer of emissions by taking the above-mentioned conditions into account. Mathematical simulation of the dispersion of air pollutants in the atmosphere with those equations is called atmospheric dispersion modeling. It is achieved with computer programs, which include algorithms to solve the numerical equations that describe the dispersion of pollutants. The dispersion models aim to calculate the atmospheric downwind concentration of air pollutants and chemicals from sources such as industrial plants, vehicles or unintentional chemical exposures. Predicting future concentrations like emission change in the source is another way to use these models in certain scenarios. Capabilities of these models allow them to be used when air quality policies are formed (Lin et al., 2009).

The USEPA recommends the use of the steady-state air dispersion model AERMOD and the Lagrangian puff model CALPUFF to demonstrate near field (<50 km) and far-field (>50 km) regulatory compliances, respectively (Rood, 2014). The AERMOD model was first developed in 1991 by the American Meteorological Society and by EPA, with the intention of integrating established concepts of planetary boundary layer into regulatory models. The model addresses the analysis of both surface and elevated sources like point sources, area sources, and volume sources in simple and complex terrain domains (Cimorelli et al., 1998). The modeling system CALPUFF is a non-steady-state Lagrangian puff model that simulates transport, transformation, and deposition of pollutants in a spatially and temporally variable wind field of three dimensions. Both local and regional scales can be applied in this model (Scire, Strimaitis, et al., 2000).

Both models consist of different modules. For AERMOD there are AERMET, a meteorological pre-processor, and AERMAP, a terrain pre-processor. AERMAP produces base elevation and hill height data for receptors and sources by examining the surrounding terrain. AERMET processes surface, upper-air and on-site meteorological data and it creates new merged data which is used as input in AERMOD (EPA, 2018, 2019a). On the other hand, CALPUFF includes CALMET, a meteorological model that creates a three-dimensional hourly wind field in a grided modeling domain.

The output of CALMET serves as an input file of the dispersion model CALPUFF (Scire, Robe, et al., 2000). CALPUFF includes various features such as dry deposition, wet deposition, and chemical transformations that can be turned on or off depending on the problem being simulated. The third module is CALPOST which reads the concentration and deposition files of CALPUFF and creates a time-averaged concentration and deposition output together with visibility impacts (Scire, 2000).

2.4. Machine Learning Models

Machine learning is a subfield of artificial intelligence (AI) that typically tries to understand the structure of the data and fit the data into models that people can understand and use. Figure 2.7 shows the sub-segments of the field of AI. Even though machine learning resides under computer science, it is different from conventional computational approaches. Algorithms are types of explicitly programmed instructions that computers use to calculate or solve problems in traditional computing. Instead, machine learning algorithms enable computers to train data inputs and use statistical analyzes to generate values within a certain range. Machine learning thus enables computers to construct models from sample data to ensure decision-making processes are automated based on data input (Tom Michael Mitchell, 2006).

More formal definition of machine learning made by Tom M. Mitchell is “A computer program is said to learn from experience *E* with respect to some class of tasks *T* and performance measure *P* if its performance at tasks in *T*, as measured by *P*, improves with experience *E*.” (Tom M Mitchell, 1997). Machine learning can be used to address a problem in broad categories, and over time, advanced methodologies like brain-inspired neural networks have been coming up with ongoing researches around the world (Sze et al., 2017).

Today, every technology user benefits from machine learning. Examples of machine learning are everywhere in our daily lives. For example, speech recognition is relied on to turn speech into text, make life easier with virtual assistants, and help to improve speaking ability in foreign languages (Hu et al., 2011). Social media platforms tag people in images to help users share photos of friends with facial recognition (Yamaguchi, 2012). Machine learning powered recommendation engines suggest what films or TV shows are to be watched next, based on user expectations (Hallinan et al., 2016). The consumer may soon have access to self-driving cars that rely on machine learning (Bojarski et al., 2016).

Forecasting is one of the important applications of machine learning. Machine learning has been used to predict air pollution. For instance, a study used neural networks with air quality data from 4 different stations to forecast maximum 1-hr average O_3 concentrations for the next day in Istanbul (Inal, 2010). The decision tree-like model random forest is shown to be more accurate than Naïve Bayes, logistic regression and neural network when predicting Air Quality Index (AQI) for the city of Shenyang in China by using historical data (Yu et al., 2016). In another study, features from images of Shanghai, Beijing (China) and Phoenix (U.S.) are used in a machine learning model alongside meteorological data to predict pollution of $PM_{2.5}$ (Liu et al., 2016).

Dispersion models require a wide range of data including emission data, land use, topography, and meteorological data. On the contrary, machine learning models require input data and corresponding output data to learn and form a mathematical model which is called supervised learning. In some cases, only input data is sufficient to train the model and this is called unsupervised learning. (Russell et al., 2016). The breakdown of machine learning into three sub-field and examples for each sub-field is illustrated in Figure 2.8.

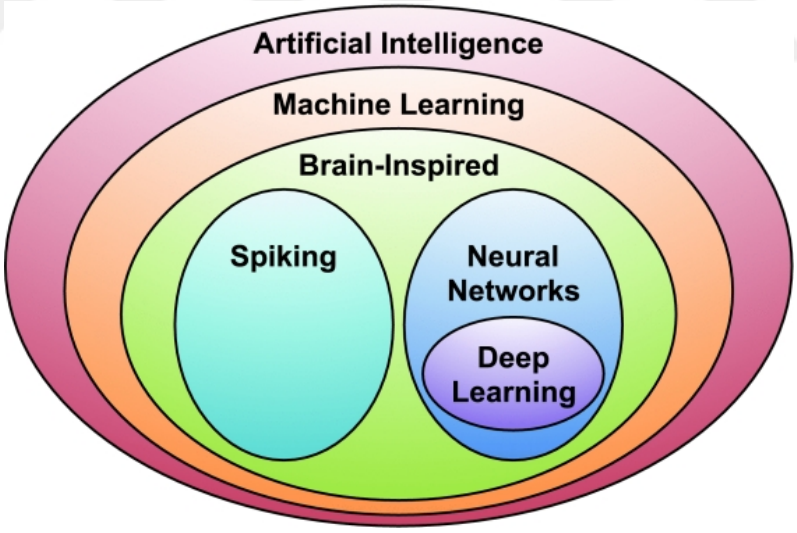

Figure 2.7. Artificial intelligence from general to the specific (retrieved from Sze et al., 2017).

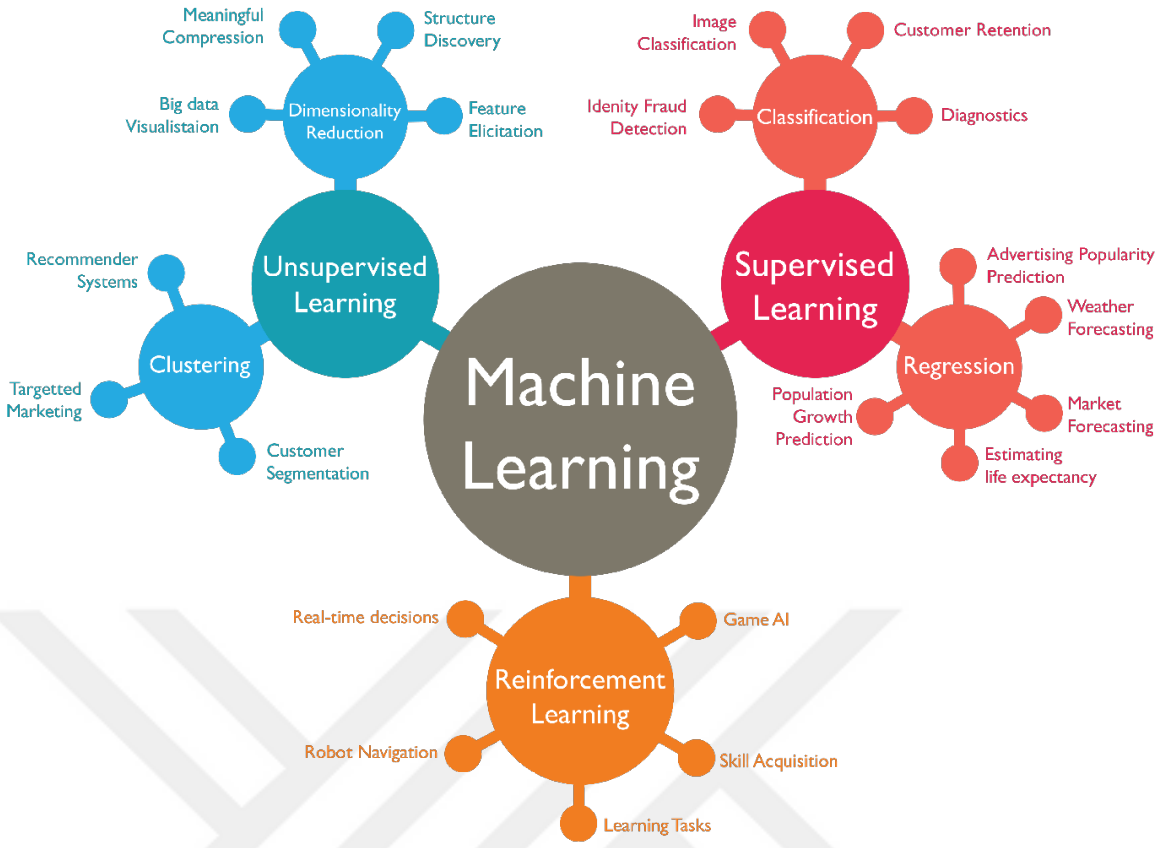

Figure 2.8. Methods of machine learning in general (retrieved from Chugh, 2018).

Machine learning models can be potentially used to predict dispersion from a source. If output training data is not available for model development, then dispersion models can be used as a ground truth. Coupled Gaussian dispersion model and neural network, which is a form of machine learning inspired by the brain, where concentrations from an atmospheric model used for training in the machine learning model, results nine times better than only the machine learning model with respect to negative mean square error (Pelliccioni et al., 2006). In a study, 1.5 years of daily data from 74 cities in China is used to calculate AQI values with WRF-Chem, a new generation regional air quality modeling system. The calculated AQI values are used as labels in machine learning models to create a classifier, and the generated model performs better than using WRF-Chem only (Xi et al., 2015). Operations data for Nanjing Lukou International Airport (ZSNJ) in 2017, the second largest airport of eastern China, is used to calculate AQI data with the Gaussian dispersion model. Calculated AQI data used as labels in training of supervised machine learning model to classify air quality and the random forest method provides the best result with around 90% accuracy (Tian et al., 2019).

In this study air dispersion modeling and machine learning will be combined to examine the problem of air pollution in the vicinity of the JFK airport in New York, one of the largest airports in North America and the world.

3. MATERIALS AND METHODS

3.1. Data Collection

Arrival and departure flight data at JFK in 2018 were gathered from a community-based receiver network called OpenSky to use in this study. Air traffic surveillance data is continuously collected and stored in a historical database to make it accessible to researchers by the OpenSky Network (Schäfer et al., 2014). The JFK dataset has around 120 million of signal data for approximately 450,000 unique flights. A sample of the collected raw data is given in Table 3.1. Thecolumns of the dataset are as follows:

- *icao24;* the 24-bit unique address of an aircraft assigned by ICAO which can be used to track that aircraft over different flights,
- *callsign:* generally indicates the airline and the flight number,
- *altitude:* measured by the barometer and depends on factors such as weather,
- *geoaltitude:* determined using the GPS sensor,
- *track:* the direction of movement as the clockwise angle from the geographic north,
- *groundspeed:* the speed over ground of the aircraft in meters per second,
- *vertrate:* the vertical speed of the aircraft in meters per second. A negative number indicates that the aircraft was descending, a positive number indicates a ascend respectively,
- *hour*,: marks the beginning of the hour to which the data belongs,
- *timestamp:* the timestamp in seconds for which data is received,
- *latitude*: aircraft coordinate
- *longitude*: aircraft coordinate
- *onground:* a flag that indicates whether the aircraft is broadcasting surface positions (true) or airborne positions (false).

The EEA, European Aviation Safety Agency (EASA) and the U.S. Federal Aviation Administration (FAA) established emission factors for calculating the emission concentrations emitted by aircraft (EASA, 2019; EEA, 2019; FAA, 2018a). These datasets include fuel consumption for each aircraft during the different LTO phases: take-off, climb out, approach and idle. The fuel consumption is in units of kg of fuel flow per second. In addition, the databases also have emission

factors expressed in g of emitted pollutants per kg fuel consumption for each engine type in each LTO phase. The focus in this study is on NO_x emissions which are a major pollutant related to the burning of fossil fuels. Sample data for fuel flow and NO_x emission index (EI) for an engine in each LTO phase is shown in Table 3.2.

Table 3.1. Preview of arrival and departure flight data on 2018 at JFK collected from OpenSky Network.

ICAO24	Callsign	Altitude	Geoaltitude	Track	Groundspeed	Vertical Rate	Hour	Timestamp	Latitude	Longitude	Onground
89901c	CAL012	-250	75	300.964	128.281	-768	1514764800	2018-01-01 00:03:59	40.64282227	-73.75726318	FALSE
400f0b	VIR9M	-325					1514764800	2018-01-01 00:03:59	40.64331572	-73.76926075	TRUE
a6c3ca	JBU2153	900	1050	286.323	145.880	1920	1514764800	2018-01-01 00:03:59	40.64396667	-73.8104248	FALSE
a6c3ca	JBU2153	950	1075	286.213	146.840	1792	1514764800	2018-01-01 00:04:00	40.64396667	-73.8104248	FALSE
400f0b	VIR9M	-325					1514764800	2018-01-01 00:04:00	40.64331572	-73.76926075	TRUE
89901c	CAL012	-275	75	300.964	128.281	-768	1514764800	2018-01-01 00:04:00	40.64282227	-73.75726318	FALSE
a6c3ca	JBU2153	1000	1125	285.627	148.489	1664	1514764800	2018-01-01 00:04:01	40.64443297	-73.81251942	FALSE
400f0b	VIR9M	-325					1514764800	2018-01-01 00:04:01	40.64331572	-73.76926075	TRUE
89901c	CAL012	-275	75	300.964	128.281	-768	1514764800	2018-01-01 00:04:01	40.64282227	-73.75726318	FALSE
89901c	CAL012	-300	75	300.964	128.281	-768	1514764800	2018-01-01 00:04:02	40.64282227	-73.75726318	FALSE
400f0b	VIR9M	-325					1514764800	2018-01-01 00:04:02	40.64331572	-73.76926075	TRUE
a6c3ca	JBU2153	1025	1150	285.054	150.153	1536	1514764800	2018-01-01 00:04:02	40.64460754	-73.81347656	FALSE
a83108	JBU2231	-325					1514764800	2018-01-01 00:04:03	40.64608383	-73.77200317	TRUE
400f0b	VIR9M	-325					1514764800	2018-01-01 00:04:03	40.64331572	-73.76926075	TRUE
a6c3ca	JBU2153	1050	1150	284.128	151.585	1408	1514764800	2018-01-01 00:04:03	40.64471229	-73.81395513	FALSE
89901c	CAL012	-300	-25	300.885	124.679	-576	1514764800	2018-01-01 00:04:03	40.64457263	-73.76114585	FALSE
a92e6e		-325					1514764800	2018-01-01 00:04:03	40.65096192	-73.76804352	TRUE
a92e6e		-325					1514764800	2018-01-01 00:04:04	40.65096192	-73.76804352	TRUE
a6c3ca	JBU2153	1050	1175	283.583	153.287	1344	1514764800	2018-01-01 00:04:04	40.64483643	-73.81481934	FALSE
400f0b	VIR9M	-325					1514764800	2018-01-01 00:04:04	40.64331572	-73.76926075	TRUE
a83108	JBU2231	-325					1514764800	2018-01-01 00:04:04	40.64608383	-73.77200317	TRUE
89901c	CAL012	-325	-25	300.885	124.679	-576	1514764800	2018-01-01 00:04:04	40.64457263	-73.76114585	FALSE
a83108	JBU2231	-325					1514764800	2018-01-01 00:04:05	40.64608383	-73.77200317	TRUE
400f0b	VIR9M	-325					1514764800	2018-01-01 00:04:05	40.64331572	-73.76926075	TRUE
89901c	CAL012	-325	-25	300.885	124.679	-576	1514764800	2018-01-01 00:04:05	40.64457263	-73.76114585	FALSE
a92e6e		-325					1514764800	2018-01-01 00:04:05	40.65096192	-73.76804352	TRUE

Table 3.2. Sample of fuel flow and NO_x emission index for aircraft engines in each LTO phase.

UID	Engine Model	NO_x EI Take-Off (g/kg)	NO_x EI Climb-Out (g/kg)	NO_x EI Approach (g/kg)	NO_x EI Idle (g/kg)	Fuel Flow Take-Off (kg/sec)	Fuel Flow Climb-Out (kg/sec)	Fuel Flow Approach (kg/sec)	Fuel Flow Idle (kg/sec)
6AL005	RR AE3007A1	19.66	16.63	7.1	3.47	0.3826	0.318	0.113	0.0461
6AL006	RR AE3007A1	22.41	19.58	7.2	3.19	0.3826	0.318	0.113	0.0461
3BR001	RR BR700-710A1-10	17.07	13.93	8.2	4	0.707	0.588	0.22	0.089
4BR008	RR BR700-710A1-10	18.79	15.07	7.68	4.69	0.713	0.594	0.214	0.089
1CM003	CFM CFM56-2-C5	18.5	16	8.2	4	0.985	0.819	0.311	0.128
1CM004	CFM CFM56-3-B1	17.7	15.5	8.3	3.9	0.946	0.792	0.29	0.114
20CM091	CFM LEAP-1A26CJ	30.8	13.38	8.75	4.61	0.861	0.71	0.244	0.091
20CM092	CFM LEAP-1A29	49.48	21.03	9.27	4.72	0.946	0.777	0.261	0.094
8GE113	GE CF34-8E2	13.6	11.82	10.29	4.45	0.591	0.485	0.168	0.062
8GE111	GE CF34-8E2A1	14.61	12.55	10.72	4.59	0.644	0.527	0.179	0.064
11GE134	GE GEnx-1B54	14.96	9.18	8.07	3.98	1.878	1.553	0.523	0.184
11GE135	GE GEnx-1B58	18.04	11.02	8.41	4.08	2.019	1.667	0.554	0.19
1IA004	IO V2528-D5	30.5	25.1	9.6	4.9	1.209	0.996	0.353	0.134
1IA005	IO V2530-A5	33.8	27.1	10.1	5	1.331	1.077	0.377	0.138
1PW011	PW JT8D-15A	18.1	13.9	6.6	3.1	1.115	0.8955	0.312	0.1372
1PW012	PW JT8D-17	19.2	15.23	6.1	3.3	1.245	0.997	0.354	0.147
3RR028	RR RB211-535E4	44.88	32.06	6.78	3.46	1.86	1.51	0.52	0.18
1RR016	RR SPEY Mk511	22.7	17.3	7.2	3.6	0.891	0.726	0.278	0.127
3RR032	RR TAY 651	17.56	13.77	5.42	2.52	0.87	0.72	0.26	0.12
2RR025	RR Trent 877	34.76	27.59	10.59	4.75	3.21	2.66	0.9	0.28
2RR026	RR Trent 884	40.05	30.63	11.07	5.04	3.56	2.89	0.97	0.31

These emission factor datasets use engine models, and more specifically unique engine identifiers (UID) to map factors with a single aircraft. In order to calculate NO_x emissions for flights, it is required to know how many engines and which engine model does an aircraft have. Therefore, datasets that contain ICAO24, aircraft model, engine model, engine count, and UID as it is shown in Table 3.3, were requested from two different aviation-related data platforms, AvDelphi and Planespotters.net (AvDelphi, n.d.; Planespotters.net, n.d.).

Table 3.3. Sample from the dataset for mapping ICAO24 to UID.

ICAO24	Aircraft Model	Engine Model	Engine Count	UID
3949ed	Boeing 777-200	GE GE90-94B	2	6GE091
424590	Airbus A320-200	CFMI CFM56-5B4/3	2	8CM055
3c6506	Airbus A340-600	RR Trent 556-61	4	6RR041
89901c	Boeing 777-300	GE GE90-115B	2	7GE099
407177	Boeing 737-800	CFMI CFM56-7B26E	2	11CM072
a5a98d	Boeing 747-400	GE CF6-80C2B1F	4	1GE023
4ca84c	Boeing 737-800	CFMI CFM56-7B26	2	3CM033
4cabb9	Boeing 737-800	CFMI CFM56-7B26E	2	11CM072
424117	Airbus A320-200	CFMI CFM56-5B4/3	2	8CM055
89639d	Airbus A380-800	GP7270	4	9EA001
4ba930	Airbus A330-200	PW PW4168A	2	4PW067
39840e	Airbus A320-200	CFMI CFM56-5B4/3	2	8CM055
3424ce	Boeing 737-800	CFMI CFM56-7B26	2	3CM033
4b8685	Airbus A321-200	CFMI CFM56-5B3/P	2	3CM025
ac25c4	Boeing 777-200	GE GE90-110B1	2	7GE097
151d41	Airbus A319-100	CFMI CFM56-5B7/3	2	8CM058
4690e3	Airbus A320-200	IAE V2527-A5	2	1IA003
a1d8fd	Boeing 777-200	PW PW4090	2	10PW099
750156	Airbus A319-100	CFMI CFM56-5B7/P	2	6CM044
76cdb0	Airbus A350-900	RR Trent XWB-84	2	14RR075
461fa3	Airbus A321-200	IAE V2533-A5	2	3IA008

Emission calculations are made by an approach that takes total activity in an LTO mode (e.g. taxi in, taxi out, approach, climb-out, and take-off) into account for all types of aircraft in the studied airport (ICAO, 2011). This approach is formulated as,

$$Ei_{jk} = TIM_{jk} \times 60 \times FF_{jk} \times Ei_{ijk} \times Ne_j \quad (1)$$

where,

Ei_{jk} total emissions of pollutant *i*, in grams, produced by aircraft type *j* for mode *k* (e.g. taxi in, taxi out, approach, climb-out, and take-off);

TIM_{jk} time-in-mode for mode *k*, in minutes, for aircraft type *j*;

FF_{jk} fuel flow for mode *k*, in kilograms per second (kg/s), for each engine used on aircraft type *j*;

Ei_{ijk} emission index for pollutant i in grams per pollutant per kilogram of fuel (g/kg of fuel), in mode *k* for each engine used on aircraft type *j*;

Ne_j number of engines used on aircraft type *j*.

