HUNGARIAN UNIVERSITY OF AGRICULTURE AND LIFE SCIENCES

THE VARIOUS TRANSMISSION AND ESTIMATION OF PM POLLUTANTS

Thesis of the doctoral (PhD) dissertation

Qor-el-aine Achraf

Gödöllő - Hungary

2023

Doctoral school denomination: Doctoral School of Mechanical Engineering

Science: Mechanical Engineering – Environmental Engineering

Head of school: Prof. Dr. Gábor Kalácska, DSc
Institute of Technology
Hungarian University of Agriculture and Life Science, Gödöllő - Hungary

Supervisor: Dr. Gábor Géczi, PhD
Institute of Environmental Sciences
Hungarian University of Agriculture and Life Science, Gödöllő - Hungary

Co-Supervisor: Dr. András Béres, PhD
University Laboratory Center, Head of Szent István Campus
Hungarian University of Agriculture and Life Science, Gödöllő – Hungary

....
.

Affirmation of head of school Affirmation of supervisor

CONTENTS

1 INTRODUCTION 5

1.1 Relevance and significance of the topic 5

1.2 Objectives 6

2 MATERIALS AND METHODS 7

2.1 PM dispersion experiments 7

2.1.1 Small scale experiments of PM10 dispersion around obstacles 7

2.1.2 Effect of small hills on PM10 and PM2.5 concentrations in short range 9

2.2 Saharan Dust storm transport 10

2.2.1 Dust Storm simulation over the Sahara Desert (Moroccan and Mauritanian regions) using HYSPLIT 10

2.2.2 The identification and evaluation of the Saharan dust storm events in Budapest, Hungary between 2018 and 2022 17

2.2.3 Case study of the Saharan dust effects on PM10 and PM2.5 concentrations in Budapest in March 2022 17

2.3 Estimation and evaluation of PM concentrations 19

2.3.1 Evaluation of PM surface concentrations simulated by Version 5.12.4 of NASA's MERRA-2 Aerosol Reanalysis over Hungary in the period between 2019 and 2021 19

2.3.2 Calibration of CAMS PM2.5 data over Hungary using machine learning 23

2.4 Data and statistics 26

2.4.1 The MERRA-2 Aerosol Reanalysis (MERRAero) 26

2.4.2 Air quality stations 27

2.4.3 Performance statistics 29

3 RESULTS 30

3.1 PM dispersion experiments 30

3.1.1 Small scale experiments of PM10 dispersion around obstacles 30

3.1.2 Effect of small hills on PM10 and PM2.5 concentrations in short range 34

3.2 Saharan Dust storm transport 37

3.2.1 Dust Storm simulation over the Sahara Desert (Moroccan and Mauritanian regions) using HYSPLIT 37

3.2.2 The identification and evaluation of the Saharan dust storm events in Budapest, Hungary between 2018 and 2022 45

3.2.3 Case study of the Saharan dust effects on PM10 and PM2.5 concentrations in Budapest in March 2022 51

3.3 Estimation and evaluation of PM concentrations 53

3.3.1 Evaluation of PM surface concentrations simulated by Version 5.12.4 of NASA's MERRA-2 Aerosol Reanalysis over Hungary in the period between 2019 and 2021 53

3.3.2 Calibration of CAMS PM2.5 data over Hungary using machine learning 59

4 Conclusions and recommendations 64

5 New scientific results 68

6 SUMMARY 74

7 Relevant publications related to the thesis 78

1 INTRODUCTION

The context and goals of this PhD thesis are described in this chapter.

1.1 Relevance and significance of the topic

Air pollution is the primary cause of the decline in air and environmental quality in many places across the world nowadays, with negative consequences for people's health. According to the most recent World Health Organization (WHO) report, more than 91% of people in urban areas are exposed to air quality levels that exceed the emission limits for air pollution (World Health Organization, 2021). Carbon monoxide (CO), particulate matter (PM), nitrogen oxides (NOx), volatile organic compounds (VOCs), ozone (O_3), and sulphur dioxide (SO_2) are the primary atmospheric pollutants. The rapid industrialization and urbanization of developing countries have increased the number of pollutants emitted (Fu and Chen, 2017). Because of the strong relationship between air pollution exposure and increased harmful short- and long-term effects on human health, the scientific community, and public opinion are both concerned about the deterioration of air quality in urban environments (Masiol *et al.*, 2014). Aside from the health dangers posed by gas and particle inhalation, urban air pollution causes other issues such as faster corrosion and deterioration of materials, damage to historical monuments and structures, and damage to plants in and around the city (Vlachokostas *et al.*, 2011).

Particulate matter (PM) is a broad word that refers to a mixture of solid particles and liquid droplets (aerosols) whose size and composition change depending on time and place. PM is composed of numerous constituents, including elemental or black carbon (BC) and organic carbon (OC) molecules, sulfate (SO_4^{-2}), nitrate (NO_{3-}), trace metals, soil particles, and sea salt. PM particles are defined based on their size variations. PM particles with a diameter that is less than or equal to 10 μm are called coarse PMs (PM10), and PM with a diameter of less than or equal to 2.5 μm are fine PMs (PM2.5). PM can be directly emitted from anthropogenic or natural sources (i.e., primary PM), or formed in the atmosphere from chemical reactions of numerous gaseous (i.e., secondary PM) (Harrison, Hester and Querol, 2016).

The research of PM pollution is crucial, in order to comprehend the causes and effects of this kind of air pollution and to create practical solutions for lowering exposure and enhancing public health. Researchers employ a range of techniques to evaluate PM pollution, including computer modelling, satellite data, and air quality monitoring stations. These techniques can offer details on the concentration and distribution of PM in various locations as well

as the pollution's origins. Chemical analysis, for instance, may be used to determine the chemical composition of PM samples and link them to particular sources such industrial activities, wildfires, or vehicle emissions.

This research, covers different aspects of the PM pollution, from the evaluation of low-cost PM sensors, the use of low-cost PM sensors in small scale experiments, the effect of dust storms on the PM concentrations in Hungary and how often they occur, and the estimation of the PM10 and PM2.5 concentrations based on Satellite, meteorological and in-situ measurements data.

1.2 Objectives

The primary goals of the present work are the following:

1) Study the effects of a simple environment on PM concentration using PM low-cost sensors.
2) Study the effects of the Saharan Dust storm on PM levels in Hungary and analyse the seasonality and frequency of recent Saharan dust events.
3) Estimating PM concentrations using satellite, meteorological, and in-situ measurement data and machine learning methods over Hungary.

2 MATERIALS AND METHODS

In this chapter I describe the tools and datasets I used, as well as the methods and relationships used in data processing.

This chapter is divided into four parts. First part contains experimental section, which is two subchapters, a small-scale experiment of PM10 dispersion around obstacles, and the effect of small hills on PM10 and PM2.5 concentrations in short range. Second is the part that deals with Saharan dust transport, divided into three subchapters, the Saharan dust event of June 2020, the identification and evaluation of the Saharan dust storm events to Hungary between 2018 and 2022, and the case study of the Saharan dust effects on PM10 and PM2.5 concentrations in Budapest in March 2022. The third section is the estimation of PM concentrations, and contains two subchapters, is the Evaluation of PM surface concentrations simulated by Version 5.12.4 of NASA's MERRA-2 Aerosol Reanalysis over Hungary in the period between 2019 and 2021 using two approaches to estimate PM ground-level concentrations using surface, satellite, and meteorological data based on machine learning algorithms, and, the calibration of CAMS PM2.5 data over Hungary using machine learning. The fourth part is subchapter that describe the common data and statistics that are used throughout the study.

2.1 PM dispersion experiments

2.1.1 Small scale experiments of PM10 dispersion around obstacles

Small scale experiments were conducted in order to investigate the effects of obstacles heights and distance from the source in the PM10 concentration. The goal was to understand the changing of the PM10 concentration around obstacles in simple environment. The experiments were done in isolated laboratory room on built table. The table had 3 PM10 sensors with 50 cm distance between each sensor. The room temperature was stable during the experiments (25 ±1˚C), the same was for the Relative Humidity (RH) (50% ±3).

2.1.1.1 Experiment set up

The experiments were done with two different wind speed (air flow speed of 2.9 and 1 m/s measured by Schiltknecht MiniAir64 vane anemometer) provided by two different ventilators. The use of the ventilators is to make sure that the PM plume will follow the wind direction toward the sensors and to avoid the spreading of plume around the room. As mentioned, three sensors were used, sensor A, B and C as shown in Figure 2.1, sensor C placed near the source, sensor B in the middle and sensor A is 1 meter away from the source.

The obstacle was placed at three different distances between sensors A and B, with changing of the obstacle height (12, 24 and 36 centimetres). The PM sensors used are NOVA PM sensors (SDS011) that use principle of laser scattering to get the particle concentration in the air, with a digital output and built-in fan that is stable and reliable.

The incense sticks were used as a source of PM10 plumes, due to the number of particles emitted from incense smoke in a short time. There were many research studies that investigated the effect of the use of incense sticks on PM10 concentrations. Numerous studies indicate that the smoke from burning incense contains particulate matter, gas products, and other organic compounds that can increase PM concentrations, CO, NOx, and SO_2 in the air (Jetter *et al.*, 2002; Ji *et al.*, 2010). Also, incense burning was found to increase PM2.5 concentrations by up to 120% (Tran *et al.*, 2021).

Each experiment took 15-20 minutes, by burning one incense stick with fixed wind speed, obstacle distance from the source and obstacle height. The total number of variations (experiments) was 18.

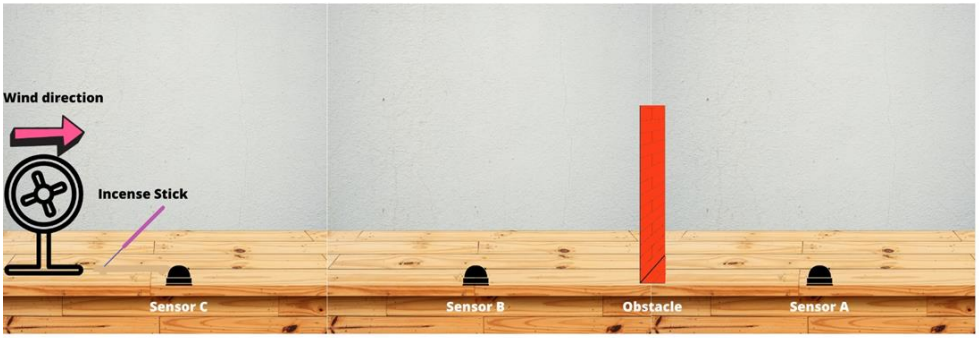

Figure 2.1. Experimental set up

2.1.1.2 Data analysis

Measurements were registered continuously in a programmed excel sheet during each experiment for every 30 seconds. The results present the average PM10 concentration in each test and presented in graphs depending of the obstacle height and distance from the source for the three sensors.

Also, we used the Multiple Linear Regression (MLR) method to calculate the PM10 concentration in sensor A ($PM10_A$) depending on the obstacle height (OH) and distance from the source (OD), PM10 concentration measured near the source (Sensor C, $PM10_C$), and wind speed (Ws).

2.1.2 Effect of small hills on PM10 and PM2.5 concentrations in short range

This part of my study aims to discover how small hills affect the PM10 and PM2.5 concentrations in short range with different wind speeds.