In order to create proper input data for both dispersion model and machine learning model, data cleaning and simplification was made on the JFK dataset. The runway of an airplane and whether it is a departure or arrival flight is determined by observing its position and altitude over time. Dataset is simplified into unique departure and arrival flights with reduced size of features. Afterward, NO_x emissions in gram for every LTO mode are calculated with Equation 1, even though arrival and departure flights consist of different phases of LTO. Sample from cleaned and simplified, final form of JFK dataset is given in Table 3.4.

Table 3.4. Sample from cleaned and simplified JFK dataset.

Date	Status	Runway	Aircraft Model	Engine Model	Take-off (g NOx)	Climb-out (g NOx)	Approach (g NOx)	Taxi-out (g NOx)	Taxi-in (g NOx)
01-01-2018 00:00	Departure	31L	F22 (Lockheed Martin F-22 Raptor)	CFMI CFM56-7B26E	2.220	4.446	1.419	1.522	0.498
01-01-2018 00:02	Departure	31L	A332 (Airbus A-330-200)	GE CF6-80E1A4	10.526	18.694	3.618	3.461	1.133
01-01-2018 00:03	Departure	31L	A319 (Airbus A-319)	IAE V2527-A5	2.344	5.181	1.363	1.985	0.650
01-01-2018 00:05	Departure	31L	B737 (Boeing 737-700)	CFMI CFM56-7B27	3.333	6.526	1.843	1.837	0.601
01-01-2018 00:06	Departure	31L	A332 (Airbus A-330-200)	PW PW4168A	10.098	20.832	5.615	3.027	0.991
01-01-2018 00:08	Departure	31L	B736 (Boeing 737-600)	CFMI CFM56-7B26	2.954	5.934	1.752	1.753	0.574
01-01-2018 00:10	Arrival	31R	A319 (Airbus A-319)	IAE V2527-A5	2.344	5.181	1.363	1.985	0.650
01-01-2018 00:11	Arrival	31R	E145 (Embraer ERJ-145ER)	AE3007A	0.650	1.453	0.437	0.619	0.203
01-01-2018 00:11	Departure	31L	B772 (Boeing 777-200)	GE GE90-85B	13.937	27.641	4.202	5.950	1.947
01-01-2018 00:13	Departure	31L	CL60 (Canadair Challenger 600)	CF34-3A1	0.397	0.895	0.392	0.625	0.205
01-01-2018 00:14	Departure	31L	CRJ9 (Canadair Regional Jet CRJ-900)	CF34-8C5	0.800	1.762	0.925	0.976	0.319
01-01-2018 00:15	Arrival	31R	A321 (Airbus A-321)	CFMI CFM56-5B3/3	3.795	6.645	1.693	1.715	0.561
01-01-2018 00:15	Arrival	31R	CRJ9 (Canadair Regional Jet CRJ-900)	CF34-8C5	0.800	1.762	0.925	0.976	0.319
01-01-2018 00:16	Departure	31L	A388 (Airbus A-380-800)	GP7270	18.487	35.926	8.805	8.093	2.649
01-01-2018 00:18	Arrival	31R	F22 (Lockheed Martin F-22 Raptor)	CFMI CFM56-7B26E	2.220	4.446	1.419	1.522	0.498
01-01-2018 00:18	Departure	31L	A319 (Airbus A-319)	CFMI CFM56-5B7/3	2.069	4.271	1.342	1.420	0.465
01-01-2018 00:19	Arrival	31L	CL60 (Canadair Challenger 600)	CF34-3A1	0.397	0.895	0.392	0.625	0.205
01-01-2018 00:20	Arrival	31R	F22 (Lockheed Martin F-22 Raptor)	CFMI CFM56-7B26E	2.220	4.446	1.419	1.522	0.498
01-01-2018 00:22	Arrival	31R	A321 (Airbus A-321)	CFMI CFM56-5B3/3	3.795	6.645	1.693	1.715	0.561
01-01-2018 00:23	Departure	31R	A319 (Airbus A-319)	IAE V2527-A5	2.344	5.181	1.363	1.985	0.650
01-01-2018 00:24	Arrival	31R	A319 (Airbus A-319)	IAE V2527-A5	2.344	5.181	1.363	1.985	0.650
01-01-2018 00:24	Arrival	31R	E145 (Embraer ERJ-145ER)	AE3007A	0.650	1.453	0.437	0.619	0.203
01-01-2018 00:24	Departure	31L	B752 (Boeing 757-200)	RR RB211-535E4	8.234	14.431	2.052	2.696	0.882
01-01-2018 00:25	Departure	31R	A319 (Airbus A-319)	IAE V2527-A5	2.344	5.181	1.363	1.985	0.650

3.2. Atmospheric Dispersion Modelling

AERMOD Gaussian dispersion model is used in this study to calculate NO_2 emission dispersion around JFK airport. The data needed for the model are atmospheric emissions, land use, geographical and meteorological data. AERMET and AERMAP are the two main pre-processor of the AERMOD model. AERMET itself has two preprocessors, AERMINUTE and AERSURFACE, the schema of AERMOD is graphed in Figure 3.1 (EPA, 2019b). Each of these modules are described in the following paragraphs.

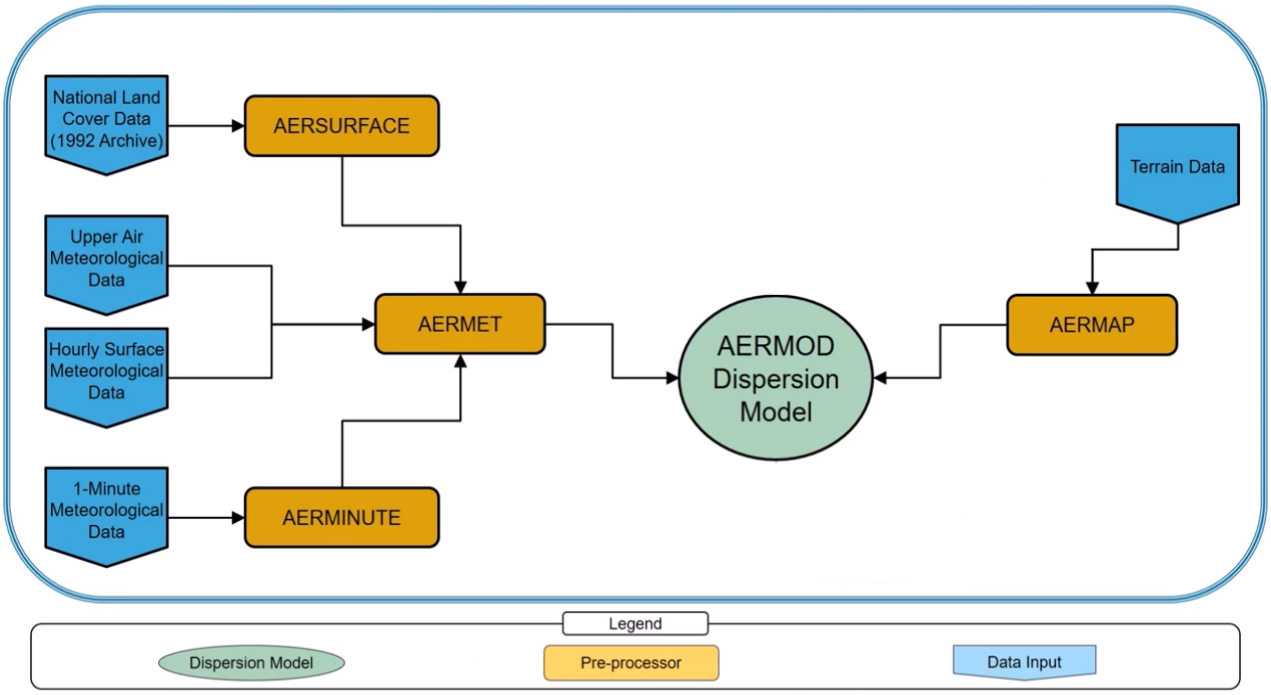

Figure 3.1. Data flow of the AERMOD Modelling System (Bajoghli et al., 2016).

AERSURFACE determines the surface characteristics needed for the air dispersion model which are obtained from the USGS National Land Cover Data 1992 archive (EPA, 2008). The data can be download from Multi Resolution Land Characteristics Consortium (MRLC, n.d.). Moreover, EPA created AERMINUTE to utilize the ASOS 1-min wind data to address a large number of calms and variable winds recorded in the hourly ASOS data files (EPA, 2015). The 1-minute ASOS wind data is available from the National Climatic Data Center (NCDC) (NOAA, n.d.). The output files of both AERSURFACE and AERMINUTE can be directly inputted into AERMET.

AERMET is EPA's meteorological data pre-processor for the AERMOD model. AERMET processes freely available National Weather Service (NWS) data or site-specific meteorological data.

AERMET generates the surface and profiles meteorological data files that are read by AERMOD (EPA, 2019a). Data taken from the station at JFK Airport are for 1-minute meteorological data and hourly surface meteorological data. On the other hand, the upper air meteorological data are taken from a station at Brookhaven, New York. Average hourly wind data at JFK airport are shown in Figure 3.2 for the year of 2018.

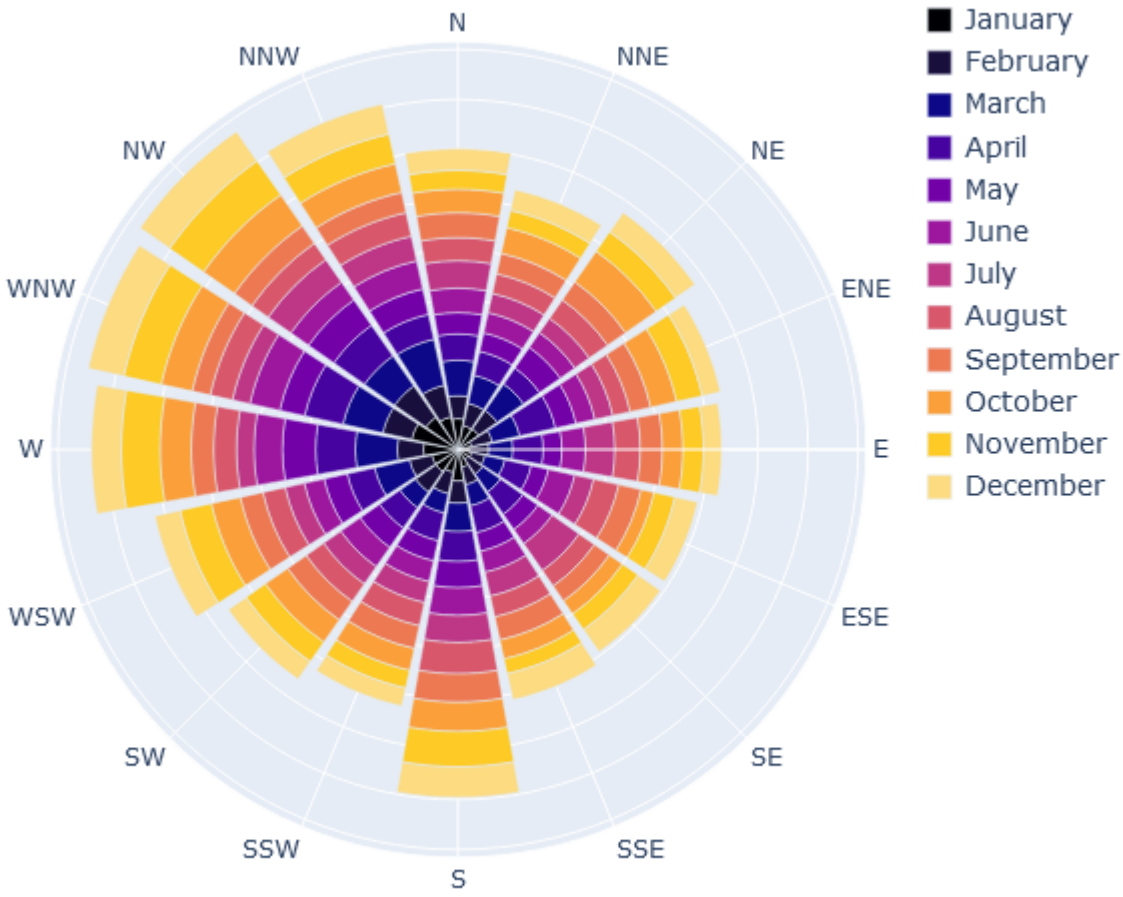

Figure 3.2. Annual wind profile of JFK airport in 2018.

AERMAP is EPA's terrain pre-processor for the AERMOD model. AERMAP determines the elevation and hill height scale for each receptor based on the US Geological Survey (USGS) digital terrain data (EPA, 2018). The digital elevation model (DEM) file is required to run AERMAP to get terrain data for receptors and sources. The 1/3 arc-second DEM file is obtained from USGS's interactive map, the National Map, by selecting the requested area on the map (USGS, 2018). Two DEM files are obtained, which, when combined, cover the desired area. Visualization of the data is possible with the software called Global Mapper using the two DEM files. Area topography of the study area can be viewed in Figure 3.3. As can be seen, elevation around the airport is up to a few

meters since it is located in the South of New York City near Jamaica Bay, which merges with the North Atlantic Ocean. New York City is mainly a flat area, however, in the North Coast, the elevation increases to between 50 - 100 meters.

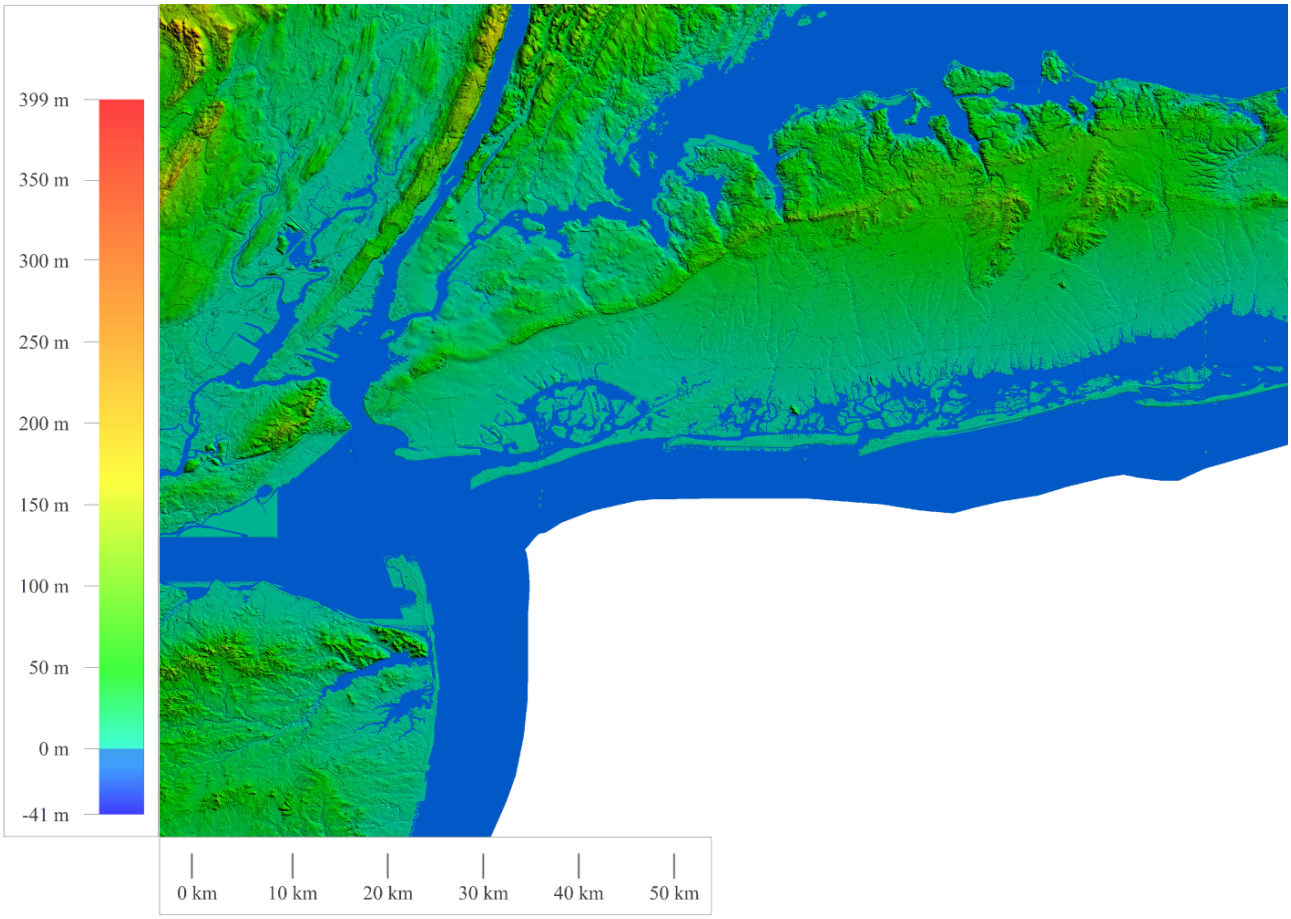

Figure 3.3. Visualization of terrain data for the study area.

AERMOD performs the calculation at selected receptors. In the current analysis, receptors are distributed uniformly over a 24 km x 24 km field around the airport with 500-meter spacing, as shown in Figure 3.4. Asides of the uniformly distributed receptors, in order to make a comparison with measured NO_2 data and show what proportion the aircraft's NO_2 emissions effect, two measurement centers in the Queens College are also included as receptors.

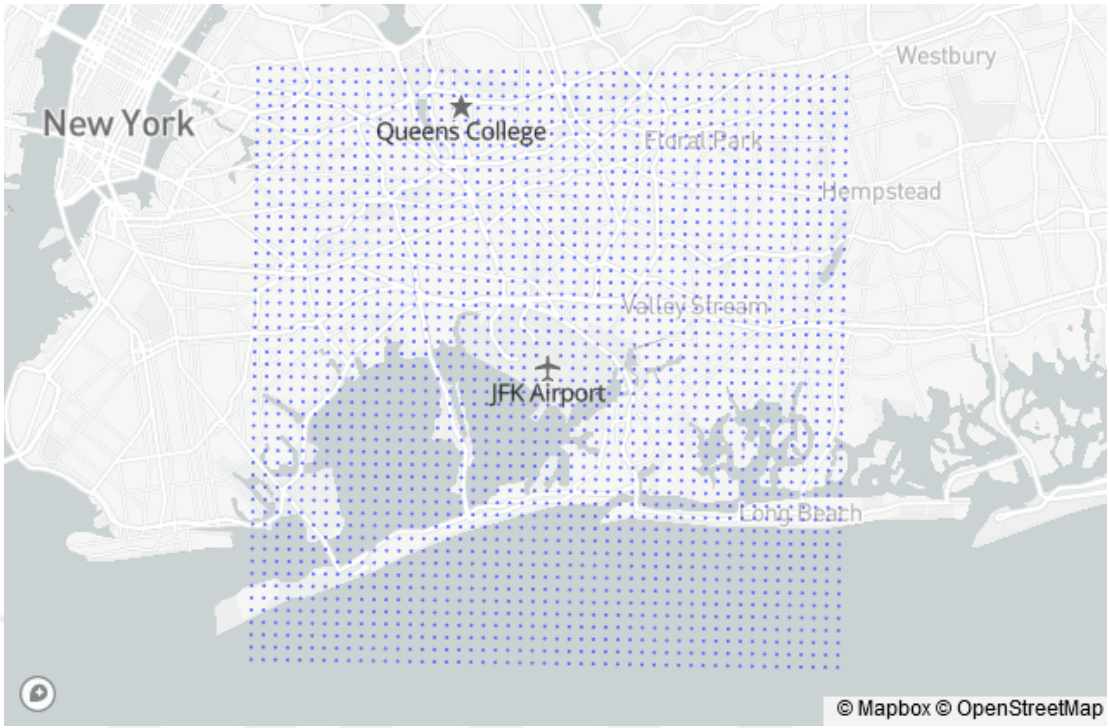

Figure 3.4. Receptor distributions around the airport and monitoring station in Queens College.

Emission sources are runways while aircraft is in take-off, climb-out and approach modes. When aircraft are in taxi-in and taxi-out, the source is the airport area itself. In both cases, the emission source type is areal. Runways are split into several areas because LTO modes can be different in different parts of the runway, also the height of emission can be different too. As illustrated in Figure 3.5, departure and arrival flights are mainly using the same part of the runway at the same elevation. For instance, the very beginning of a runway is an area source with a height of 50 m because a flight approaching the airport, or a flight just took off passes the same part in approximately the same elevation.

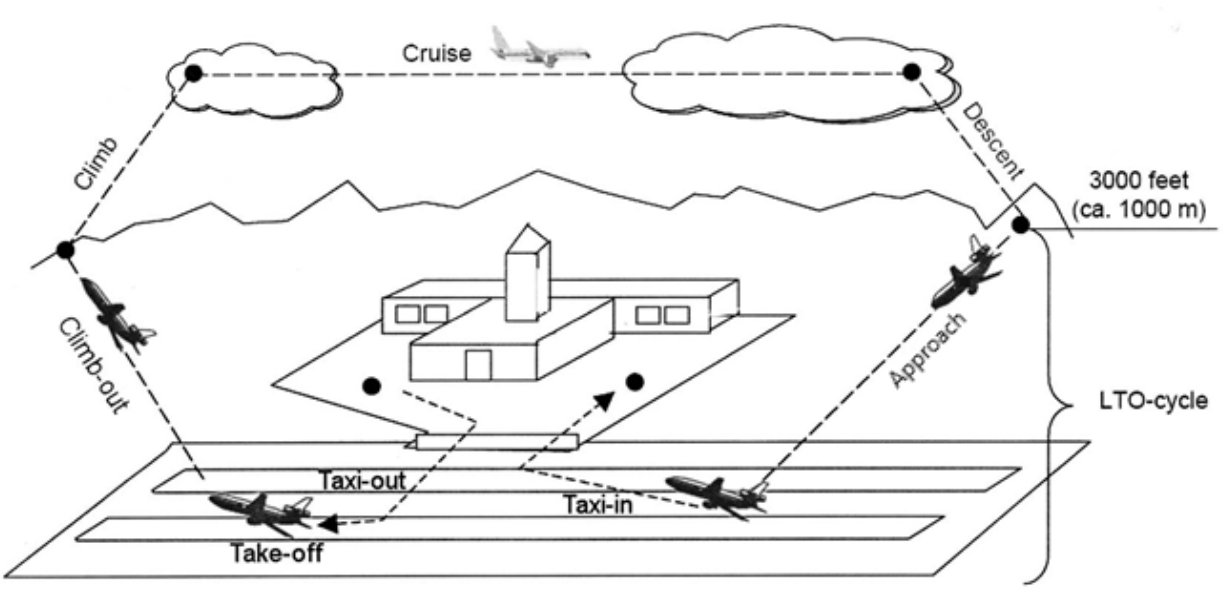

Figure 3.5. Movement of an aircraft during the LTO cycle (Elbir, 2008).

Therefore, runway sources are split into areas with a length of 500 m and the same width as the runway, which is mostly 60 m, only runway 13L and 31R have a width of 45m. Emission sources are not limited to the length of the runway, there are 3 to 5 more area sources defined before a runway starts or after a runway ends, to account for the climb-out or approach phase until 800 m elevations. Figure 3.6 shows the taxi area as a polygon. It also shows all the runways, with the runways split into smaller areas. The figure also shows the length of each runway. Since they are longer than the other runways, 13R and 31L are split to 15 areas, while the other runways are split into 14 areas.

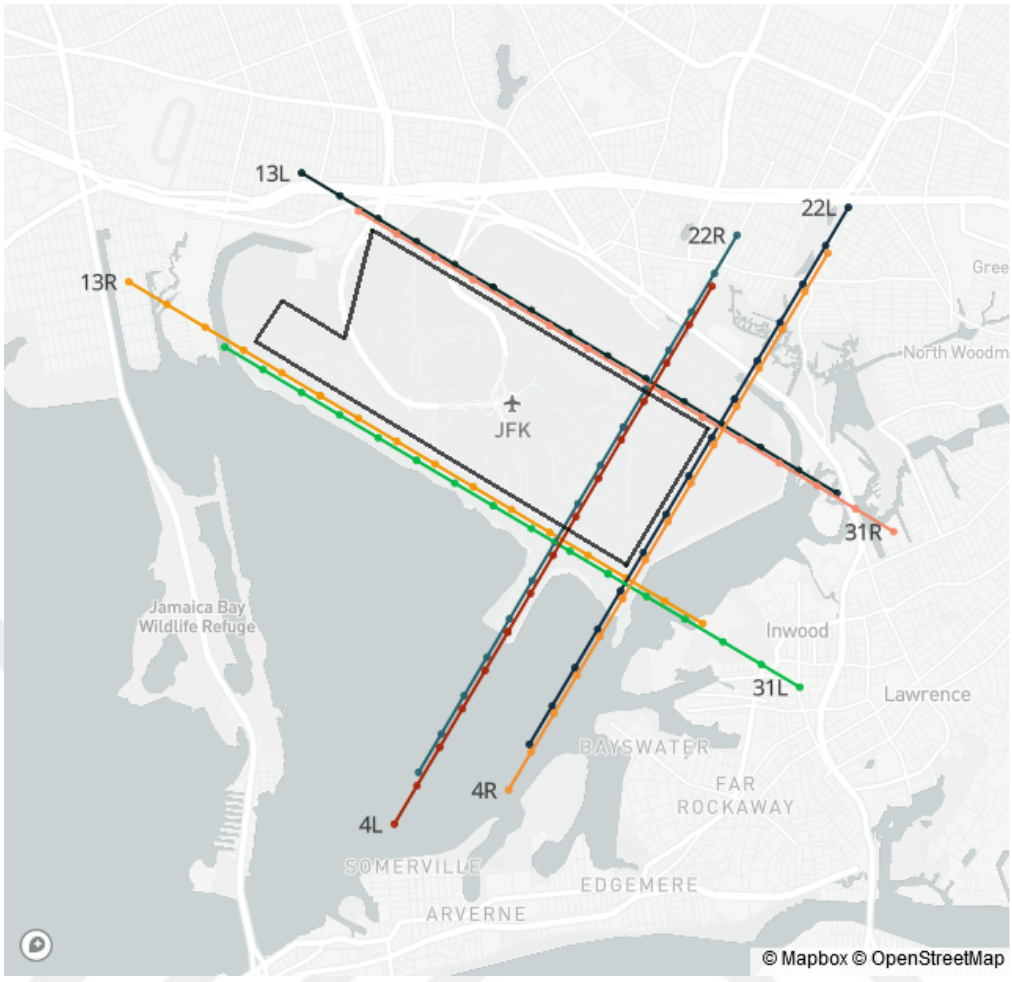

Figure 3.6. Taxi area, runways and where the runways are split into areas.

After AERMAP and AERMET runs are finished, hourly emission data is prepared using previously mentioned NO_x emissions ready JFK dataset. First, the area source of every flight is changing according to which runway is used and whether it is departure or arrival flight. Therefore, every flight's emissions corresponding to their flight status are considered, for instance, an arrival flight does not have take-off, climb-out, and taxi-out emissions; it only has approach and taxi-in. Then emission in an LTO mode is divided by how many area sources are defined for that mode in that runway. After that, it is divided by the area of source in m^2 and time period in that source in seconds to obtain emission rate of NO_x produced by an aircraft during an LTO mode in $g/(s \cdot m^2)$. Finally, every calculated emission for every area source is summed up in hourly blocks. The following Table 3.5 shows a preview of hourly emission data for every source, it is defined as HOUREMIS in AERMOD. The first two letters of source names other than 'AIRPORT', annotates whether it's departure space (DS), climbing space (CS) or arrival space (AS), the last two digits annotate which part of the runway it is, and the middle part is runway's name.

Table 3.5. Hourly emission data for every area source.

Year	Month	Day	Hour	Source	Concentration
18	5	7	24	AIRPORT	0.000005848059
18	5	7	23	DS22R09	0.000000000000
18	5	7	23	DS22R10	0.000000000000
18	5	7	23	CS22R11	0.000000000000
18	5	7	23	CS22R12	0.000000000000
18	5	7	23	CS22R13	0.000000000000
18	5	7	23	CS22R14	0.000000000000
18	5	7	23	AS4L01	0.000036151233
18	5	7	23	AS4L02	0.000036151233
18	5	7	23	AS4L03	0.000036151233
18	5	7	23	AS4L04	0.000036151233
18	5	7	23	AS4L05	0.000036151233
18	5	7	23	AS4L06	0.000036151233
18	5	7	23	DS4L06	0.000121206067
18	5	7	23	DS4L07	0.000121206067
18	5	7	23	DS4L08	0.000121206067
18	5	7	23	DS4L09	0.000121206067
18	5	7	23	DS4L10	0.000121206067
18	5	7	23	CS4L11	0.000278872024
18	5	7	23	CS4L12	0.000278872024
18	5	7	23	CS4L13	0.000278872024
18	5	7	23	CS4L14	0.000278872024
18	5	7	23	AS4R01	0.000022527961
18	5	7	23	AS4R02	0.000022527961
18	5	7	23	AS4R03	0.000022527961
18	5	7	23	AS4R04	0.000022527961
18	5	7	23	AS4R05	0.000022527961
18	5	7	23	AS4R06	0.000022527961
18	5	7	23	DS4R06	0.000115564668

The NO_x conversion method is one of the required inputs to define when running AERMOD to calculate NO_2 emission dispersion. There was a method called ARM which uses a fixed conversion rate to calculate how much of the NO_x is converted to NO_2 (Kimbrough et al., 2017). However, it is reported that defining a fixed rate is not a valid way to calculate the conversion since variability over time is very high. Therefore, a more advanced version of ARM comes into play, ARM Version 2 (ARM2). A large dataset of observed NO_2/NO_x conversion ratios in diverse conditions of source-monitoring distance, atmospheric dispersion conditions, and atmospheric ozone concentrations are used to create an empirical equation in which the upper bound of the conversion factor can be estimated as a function of NO_x (Podrez, 2015). ARM2 is used as a conversion method to run AERMOD in this study.

3.3. Machine Learning Modeling

Tree-based learning algorithms are regarded as one of the best and frequently used supervised learning methods. Tree-based approaches allow for high precision, consistency and ease of analysis for predictive models, in contrast to linear models, non-linear relationships are quite well mapped. They can be modified to solve any type of problem such as classification or regression problems (Clark et al., 2017). The evolution of the tree-based algorithm is shown in Figure 3.7.

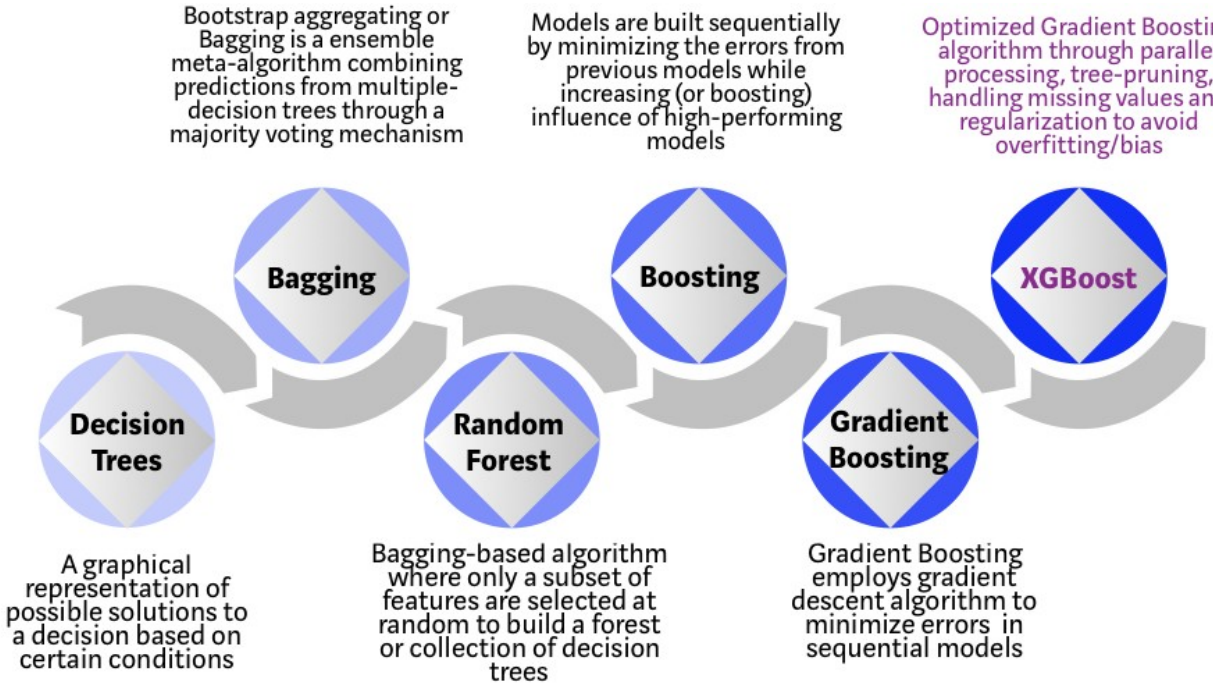

Figure 3.7. Evolution of tree-based models from decision trees to extreme gradient boosters (XGBoost) (Morde, 2019).

The decision tree is a graphical instruction of the decision-making process. It is an “if … then … else” type of process which makes it a good fit for the programmatic structure. A sample decision tree is illustrated in Figure 3.8.

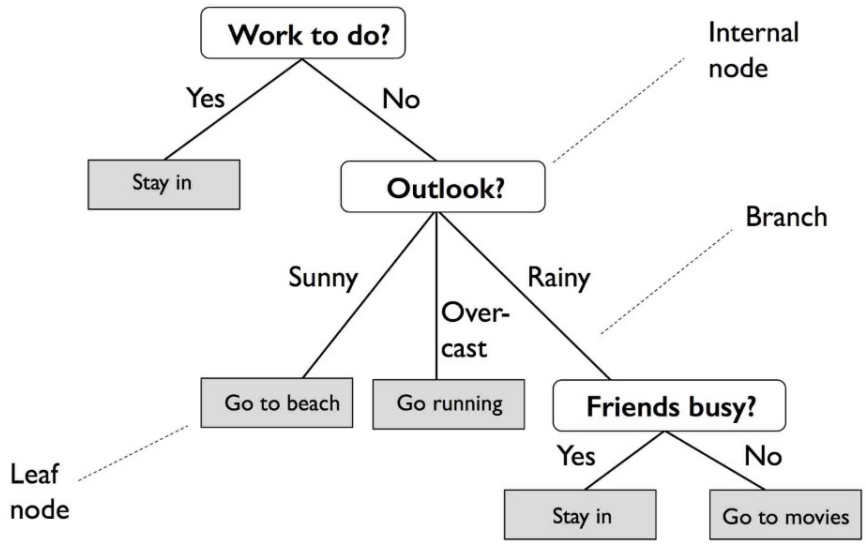

Figure 3.8. An example visualization of a decision tree (Li, 2019).

Like any other model, a tree-based model often suffers from bias and variance. Bias is the difference between predicted and actual values; if the difference is high then the bias is high. A high biased model can not effectively learn relationships between features and oversimplifies it. Therefore, it would underfit the dataset which leads to high error in training and test data. Variance is the amount of change in model prediction if different training datasets were used. High variance causes the model to memorize the training data and does not generalize well on the data that it has not seen before. In other words, models with high variance show good performance on training data but show low performance and high error rates on test data (Geurts, 2002). Figure 3.9 shows the contribution of bias and variance on total error, while Figure 3.10 shows the graphical illustration of bias and variance.

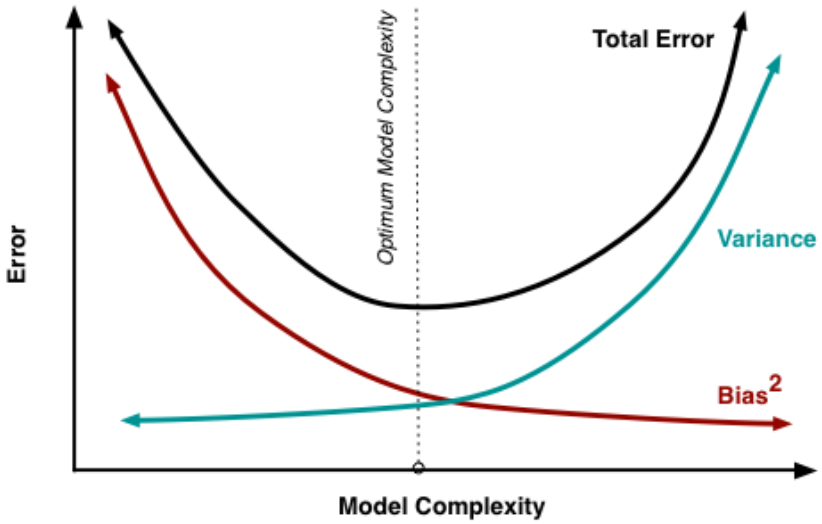

Figure 3.9. Contribution of bias and variance on total error (Fortmann-Roe, 2012).

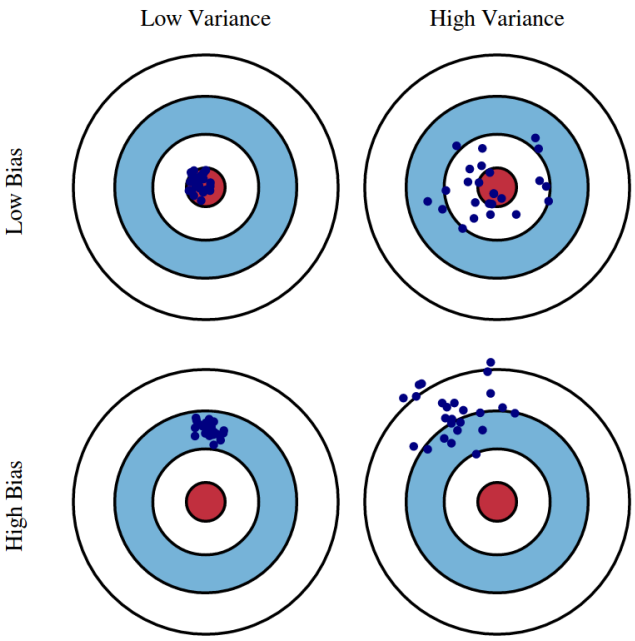

Figure 3.10. Bullseye diagram to show bias variance trade-off (Fortmann-Roe, 2012).

In order to balance the trade-off between bias and variance, model complexity can be increased, which will reduce the bias. However, making the model more complex after a level will cause overfitting, which makes the model suffer from high variance. One way to manage bias variance trade-off is ensemble learning (Zhang et al., 2012). Bagging is an ensemble learning technique to reduce variance by taking multiple simple classifiers (weak learners) created by sub-samples of the same data. It combines their result for a given input and takes the most frequent result as a prediction among them (Prasad et al., 2006).

The random forest is one of the algorithms that use the bagging method to make predictions over many different decision trees. It also uses feature randomness when to build each tree to try to create a forest of uncorrelated trees whose prediction by the committee is more precise than that of each individual tree (Breiman, 2001). Furthermore, boosting is another method that trees are built sequentially and each subsequentially built tree aims to reduce the errors of the previous trees. Each tree learns from residual errors that are updated by the tree that came before. Trees created in boosting have fewer splits when compared to the bagging technique used model types like the random forest. Initial weak learners in boosting have a high bias; however, being interpretable because of their small size makes the boosting's final learner have low bias and variance at the end of the iterative learning process (Schapire, 2003). Below in Figure 3.11, the illustration of comparison bagging and boosting is given.

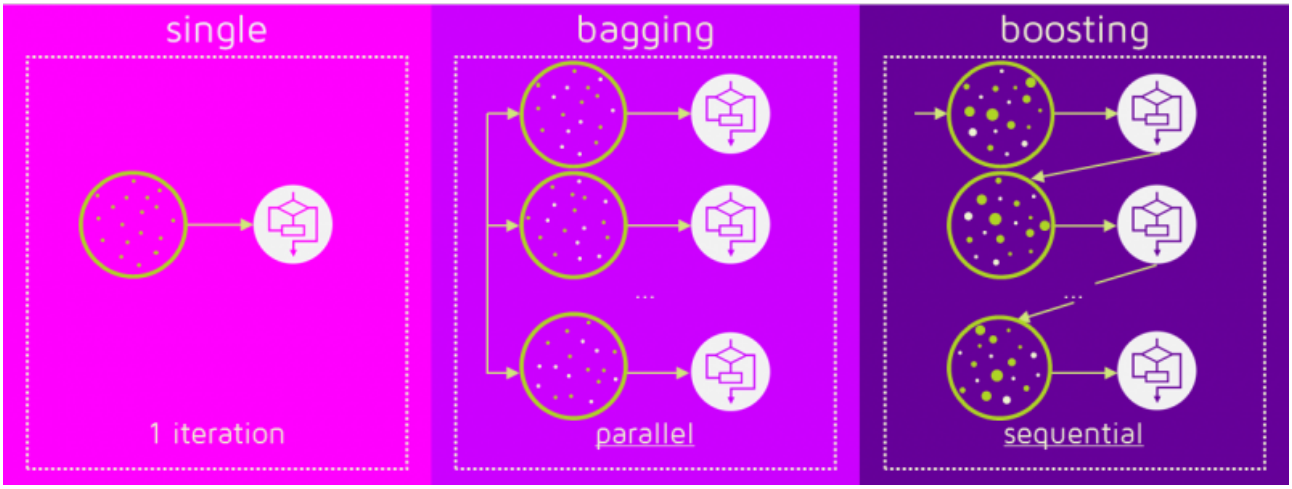

Figure 3.11. Comparison of bagging and boosting methods (Aporras, 2016).

Gradient boosting, as a supervised learning algorithm, tries to predict a target variable accurately by combining predictions of a series of simpler and weaker models. If regression is made by using gradient boosting, regression trees are used as weak learners, and every weak learner takes input data to calculate a score to pass on its leaves continuously. Training in gradient boosting is an iterative process. Errors of previous trees are calculated after adding new trees, then both trees are combined to make a new prediction. It is referred to as gradient boosting because a gradient descent algorithm is used to minimize loss when adding new models (Taieb et al., 2014).

Gradient boosting machines (GBM) differ from boosting algorithms while handling the shortcoming of weak learners. Similar to how neural networks use gradient descent to optimize weights, GBM also uses gradients to minimize a loss function. Every training round builds the weak learner and compares its predictions to the correct result that is expected. The distance between prediction and the true value represents the model's error rate, and this error can be used to calculate the gradient. The main focus in GBM is to make changes in the model's predictions by using gradient descent to get a sum of residuals closer to 0 so that predictions will be close to actual values as well (Natekin et al., 2013).

XGBoost is an improved implementation of the gradient boosted trees algorithm by systems optimization and algorithmic enhancements. It is also an open-source project optimized by implementing parallelization, adding stopping criterion for tree splitting, handling big sized data frames that do not fit into memory. Furthermore, the algorithm is enhanced by adding L1 and L2 regularization method to have better control on overfitting, also with handling missing data with sparsity awareness and being able to make cross-validation (T. Chen et al., 2016). The superiority of

XGBoost then the other algorithms like GBM, random forest and even logistic regression are shown in Figure 3.12, a benchmark test made on a classification dataset. Even though XGBoost and GBM are almost identical in terms of accuracy, XGBoost is 86 times faster than GBM.