Figure 3.2 shows the experimental setting environment, where the PM sensors are hung on steel infrastructure where sensor 1 (S1) is close to the source (smoke machine), and sensor 2 (S2) is a sensor placed in the middle of the slope. Sensor 3 (S3) is at the top of the hill. The sensors used are low-cost sensors that are calibrated and used in other cities in Hungary to monitor the PM concentrations. The sensors were developed for a project called HUNGAIRY, which is a project that aim to improve air quality at 8 Hungarian regions through the implementation of air quality plan measures, where until this moment PM sensor (LIFE IP HUNGAIRY project sensor) is used in 60 PM monitoring stations in Miskolc and 20 PM monitoring stations in Kaposvár, and the network of the PM monitoring stations will be expended to other Hungarian cities (LIFE IP HungAIRy, 2019). The sensor is based on low-cost, laser scattering PM sensor (Plantower PMS7003), and an auxiliary sensor (Bosch BME680) for measuring humidity, temperature and pressure coupled with a Raspberry Pi 3 single-board computer to collect and store measurements (Báthory *et al.*, 2022). The smoke machine (Haze machine hs-600) is a machine used in concerts and festivals to generate smoke, and the wind machine is controlled via a variable frequency drive (VDF). In contrast, wind speed and directions were used to ensure that no wind was disturbing the experiments and that the wind was going in the right direction. In addition, the sensors register the PM2.5 and PM10 concentrations each minute.

Three cases were adopted in this study. Case 1 is where the ground is almost flat, case 2 is where there is a small hill with an elevation of 0.8 m (shown in Figure 2.2), and case 3 is where a higher elevation is 1 m. The place where the experiments took place was the backyard of a laboratory.

The experiments were done many times, and each time, the smoke machine was on for one hour because, after one hour, the performance of the smoke machine was not stable. The first 10 minutes are without any wind, and then every 10 minutes of wind, the frequency is increased via VDF until we have the maximum frequency possible (here, five frequencies were used). Each wind frequency corresponds to wind speed measured by the wind speed sensor (presented in Table 1).

Also, we used the Multiple Linear Regression (MLR) method to calculate the PM10 concentration in sensor 3 ($PM10_{S3}$ in $\mu g/m^3$) based on the PM10

concentration near source (concentration registered by Sensor 1, $PM10_{S1}$), PM10 concentration at the bottom of the hill (concentration registered by Sensor 2, $PM10_{S2}$), the wind speed (Ws in m/s), and the height of the hill (H in m).

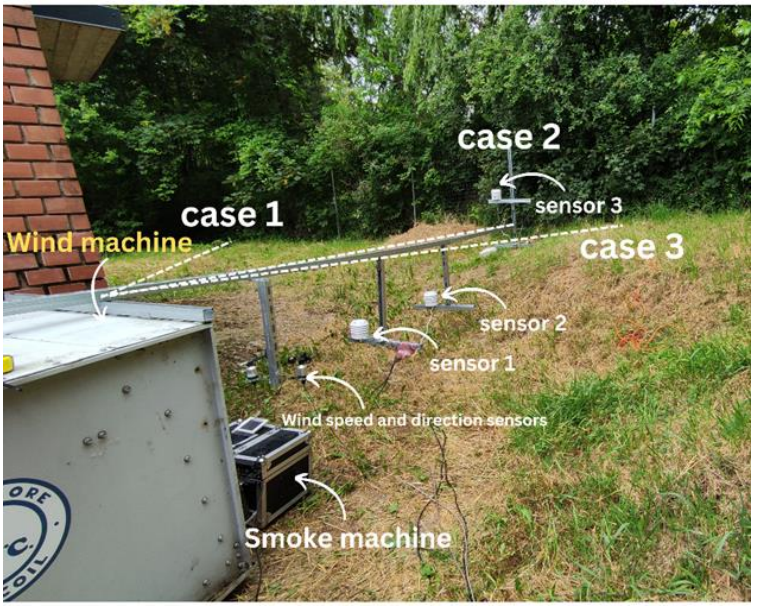

Figure 2.2. Experimental setting environment where dashed lines represent the placement of the metal structure in each of the three cases

Table 1: Wind speed depending on VDF output frequency

VDF output frequency (Hz)	wind speed (m/s)
10	0.7
20	2.4
30	3.7
40	5.1
50	6.1

2.2 Saharan Dust storm transport

2.2.1 Dust Storm simulation over the Sahara Desert (Moroccan and Mauritanian regions) using HYSPLIT

June 2020 was a month where a breaking record dust storm occurred over the Sahara and transported toward the Americas. According to Francis *et al.* (2020), the dust clouds that were generated in this event registered the highest

record of Aerosol Optical Depth (AOD). The dust emission was continuous for 4 days, and uplifted to 5-6 km above the ground surface, and transported across the tropical Atlantic oceans by the powerful mid-atmospheric winds that had a speed higher than 20 m/s (Francis *et al.*, 2020).

For (Francis *et al.*, 2020) the primary objective was to determine the processes responsible for the lifting and transport of dust during the dust event, as well as their relationship to large-scale circulation, and focus on the characteristics of the atmospheric mechanisms that led to massive transport of the Saharan dust. While in this research the central goal is to locate the most active regions in Western Sahara during that event, and the contribution of those regions in increasing the level of PM10 concentration in some regions that are far away from the source place like the US coastal part of the Gulf Mexico and the Martinique islands.

The dust clouds generated covered a huge space as shown in the true colour images of MODIS-Aqua satellite on the 14, 15, 16, 17, 18, and 19 June 2020 (Figure 2.3), where the dust in yellow colour is spreading from the Western Saharan region to the Atlantic Ocean.

2.2.1.1 HYSPLIT model description

The Hybrid Single-Particle Lagrangian Integrated Trajectory model (HYSPLIT) is a software developed by the Air Resources Laboratory (ARL) of the National Oceanic and Atmospheric Administration (NOAA) of USA (Draxler and Hess, 1998). The model is a comprehensive system for simulating basic air parcel trajectories as well as complicated transport, dispersion, chemical transformation, and deposition scenarios. The model calculation method is a hybrid of the Lagrangian approach, which uses a moving frame of reference to calculate advection and diffusion as trajectories or air parcels move away from their initial location, and the Eulerian methodology, which uses a fixed three-dimensional grid as a frame of reference to compute pollutant air concentrations. Over more than 30 years, the HYSPLIT model has developed from predicting simplistic single trajectories based on radiosonde measurements to a system that accounts for numerous interacting pollutants carried, dispersed, and deposited on local to global scales. In addition, HYSPLIT was used to assess the consequences of the accidental release of nuclear material into the atmosphere from the Fukushima Daiichi nuclear power plant after an earthquake and tsunami in March 2011. NOAA's interest since the middle of the last century at the latest, and modelling the movement of smoke from large wildfires has been an ongoing development activity at ARL since 1998. Today, in addition to the United States, smoke forecasts for Alaska and Hawaii are conducted daily to

provide to air quality forecasters and the public information on fine particles (PM2.5) in the air (http: //airquality.weather.gov/) (Stein *et al.*, 2015).

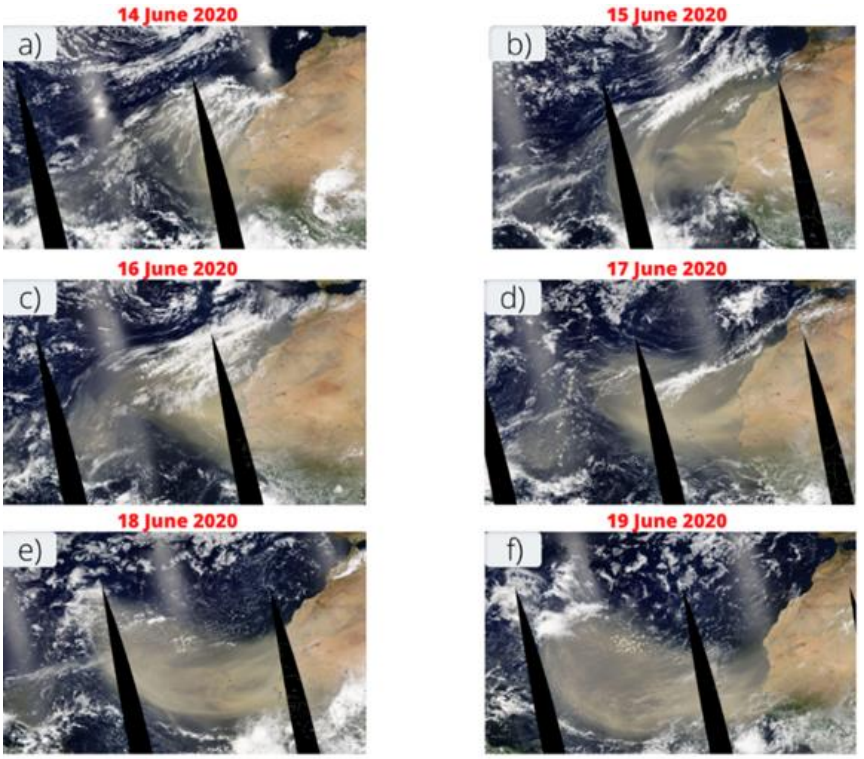

Figure 2.3. MODIS-Aqua true colour images on the (a) 14th, (b) 15th, (c) 16th, (d) 17th, (e) 18th, and (f) 19th of June 2020 over western Africa and the northern tropical Atlantic Ocean. The white colour represents the Clouds and the yellow the dust.

Dust storm Model

HYSPLIT dust storm model is a model for the emission of PM10 dust that has been built using the theory of a surface-roughness-dependent threshold friction velocity (Draxler *et al.*, 2001). When the local wind velocity exceeds the threshold velocity for the soil properties of that emission cell, a dust emission rate is computed from that model grid cell. The predominant mechanism for PM10 emission is "sandblasting," in which larger particles that cannot go airborne bounce along the surface (saltation), allowing additional smaller particles to become airborne (Draxler, Ginoux and Stein, 2010). This emission module makes use of HYSPLIT's 1° land-use file, assuming that a "desert" land-use grid cell corresponds to the roughness identification class

"active sand sheet." Only on dry days and when the friction velocity exceeds the threshold value (0.28 m/s for an active sand sheet) do dust emissions occur. Once the emission strength is determined, the model emits Lagrangian particles with a mass calculated by multiplying the PM flow by the 1° area corresponding to a desert category in HYSPLIT's land-use file. These Lagrangian particles are distributed and moved forward in time in response to NOAA's GFS model's meteorological fields with a horizontal resolution of 1°. A more specific description of particle dispersion and transport can be found in (Draxler *et al.*, 2001; Escudero *et al.*, 2006).

The meteorological data fields needed for the model can be accessed from the National Climatic Data Centre (NCDC) website which is NOAA's National Centers for Environmental Information (NCEI) that provides public access to remarkable archives for environmental data on Earth. In this study, we used the GDAS (Global Data Assimilation System) meteorological data (GDAS1) with a horizontal resolution of 1˚×1˚ corresponding to approximately 100 km x 100 km and 23 vertical layers. GDAS1 is chosen to match the resolution of the HYSPLIT land-use file resolution. GDAS is a system used by the Global Forecast System (GFS) model to insert observations into a gridded model space to begin or initialize, weather predictions using observed data. Surface observations, balloon data, wind profiler data, airplane reports, radar observations, and satellite observations are all added to a gridded, 3-D model space by GDAS. GDAS data are provided as both GDAS input observations and GDAS gridded output fields. The GFS model can be started using gridded GDAS output data. Input data are accessible in a number of data formats due to the varying nature of the assimilated data types, notably Binary Universal Form for the Representation of Meteorological Data (BUFR) and Institute of Electrical and Electronics Engineers (IEEE) binary. World Meteorological Organization (WMO) Gridded Binary is the GDAS output (GRIB) (Kleist *et al.*, 2009). The GDAS dataset covers the entire globe and is freely available.

In the dust storm model, the study domain is defined from 15.0N -18.0E to 32.0N -05.0E (Domain covered with stars in Figure 2.4) which covers the Western Sahara of Morocco Mauritania, and a small part from Algeria. While the PM10 concentrations are averaged over every 12h. The dust simulation Started on the 14th of June 2020 at 00UTC until the 19th of June 2020 at 00UTC. HYSPLIT dust storm modelling was set for 0.5°x0.75° resolution for desert dust sources, with a total of 10 million particles or puffs released during one release cycle and a maximum of 5 million particles permitted to be carried at any time during the simulation. The release mode is sampled using three-dimensional particles in both horizontal and vertical orientations.

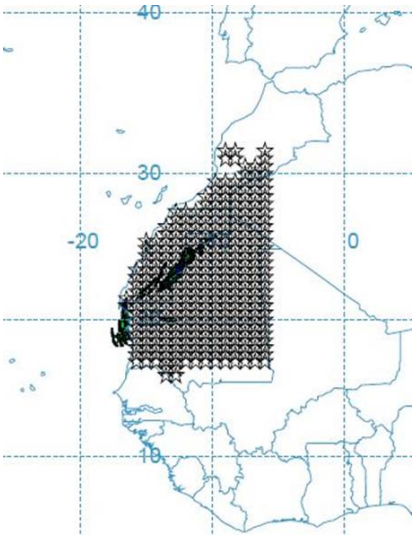

Figure 2.4. Map of the study domain

Trajectory cluster analysis

Forward and Backward trajectory analysis are reliable methods to identify the long-range transport patterns with the use of archived meteorological data (Baker, 2010). However, considering the benefits of the trajectory model, individual trajectories are subject to errors due to the precision and quality of the meteorological data, as well as the simplifying assumptions employed in the trajectory model, which ultimately limits their utility. This problem was solved by computing a large number of trajectories and then subjecting them to cluster analysis. The large number of trajectories computed in HYSPLIT trajectory cluster analysis refers to the number of individual trajectories generated and then subjected to cluster analysis. The benefits of computing a large number of trajectories include minimising the effects of individual trajectory errors, providing a more comprehensive picture of the atmospheric conditions, and identifying rare or unusual events that a smaller number of trajectories may not capture (Baker, 2010). The exact number of trajectories computed will depend on the specific analysis being conducted and the available computing resources. The differences among these trajectories are determined by calculating the distance between clusters, with smaller distances indicating higher similarity. The clustering computation minimises the differences between trajectories within a cluster while maximising the differences between clusters. Trajectories are combined until the total variance of the individual trajectories about their cluster mean starts to increase. This

occurs when disparate clusters are combined. The clustering computation is described in more detail in the literature (Shaw and Gopalan, 2014).

HYSPLIT forward trajectory cluster analysis was performed for the regions that are considered the most active sources of dusts and particles in the region of study as long as some surrounding regions. The list of the regions is presented in Table 2 with names, latitudes, longitudes, and time periods.

Table 2. Cluster analysis location lists

Location	**Latitude**	**Longitude**	**Simulation period**
Dakhla, Morocco	23.8	-15.6	10-30 June 2020
Bir Anzarane, Morocco	23.88	-14.53	10-30 June 2020
Oum Dreyga, Morocco	24.1	-13.25	10-30 June 2020
Aousserd, Morocco	22.5	-14.3	10-30 June 2020
Nouakchott, Mauritania	18.09	-15.95	10-30 June 2020
Atar, Mauritania	20.5	-13.05	10-30 June 2020
Tichit, Mauritania	18.45	-9.5	10-30 June 2020
Toumbouctou cercle, Mali	20.0	-3.0	10-30 June 2020
Bordj Badji Mokhtar, Algeria	22.62	0.12	10-30 June 2020
Tamanrasset, Algeria	24.37	4.32	10-30 June 2020

2.2.1.2 Satellite Observations

Satellites are increasingly being utilized to collect data on aerosol features such as aerosol optical depth (AOD), the columnar concentration of particles, and particle sizes, taking advantage of technological and scientific advances over the previous years. There are various Earth Observing satellite instruments that developed many aerosols remote sensing algorithms for the retrieval of the AOD. One of those instruments is the Moderate Resolution Imaging Spectroradiometer (MODIS). The MODIS instrument, which is mounted on both the Terra and Aqua satellites, measures upwelling radiances in 36 bands with wavelengths ranging from 0.4 to 14.5µm. MODIS data, with

spatial resolutions of 250, 500 m, or 1 km, have been used to construct the most detailed aerosol products, including AOD (Lee *et al.*, 2009). The most recent MODIS collection 6 (C6) aerosol products feature enhanced Dark-Target (10 km DT) and Deep-Blue (10 km DB) AOD. The MODIS science team has carried out a few worldwide validation tests to demonstrate the cumulative impact of these adjustments and the discrepancies between the various parameters (Belle and Liu, 2016). Dark-Target (DT) was created to provide coverage over dense, dark vegetation, whereas Deep Blue (DB) was created to fill in the gaps in DT by providing coverage over bright surfaces (such as deserts) (Sayer *et al.*, 2014). In this study, the MODIS-Aqua Deep Blue AOD 550nm with a spatial resolution of 1˚ was retrieved as an average daily map from the https://giovanni.gsfc.nasa.gov, which is an online platform created by NASA for displaying and analysing geophysical parameters, with easy access to provenance.

In addition to the MODIS-Aqua AOD product, another product from another instrument is used also in this study, which is the CALIPSO (Cloud-Aerosol Lidar and Infrared Pathfinder Satellite Observations). CALIPSO's mission is an ongoing collaboration between NASA Langley Research Center (LaRC) and the Centre National D'Etudes Spatiales (CNES) to explore the global radiative effects of aerosols and clouds on climate. CALIPSO has been providing nearly continuous measurements of the vertical structure and optical properties of clouds and aerosols since its launch on April 28, 2006, to improve our understanding of their role in the Earth's climate system and the performance of a variety of models ranging from regional chemical transport to global circulation models used for climate prediction (Winker *et al.*, 2010). CALIPSO Lidar Level 1 532nm Total Attenuated Backscatter version 4.10 is the product used in this study, which describes the vertical aerosol profile and provides a clear vision about the altitude of the existing aerosols (including dust) in the troposphere and stratosphere level, more in-depth literature can be found in (Getzewich *et al.*, 2018; Kar *et al.*, 2018; Kim *et al.*, 2018). The CALIPSO 532nm Total Attenuated Backscatter images were retrieved from the official website of CALIPSO (https://www-calipso.larc.nasa.gov).

MODIS-Aqua AOD average maps were used to compare them with the average PM10 concentration maps between 0 and 100m from the HYSPLIT dust simulation results, due to the lack of PM10 ground measurements in the area of study. While CALIPSO Lidar Level 1 532nm Total Attenuated Backscatter was used to get the altitude top layer of the dust transported from the Saharan region as well as the thickness of the dust cloud over the Caribbean Sea and the South-eastern region of the United States. Also,

MERRA-2 AOD data (Description of MERRA-2 AOD data can be found in section 3.4) was used for specific places where the Saharan dust particles shown to be transported to the South-eastern region of the United States to identify the intensity of the Saharan dust storm at that time in those chosen regions.

2.2.2 The identification and evaluation of the Saharan dust storm events in Budapest, Hungary between 2018 and 2022

The dust aerosol loading within the whole atmospheric column is represented by the MERRA-2 dust column mass concentration. We utilized data from the MERRA-2 Visualization tool's atmospheric composition (2D) maps (https://gmao.gsfc.nasa.gov/reanalysis/MERRA-2/data_access), as well as, hourly data (MERRA-2 M2T1NXAER V5.12.4) Obtained from NASA's Earth data website (https://disc.gsfc.nasa.gov/datasets/M2T1NXAER_5.12.4/summary).

MERRA-2 dust column mass concentration is a good measure of the intensity of the Saharan dust storm alongside with PM concentrations (Wang, Gu and Wang, 2020). In our case, to identify and evaluate Saharan dust events transported to Hungary, we used MERRA-2 dust column mass concentration data, 2D maps of dust column mass concentration, and hourly-mean PM10 mass concentrations, retrieved from Budapest Gilice tér station.

2.2.3 Case study of the Saharan dust effects on PM10 and PM2.5 concentrations in Budapest in March 2022

In 2022, Europe suffered from two severe Saharan Dust Events (SDE) during March. Large storms in March 2022 sent clouds of Saharan dust to Europe. One of them also brought long-lasting, dusty, high-altitude cirrus clouds, which caused widespread cloud cover for more than a week, from Iberia to the Arctic. It was a rare kind of storm that researchers have only recently learned to comprehend. Its characteristics include icy clouds that are infused with dust, hence the name dust-infused baroclinic storm (DIBS). A DIBS entrained and lifted an atmospheric river of Saharan dust into the troposphere in the middle of March, attaining an altitude of 10 kilometres. Dust-infused, high-altitude cirrus clouds formed as a result of the dust acting as ice nucleation particles. They continued for almost a week, covering a sizable portion of Europe. On March 15, 2022 (SDE1), the first storm developed over north-central Europe and moved south through Poland, the Czech Republic, and Austria to the eastern Mediterranean (Figure 2.5). 13 days (28th March 2022-SDE2) after the first Saharan dust storm, another wave of Saharan dust hit the south of Europe a spread to reach the Eastern European countries.

To evaluate the effect of the Saharan dust load transported to Budapest Hungary, measurements of PM10 and PM2.5 from the Hungarian Air Quality Network platform for Budapest Gilice tér air quality station.

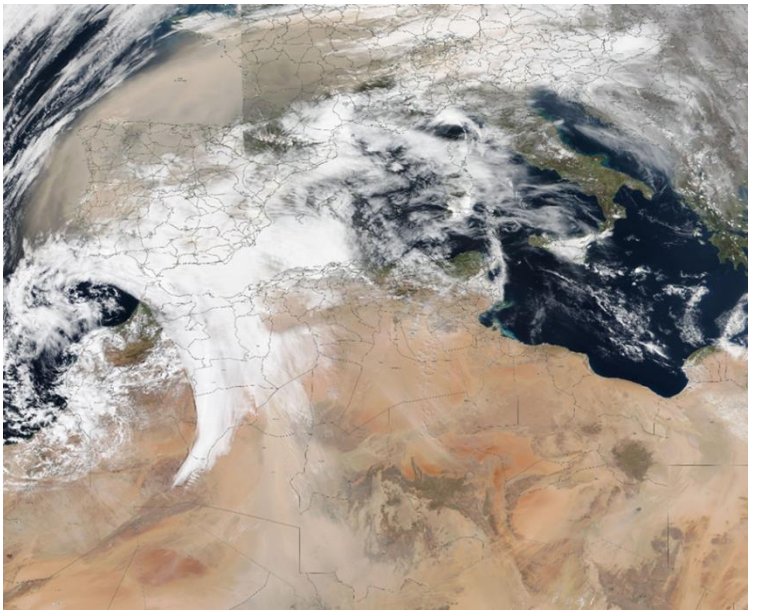

Figure 2.5. Suomi NPP / VIIRS true color image on the 15th of March 2022. Clouds appear in white and Saharan dust in pale yellow/brown

PM10 and PM2.5 concentrations are always higher in winter and fall seasons due to the alternating variability of the weather conditions and the emission source.