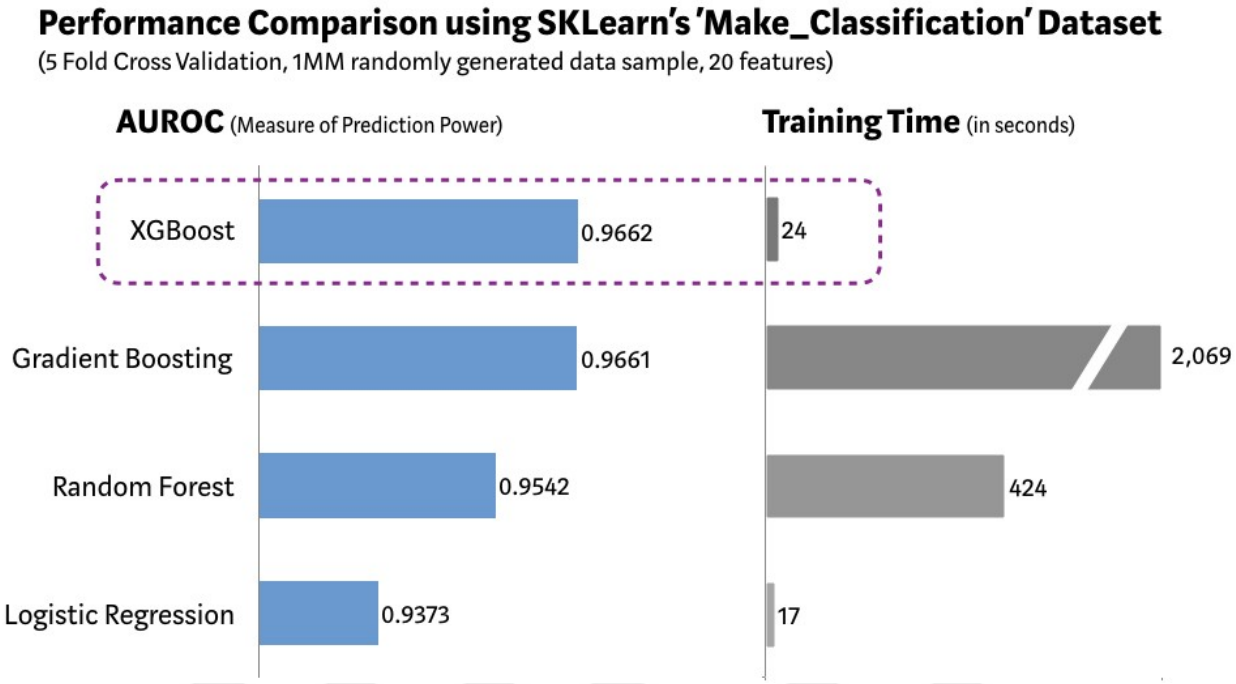

Figure 3.12. Benchmark of XGBoost and other machine learning algorithms on a classification dataset (Morde, 2019).

XGBoost is a supervised learning method where the training data x_i is used to predict target variable y_i. The supervised learning model generally constructed on a mathematical model, like the linear model, where prediction is a linear combination of weighted input features. Classification or regression can be the task that the prediction by the model is interpreted. During the model training, parameters θ, which are coefficients in the linear model, are expected to be learned from the data.

$$\hat{y}_i = \sum_j \theta_j x_{ij} \tag{2}$$

In order to measure how well the model finds the best parameters θ that best fit the training data x_i and labels y_i, an objective function is defined. Objective functions are consisting of two parts, training loss functions L and regularization term Ω.

$$L(\theta)=L(\theta)+\Omega(\theta) \tag{3}$$

One of the common loss functions is mean absolute error (MAE), which measures, without considering their direction, the average magnitude of the errors in a sequence of predictions.

$$\mathrm{MAE} = \frac{1}{n}\sum_{j=1}^{n}\left|y_j - \widehat{y_j}\right| \tag{4}$$

Another loss function that is commonly used is the root mean squared error (RMSE), which also measures the average magnitude of the error as a quadratic scoring method.

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{j=1}^{n}\left(y_j - \widehat{y_j}\right)^2} \tag{5}$$

The model will not generalize well and overfit if there is noise in the data, in order to avoid this, regularization term shrinks or regularizes these learned estimates towards zero to discourage learning a more complex or flexible model. It is used to find the optimum model complexity in the bias-variance tradeoff as it is illustrated previously in Figure 3.10.

XGBoost which uses classification and regression trees (CART) as weak learners, is different than the regular decision trees which only give the decision values in the leaf. CART gives real prediction scores associated with each leaf, which makes possible much better interpretations. In Figure 3.13, tree ensemble can be seen as, prediction scores of two trees summed up to get the final result.

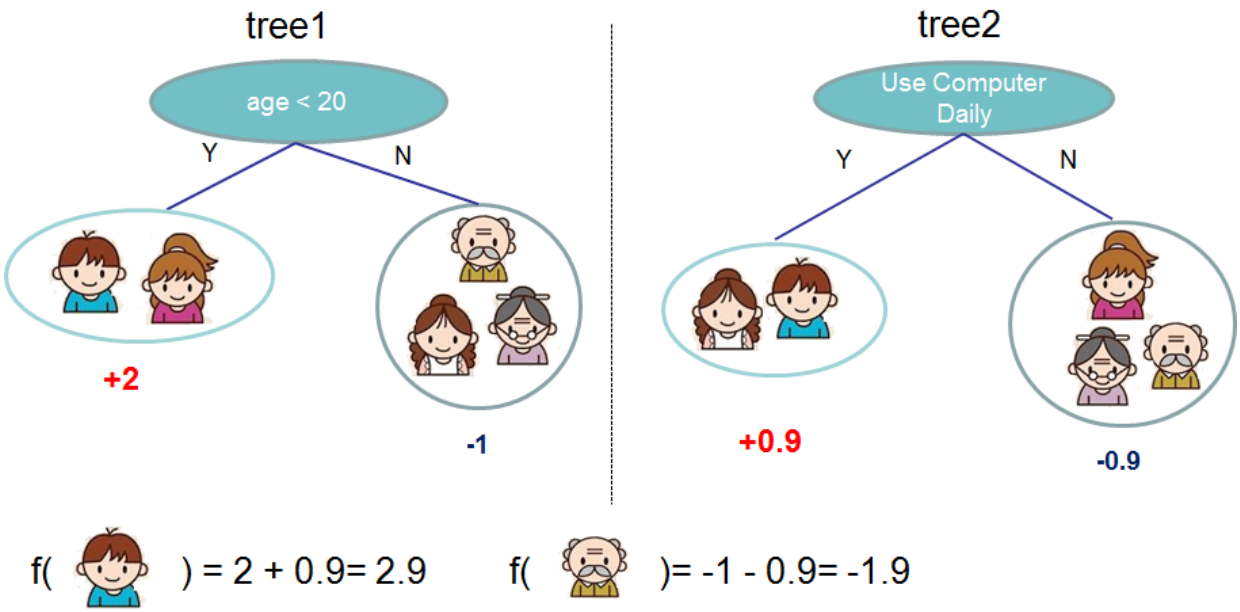

Figure 3.13. Tree ensemble of two trees (T. Chen, 2016).

Here is the dataset $\mathcal{D} = \{(x_i, y_i): i = 1 \cdots n, x_i \in \mathbb{R}^m, y_i \in \mathbb{R}\}$, n observations with m features each and with a true label y. The model uses this data to generalize on and give a result of $\hat{y}_i$, it can be represented as in follows,

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i), f_k \in \mathcal{F} \tag{6}$$

where K is the number of trees, f_k is a regression tree among $\mathcal{F}$, which is all possible CARTs, and score given to i-th observation by the k-th tree is represented as $f_k(x_i)$. The following regularized objective function should be minimized to get an accurate result for $\hat{y}_i$,

$$\mathcal{L}(\theta) = \sum_{i}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k) \tag{7}$$

$$\text{where } \Omega(f) = \gamma T + \frac{1}{2}\lambda |w|^2 \tag{8}$$

l is the loss function and Ω is a regularization term, where T is the number of leaves, w is the magnitude of leaf weights and the penalty is controlled by parameters γ and λ to prevent the increase in model complexity. Optimization of the model cannot be made as in the traditional methods which simply take the gradient since it is difficult to learn all the trees at once. Therefore, an additive strategy

is used by adding a new tree to the prediction of the *i*-th instance at the *t*-th iteration, $\hat{y}_i^{(t)}$ that is already made.

$$\mathcal{L}^{(t)} = \sum_{i=1}^{n} l\left(y_i, \hat{y}_i^{(t-1)} + f_t(\boldsymbol{x_i})\right) + \Omega(f_t) \tag{9}$$

Simplification of the objective function Eq. 9 can be made by using the Taylor expansion. A tree greedily splits in every possible direction by taking the remaining features in the tree into account. Then the new loss is calculated for each split, and with the loss reduction formula (Eq. 10), it's decided to pick the tree which reduced the loss most.

$$\mathcal{L}_{split} = \frac{1}{2}\left[\frac{\left(\sum_{i\in I_L} g_i\right)^2}{\sum_{i\in I_L} h_i + \lambda} + \frac{\left(\sum_{i\in I_R} g_i\right)^2}{\sum_{i\in I_R} h_i + \lambda} - \frac{(\sum_{i\in I} g_i)^2}{\sum_{i\in I} h_i + \lambda}\right] - \gamma \tag{10}$$

I_L and I_R are the scores on the new left and right leaves after the split, I is the score on the original leaf. In order to find the best tree, it is intractable to sort all the trees possible and choose the best one. Therefore, evaluation of the split candidates by one level at a time is made by Eq. 10.

3.4. Implementation of the Machine Learning Model to Aircraft Emissions

In this study, a model is trained to predict using XGBoost pollutant atmospheric concentrations at a given location near the JFK airport due to aircraft emissions. Inputs are hourly meteorological data at the airport and hourly emission data calculated using engine emission factors. Hourly concentration distributions calculated from AERMOD for receptors distributed with 4000 m spacing are used as labels to train the model (Figure 3.14). Data from 49 evenly distributed receptors are used for training, and data from 2 receptors at Queens College are used as test data.

Figure 3.14. Receptors used in model training.

The dataset used in model training consists of 9 features and 1 label. The 9 features are:

- distance of a receptor to the center of the airport (latitude: 40.641, longitude: -73.779) in meters;
- direction, heading, of the receptor to the center of the airport in degrees;
- hourly meteorological values around the airport like wind speed, wind direction, temperature, precipitation, relative humidity, and pressure;
- hourly total NO_x emitted by aircraft in the airport area in grams;

The label is hourly NO_2 concentrations at every receptor calculated by AERMOD. Since 1 year-long data from 49 receptors are used to train the model, this means that the length of our training data is 429,240. The sample from the training dataset is shown in Table 3.6.

Pollutant concentrations at receptors acquired from AERMOD to use as labels in the machine learning model contains clusters of data at 0, especially in the farther areas to the source. In order to handle this kind of data, the Tweedie distribution is used as a learning objective in XGBoost (Jørgensen, 1987). It is a special case of an exponential distribution. It is useful for modeling when there is a mixture of zeros and non-negative data points, like medicine and genetics, biology research, insurance claims, and rainfall data (Bonat et al., 2016; Hasan et al., 2012; Kendal, 2004; Kendal et al., 2000; Smyth et al., 2002). The Tweedie model is a good candidate in the case of having a spike at zero when a histogram is drawn with the owned dataset. With the changing power parameter, some well-known distributions fall into the Tweedie distribution as illustrated in Figure 3.15. With the power parameter close to 1, the distribution is a Poisson distribution; if the power parameter is close to 2, it gives gamma distribution (Gilchrist et al., 2000).

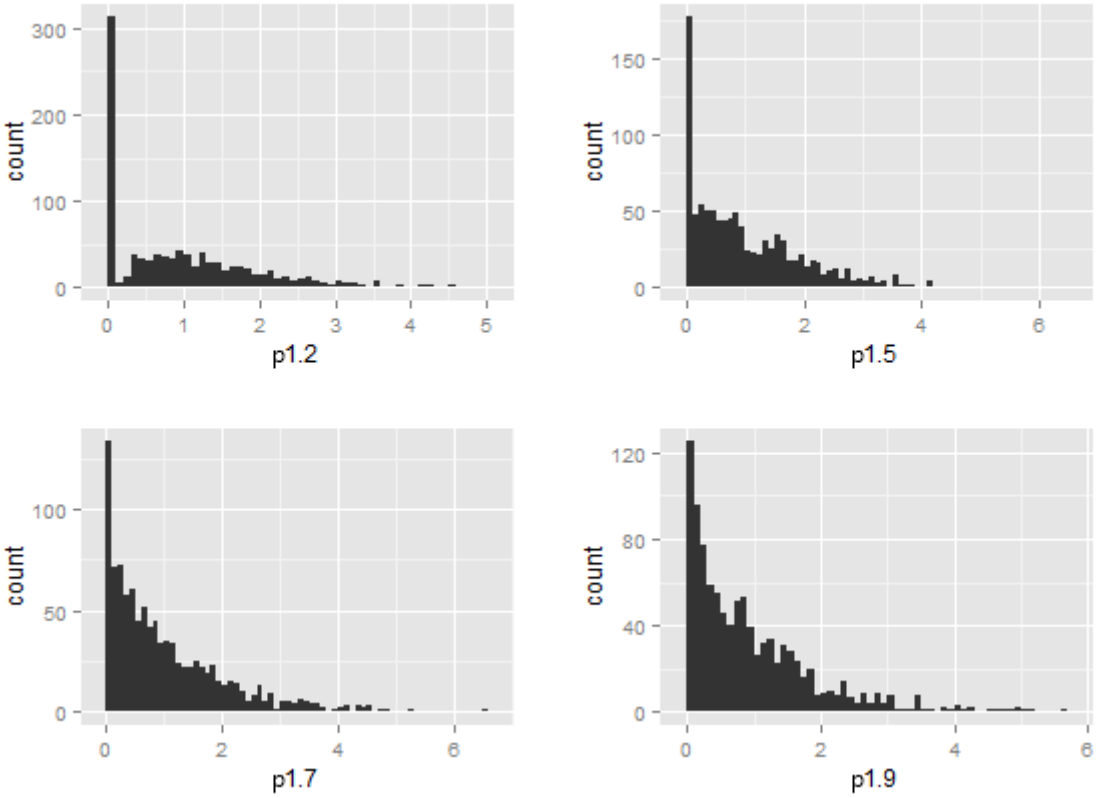

Figure 3.15. Four Tweedie histograms with the power parameter from 1.2 to 1.9.

There are sets of parameters that can be adjusted according to the problem that is being to be solved with the available dataset and re-tuned with the machine learning results. Bias variance tradeoff is the main concern of parameters in XGBoost. The predictive power of the model and model complexity are the subjects that the best model should trade. Parameters of XGBoost are defined as follows:

- *num_boost_round* is the number of iterations for boosting or building trees. Other parameters have an effect on it; therefore, it might need to be re-tuned after a parameter is updated.
- *early_stopping_rounds* are to stop training when there is no improvement after N rounds (defined by *early_stopping_round*) after the training stopped, the best number of boosting rounds are kept.
- *evals* is a list of pairs to evaluate the model on, it's usually the training dataset and an evaluation dataset together.
- *max_depth* is the maximum number of nodes permitted from the root to the tree's farthest point. Deepening trees can form more complex relationships by adding additional nodes,

but as we deepen, tree splits become less important and sometimes due to noise, causing the model to overfit (default=6).

- *min_child_weight* is the required minimum weight to create a new node in the tree, smaller weight results in creating a child, which is a leaf node created after a partition, by using fewer samples. By default, all the samples are used to create a child (default=1).
- *subsample* is the fraction of how much of the rows or in other words observations used to subsample at each step. In default, all the rows are used (default=1).
- *colsample_bytree* is the fraction that determines how much of the features that will use, by default value 1, it uses all the features (default=1).
- *ETA* is also called the learning rate. It is the amount of shrinkage of the feature weights after each round, it can also be defined as the correction amount at each step (default=0.3).
- *gamma* is the minimum loss reduction needed to achieve by the previous split in order to make a new split (default=0).
- *lambda* is an L2 regularization term on weights, it is used to handle the regularization part of XGBoost (default=1).
- *alpha* is the L1 regularization term on weight, it is used to run the algorithm faster when the dimensionality is very high (default=0).
- *objective* defines the loss function to be minimized (default=reg: squared error).
- *eval_metric* is the metric to be used for validation data (default according to objective).
- *tweedie_variance_power* is the parameter that controls the variance of the Tweedie distribution, it's in range of between 1 and 2. Shifts towards a Poisson distribution if it's set close to 1, otherwise, it shifts towards a gamma distribution (default=1.5).

After a series of iteration, only the following parameters are optimized and changed to other than their default values: *num_boost_round* is set to 100,000 and *early_stopping_rounds* is set to 50 to stop the training after the training and evaluation stops converging; *subsample* is set to 0.8, for each tree to take the 80% of the rows to build the tree; *ETA* is set to 0.01 to make the boosting process more conservative; *alpha* is set to 0.8 and *gamma* is set to 1 to make the model more conservative and prevent overfitting; *objective* is set to previously mentioned Tweedie distribution, therefore, *eval_metric* is set to negative log-likelihood as a default of Tweedie distribution; *tweedie_variance_power* is set to 1.2 which shifts the model towards to Poisson distribution.

Table 3.6. Input data used in XGBoost for training.

Distance (m)	Heading (degrees)	Wind speed (m/s)	Wind direction (degrees)	Temperature (K)	Precipitation (mm/h)	Relative humidity (%)	Station pressure (millibars)	Total NOx (g/h)	NO_2 ($\mu g/m^3 \cdot h$)
8946.751	334.2005	2.08	166	278.8	0	64	1035	683.5453	7.80073
12003.2	180.7929	4.41	128	296.4	0	81	1019	773.8253	0
11325.43	343.4379	8.61	188	304.2	0	72	1015	592.0355	0.14763
12003.03	90.88727	9.3	185	290.9	0	96	1006	112.9384	0
12652.7	252.2671	4.03	222	297	0	100	1016	175.8837	0
14426.26	327.0453	7.5	137	296.4	2	100	1016	368.6152	2.57095
12652.72	289.1382	6.47	233	285.9	0	50	1018	536.3279	0
4001.045	90.82541	9.03	325	266.4	0	47	1029	800.305	0
4001.049	180.7944	4.02	222	298.1	0	90	1015	823.092	0
16975.36	225.7008	3.86	129	287.5	0	80	1021	626.877	0
5658.32	135.8249	6.28	262	269.9	0	42	1027	471.9164	0
8002.058	90.85641	6.75	253	284.9	0	50	1024	189.6323	0.7907
5658.32	135.8249	8.03	293	279.9	0	48	1022	902.4589	0.52766
14425.83	124.5758	4.39	186	296.4	0	100	1011	402.5945	0
4001.049	180.7944	3.08	7	270.9	0	74	1022	305.1751	117.7305
12652.7	252.2671	5.72	186	293.1	0	83	1021	712.1831	0
5658.358	45.82628	6.92	183	303.8	0	68	1016	427.1357	0
8002.179	0.796357	9.85	310	291.4	0	53	1012	487.0007	0
12652.55	199.1972	8.79	254	266.4	0	54	1027	497.3833	0
8946.788	297.2992	2.3	127	298.8	0	90	1022	116.8668	10.02926

A summary of the overall procedure followed in this study is given in Figure 3.16. Firstly, pollution sources were identified, and then pollution emissions were calculated by using specific emission factors. Afterward, pollutant concentrations were calculated from these emissions by using dispersion modeling, and later the model was trained using machine learning. Lastly, the ability of the model to predict aircraft emission was evaluated.

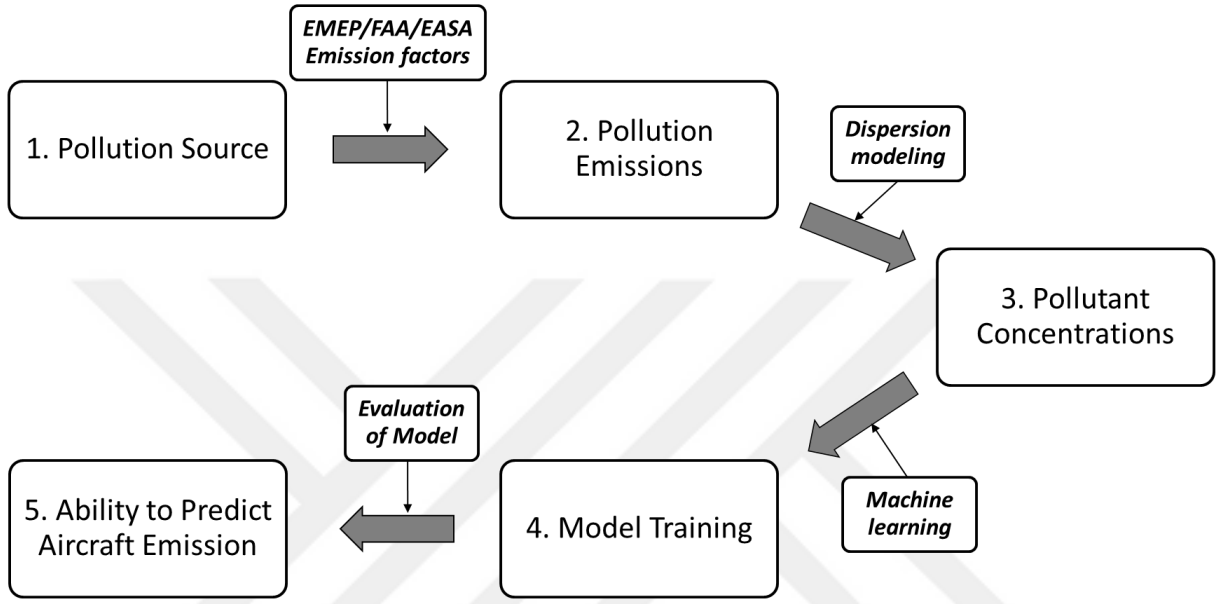

Figure 3.16. Overall Methodology for Emission Prediction with ML.

4. RESULTS AND DISCUSSION

4.1.LTO Cycle Emissions

The NOx emissions for LTO cycles of all aircrafts and all runways combined at JFK airport were calculated using Eq. 1 mentioned in section 3.1. The calculated emissions of every phase of the LTO cycle and total annual emission are given in Figure 4.1. Examination of the calculated NO_x emissions for JFK airport shows that the climb-out phase of the LTO cycle has the highest annual emission with 1951 tons (t), followed by a take-off phase with 1047 t. The total number of flights in 2018 was 455,000. Contributions to the NO_x emissions of take-off, climb-out, approach, and taxi phases are 25.3%, 47.1%, 12%, and 15.6%, respectively. In other words, when the engine thrust is higher, then the NO_x emission is noticeably increased.

The calculated NO_x emissions for JFK airport are consistent with the reported values from the literature. Annual NO_x emissions in another busy airport, Istanbul Ataturk Airport, is calculated as 4249 t based on the 465,000 flights' data in 2015 (Kuzu, 2018). In Chania Airport, Greece, which is a relatively small airport with 10,324 flights during the year 2016, annual NO_x emission is calculated as 137 t (Makridis et al., 2019). The average NOx emission per flight at JFK is 9.09 kg/flight which is close to that reported for Istanbul (9.14 kg/flight) and Chania (13.3 kg/flight).

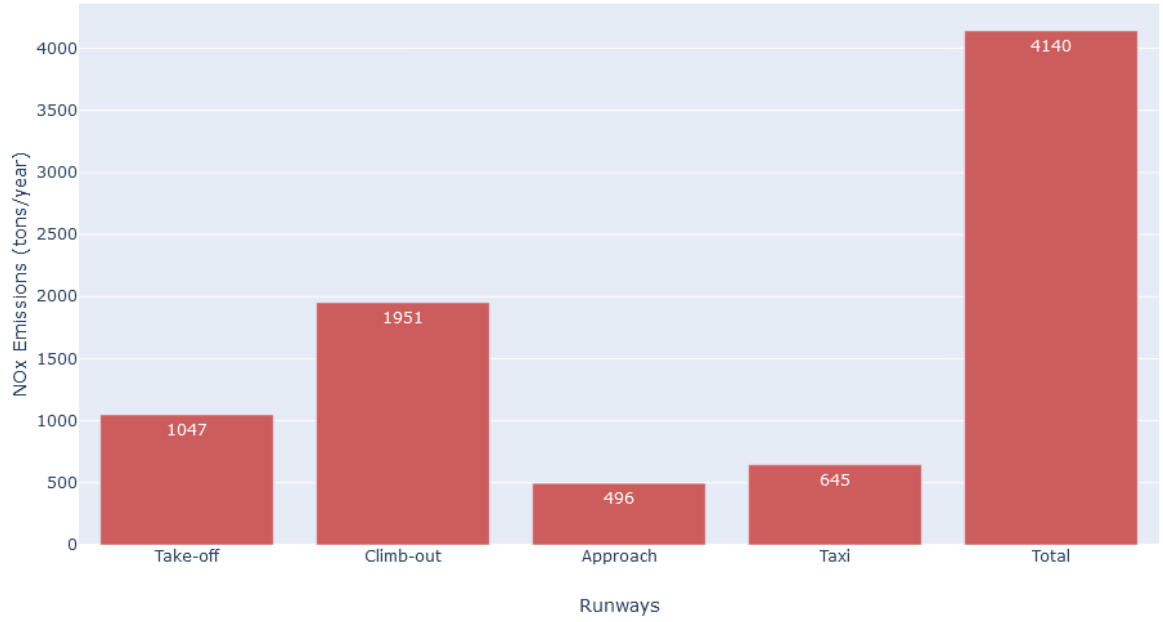

Figure 4.1. Annual NO_x emissions of phases of the LTO cycle and total annual emission.

Some variations in emissions between runways are also observed since their usage varies throughout the year. As can be seen in Figure 4.2, 22L is the busiest runway in the airport with both arrival and departure flights, followed by 31R. Therefore, they have the two highest emissions. Runway 22R and 31L also have high NO_x emissions, however, they mostly have departure flights.

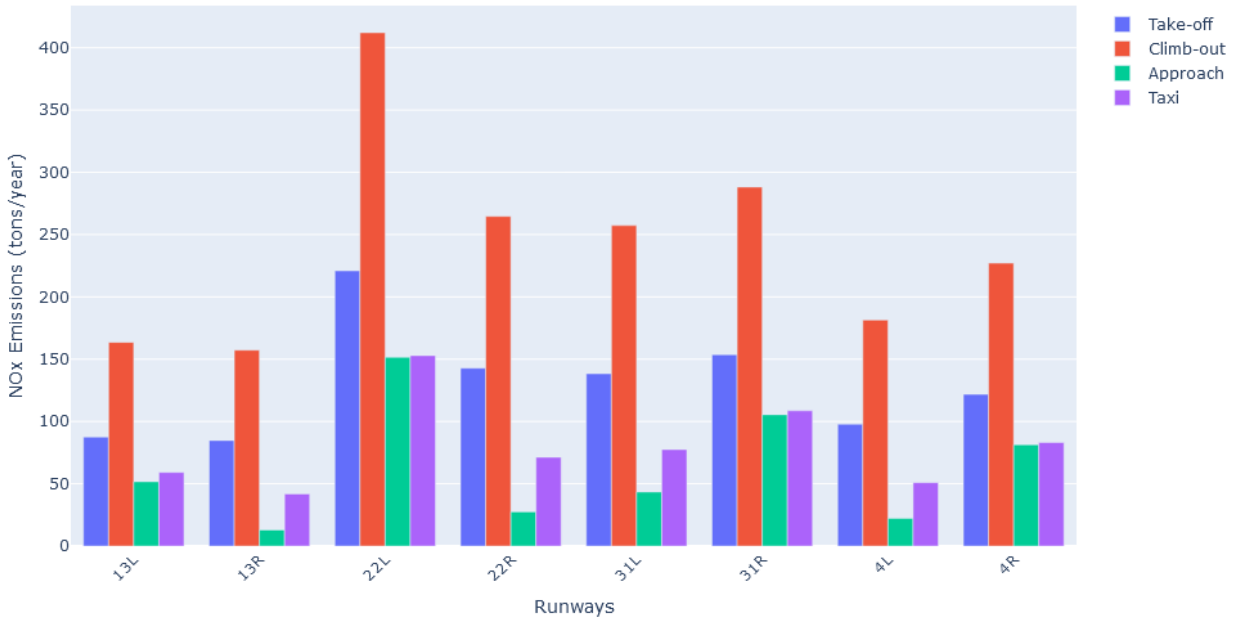

Figure 4.2. Annual NO_x emissions by phases of the LTO cycle for every runway.

The monthly number of flights varies throughout the year. As a result, NO_x emissions also vary accordingly (Figure 4.3). As can be seen from the graph, NO_x emissions are the lowest level in February, when the minimum number of flights took place. By contrast, NO_x emissions are at the highest level in July, when the maximum number of flights took place. In general, the number of flights performed in holiday seasons like summer and spring is higher than in other months, and thus, the NO_x emissions measured in these months are also higher than the monthly average, which is estimated as 344.42 tons/month.

It is apparent that NO_x emissions are strongly dependent on the number of flights, but there are slight variations in some months. The reason for this discrepancy can be attributed to the type of aircraft, such as the increasing number of charter flights in certain months which tends to increase the NO_x emissions.

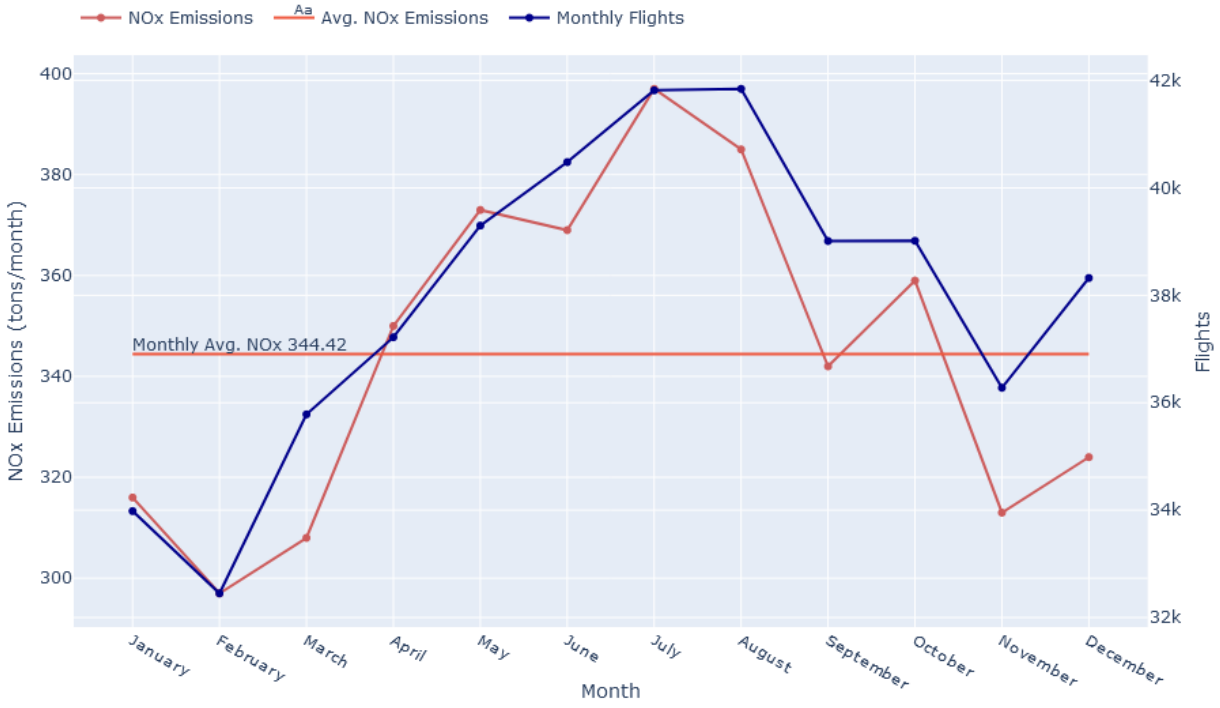

Figure 4.3. Monthly flight counts and NO_x emissions for all aircrafts using JFK airport.

As shown in Figure 4.4, NO_x emissions can also change hourly throughout the day depending on the rush hours. The ratio of peak hourly emissions to minimum hourly emissions is a factor of 10 approximately. The least NO_x emissions were measured between 2 and 5 a.m, while the highest NO_x emissions were measured between 5 and 10 p.m. There was a rapid increase in NOx emissions from 4 a.m to 7 a.m and a significant decline in NO_x emissions from 9 p.m to 2 am of the following morning.

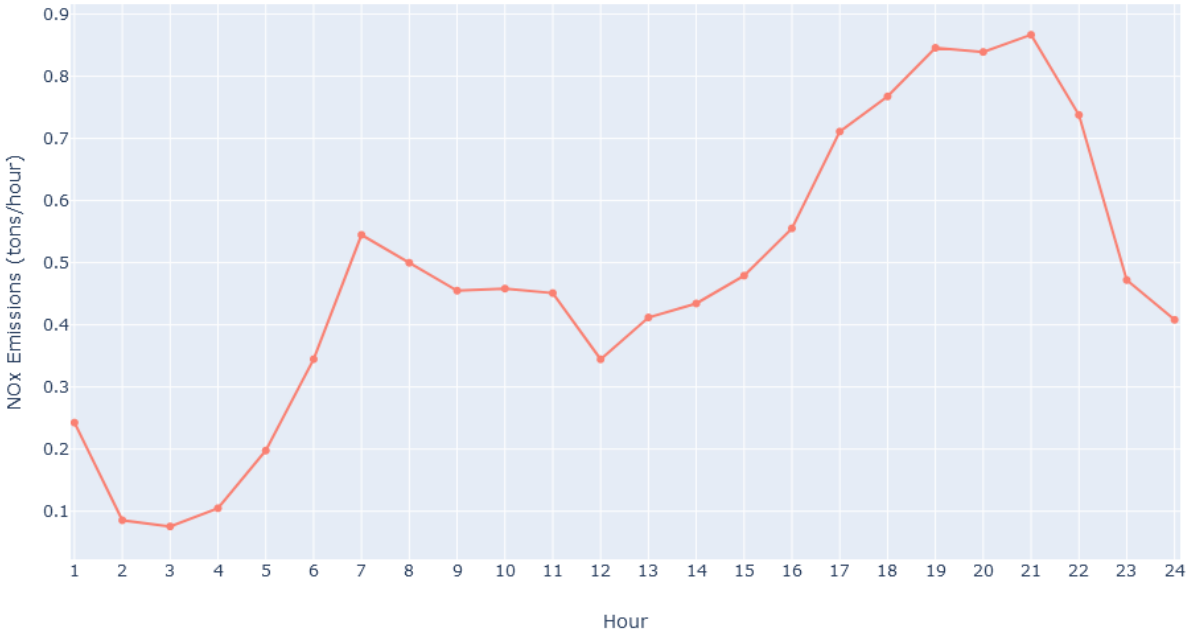

Figure 4.4. Hourly average NO_x emission distribution throughout the day.

4.2. Atmospheric Dispersion Modelling

After the emissions are calculated for the different LTO phases of every arrival and departure flight, emission sources were defined and hourly emission data for use as input in AERMOD were created. Since an aircraft either ascends or descends in the LTO phases of take-off, climb-out, and approach, emission sources for those phases are defined as areal sources shown in Figure 4.5 (Wayson et al., 2003). By dividing the calculated total emissions of an aircraft for LTO phases by source areas, the emissions are converted from units of g/h to g/s*m^2. Subsequently, they are evenly distributed over the corresponding area sources (Figure 4.5) to define the hourly emissions file for AERMOD.

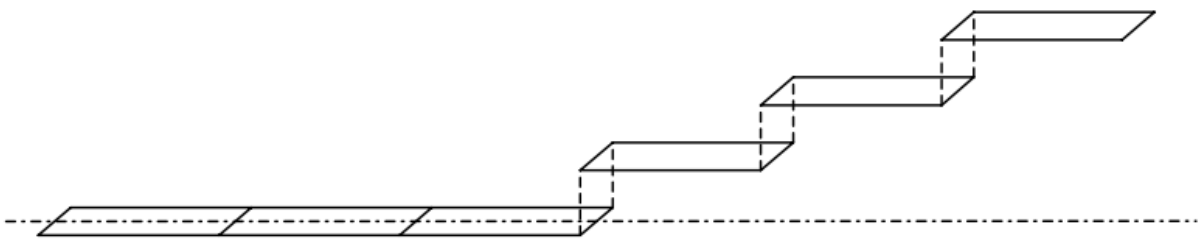

Figure 4.5. Aircraft emission sources in take-off, climb-out and approach phases (Wayson, 2003).

AERMOD was used to calculate the hourly ground-level NO_2 concentrations at all receptors (Figure 3.4). ARM2 is used to calculate NO_x/NO_2 conversion in AERMOD. It is a sixth-order polynomial regression based on 1-hr NO_x measurements from 580 stations between 2001-2010 (EPA, 2013).

In order to compare results with the hourly air quality standard, the 19th highest hourly concentration are taken, since the EPA and WHO regulations allow for the exceedance of the hourly standard 18 times in a year. Moreover, the annual average of NO_2 concentrations at every receptor was also computed and compared with annual air quality standards. On the other hand, a machine learning model requires hourly data to train a model. Therefore, hourly concentration data for the receptors that are shown in Figure 3.14 were used for model learning.

The WHO hourly and annual NO_2 air quality standards are 40 µg/m^{3}and 200 µg/m^3. Throughout the year, the hourly threshold of NO_2 can be exceeded up to 18 times. Figure 4.6 shows the receptors (in red) where the air quality standard was exceeded 19 or more times. Figure 4.7 depicts the receptors (designated in red) that have higher annual average emission than the threshold.

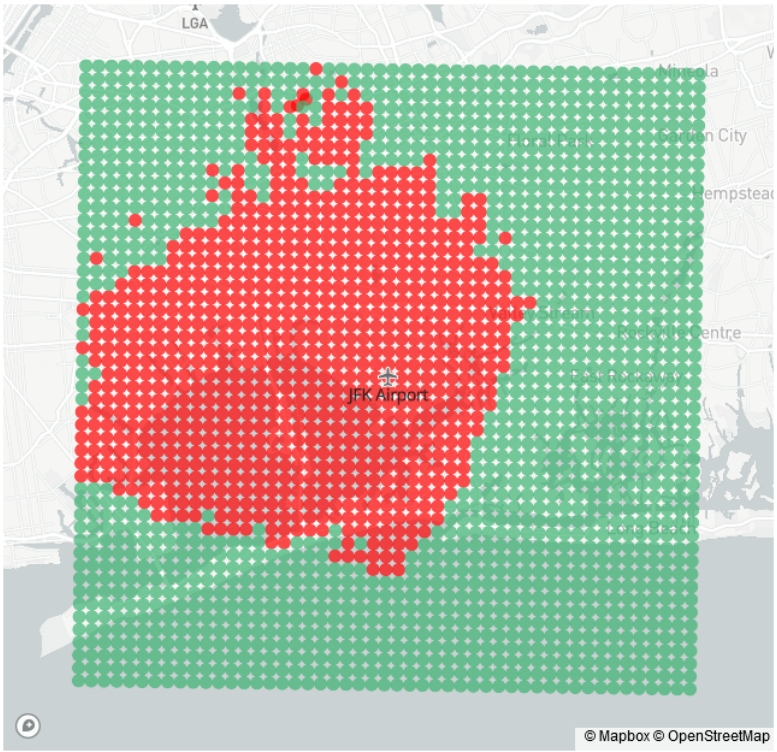

Figure 4.6. Receptors (shown in red) where the 19th highest hourly NO_2 concentrations exceed the allowable hourly air quality standard of 40 µg/m^3.

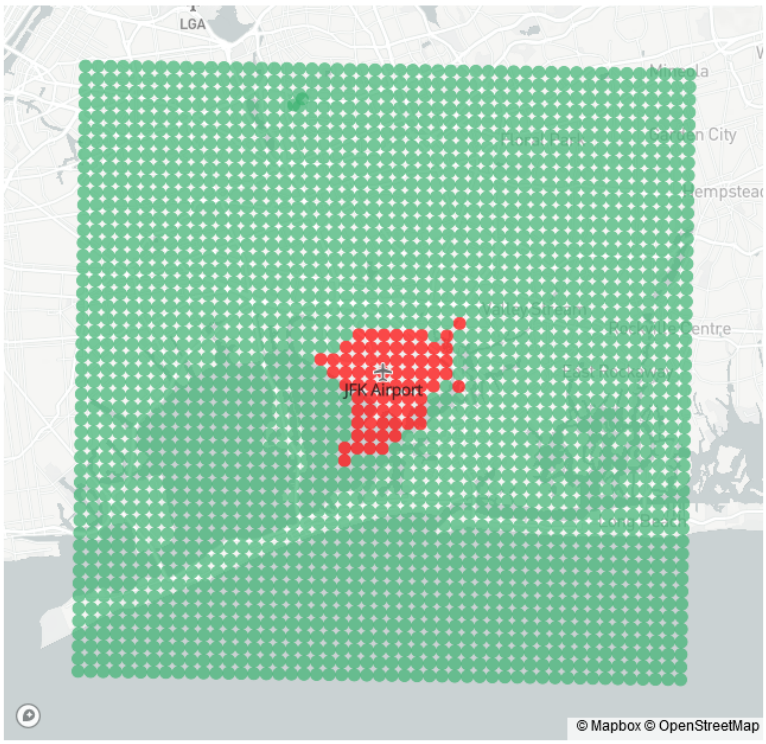

Figure 4.7. Receptors (shown in red) where the average annual NO_2 concentrations exceed the annual air quality standard of 200 μg/m^3.

According to the air dispersion modeling results, 853 of the 2403 receptors are exceeding the hourly limit more than 18 times. NO_2 Concentrations around the airport area decrease rapidly with distance due to dispersion. It can be seen that the 19th highest hourly concentrations are correlated with the annual wind profile shown in Figure 3.2. On the other hand, the annual air quality standard is exceeded only at 74 receptors, all near to the airport. Concentrations are high near the runways which are the major emission sources. Contour plots of annual average concentrations and 19th highest hourly concentrations are illustrated in Figure 4.8 and Figure 4.9, respectively, It is observed that concentrations started to increase from both ends of the runways, reaching a peak value where the two runways intersect.

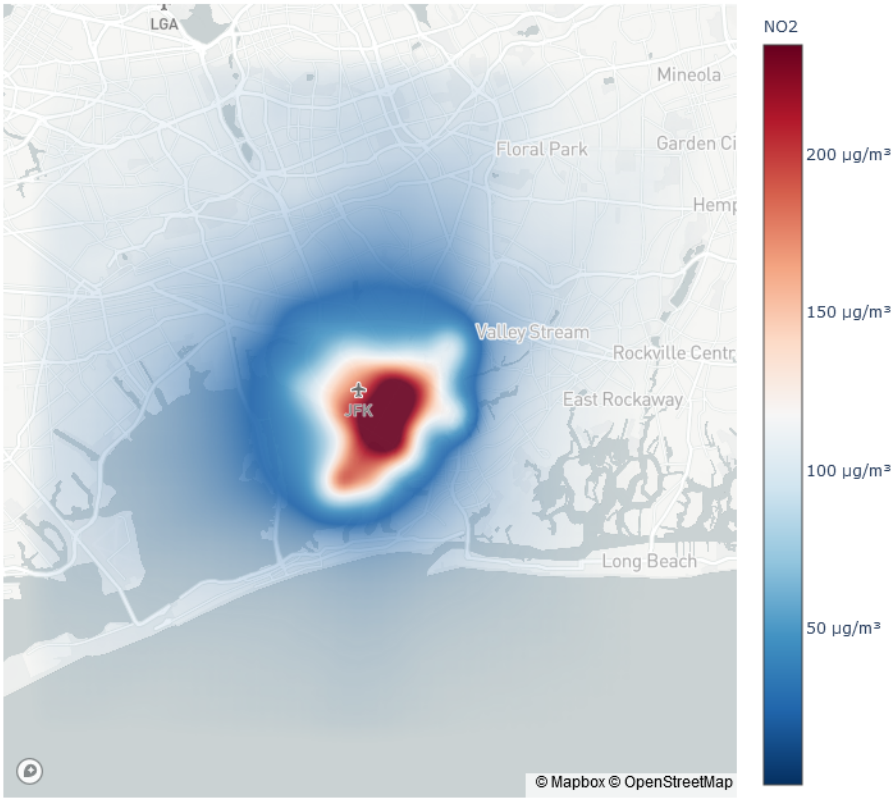

Figure 4.8. Contour plot of the annual average NO_2 concentrations around the airport.