Figure 2.6 shows the PM10 and PM2.5 in Gilice tér air quality station during March, the first 10 days of April 2022. During March, the PM10 and PM2.5 concentrations are usually high, however, in March 2022 the PM10 concentration was below the daily EU limit value of 50 $\mu g/m^3$, alternating between 14 and 47 $\mu g/m^3$, and registering lower values in April and May.

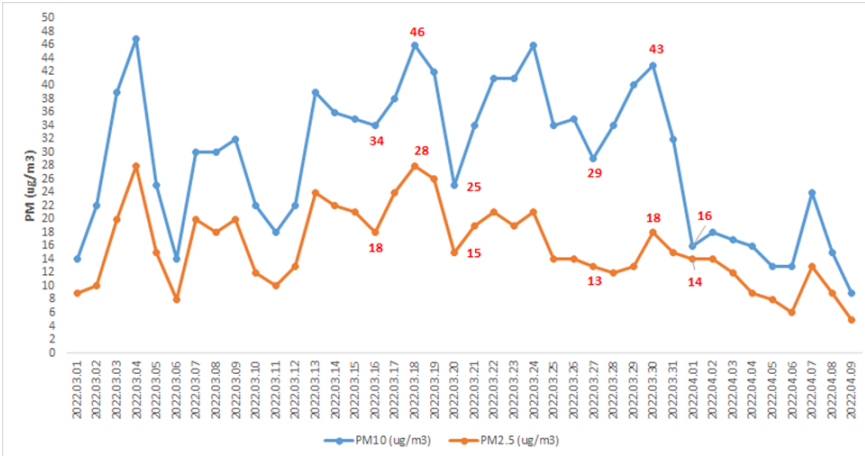

Figure 2.6. PM10, and PM2.5 concentrations (µg/m^3) in Gilice tér air quality station during March and the first 10 days of April 2022

2.3 Estimation and evaluation of PM concentrations

2.3.1 Evaluation of PM surface concentrations simulated by Version 5.12.4 of NASA's MERRA-2 Aerosol Reanalysis over Hungary in the period between 2019 and 2021

In the following sections, I describe the different methods used in Evaluation of estimated PM surface concentrations using NASA's MERRA-2 Aerosol Reanalysis over Hungary in the period between 2019 and 2021.

2.3.1.1 Description of the study

In this of the study I used two approaches. The 1st approach is estimating the PM10 and PM2.5 based on equations 1 and 2 that will be presented later and based on concentrations of components from MERRAero and compare the results with results of equations 1 and 2 with results of machine learning algorithms that will be used also to estimate PM10 and PM2.5, but based on the same concentrations of components used in equations 1 and 2 coupled with meteorological data. The second approach is estimating PM2.5 using machine learning algorithms but this time based on AOD coupled with NO_2, O_3, SO_2 and meteorological data.

The following sections will describe the data and machine learning algorithms chosen for the two approaches described before.

2.3.1.2 The MERRA-2 Aerosol Reanalysis (MERRAero)

A detailed description of the MERRA-2 Aerosol Reanalysis (MERRAero) data is provided in section 3.4.

The five PM species simulated by the MERRAero data collection every hour (SO_4, OC, BC, DS, and SS) allow for an estimation of the total PM10 concentration (Buchard *et al.*, 2016) as follows:

$$[PM_{10}] = 1.375 * [SO_4] + 1.8\,[OC] + [BC] + [DS] + [SS] \qquad (1)$$

Concentration is shown by brackets. [SO_4] is multiplied by 1.375 because it is assumed that SO_4 is completely neutralized by ammonium (NH_4) in the form of ammonium sulphate ($(NH_4)_2SO_4$). The particulate organic matter (POM) is estimated from modelled OC multiplied by a factor that takes into account contribution from other elements associated with the organic matter. This factor has values ranging from 1.2 to 2.6 and is spatially and temporally variable (Malm *et al.*, 1994). In our simulation, a constant value of 1.8 is utilized.

Moreover, since [PM2.5], can be estimated as follows using MERRA-2 Aerosol Reanalysis data collection (Buchard *et al.*, 2016), which separates PM sizes of DS and SS:

$$[PM_{2.5}] = 1.375 * [SO_4] + 1.8\,[OC] + [BC] + [DS_{2.5}] + [SS_{2.5}] \qquad (2)$$

Equations 1 and 2 assume that SO_4, OC, and BC are all in the form of PM2.5 and do not contain nitrate particles, which can account for a sizable portion of the total [PM2.5] (Provençal *et al.*, 2017).

In our case we used AOD retrieved from MERRA-2 global atmospheric reanalysis platform for Budapest, Kecskemét and Kazincbarcika as well as in-situ measurements of PM10 and PM2.5 for the period of 2019 and 2021.

2.3.1.3 Meteorological datasets

Meteorological data were retrieved from NASA Power (Prediction of Worldwide Energy Resources) platform. The platform's list of POWER meteorological characteristics is based on the MERRA-2 assimilation model developed by NASA Goddard's Global Modeling and Assimilation Office (GMAO). Each of the parameters is either estimated using meteorological parameters acquired from NASA's MERRA-2 assimilation model, or it is directly retrieved from those values. The period from January 1, 1981, through a few months in near-real time is covered by the MERRA-2 meteorological data that is accessible through POWER. A time series of hourly (or longer time scale) values is supplied for each parameter of the POWER MERRA-2 model. The average value over the whole geographic grid is represented by each MERRA-2 parameter. The wind speed is at 10 meters, and 50 meters

above the grid's average elevation and its averaged precipitation surface value. The following parameters are derived from the model at a height of 2 m above the grid box's typical elevation. The MERRA-2 parameters are computed in hourly increments and transformed to local time by the POWER project. The 24-hourly temperature measurements, not an average of those values, are used to determine the daily maximum and minimum temperatures.

In estimating PM concentrations, we used hourly temperature at 2 m (T in ˚C), wind speed at 10 m and 50 m (Ws_{10}, and Ws_{50} in m/s), Relative Humidity (RH), surface pressure (P in kPa) from NASA Power, and Planetary Boundary Layer Height (PBLH in m) from MERRA-2 global atmospheric reanalysis platform.

2.3.1.4 Machine learning algorithms

One of the finest approaches to address the complicated interaction between AOD, PM, and associated factors, such as the meteorological parameters, and typically obtain amazing predicted outcomes, is machine learning, a branch of artificial intelligence. The machine learning models were created using Python 3 and the scikit-learn library in JupyterNotebook 6.4.12.

Data preprocessing

Before applying the machine learning algorithm to the data, all data were integrated and matched by time using Microsoft Excel, and cleaned from non-values, in order to generate clean CSV file that will be loaded to the algorithm (Figure 2.7).

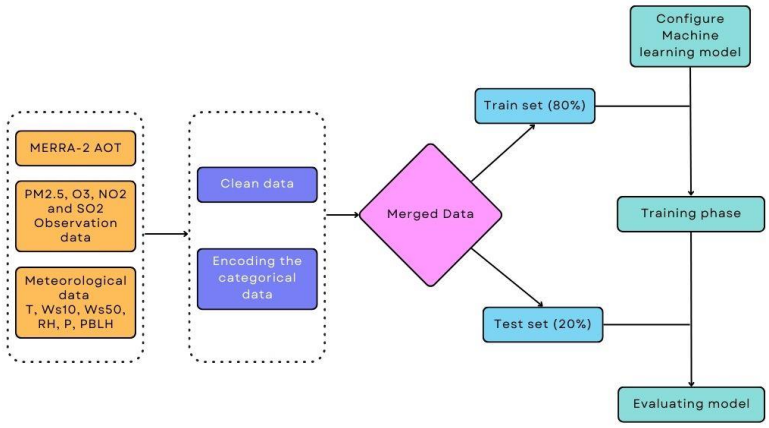

Figure 2.7. Algorithm flowchart

Multiple Linear Regression

The multiple linear regression (MLR) model is the most often used statistical tool for determining the relationship between two or more variables. MLR is an extended model of simple linear regression, where instead of using one variable to predict one outcome, multiple variables are used to predict one outcome.

MLR can be described using the following equation:

$$y = \beta_0 + \sum_{i=1}^{n} \beta_i x_i + \varepsilon \quad (3)$$

Where: y is the outcome value, x_i are the different variables, β_0 is the intercept term, β_i are regression coefficients, and ε is the error term.

Ordinary least squares regression

Ordinary least squares (OLS) models assume that the researcher is fitting a model of a relationship between one or more independent variable and a continuous or at least increment outcome variable that reduces the sum of square errors, in which an error is the difference between the real and predicted value of the outcome. Linear regression (with a single or many predictor variables) is the most frequent analytical approach that employs OLS models (Michalos, 2014).

OLS regression is increasingly employed in many scientific disciplines, including physics, economics, and psychology, and a variety of textbooks have been created to teach this approach and its applications in many fields of inquiry (Cohen, 2013; Kleinbaum *et al.*, 2013; Montgomery, Peck and Vining, 2020).

Random Forest regression

One of the most well-liked ensembles learning strategies based on decision tree predictors is Random Forest (RF), which is a straightforward, effective, and understandable strategy. The trees are bagged in the first stage, and then the tree is divided in the second step using the random subspace technique or the random split selection, applied at each node of the algorithm, and utilizing just a subset of the characteristics to split the node. The benefits of RF included completing regression and classification tasks as well as generating accurate predictions and outcomes that can be simply explained (Breiman, 2001).

Extra Tree regression

The Extra-Trees approach (XT) employs the traditional top-down construction method to construct an ensemble of unpruned decision or regression trees. It separates nodes by selecting cut-points completely at random, which sets it apart from previous tree-based ensemble approaches. It also grows the trees using the entire learning sample rather than a bootstrap replica. In the worst situation, XT creates completely random trees, whose architectures are independent of the learning sample's output values. By selecting the right parameter, the power of the randomization may be adjusted to the particulars of the situation. The algorithm's biggest advantage, aside from accuracy, is computational speed (Geurts, Ernst and Wehenkel, 2006).

In the first approach, we estimated the PM10 and PM2.5 using MLR, OLS, RF and XT machine learning algorithms, based on BC, OC, DS, SS, SO_4, AOD, and PBLH from MERRAero data, and T, RH, Ws_{10}, Ws_{50} and P from NASA Power platform. While in the second approach, we estimated the PM2.5 using MLR, OLS, RF and XT machine learning algorithms, based on AOD, and PBLH from MERRAero data, and T, RH, Ws_{10}, Ws_{50} and P from NASA Power platform, and measurements from Hungarian air quality network of NO_2, SO_2 and O_3. In addition, both approaches were done for Budapest, Kecskemét and Kazincbarcika for the period of 2019 and 2021.

2.3.2 Calibration of CAMS PM2.5 data over Hungary using machine learning

The purpose of this study is to calibrate CAMS PM2.5 data using the LightGBM algorithm and evaluate its impact on improving the accuracy and correlations with in-situ measurements in Hungary. The study aimed to address the limitations of raw CAMS data and provide more reliable information for air quality assessments.

2.3.2.1 CAMS global reanalysis (EAC4)

The Copernicus Atmosphere Monitoring Service (CAMS) reanalysis is the most recent global reanalysis dataset of atmospheric composition produced by the European Centre for Medium-Range Weather Forecasts (ECMWF), and it consists of three-dimensional time-consistent atmospheric composition fields, which include aerosols and chemical species (Inness *et al.*, 2019). The CAMS reanalysis expands on the knowledge gathered during the previous Monitoring Atmospheric Composition and Climate (MACC) reanalysis and interim CAMS reanalysis. Total column CO, tropospheric column NO_2, aerosol optical depth (AOD), and total column, partial column, and profile ozone retrievals from satellites were used in the CAMS reanalysis with the ECMWF's Integrated Forecasting System. The CAMS forecasts air pollution

levels throughout the world over the following few days. The CAMS, in particular, generates a forecast of global atmospheric composition with time horizons as long as the next 120 hours, consisting of 56 reactive trace gases in the troposphere, stratospheric ozone, and five different types of aerosols (i.e., desert dust, sea salt, organic matter, black carbon, and sulphate) (Wagner *et al.*, 2021).

EAC4 (ECMWF Atmospheric Composition Reanalysis 4) is the fourth generation of ECMWF global atmospheric reanalysis. Reanalysis integrates model data with observations from throughout the world to create a globally complete and consistent dataset using an atmosphere model based on physical and chemical rules. This data assimilation principle is based on the method employed by numerical weather prediction centers and air quality forecasting centers (Peuch *et al.*, 2018). The assimilation system can estimate biases between observations and separate high-quality data from low-quality data. Estimates can be made using the atmosphere model in areas with limited data coverage or for atmospheric contaminants for which no direct observations are available. Reanalysis is a very convenient and popular dataset to work with since it provides estimates for each grid point around the world for each regular output time over a long period of time, always in the same format. The observing system has evolved significantly over time, and while the assimilation system can fill data gaps, the initially more sparser networks result in less accurate estimations. As a result, EAC4 is only available since 2003 (Copernicus, 2020). CAMS gives global estimates every 3 h, with a horizontal resolution of 0.75°x 0.75° and a vertical structure of 60 hybrid model levels, with a top-level at 0.1 hPa.

In the current study we used a single-level PM2.5 data downloaded from CAMS website (https://ads.atmosphere.copernicus.eu) for the years 2019 and 2020.

2.3.2.2 ERA5 Meteorological datasets

ECMWF prepared the ERA5 reanalysis as part of the Copernicus Climate Change Service (C3S), which will contain a full record of the global atmosphere, land surface, and ocean waves from 1950 onwards. This new reanalysis will take the place of the ERA-Interim reanalysis, which began in 2006 (Hersbach *et al.*, 2020). ERA5 produces hourly estimates for a wide range of atmospheric, oceanographic, and land-surface variables. An underlying 10-member ensemble samples an uncertainty estimate at three-hourly intervals. For your convenience, the ensemble mean and spread have been pre-calculated. Such uncertainty estimations are intimately tied to the available observing system's information content, which has developed

significantly over time. They also show flow-dependent sensitivity zones. Monthly-mean averages have also been pre-calculated to help with many climatic applications, while monthly means for the ensemble mean and spread are not available. For the reanalysis, data was regridded to a standard lat-lon grid of 0.25 degrees and 0.5 degrees for the uncertainty estimate (0.5 and 1 degree for ocean waves, respectively). There are four major subsets: hourly and monthly products on pressure levels (upper air fields) as well as single levels (atmospheric, ocean-wave, and land surface values). ERA5 hourly data on single levels starts from 1940 to the present (Hersbach *et al.*, 2020).

Data was downloaded from Copernicus climate data platform website (https://cds.climate.copernicus.eu). In this study we used the temperature of air at 2 m above the surface (T in ˚C), relative humidity (RH), Planetary boundary layer height (PBLH in m), 10 m u and v components of wind (u10 and v10 in m/s), surface pressure (P in Pa) and total could cover (tcc).

2.3.2.3 LightGBM algorithms

LightGBM is a highly effective and scalable gradient boosting decision tree technique that benefits from its histogram-based approach, leaf-wise tree development strategy, and proprietary feature bundling (Ke *et al.*, 2017). LightGBM algorithms are a type of gradient boosting framework that have received a lot of attention due to their remarkable performance and efficiency when dealing with large-scale datasets (Sheridan, Liaw and Tudor, 2021). LightGBM algorithms help to advance cutting-edge technology by boosting our understanding of complicated data patterns and, eventually, decision-making processes across numerous industries. Overall, LightGBM algorithms offer extraordinary societal benefit by expanding the field of machine learning and enabling more accurate and efficient data processing (Xia *et al.*, 2021).

LightGBM can process massive amounts of high-dimensional big data with greater efficiency and performance than traditional machine learning approaches. In our study, LightGBM is an appropriate choice. The mathematical equations for PM2.5 calibration schemes are as follows:

$$PM_{2.5,C} = f_{model}(\,CAMSPM2.5, T, RH, blh, u10, v10, P, tcc, hour, day, month) \qquad (4)$$

The data preprocessing and data matching phase involved preparing and aligning the CAMS and ERA5 datasets for further analysis. The resolution of the CAMS dataset is 0.75x0.75, while the resolution of the ERA5 dataset is 0.25x0.25. A geographic matching procedure was used to match the air quality stations with the relevant grid points in each dataset. The purpose was to find the CAMS and ERA5 grid point that was nearest to each air quality station.

The geographic coordinates of the air quality stations were matched with the grid points in both the CAMS and ERA5 datasets during the data matching process. The closest grid point to each station was obtained by computing the distances between the station coordinates and the grid point coordinates.

After the data preprocessing and matching phase, the datasets were further divided into training and test sets (80 x 20% split) with 5-fold cross validation. The Pearson correlation R is calculated between the raw CAMS and in-situ PM2.5 before the training of the model for the test data, and after of the training of the model between calibrated and in-situ PM2.5.

2.4 Data and statistics

2.4.1 The MERRA-2 Aerosol Reanalysis (MERRAero)

The Goddard Earth Observing System Model, Version 5 (GEOS-5) is the foundation of the MERRA-2 assimilation system (Molod et al., 2015). MEERA-2 incorporates spaceborne aerosol products from Moderate Resolution Imaging Spectroradiometer (MODIS), Multi-angle Imaging Spectro Radiometer (MISR), and the ground-based remote sensing network AErosol RObotic NETwork (AERONET) as data for its aerosol dataset. The optical characteristics, emissions, deposition, and aerosol mixing ratios of the five different types of aerosols are all included in the MERRA-2 aerosol dataset vertically (Buchard *et al.*, 2017; Randles *et al.*, 2017). The data from MERRA-2 comprise 21 different types of products, such as atmospheric aerosols, radiation, temperature, water vapor, precipitation, etc. The data span the years 1980 to the present, and are saved in a standard grid of 0.5° × 0.625° (Randles *et al.*, 2017).

The GOCART (the Goddard Chemistry Aerosol Radiation and Transport model) chemistry module, which simulates five different forms of aerosols, is integrated with the MEERA-2 model (sulfate (SO_4), organic carbon (OC), black carbon (BC), sand dust (DS), and sea salt (SS)). These aerosols are considered as external mixes that do not interact with one another. While the surface wind speed affects the emissions of dust and sea salt, other aerosol types are predicted from potential natural and anthropogenic sources. Convective large-scale wet removal, dry deposition, sedimentation, and chemical processes to generate sulphate aerosol from Sulphur dioxide (SO_2) oxidation are all included within the GOCART model (Randles *et al.*, 2017).

The parameterizations of natural and anthropogenic emissions in MERRAero have got numerous significant modifications from the previous edition of the GEOS-4 modelling system (Colarco et al., 2010). The Edgar-4.1 inventory

was used to calculate SO2 emissions from anthropogenic sources, and the injection scheme was changed to account for changes in the injection profiles of emission sources from the energy and non-energy sectors (Buchard *et al.*, 2014). The emissions from biomass burning are from the NASA Quick Fire Emission Dataset (QFED) Version 2.1. QFED is a worldwide fire radiative power-based inventory of daily aerosol precursor and trace gas emissions (Koster, Darmenov and da Silva, 2015). According to the study of Jaeglé *et al.* (2011) a novel independently obtained sea surface temperature (SST) adjustment term was used to modify the intensity of sea-salt emissions. Dust emission is predicated on the correlation of reported dust source sites with large-scale topographic depressions, as proposed by Ginoux *et al.* (2001).

MERRA-2 coupled AOD at 550 nm, from a variety of ground- and space-based remote sensing platforms, including (i) bias-corrected AOD from Moderate Resolution Imaging Spectroradiometer (MODIS) Terra and Aqua, (ii) the Advanced Very High Resolution Radiometer (AVHRR) instruments, (iii) AOD retrievals from the Multiangle Imaging SpectroRadiometer (MISR) over bright surfaces, and (iv) ground-based Aerosol Robotic Network (AERONET) direct measurements of AOD (Level 2) (Randles *et al.*, 2017).

2.4.2 Air quality stations

The Hungarian Air Quality Monitoring Network provides real-time and historical air quality monitoring data throughout Hungary. The network is divided into two main parts: automatic monitoring stations that continuously measure a wide range of air pollutants in the ambient air, and a manual system with sample points and subsequent laboratory examination. The existing network in Hungary comprises 37 fully automatic monitoring stations. The National Air Quality Reference Centre and Laboratory's primary responsibilities are as follows: Oversight of the operation of the Hungarian Air Quality Monitoring Network (HAQM) in accordance with Ministry of Agriculture standards, coordination and regulation of HAQM methods and procedures in accordance with EU regulations, maintain measurement traceability by running an approved Calibration Laboratory, and participation in national and worldwide standards development. A CO analyser, PM10 / PM2.5 monitors, a calibration tower, and a mass flow meter calibration system were added to the calibration laboratory instrument fleet (Weidinger *et al.*, 2010).

Among the monitoring sites in Budapest, the Gilice tér urban background station (located in the SE part of the city) was chosen for our analysis because it is a standard meteorological and air quality monitoring station that provides

PM10 and PM2.5 concentrations and detailed meteorological observations with good data coverage.

Kecskemét is located 86 kilometers from both the capital Budapest and the country's third-largest city, Szeged, and is almost equal distance from the country's two major rivers, the Danube and the Tisza. Kecskemét is the city most vulnerable to climate change, with a slew of environmental issues in the Danube-Tisza Interfluve. The most significant changes include the degradation of air quality, the influence of urban heat islands, and water management (Hoyk, Kanalas and Farkas, 2020). The air quality station in Kecskemét is an urban background station.

Kazincbarcika is a town in the county of Borsod-Abaj-Zemplén in Northern Hungary. It is located in the valley of the Sajó River, 20 km away from Miskolc, the county capital. The air quality station in Kazincbarcika is an international urban background station. Table 3 presents the list of air quality stations used throughout the different studies as well as their geographical coordinates.

All PM10 and PM2.5 data were retrieved from the Hungarian Air Quality Network platform (Országos Légszennyezettségi Mérőhálózat (OLM), https://legszennyezettseg.met.hu), which is a platform that provides actual and historical air quality monitoring data throughout Hungary.