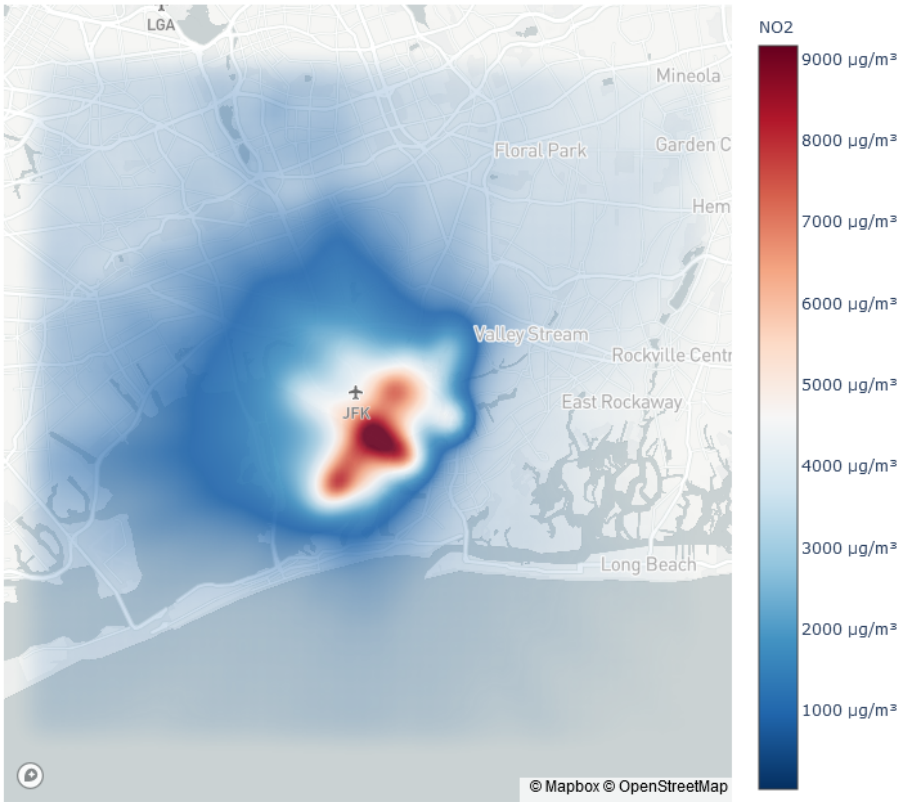

Figure 4.9. Contour plot of the 19th highest hourly NO_2 concentrations around the airport.

AERMOD results for Queens College and Queens Near Road air quality stations are compared with the real data in Figure 4.10 and Figure 4.11, respectively. It is important to note that the observed concentrations are the total NO_2 concentrations due to all sources including vehicle traffic and heating while the simulated concentrations are due to aircraft emissions from JFK airport only. The average simulated concentrations are 1.79 $\mu g/m^3$ and 1.94 $\mu g/m^3$ at Queens College and Queens Near Road, respectively, while the average observed concentrations are 26.05 $\mu g/m^3$ and 29.02 $\mu g/m^3$ at Queens College and Queens Near Road. This suggests that NO_2 concentrations due to aircraft emissions are about 6.7% of total observed concentrations. In both stations, at some points, modeled data are higher than the measured data, which is unexpected in normal conditions. Regulatory dispersion models such as AERMOD can provide good long-term estimates of atmospheric concentrations; however it can not necessarily produce accurate concentration vs time data. This may be attributed to the meteorological data used in the model. With a 1 km by 1 km grid, it is not possible to accurately simulate the air flow field. Another reason for this discrepancy is the adjustments in land use data. For instance, NLCD1992 land cover data which is required for AERSURFACE could not be found in MLRC, therefore NLCD2011 data is used in the model (EPA, 2008). The emission factors used in the calculation of the emissions might be causing this overestimation since they are accepted as constants but ideally, a single value for all seasons and aircraft operation conditions may not accurately represent an aircraft's actual emissions.

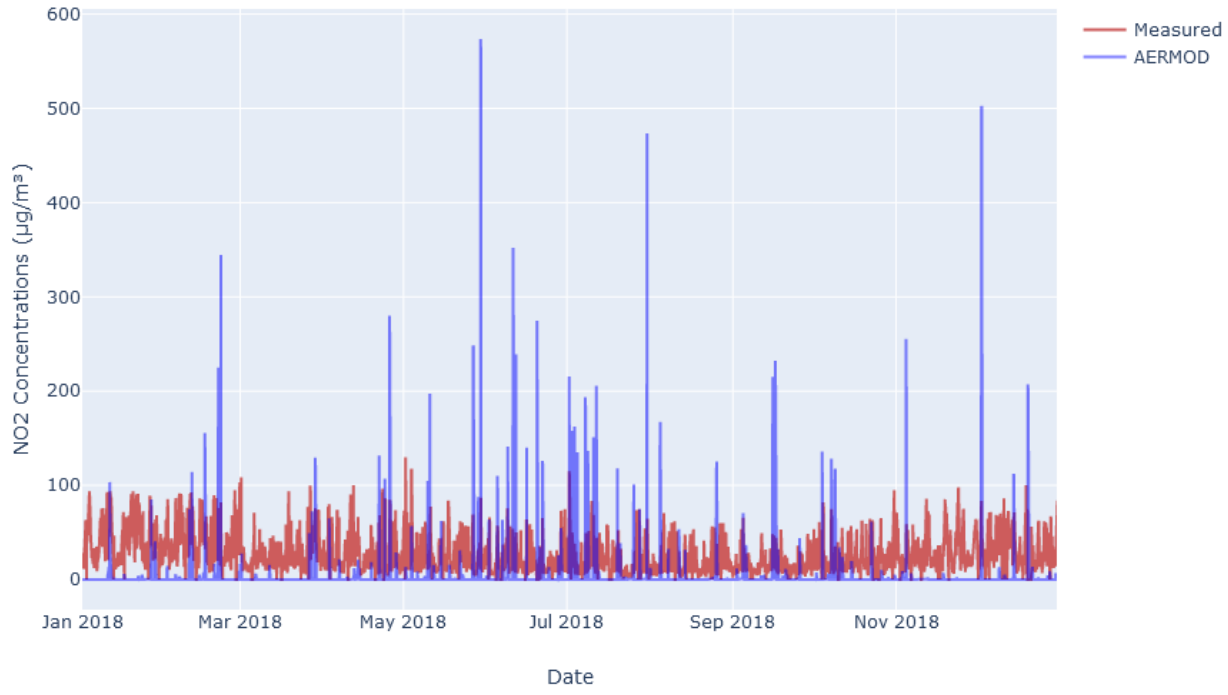

Figure 4.10. Comparison of modelled and observed NO_2 concentrations at Queens College air quality station.

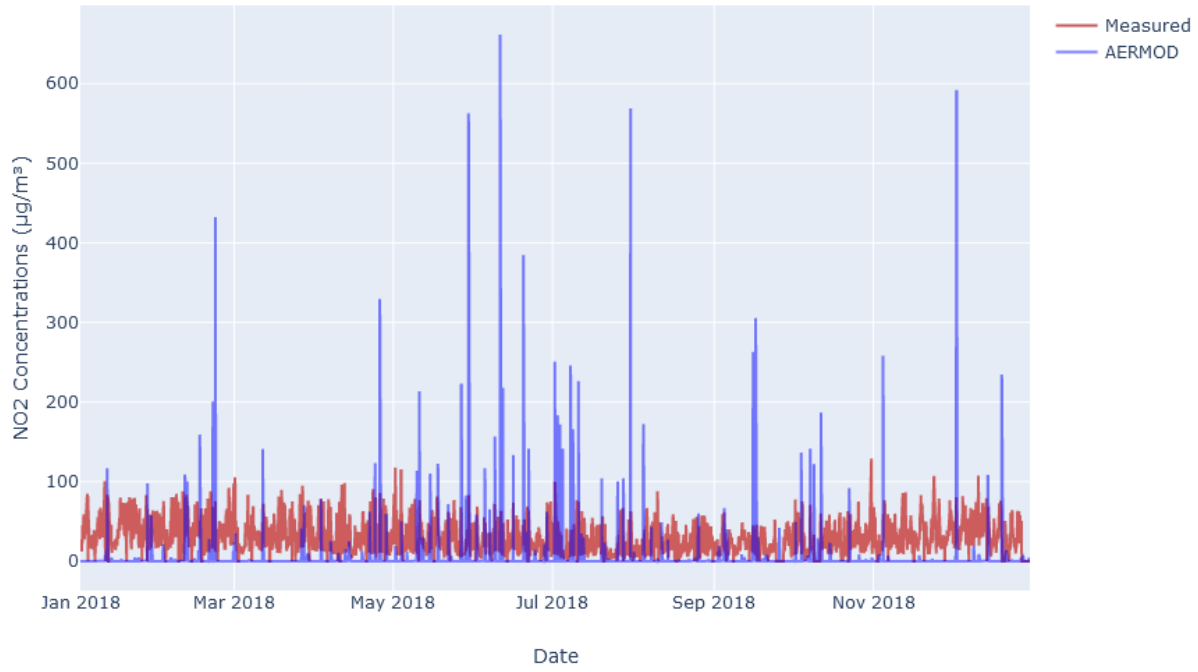

Figure 4.11. Comparison of modelled and observed NO_2 concentrations at Queens Near Road air quality station.

Additionally, scatter plots between observed and modeled data for both stations were constructed to evaluate the correlation between the two data sets (Figure 4.12 and Figure 4.13). Even though modeled data come up with high values at some points, generally, it is either zero or close to zero. Since observed data represents NO_2 concentrations resulted from every possible source, the correlation between two data for both stations is low as shown in Figure 4.12 and Figure 4.13. Another important feature of the AERMOD is the steady and spatially uniform wind distribution which in reality is much more complex.

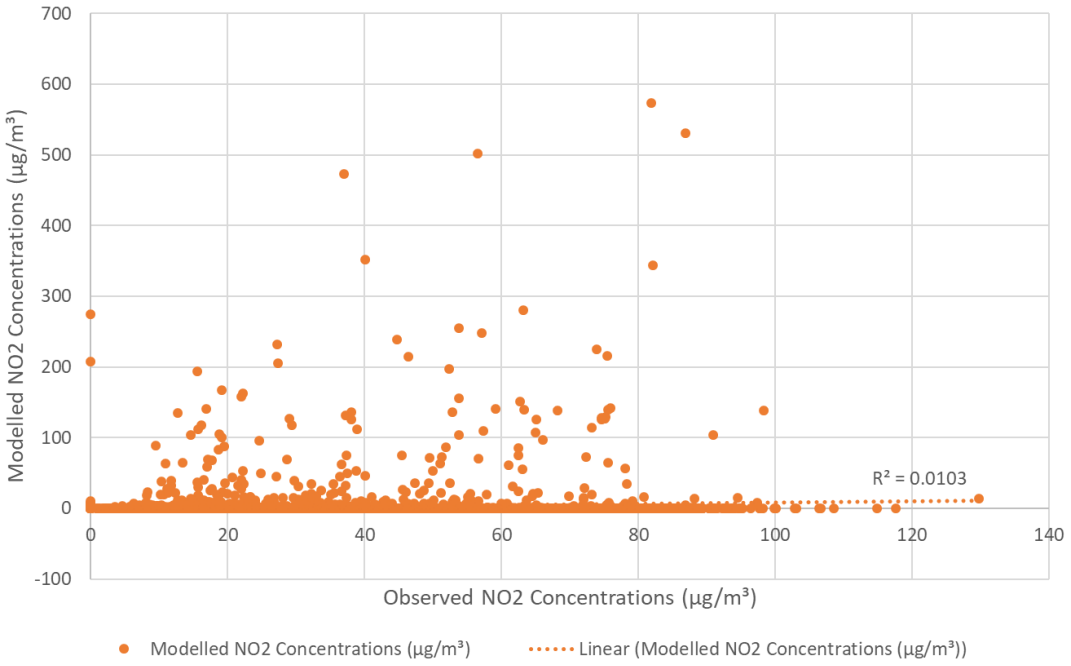

Figure 4.12. Correlation between observed and modeled concentrations in Queens College station.

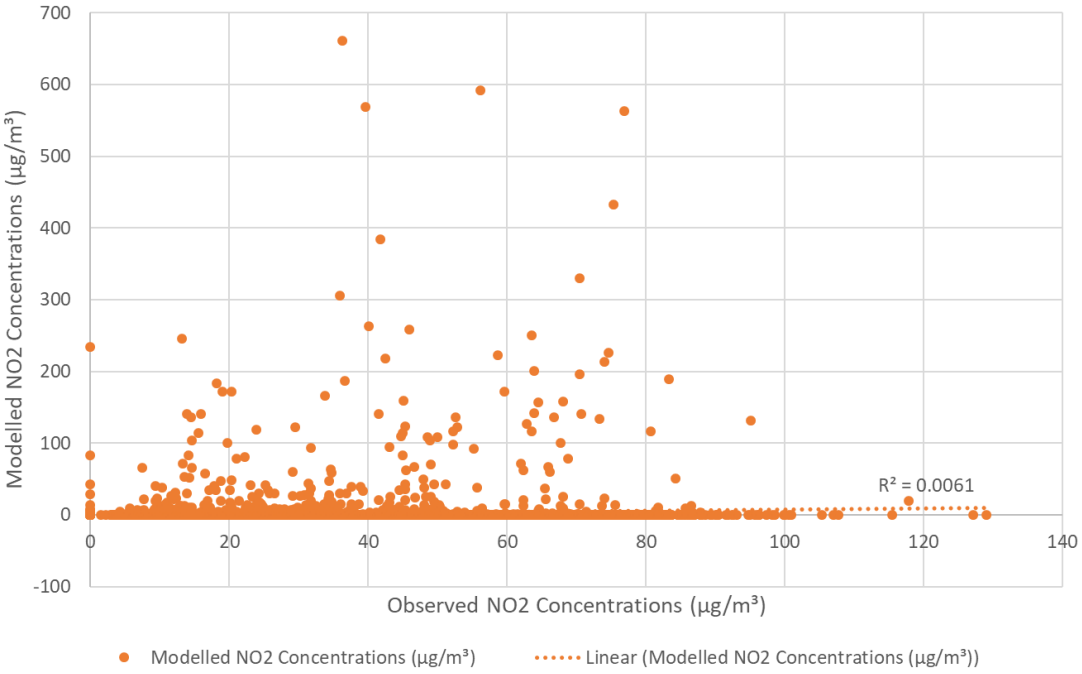

Figure 4.13. Correlation between observed and modeled concentrations in Queens near road station.

Analysis of daily average observed and modeled NO_2 concentrations of both Queens College and Queens near road air quality stations is presented in Figure 4.15 and Figure 4.16, respectively. Contrary to the hourly concentrations, modeled data are lower than the observed data as expected. In Figure 4.16, the correlation between observed and modeled data in both stations are given, again measured data contains concentrations from every other source, therefore, the correlation coefficient R^2 values are low.

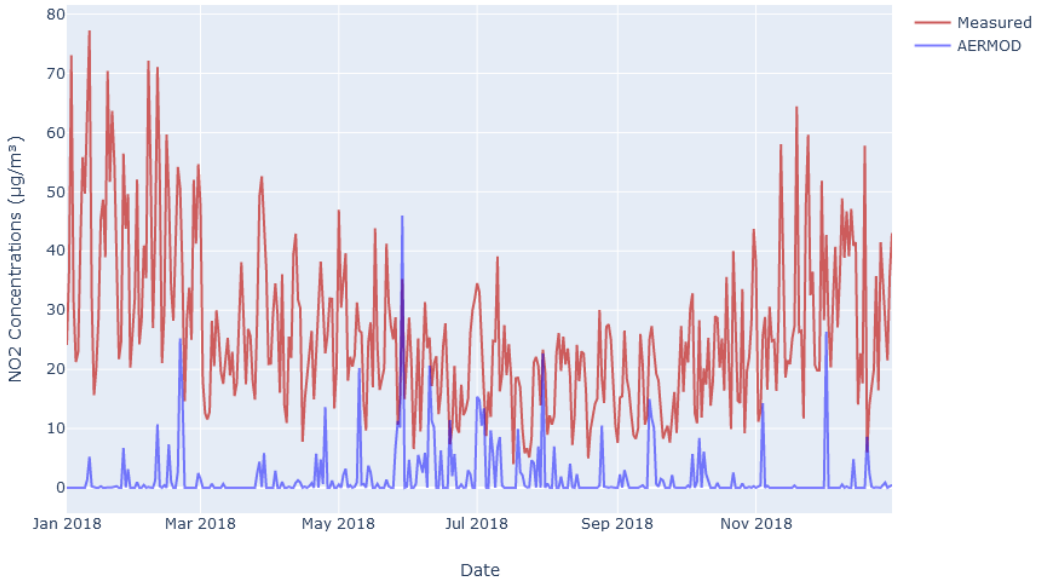

Figure 4.14. Daily average observed and modeled concentrations at Queens College.

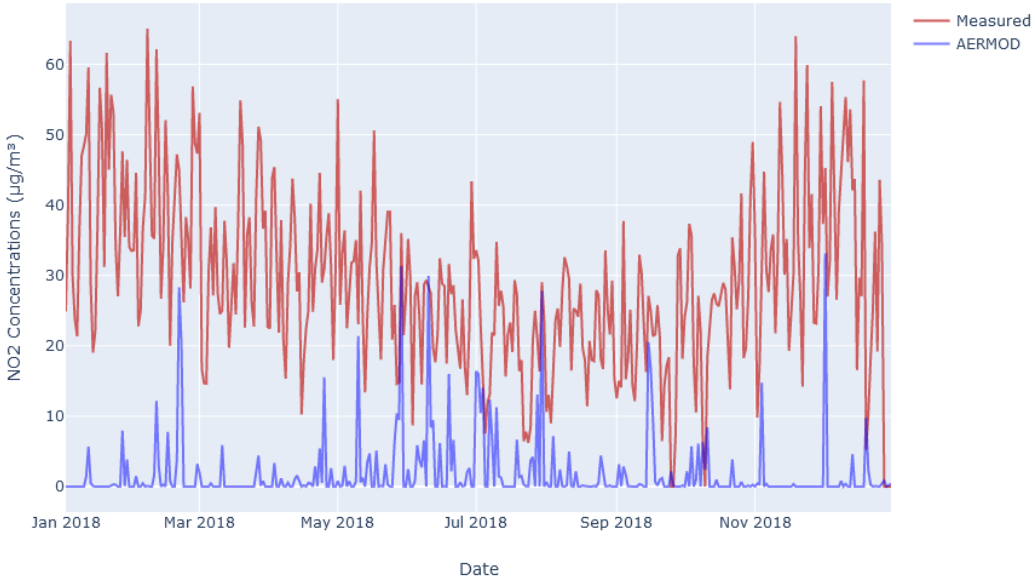

Figure 4.15. Daily average observed and modeled concentrations at Queens near road.

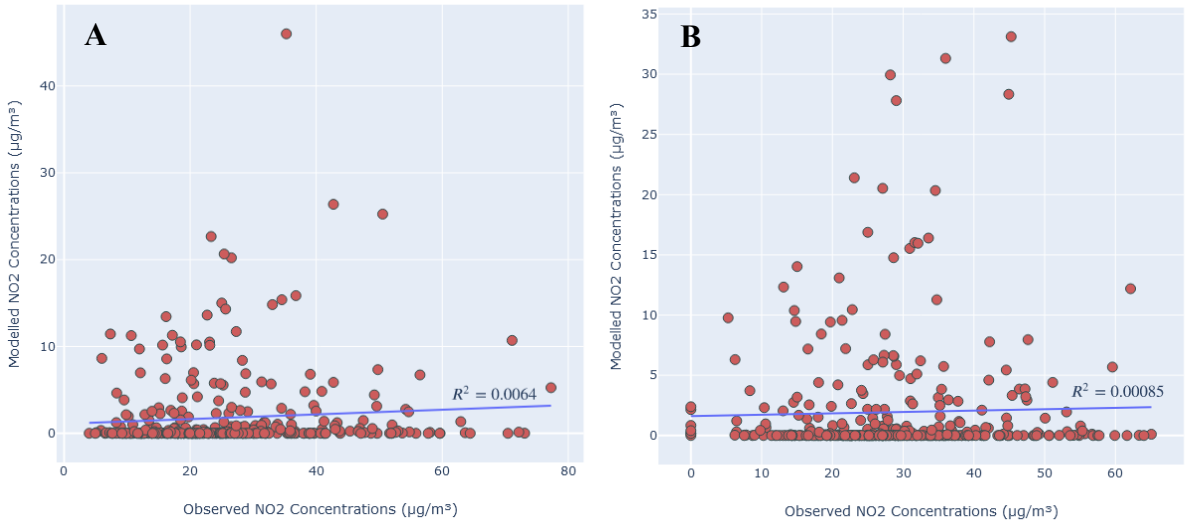

Figure 4.16. Correlation between daily average observed and modeled concentrations in A) Queens College and B) Queens near road.

4.3. Machine Learning Modelling

As noted above, hourly emission results for the 51 receptors displayed in Figure 3.4 were obtained from the atmospheric dispersion model. These 51 receptors, located at 4000 m spacing, are a subset of the receptors used in the AERMOD simulations. Two of the 51 receptors, Queens College and Queens Near Road, are separated as evaluation and test set; the remaining 49 receptors, which include 429,240 data points, are used as the training set in the machine learning model. During training, the model use input data and Eq. 6 to generate predictions. The generated results are subsequently compared with the AERMOD concentrations. The agreement between the two sets is evaluated in terms of the mean absolute error (MAE), the root-mean-square of the error (RMSE) and the negative log-likelihood. Eq. 10 is used to select the tree which reduced the loss the most. When training is performed, it is seen that the error in both training and evaluation datasets decreased with a correlation. Early stopping stepped in to avoid overfitting when there are 50 steps with no improvements and stopped the training in the 83150th step. Negative log-likelihood and RMSE of both training and evaluation sets are decreased and converged during the training as illustrated in Figure 4.17 and Figure 4.18, respectively.