Table *3*: List of air quality stations with latitudes and longitudes

Station	**latitude**	**longitude**
Ajka	47.10	17.55
Budapest Gilice	47.43	19.18
Kazincbarcika	48.24	20.61
Kecskemet	46.90	19.68
Miskolc_Alfoldi	48.09	20.81
Nyiregyhaza	47.96	21.71
Pecs Nevelesi Kozpont	46.04	18.22
Szazhalombatta_Buzavirag_ter	47.31	18.92
Szeged_Rozsa	46.27	20.15
Szolnok	47.18	20.2
Veszprem	47.09	17.9

2.4.3 Performance statistics

The performance of the air quality forecast models (in sections 3.3.1 and 3.3.2) using the testing dataset was assessed using model performance metrics, such as R^2 computed by Equation (5), RMSE calculated by Equation (6), MAE calculated by Equation (7), and Pearson correlation R calculated by Equation (8)

$$R^2 = \frac{\left[\int_{i=1}^{n}(p_i - \bar{p}) - (o_i - \bar{o})\right]^2}{\left[\int_{i=1}^{n}(p_i - \bar{p})^2\right]\left[\int_{i=1}^{n}(o_i - \bar{o})^2\right]}, \quad (5)$$

$$RMSE = \sqrt{\sum_{i=1}^{n}\frac{(p_i - o_i)^2}{n}}, \quad (6)$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|p_i - o_i|, \quad (7)$$

$$R = \frac{\sum(p_i - \bar{p}) - (o_i - \bar{o})}{\sqrt{\sum(p_i - \bar{p})^2 \sum(o_i - \bar{o})^2}} \quad (8)$$

Where:

p_i the predicted value of the sample, and $\bar{p}$ is the predicted average.

o_i the observation value, and $\bar{o}$ is the observation average.

n the number of the samples.

3 RESULTS

In this chapter, I present the results of all the 3 main chapters presented in Material and Method section.

3.1 PM dispersion experiments

3.1.1 Small scale experiments of PM10 dispersion around obstacles

The results of the experiments showed some interesting aspects for the understanding of the PM10 dispersion around simple obstacle (Wall).

3.1.1.1 Sensor A

The sensor A is the sensor behind obstacle. Figure 3.1 shows the average concentration of PM10 during each experiment in function of Obstacle heights (OH) and distance from the source (OD). The average PM10 concentration increase with increasing of the obstacle distance from the source at higher wind speed while in low wind speed it is almost stable. At wind speed of 2.9 m/s the average PM10 concentration was the same for obstacle height 240 and 360 mm while it was at its peak when obstacle height was 120 mm. while, for wind speed of 1m/s the peak average PM10 concentration was at obstacle height of 360 mm and almost the same in the other two heights.

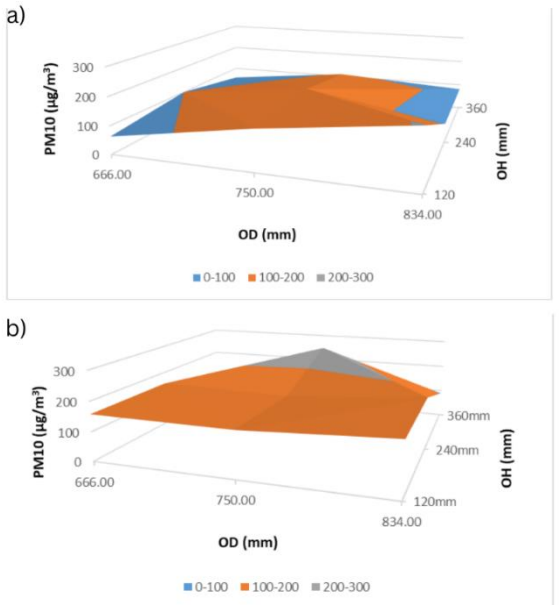

Figure 3.1. graphs of Average PM10 concentration registered by Sensor A in function of Obstacle heights and distance from the source in case of a) wind speed 2.9 m/s and b) wind speed 1 m/s

3.1.1.2 Sensor B

For the sensor B (Figure 3.2), which is the sensor placed before the wall, the PM10 average concentration was higher in case of wall height of 240 and 360 mm, and wall distance of 750 mm at wind speed of 1 m/s. While it reaches the maximum when obstacle distance from the source is 834mm, obstacle height is 120 mm and wind speed of 2.9 m/s.

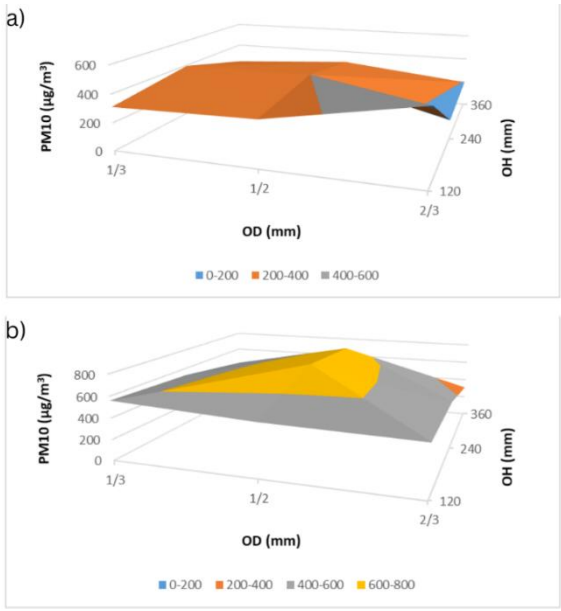

Figure 3.2. Graphs of Average PM10 concentration registered by Sensor B in function of Obstacle heights and distance from the source in case of a) wind speed 2.9 m/s and b) wind speed 1 m/s

3.1.1.3 Sensor C

The sensor C placed near the source registered almost same average concentration of PM10 at wind speed of 1m/s with decrease in concentration in case of obstacle height 360 mm and distance from source 834 mm (Figure 3.3). In the other hand it was changing at wind speed of 2.9 m/s. The peak average PM10 concentration was as the same as sensor B, when obstacle distance from the source is 834 mm and obstacle height is 120 mm.

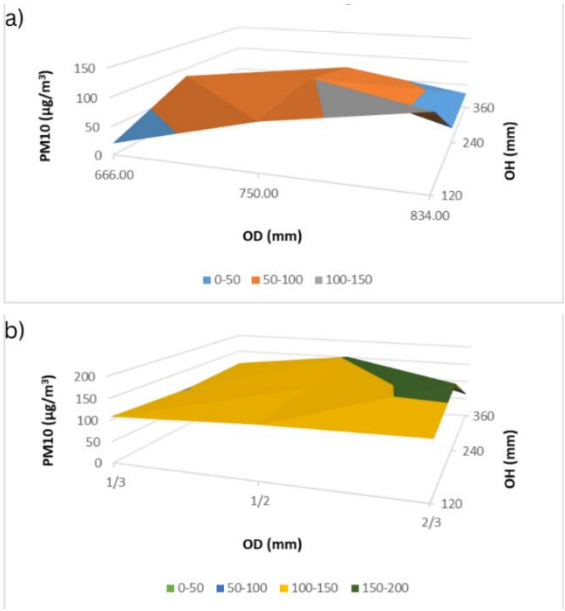

Figure 3.3. Graphs of Average PM10 concentration registered by Sensor C in function of Obstacle heights and distance from the source in case of a) wind speed 2.9 m/s and b) wind speed 1 m/s

The result of a multiple regression analysis aims to predict $PM10_A$ concentration using four independent variables: "OH", "OD", "$PM10_C$," and "Ws". The resulting the equation of the regression model (Eq.9) can be written as follows:

$$PM10_A = 143.07 - 71.86 * OH - 171.42 * OD + 1.23 * PM10_C + 12.34 * Ws \quad (9)$$

The correlation coefficient (R) was 0.89, suggesting that the dependent variable and the set of independent factors had a moderately significant positive connection. Furthermore, the coefficient of determination (R^2) was 0.79, indicating that the independent variables in the model explain roughly 79% of the variation in the dependent variable.

The results of this research concluded that there is a positive significant effect of Obstacle heights, distance of the obstacle from the source, and the wind speed. The PM10 average concentration decrease significantly in the sensor A (behind the wall) when the obstacle height increases and also when the obstacle distance from the source increase also in case of the two-wind speed (1 m/s and 2.9 m/s) with higher concentrations registered in case of wind speed is 1 m/s. while, changes in the PM10 average concentration was also seen in

case of Sensor B (in the middle) and sensor C (near the source) especially in case of high wind speed (2.9 m/s) and that is due to the turbulence created before and after the walls when the wind hits it, in addition to the reflexing of PM plumes by the obstacle. Also, maximum PM10 concentration sensor A (after wall) and sensor B (before wall) at obstacle distance 834 mm, and obstacle height 120 mm, while at low wind speed (1 m/s) the PM10 concentrations does not change with effects of obstacle height and distance from source. In contrast, at higher wind speed (2.9 m/s), the obstacle height and distance affect the PM10 concentration before and after the obstacles in the same way, meaning that the concentrations tend to decrease with obstacle height increases, and as close as the obstacle to the sensor the concentration increases with low obstacle height, due to the turbulence created near the obstacle which trap the PM10 particles near the obstacle. Thus, the experiments results prove the same effects of simple obstacle presence as larger scale study where complex urban landscape and structure are involved. The experiments proved that also in small scale experiments the transportation of the PM particles are the same as in real scale transportation of PM.

Generally, the PM10 average concentration tends to decrease when obstacle heights increase but also combined with position of the obstacle far from the source. In our case, the experiment is a simplification of the dispersion of PM concentration (PM10 specifically) in an austere environment. It represents the basis for understanding the PM pollutant source interaction with the barrier and how it affects PM10 concentrations. The results may change in a complex urban setting, where many parameters can intervene to change the dispersion of air pollutants. Our case study's results are valid but subject to investigation in other experimental settings.

Moreover, using Incense sticks as source of PM pollution showed that while the stick is burning it continues to spike the PM10 concentration, as before the experiments the background concentration of PM10 was 7±3 $\mu g/m^3$ and during the experiments it can reach 700 $\mu g/m^3$, which manifest the short-term effect of burning the incense stick and its risk of affecting the indoor air quality if used in excess. Finally, the experiment is representation of trying to find simple obstacle placement that can reduce significantly PM plume coming from source that could be industrial or traffic source. The results show that the higher the obstacle is better but also closer is better also, but in real situation simple obstacle can be put in the way of PM plumes and as closer as possible to the area that is subject to be defended from high PM concentrations. And one of the best options is to combine simple obstacle (solid barrier) with vegetated/tree barrier as the last was proven to improve air exchange, and The

tree planting and trunk height have a considerable impact on the air flow and pollution dispersion (Buccolieri *et al.*, 2022).

3.1.2 Effect of small hills on PM10 and PM2.5 concentrations in short range

The average concentrations registered by sensor 3 (S3) of PM10 and PM2.5 are higher in the case of the 1m height and 0.8m height compared to the concentrations recorded during flat case Figure 3.4.

At low wind speeds (0 and 0.7 m/s), the average concentrations of PM10 and PM2.5 registered by S3 are almost the same in all the 3 cases. At wind speeds of 2.4, 3.7, and 5.1 m/s, the average concentration of PM10 and PM2.5 are higher in the case of the two different heights compared to flat areas. The peak concentration of PM10 and PM2.5 in case of 1m height registered when the wind speed was 3.7 m/s, while in case of 0.8 m height was at a wind speed of 5.1 m/s, while in a flat area, average concentrations registered were almost the same when wind speed was higher than 2.4 m/s. In addition, the same in the case of 0.8 height, but the average concentration was 2.5 to 3 times higher than in the flat case with a slight decrease at high speed (6.1 m/s). While in the case of 1m height, the average PM concentration was 2 to 3 times with wind speeds of 3.7 and 5.1m/s, and almost the same at wind speeds of 2.4 and 6.1 m/s.

The difference in the ground surface elevation between case 2 and case 3 is just 0.2 m, but the effect on the dispersion of the PM plumes can be seen from the average PM concentrations. In the case of a flat ground surface, the spread of PM pollutants is parallel to the wind direction. In contrast, high ground (in our case, in the form of a hill) at different elevations changes the dispersion pathway of the PM particles. The different slopes of the hills create other flows of the PM dispersion; in case two, the approximate same PM concentrations registered in different ranges of wind speed means the PM particles are trapped in the same way regardless of the wind speed.

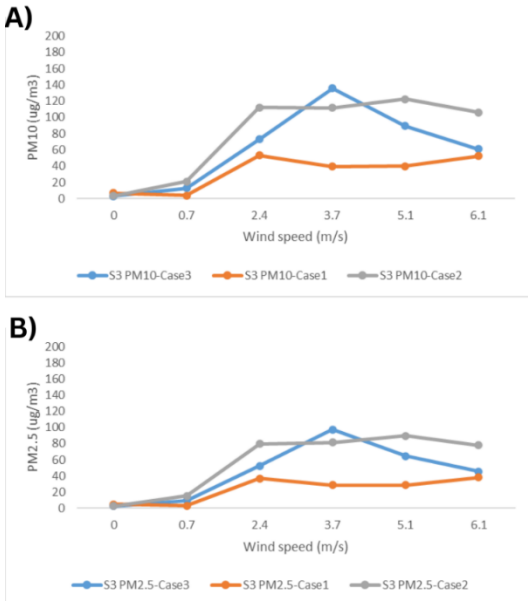

Figure 3.4. Average S3 concentrations in all cases of A) PM10 and B) PM2.5

Comparing the three PM10 concentrations registered by the 3 sensor (Figure 3.5), it's seen that S1 registered higher concentrations in case 2 and 3 than in case 1, especially at wind speeds less than 3 m/s, and that is due to reflective effect of the hill and also low wind speed. While for S2 the PM particles are trapped before the hill which promote higher PM concentrations.

In this study also, multiple linear regression method was used to estimate PM10 concentration at the top of the hill ($PM10_{S3}$ in µg/m3) based on the PM10 concentration near source (concentration registered by Sensor 1, $PM10_{S1}$), PM10 concentration at the bottom of the hill (concentration registered by Sensor 2, $PM10_{S2}$), the wind speed (Ws in m/s), and the height of the hill (H in m).

The result of the multiple linear regression is the following equation:

$$PM10_{S3} = 5.92 - 0.173 * PM10_{S1} + 0.580 * PM10_{S2} + 4.29 * Ws - 11.29 * H \quad (10)$$

The correlation coefficient (R) was 0.9, indicating a relatively strong positive correlation between the dependent variable and the combination of independent variables. In addition, the coefficient of determination (R^2) was 0.82, which means that approximately 82% of the variance in the dependent variable is explained by the independent variables in the model.

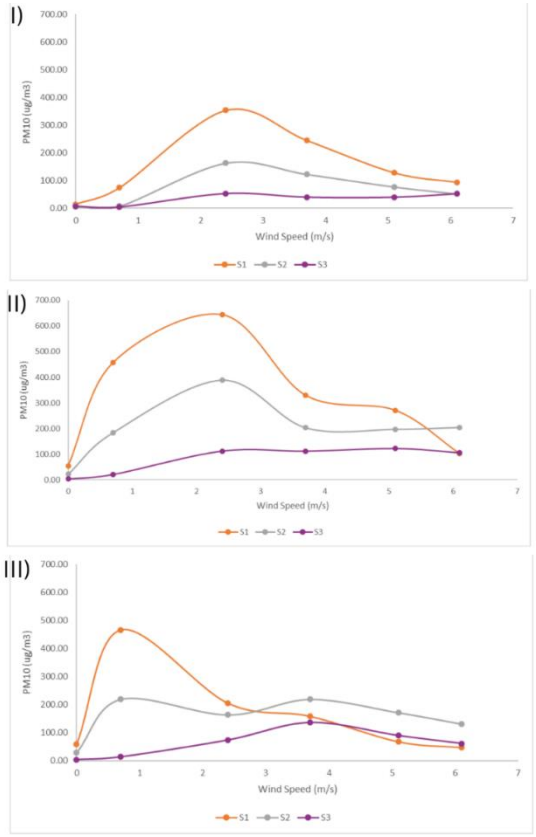

Figure 3.5. PM10 concentrations of the 3 sensors in I) case 1, II) case 2, and III) case 3

Analysing the correlations between PM10 concentrations of sensors S1, S2 and S3 (Figure 3.6), it shows that for S1 correlation was low and positive in case 1 (0.2), but it changes to negative in cases 2 and 3 (-0.18 and -0.5 respectively), which show the effects of the height of the hills. For S2, the correlation between PM10 concentration and wind speed decrease as the height of the hill increase, while for S3, a strong correlation is observed in case 1 and 2 (0.8 and 0.84, respectively), and it decreases in case 3. Thus, the decrease in the correlation due to the higher elevation of the hill could be because of the changes in the wind flow created by different elevations of the hill. The results underscore the significant influence of hill elevation on the correlation between PM10 and wind speed at various sensor locations, emphasizing the role of local topography in shaping air pollution patterns during the experiments.

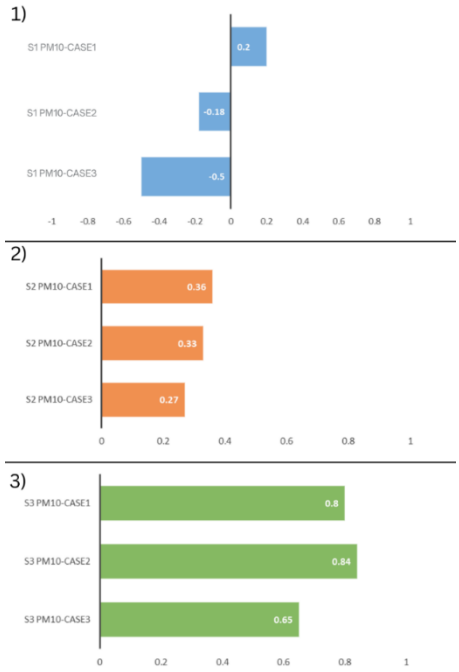

Figure 3.6. Correlations between PM10 concentrations and wind speed in all cases for 1) Sensor 1, 2) Sensor 2, and 3) Sensor 3

3.2 Saharan Dust storm transport

3.2.1 Dust Storm simulation over the Sahara Desert (Moroccan and Mauritanian regions) using HYSPLIT

In this chapter I describe the results Saharan Dust simulation study using HYSPLIT model.

3.2.1.1 Dust Simulation and cluster analysis results

Hysplit PM10 emission modelling results (Figures 3.7 and 3.8) show that the PM10 emission on the 14th of June 2020 started from the region of Tinduf, Algeria (Close to the Moroccan borders), Adrar, Tiris Zemmour, and Tagant in Mauritania. While the dust storm was continuous for 4 days and the dust was transported to the North Atlantic Ocean, the average PM10 concentration between 0 and 100m was between 100 µg/m^3 and 10000 µg/m^3 in some critical regions like Tinduf, Algeria on the 14th and 17th of June 2020, Adrar, Mauritania on the 15th, Bir Anzarane, Morocco on the 16th, Tiris Zemmour, Mauritania on the 17th, Goundam Cercle, Mali on the 18th of June 2020. Comparing the average PM10 concentration maps between 0 and 100m from the HYSPLIT modelling results and the MODIS Aqua Deep Blue AOD maps

(Figure 3.9), it can be seen that in most of the regions where the PM10 concentration are high, the AOD index is also at a high level, which indicates a positive correlation between the PM10 concentration and the AOD index of MODIS Aqua. Regions like Tiris Zemmour in Mauritania, Western Sahara of Morocco, Western and Southern regions of Algeria, are also characterized as source regions that influence the level of the PM10 concentration over the western Mediterranean Basin (Salvador et al. 2014; Russo et al. 2020). Moreover, all the areas that had a high concentration of PM10 in the HYSPLIT dust simulation results and high AOD values (between 0.7 and 1) according to time-averaged maps of MODIS-Aqua are considered as primary dust natural source regions (Ginoux et al. 2012), and they are active throughout the year, although their peak activity is between April and September (Prospero et al. 2002).

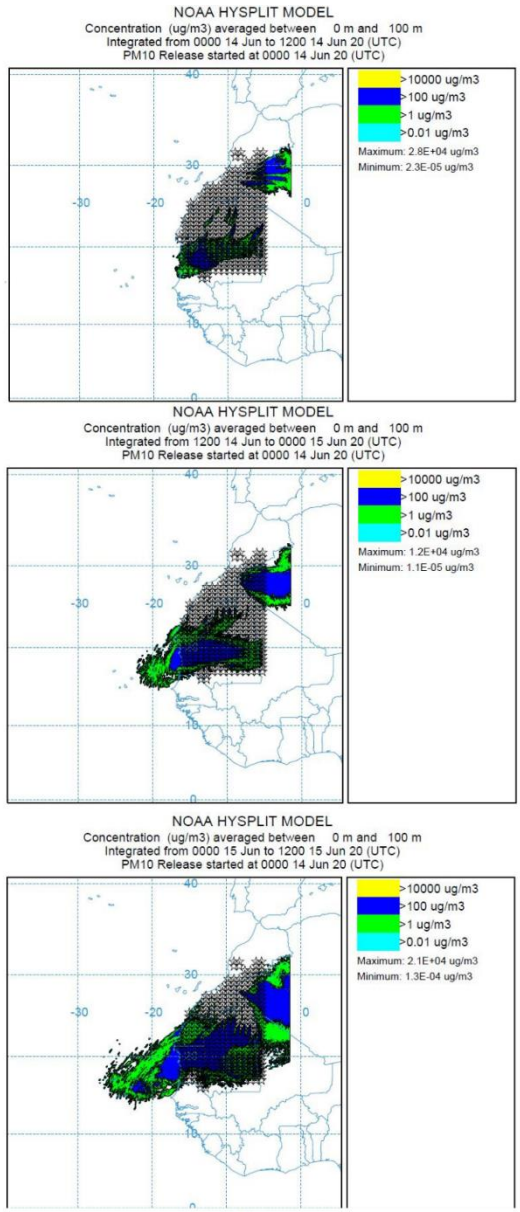

Figure 3.7. Modelling results for the concentration of PM10 averaged across the 0-100m altitude range in June a) 14th from 00 UTC to 12 UTC, b) 14th from 12 UTC to 15th 00 UTC c) 15th from 00 UTC to 12 UTC