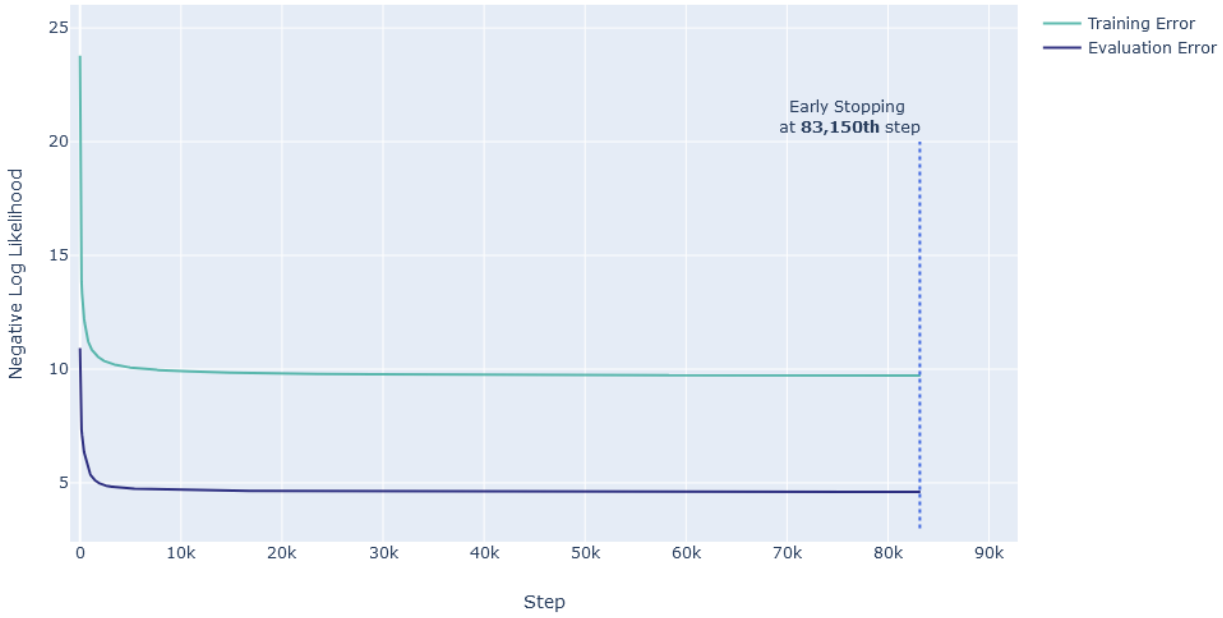

Figure 4.17. Negative log-likelihood errors of training and evaluation during model training.

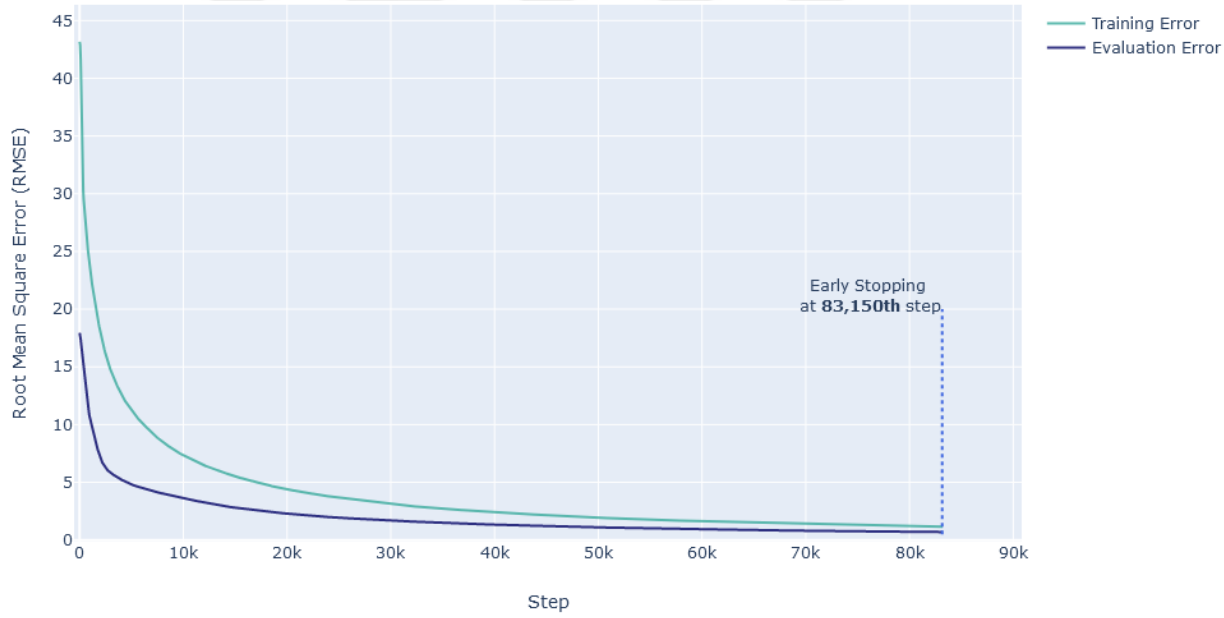

Figure 4.18. Root mean square error (RMSE) of training and evaluation during model training.

After the training step, error calculations for both evaluation and test datasets were made and were found to be close to each other in terms of error metrics, negative log-likelihood, and RMSE. The evaluation and test errors were computed as 4.60 and 4.92 for negative log-likelihood, and 0.70 and 0.78 for RMSE, respectively. A plot of prediction and label of the training dataset is given in Figure 4.19. Orange lines show true concentration values at the given data points, while blue lines are the predictions made by the model. The y-axis represents every single hourly emission data point from 49 receptors. It can be seen that the machine learning model achieved to learn pollutant concentration correlations at the 49 receptors distributed around the airport.

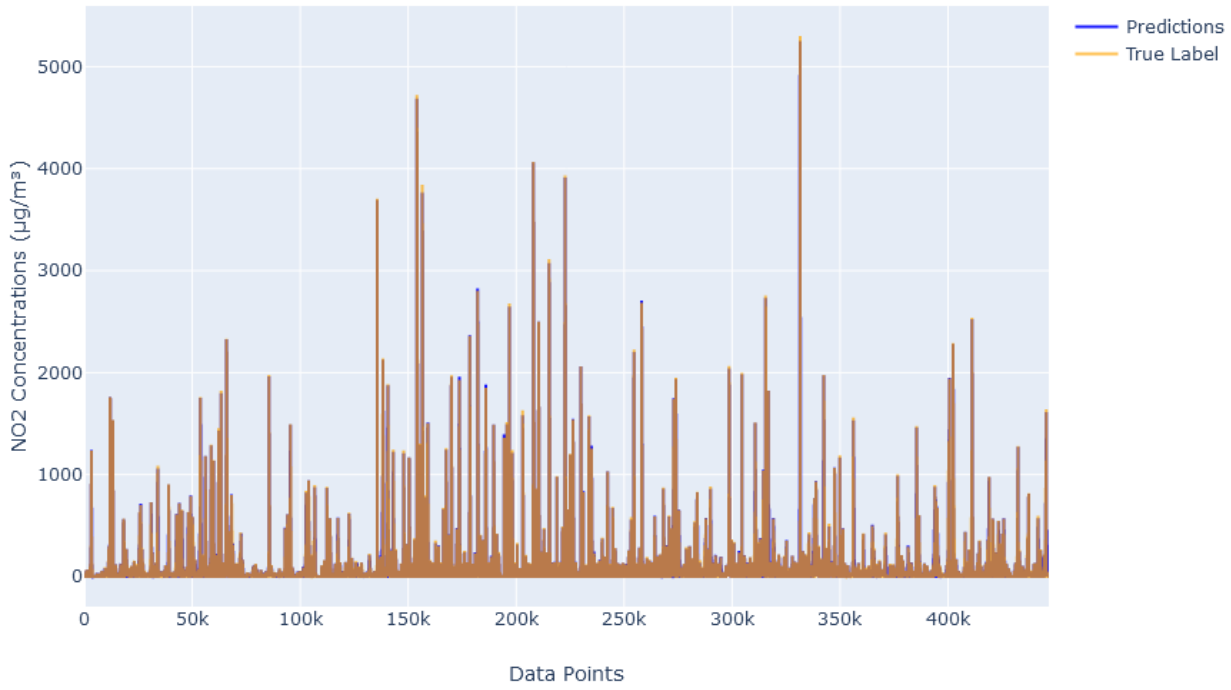

Figure 4.19. Training results of the model.

The agreement of the training results is not sufficient to evaluate the quality of the machine learning model as the model might memorize the dataset and this can lead to overfitting. In order to check whether the model generalizes well in other datasets, predictions on Queens College and Queens Near Road are made. Both Figure 4.20 and Figure 4.21 show that the model is trained to generalize on datasets that it did not see before. Despite the fact that the model is applied to highly turbulent data like emission dispersion and meteorology together, it is seen that it is capable of producing small deviations while predicting at some points.

Further improvements can be made by gathering more than 1 year of air traffic data and/or data from more than 1 airports to let the model learn how emissions are dispersed under different geographical and meteorological conditions. On the other hand, for the labeling of the dataset, more advanced atmospheric models like CALPUFF or FAA's modified AERMOD model, AEDT, which can more accurately simulate the three-dimensional wind distribution, can be used (Kenney, 2017; Scire, 2000). Also, more complex neural network models like recurrent neural networks can be used to learn emission dispersion with respect to change in time (Arsie et al., 2010).

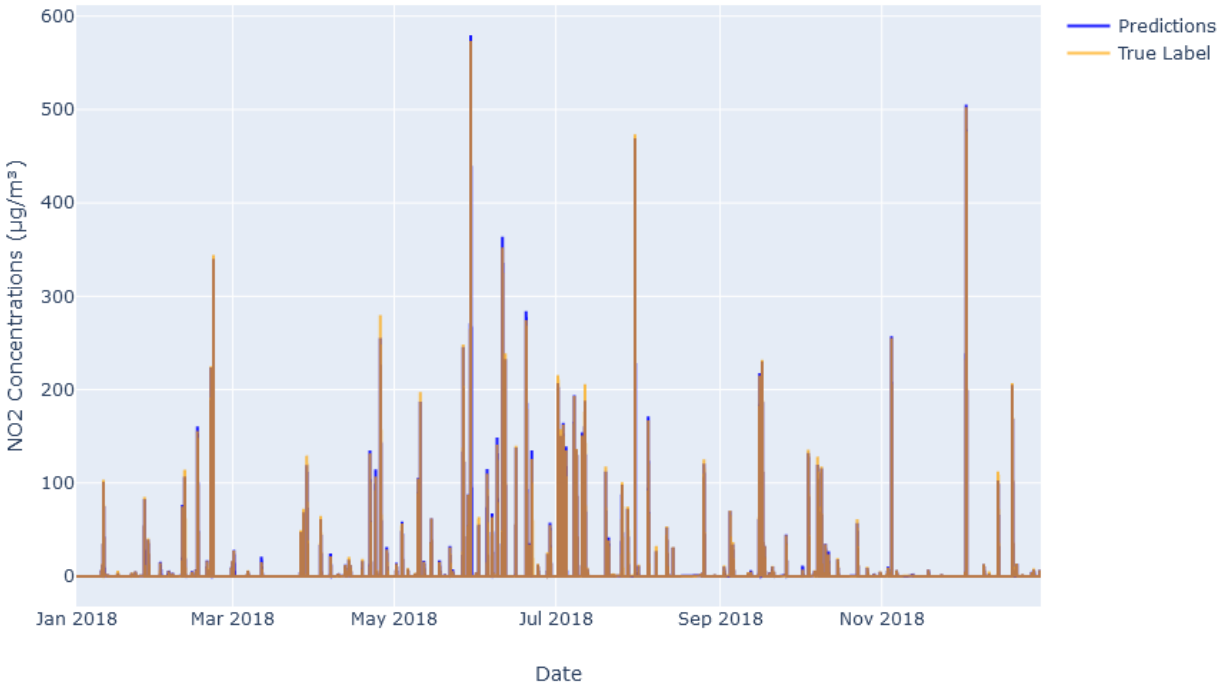

Figure 4.20. Predictions on Queens College datasets with its true labels.

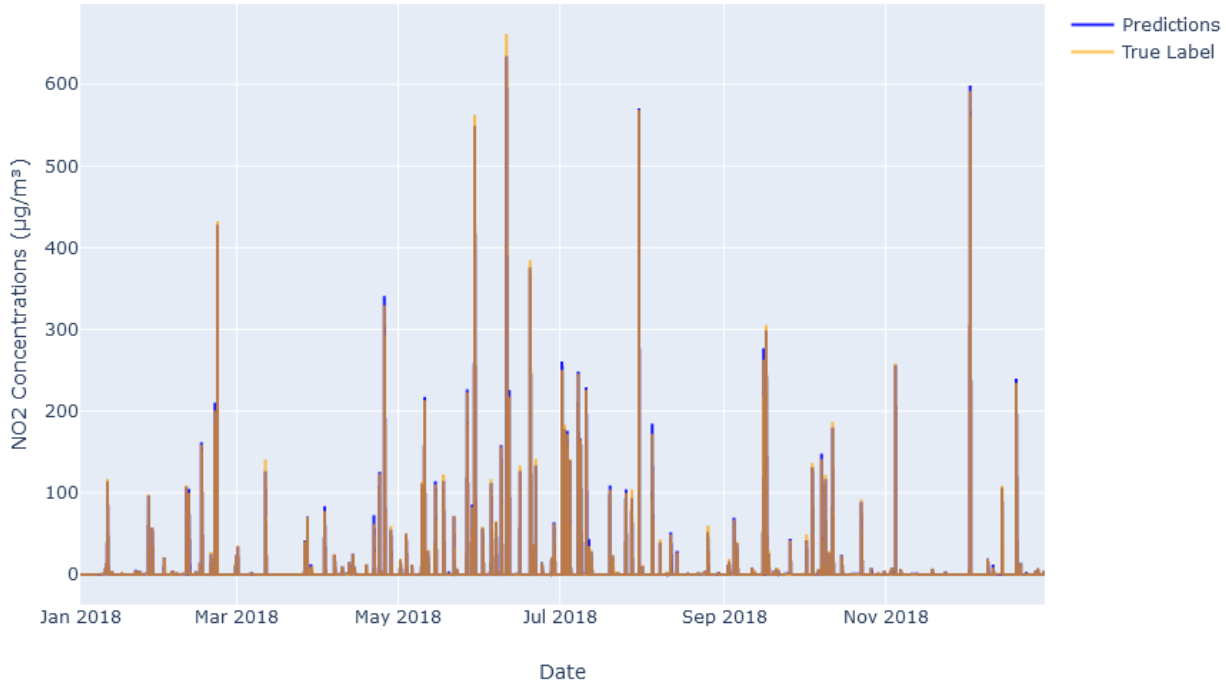

Figure 4.21. Predictions on Queens Near Road datasets with its true labels.

Model performance can also be evaluated by looking into the correlation between predicted and true values in training, evaluation and test datasets. In Figure 4.22, it is illustrated that all three datasets have an R^2 value above 0.99, which means that the model almost perfectly predicted true values with the corresponding input data.

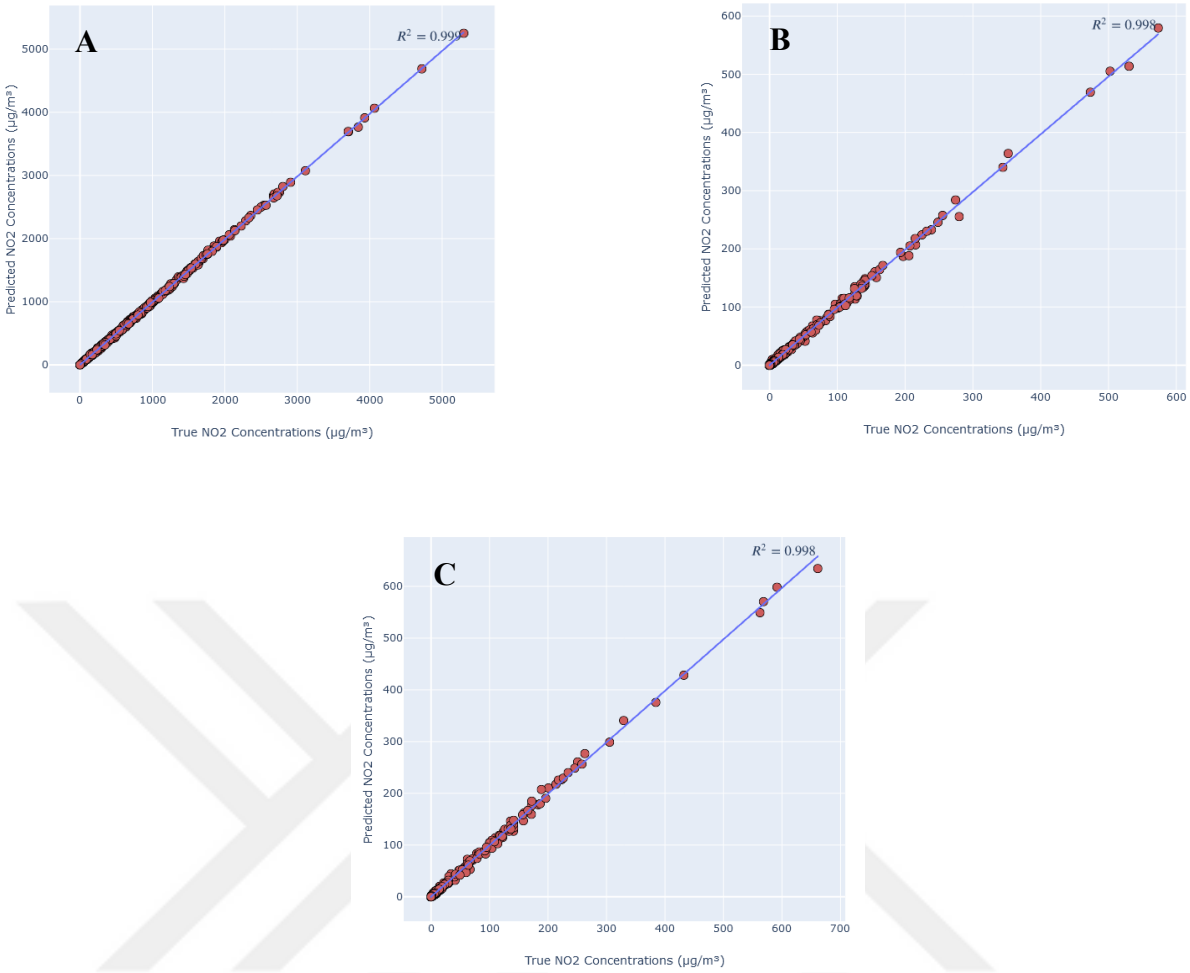

Figure 4.22. Correlation between predicted and true concentration in A) training, B) evaluation and C) test datasets.

Figure 4.23 presents the feature importance scores which indicates which features (or variables) are most useful and has more value when building boosted decision trees during the model training. The relative importance of a feature increases depending on whether it contributes to making an important decision in the decision trees. In order to compare and rank features, importance value for every feature in the dataset is calculated. The feature importance was given by XGBoost at the end of the training. Based on this figure, it can be said that wind direction and magnitude along with NO_x source have the most dominant effect on predicting NO_2 concentrations at the given location. On the other hand, precipitation is by far the less important feature.

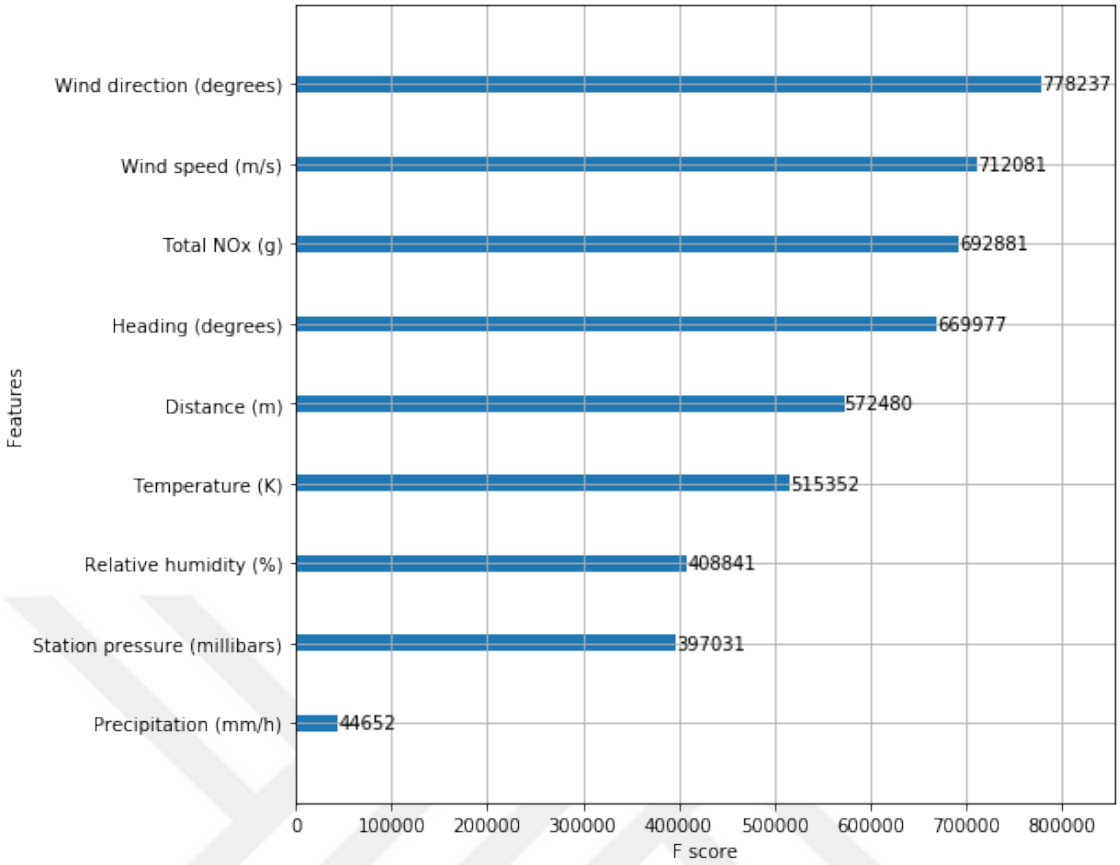

Figure 4.23. Feature importance values of every attribute in the trained model.

5. CONCLUSIONS

Over the years, there has been an increasing trend in air pollution caused by anthropogenic activities like industries, household heating, landfills, and transportations. Numerous pollutants, NO_2, SO_2, CO, O_3 $PM_{2.5-10}$, and Pb, which are released into the atmosphere from these activities are identified as major pollutants by health and environmental organizations like WHO, EPA, and EEA. Emitted pollutants are dispersed through the atmosphere to wide areas primarily due to meteorological conditions, which can lead to the contamination of the air, water and food production at large distances from the release point. Therefore, inhaling or ingesting the pollutants can result in serious adverse health effects, such as cardiovascular, cerebrovascular and respiratory diseases. Some of these pollutants are also carcinogenic. (Paul, 2019).

One of the important sources of urban air pollution is transportation, which includes land, air, and marine transport. Even though there are significant efforts to lower the pollution levels of major pollutants through advanced technologies and stricter regulations and standards, aircraft emissions continue to increase due to increasing air traffic demand since the regulations on aircraft pollutions are not as strict as road vehicles. Due to the high air traffic in metropolitan areas, aircraft result in a high amount of emissions during the LTO cycle. The latest flight counts in the world per year are around 300 million; at JFK Airport only, over 450,000 flights took place in 2018. Due to non-ideal combustion, aircraft result in a high concentration of $PM_{2.5}$ and NO_x emissions.

Atmospheric dispersion modeling gained significant importance year after year with models like AERMOD and CALPUFF. These and similar models have undergone significant improvements in recent years to reflect the latest scientific developments in the field. These models can be readily used to compute the impacts of different sources on air quality. The focus of this study is to estimate the NO_2 emissions from air traffic at JFK Airport and the resulting impact on air quality. To achieve this goal, air traffic data in JFK Airport for the year 2018 were gathered from air traffic surveillance platform called OpenSKY, in order to calculate NO_2 emissions. The emissions were calculated using emission factors gathered from environmental and aviation agencies like EEA, FAA, and EASA. The emissions data were then used as input in AERMOD to calculate NO_2 concentrations in the vicinity of the JFK airport. From the results of the AERMOD dispersion model, it is seen that the hourly NO_2 air quality standard is exceeded at 853 of evenly distributed 2403 receptors distributed within the airport and the surrounding area. Moreover, the annual average concentrations exceeded the limit at 74 receptors. It is important to note that these exceedances are due to a single source of NO_x, aircraft

emissions during the LTO cycle. There are also various sources of NO_x such as heating and vehicular emissions that would contribute further to NO_2 concentrations in the atmosphere.

When the calculated hourly emissions of two air quality stations are compared with the measurements, it is seen as the dispersion model calculated high values at some points. This is attributed to the lack of detailed meteorological data at a fine grid and some limitations of defining land-use types in AERMET, the meteorological pre-processor of AERMOD. Another limitation of AERMOD to calculate accurate concentrations on farther points is that it uses steady and uniform hourly wind distribution, instead of a fully transient 3-dimensional wind profile. Further investigations on improving the model results will be made in future studies.

Hourly emission results for 51 receptors from AERMOD were used alongside with the meteorology data in XGBoost, a decision tree based gradient boosting model. A model trained on the data of 49 receptors while it is evaluated and tested by the other 2 receptors. Training results show that the model can learn to make a correlation between emissions at the airport, meteorology and dispersion model results. Meanwhile, the model's predictions on evaluation and test datasets show that even though the model generalizes well, there is still room for improvement. The advantage of the machine learning models is that they can potentially predict air pollution concentrations when sufficient data are available from meteorological stations and air dispersion models.

For future studies, advanced atmospheric dispersion models like CALPUFF or AEDT (Aviation Environmental Design Tool), or even more advanced Eulerian models like CMAQ (Community Multiscale Air Quality Model) or CAMx (Comprehensive Air Quality Model with Extensions) coupled with a meteorological forecasting model like WRF (Weather Research and Forecasting Model) should be used to get more accurate labels for machine learning modeling.

Furthermore, increasing data variety by increasing time interval and location may increase the model's performance by adapting it to different geographical and meteorological conditions. Moreover, similar modeling efforts can also be applied to other airports such as the new Istanbul airport, which is one of the busiest airports in the world. Future studies can develop and test machine learning models that also include more pollution sources from industry, land transport, and heating. Last but not least, using advanced neural network models, like a recurrent neural network that is favored by sequential data, could also be examined in future studies to obtain improved model performance.

REFERENCES

Aporras. 2016. What is the difference between Bagging and Boosting? Retrieved from https://quantdare.com/what-is-the-difference-between-bagging-and-boosting/

Arsie, I., Pianese, C., & Sorrentino, M. 2010. Development of recurrent neural networks for virtual sensing of NOx emissions in internal combustion engines. *SAE International Journal of Fuels and Lubricants*, *2*(2), 354–361.

Arya, S. P. 1999. *Air pollution meteorology and dispersion* (Vol. 6). Oxford University Press New York.

AvDelphi. n.d.. AvDelphi - One place for all your aviation data requirements. Retrieved September 20, 2008, from https://www.avdelphi.com/

Bajoghli, M., Abari, M., & Radnezhad, H. 2016. Dispersion Modeling of Total Suspended Particles (TSP) Emitted from a Steel Plant at Different Time Scales Using AERMOD View. *Journal of Earth, Environment and Health Sciences*, *2*(2), 77. https://doi.org/10.4103/2423-7752.191399

Baldasano, J. M., Valera, E., & Jiménez, P. 2003. Air quality data from large cities. *Science of the Total Environment*, *307*(1–3), 141–165. https://doi.org/10.1016/S0048-9697(02)00537-5

Bojarski, M., Del Testa, D., Dworakowski, D., Firner, B., Flepp, B., Goyal, P., … others. 2016. End to end learning for self-driving cars. *ArXiv Preprint ArXiv:1604.07316*.

Bonat, W. H., & Jørgensen, B. 2016. Multivariate covariance generalized linear models. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *65*(5), 649–675.

Breiman, L. 2001. Random forests. *Machine Learning*, *45*(1), 5–32.

Brimblecombe, P. 2006. The Clean Air Act after 50 years. *Weather*, *61*(11), 311–314. https://doi.org/10.1256/wea.127.06

Brunekreef, B., & Holgate, S. T. 2002. Air pollution and health. *Lancet*, *360*(9341), 1233–1242. https://doi.org/10.1016/S0140-6736(02)11274-8

Carslaw, D. C., Beevers, S. D., Ropkins, K., & Bell, M. C. 2006. *Detecting and quantifying aircraft and other on-airport contributions to ambient nitrogen oxides in the vicinity of a large international airport*. *40*(December 2003), 5424–5434. https://doi.org/10.1016/j.atmosenv.2006.04.062

Chen, T., & Guestrin, C. 2016. Xgboost: A scalable tree boosting system. *Proceedings of the 22nd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining*, 785–794.

Chen, T. M., Gokhale, J., Shofer, S., & Kuschner, W. G. 2007. Outdoor air pollution: Nitrogen dioxide, sulfur dioxide, and carbon monoxide health effects. *American Journal of the Medical Sciences*, *333*(4), 249–256. https://doi.org/10.1097/MAJ.0b013e31803b900f

Chugh, J. 2018. Types of Machine Learning and Top 10 Algorithms Everyone Should Know. Retrieved from https://blogs.oracle.com/ai/types-of-machine-learning-and-top-10-algorithms-everyone-should-know

Cimorelli, A. J., Perry, S. G., Venkatram, A., Weil, J. C., Paine, R. J., & Peters, W. D. 1998. *AERMOD--Description of model formulation*.

Clark, L. A., & Pregibon, D. 2017. Tree-based models. In *Statistical models in S* (pp. 377–419). Routledge.

EASA. 2019. ICAO Aircraft Engine Emissions Databank. Retrieved September 20, 2009, from https://www.easa.europa.eu/easa-and-you/environment/icao-aircraft-engine-emissions-databank

EEA. 2019. *EMEP/EEA air pollutant emission inventory guidebook 2019*. https://doi.org/10.2800/293657

Elbir, T. 2008. Estimation of engine emissions from commercial aircraft at a midsized Turkish airport. *Journal of Environmental Engineering*, *134*(3), 210–215. https://doi.org/10.1061/(ASCE)0733-9372(2008)134:3(210)

EPA. 2008. *AERSURFACE User's Guide*.

EPA. 2013. *Ambient Ratio Method Version 2(ARM2) for use with AERMOD for1-hr NO2 Modeling*.

EPA. 2015. *AERMINUTE User's Guide*.

EPA. 2018. *User's Guide for the AERMOD Terrain Preprocessor (AERMAP).*

EPA. 2019a. *User's Guide for the AERMOD Meteorological Preprocessor (AERMET).*

EPA. 2019b. *User's Guide for the AMS/EPA Regulatory Model (AERMOD)* (No. EPA-454/B-19-027).

FAA. 2018a. Aircraft Registry. Retrieved September 20, 2009, from https://www.faa.gov/licenses_certificates/aircraft_certification/aircraft_registry/releasable_aircraft_download/

FAA. 2018b. New York - John F. Kennedy International Airport. Retrieved September 20, 2010, from https://www.faa.gov/nextgen/snapshots/airport/?locationId=34

Fortmann-Roe, S. 2012. Understanding the Bias-Variance Tradeoff. Retrieved from http://scott.fortmann-roe.com/docs/BiasVariance.html

Gauderman, W. J., Avol, E., Gilliland, F., Vora, H., Thomas, D., Berhane, K., … others. 2004. The effect of air pollution on lung development from 10 to 18 years of age. *New England Journal of Medicine*, *351*(11), 1057–1067.

Geurts, P. 2002. *Contributions to decision tree induction: bias/variance tradeoff and time series classification*. University of Liège Belgium.

Gilchrist, R., & Drinkwater, D. 2000. The use of the Tweedie distribution in statistical modelling. *COMPSTAT*, 313–318.

Giussani, V. 1994. *The UK clean air act 1956: an empirical investigation*. CSERGE Norwich.

Hallinan, B., & Striphas, T. 2016. Recommended for you: The Netflix Prize and the production of algorithmic culture. *New Media & Society*, *18*(1), 117–137.

Harrison, R. M., Masiol, M., & Vardoulakis, S. 2015. Civil aviation, air pollution and human health. *Environmental Research Letters*, *10*(4), 8–11. https://doi.org/10.1088/1748-9326/10/4/041001

Hasan, M. M., & Dunn, P. K. 2012. Understanding the effect of climatology on monthly rainfall amounts in Australia using Tweedie GLMs. *International Journal of Climatology*, *32*(7), 1006–1017. https://doi.org/10.1002/joc.2332

Hu, R., Zhu, S., Feng, J., & Sears, A. 2011. Use of speech technology in real life environment. *International Conference on Universal Access in Human-Computer Interaction*, 62–71.

ICAO. 2011. *Airport Air Quality Manual*. International Civil Aviation Organization Montréal.

ICAO. 2019. *The ICAO Environmental Report 2019*.

Inal, F. 2010. Artificial neural network prediction of tropospheric ozone concentrations in Istanbul, Turkey. *CLEAN--Soil, Air, Water*, *38*(10), 897–908.

Jain, R. K., Cui, Z. "Cindy," & Domen, J. K. 2016. Chapter 4 - Environmental Impacts of Mining. In R. K. Jain, Z. "Cindy" Cui, & J. K. Domen (Eds.), *Environmental Impact of Mining and Mineral Processing* (pp. 53–157). https://doi.org/https://doi.org/10.1016/B978-0-12-804040-9.00004-8

Jørgensen, B. 1987. Exponential dispersion models. *Journal of the Royal Statistical Society: Series B (Methodological)*, *49*(2), 127–145.

Kendal, W. S. 2004. A scale invariant clustering of genes on human chromosome 7. *BMC Evolutionary Biology*, *4*(1), 3.

Kendal, W. S., Lagerwaard, F. J., & Agboola, O. 2000. Characterization of the frequency distribution for human hematogenous metastases: Evidence for clustering and a power variance function. *Clinical & Experimental Metastasis*, *18*(3), 219–229.

Kenney, M. 2017. Development of a NOx Chemistry Module for EDMS/AEDT to Predict NO2 Concentrations. In *Development of a NOx Chemistry Module for EDMS/AEDT to Predict NO2 Concentrations*. https://doi.org/10.17226/24706

Kimbrough, S., Owen, R. C., Snyder, M., & Richmond-Bryant, J. 2017. NO to NO2 conversion rate analysis and implications for dispersion model chemistry methods using Las Vegas, Nevada near-road field measurements. *Atmospheric Environment*, *165*, 23–34.

Kuzu, S. L. 2018. Estimation and dispersion modeling of landing and take-off (LTO) cycle emissions from Atatürk International Airport. *Air Quality, Atmosphere and Health*, *11*(2), 153–161. https://doi.org/10.1007/s11869-017-0525-5

Last, J. A., Sun, W. M., & Witschi, H. 1994. Ozone, NO, and NO2: Oxidant air pollutants and more. *Environmental Health Perspectives*, *102*(SUPPL. 10), 179–184.

Li, L. 2019. Classification and Regression Analysis with Decision Trees. Retrieved from https://towardsdatascience.com/https-medium-com-lorrli-classification-and-regression-analysis-with-decision-trees-c43cdbc58054

Lin, H., Zhu, J., Xu, B., Lin, W., & Hu, Y. 2009. A virtual geographic environment for a simulation of air pollution dispersion in the Pearl River Delta (PRD) region. In *3D Geo-Information Sciences* (pp. 3–13). Springer.

Liu, C., Tsow, F., Zou, Y., & Tao, N. 2016. Particle pollution estimation based on image analysis. *PloS One*, *11*(2), e0145955.

Mage, D., Ozolins, G., Peterson, P., Webster, A., Orthofer, R., Vandeweerd, V., & Gwynne, M. 1996. Urban air pollution in megacities of the world. *Atmospheric Environment*, *30*(5), 681–686. https://doi.org/10.1016/1352-2310(95)00219-7

Makridis, M., & Lazaridis, M. 2019. Dispersion modeling of gaseous and particulate matter emissions from aircraft activity at Chania Airport, Greece. *Air Quality, Atmosphere and Health*, 933–943. https://doi.org/10.1007/s11869-019-00710-y

Mayer, H. 1999. Air pollution in cities. *Atmospheric Environment*, *33*(24–25), 4029–4037. https://doi.org/10.1016/S1352-2310(99)00144-2

Mazaheri, M., Johnson, G. R., & Morawska, L. 2009. *Particle and Gaseous Emissions from Commercial Aircraft at Each Stage of the Landing and Takeoff Cycle*. 441–446.

Mitchell, Tom M. 1997. Machine learning. 1997. In *Burr Ridge, IL: McGraw Hill* (Vol. 45).

Mitchell, Tom Michael. 2006. *The discipline of machine learning* (Vol. 9). Carnegie Mellon University, School of Computer Science, Machine Learning~….

Morde, V. 2019. XGBoost Algorithm: Long May She Reign! Retrieved from https://towardsdatascience.com/https-medium-com-vishalmorde-xgboost-algorithm-long-she-may-rein-edd9f99be63d

MRLC. n.d.. National Land Cover Dataset (NCDC). Retrieved from https://www.mrlc.gov/viewer/

Natekin, A., & Knoll, A. 2013. Gradient boosting machines, a tutorial. *Frontiers in Neurorobotics*, *7*, 21. https://doi.org/10.3389/fnbot.2013.00021

NOAA. n.d.. National Climatic Data Center. Retrieved from https://www.ncdc.noaa.gov/isd/data-access

Paul, K. C., Haan, M., Mayeda, E. R., & Ritz, B. R. 2019. Ambient Air Pollution, Noise, and Late-Life Cognitive Decline and Dementia Risk. *Annual Review of Public Health*, *40*(1), 203–220. https://doi.org/10.1146/annurev-publhealth-040218-044058

Peden, D. B. 2008. Air pollution: indoor and outdoor. In *Middleton's Allergy: Principles and Practice* (7th ed., pp. 495–508). Elsevier.

Pelliccioni, A., & Tirabassi, T. 2006. Air dispersion model and neural network: A new perspective for integrated models in the simulation of complex situations. *Environmental Modelling and Software*, *21*(4), 539–546. https://doi.org/10.1016/j.envsoft.2004.07.015

Planespotters.net. n.d.. Civil aviation database.

Podrez, M. 2015. An update to the ambient ratio method for 1-hNO2 air quality standards dispersion modeling. *Atmospheric Environment*, *103*(2), 163–170. https://doi.org/10.1016/j.atmosenv.2014.12.021

Prasad, A. M., Iverson, L. R., & Liaw, A. 2006. Newer classification and regression tree techniques: bagging and random forests for ecological prediction. *Ecosystems*, *9*(2), 181–199.

Rood, A. S. 2014. Performance evaluation of AERMOD, CALPUFF, and legacy air dispersion models using the Winter Validation Tracer Study dataset. *Atmospheric Environment*, *89*, 707–720. https://doi.org/10.1016/j.atmosenv.2014.02.054

Ruijgrok, G. J. J., & Van Paassen, D. M. 2005. *Elements of aircraft pollution*. Delft University Press The Netherlands.

Russell, S. J., & Norvig, P. 2016. *Artificial intelligence: a modern approach*. Malaysia; Pearson Education Limited,.

Schäfer, M., Strohmeier, M., Lenders, V., Martinovic, I., & Wilhelm, M. 2014. Bringing up OpenSky: A large-scale ADS-B sensor network for research. *IPSN 2014 - Proceedings of the 13th International Symposium on Information Processing in Sensor Networks (Part of CPS Week)*, 83–94. https://doi.org/10.1109/IPSN.2014.6846743

Schapire, R. E. 2003. The boosting approach to machine learning: An overview. In *Nonlinear estimation and classification* (pp. 149–171). Springer.

Scire, J. S., Robe, F. R., Fernau, M. E., & Yamartino, R. J. 2000. A user's guide for the CALMET Meteorological Model. *Earth Tech, USA*, *37*.

Scire, J. S., Strimaitis, D. G., & Yamartino, R. J. 2000. A user's guide for the CALPUFF dispersion model. *Earth Tech, Inc. Concord, MA*, *10*.

Smyth, G. K., & Jørgensen, B. 2002. Fitting Tweedie's compound Poisson model to insurance claims data: dispersion modelling. *ASTIN Bulletin: The Journal of the IAA*, *32*(1), 143–157.

Sze, V., Chen, Y.-H., Yang, T.-J., & Emer, J. S. 2017. Efficient processing of deep neural networks: A tutorial and survey. *Proceedings of the IEEE*, *105*(12), 2295–2329.

Taieb, S. Ben, & Hyndman, R. J. 2014. A gradient boosting approach to the Kaggle load forecasting competition. *International Journal of Forecasting*, *30*(2), 382–394.

Thron, R. W. 1996. Direct and indirect exposure to air pollution. *Otolaryngology - Head and Neck Surgery*, *114*(2), 281–285. https://doi.org/10.1016/S0194-5998(96)70184-5

Tian, Y., Huang, W., Ye, B., & Yang, M. 2019. A New Air Quality Prediction Framework for Airports Developed with a Hybrid Supervised Learning Method. *Discrete Dynamics in Nature and Society*, *2019*.

Tiwary, A., & Williams, I. 2018. *Air pollution: measurement, modelling and mitigation*. CRC Press.

USGS. 2018. The National Map. Retrieved from https://viewer.nationalmap.gov/basic/

Wayson, R. L., Kim, B., Fleming, G. G., Hall, C., Thrasher, T., Colligan, B., … others. 2003. *Integration of AERMOD into EDMS*.

WHO. 2006. *WHO Air quality guidelines for particulate matter, ozone, nitrogen dioxide and sulfur dioxide: global update 2005: summary of risk assessment*.

Wilkins, E. T. 1954. Air pollution and the London fog of December, 1952. *Journal of the Royal Sanitary Institute*, *74*(1), 1–21.

Wormhoudt, J., Herndon, S. C., Yelvington, P. E., Miake-Lye, R. C., & Wey, C. 2007. Nitrogen oxide (NO/NO 2/HONO) emissions measurements in aircraft exhausts. *Journal of Propulsion and Power*, *23*(5), 906–911. https://doi.org/10.2514/1.23461

Xi, X., Wei, Z., Xiaoguang, R., Yijie, W., Xinxin, B., Wenjun, Y., & Jin, D. 2015. A comprehensive evaluation of air pollution prediction improvement by a machine learning method. *2015 IEEE International Conference on Service Operations And Logistics, And Informatics (SOLI)*, 176–181.

Yamaguchi, O. 2012. Face recognition technology and its real-world application. *Indo-Japanese Conference on Perception and Machine Intelligence*, 28–34.

Yang, X., Cheng, S., Lang, J., Xu, R., & Lv, Z. 2018. Characterization of aircraft emissions and air quality impacts of an international airport. *Journal of Environmental Sciences*, 1–10. https://doi.org/10.1016/j.jes.2018.01.007

Yu, R., Yang, Y., Yang, L., Han, G., & Move, O. 2016. RAQ--a random forest approach for predicting air quality in urban sensing systems. *Sensors*, *16*(1), 86.

Zhang, C., & Ma, Y. 2012. *Ensemble machine learning: methods and applications*. Springer.