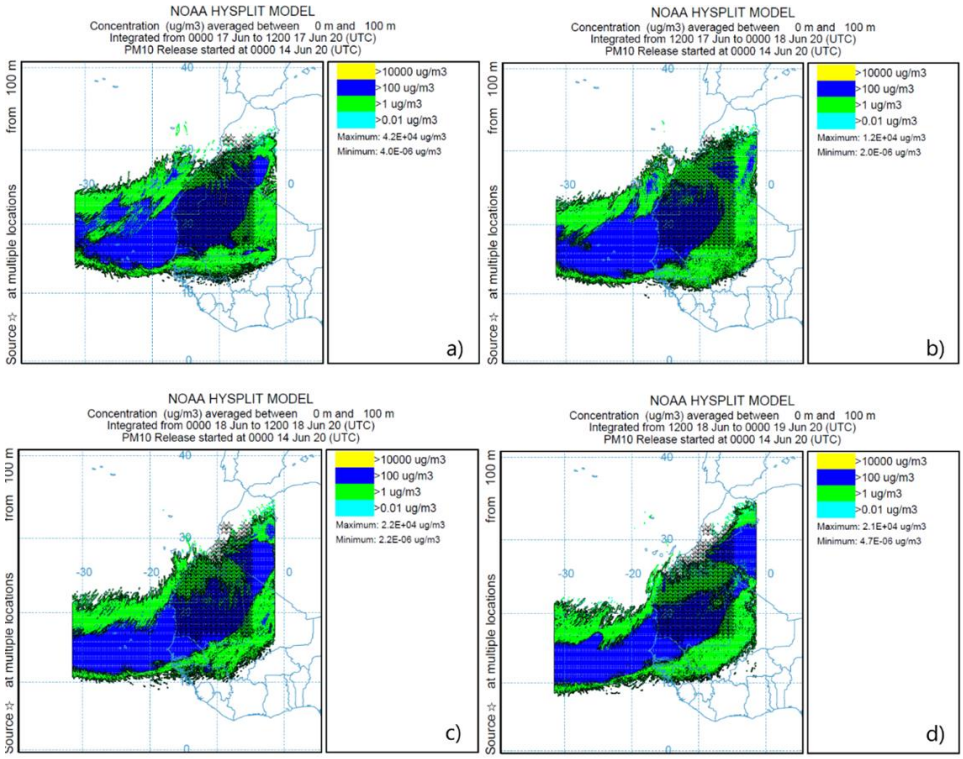

Figure 3.8. Modeling results for the concentration of PM10 averaged across the 0-100m altitude range in June a) 17th from 00 UTC to 12 UTC, b) 17th from 12 UTC to 18th 00 UTC c) 18th from 00 UTC to 12 UTC d) 18th from 12 UTC to 19th 00

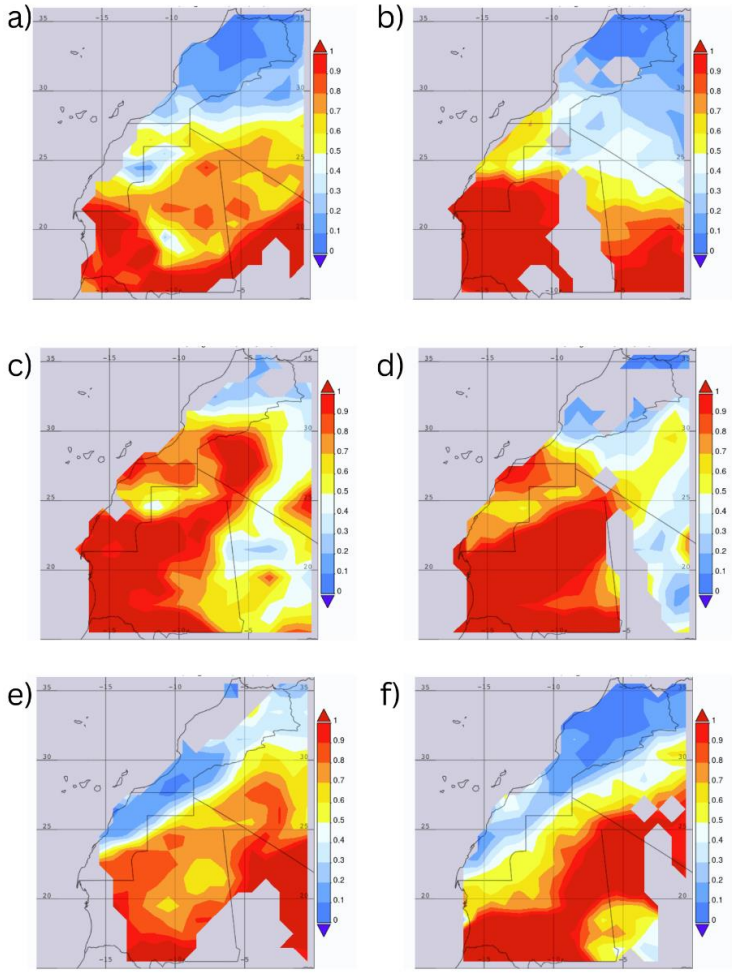

Figure 3.9. MODIS Aqua time Averaged Map of Aerosol Optical Depth 550 nm (Deep Blue, Land-only) daily 1 deg in the region of western Sahara on the a) 14, June 2020 b) 15, June 2020 c) 16, June 2020 d) 17, June 2020 e) 18, June 2020 f) 19, June 2020

The analysis of the trajectories of the PM10 particles emitted from numerous locations in the western Sahara during the June dust storm event using the HYSPLT cluster analysis method is shown in Figures 4.10 and 4.11. A large percentage of the PM10 trajectories analysed in the period between 10th and 30th of June 2020, reached the middle-upper troposphere of the Caribbean Sea and the Gulf of Mexico. 80%, 51%, 76%, and 70% of the PM10 particle

trajectories from Bir Anzarane Morocco, Nouakchott, and Tichit Mauritania, and Bordj Badji Mokhtar Algeria arrived to the Gulf of Mexico respectively.

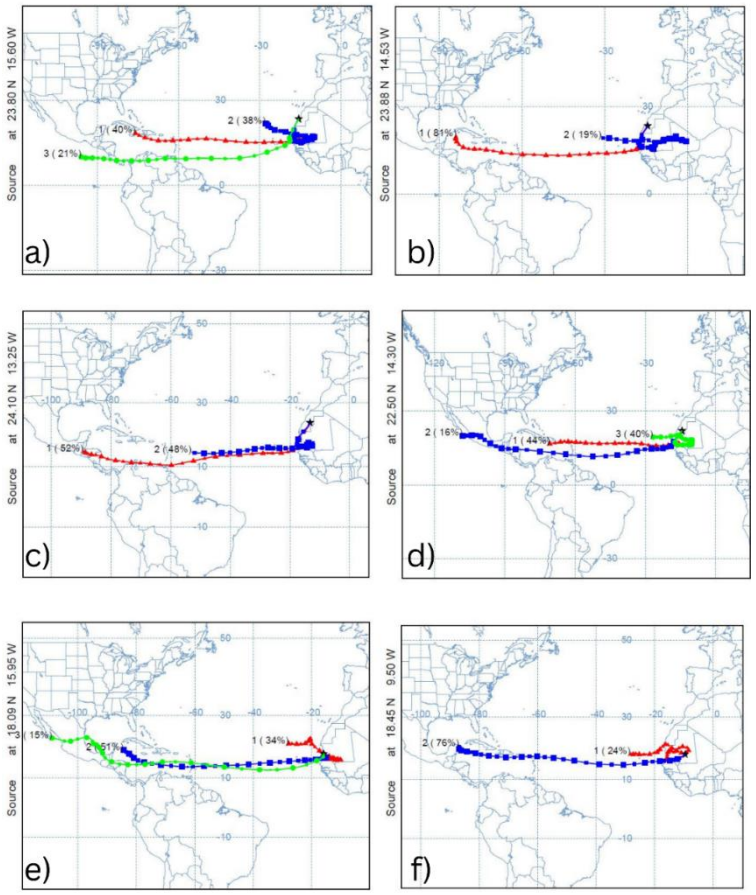

Figure 3.10. Forward trajectory cluster analysis results, each picture shows Cluster mean trajectories with the percentage of trajectories in each cluster from a) Dakhla, b) Bir Anzarane, c) Oum Dreyga, d) Aousserd, e) Nouakchott, and f) Tichit.

Many studies state that during summer, and especially during Saharan dust events, the level of PM10 and PM2.5 concentration increased dramatically. (Bozlaker *et al.*, 2013) state that during the Saharan episode in 2008, the total dust contribution for PM10 increased by 85% in Houston, Texas, which shows a dominance of the transported PM10 particles from Sahara during dust episodes. Also, (Bozlaker *et al.*, 2019) found dust contributions of 19% to 48% of PM2.5 during the 9-day dust episode in 2014 to African dust. Additionally, the results of the cluster analysis point out a number of source regions in the western Sahara that contribute to the rise in PM10 concentrations in the

Southern Coast of the United States, such as Bir Anzarane Morocco, Nouakchott, and Tichit Mauritania, and Bordj Badji Mokhtar Algeria.

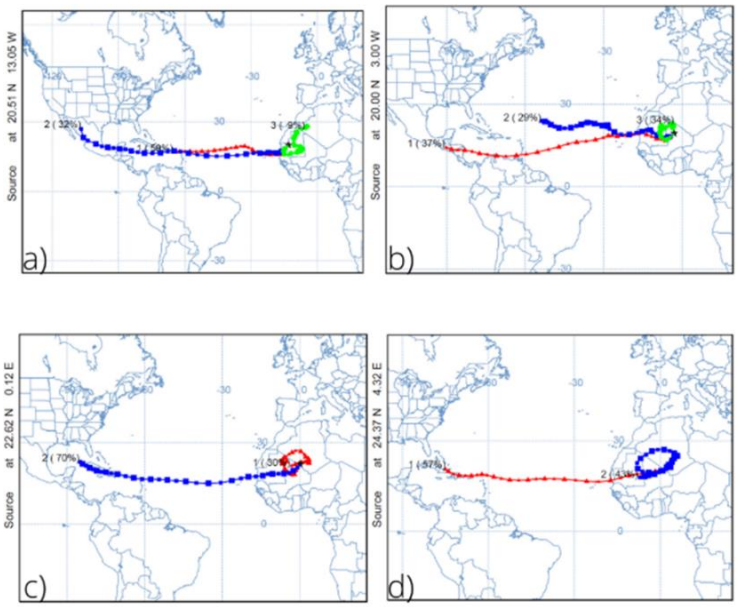

Figure 3.11. Forward trajectory cluster analysis results, each picture shows Cluster mean trajectories with the percentage of trajectories in each cluster from a) Atar, b) Toumbouctou cercle, c) Bordj Badji Mokhtar and d) Tamanrasset

3.2.1.2 PM concentration levels and AOD

In addition, the backscatter vertical profile as measured by CALIPSO on June 21 and 23, 2020 (Figure 3.12) shows evidence of the high altitude of the dust particles transported from the Saharan region. The top layer altitude of the dust on June 21 and 23 were between 4 and 4.5 km, forming a massive dust cloud (more than 2 km of thickness) over the Windward and Leeward islands in the Caribbean Sea, and the effect was seen in the hourly measurements of the PM10 concentrations of the Fort de France station in Martinique Island where the PM10 daily average concentration was 181, 264, and 183 µg/m3 on the 21, 22 and 23 of June consecutively with an hourly concentration that reached 372 µg/m3, comparing to 42 µg/m3 that was registered at the beginning of that month. Furthermore, and after 11days of the starting of the Saharan dust storm, the effect of the transported particles was clear in the US coastal cities of the Gulf of Mexico. Texas and Florida states were the most affected by having an Unhealthy level of PM10 and PM2.5 concentrations, followed by Georgia, Alabama, Mississippi, and Louisiana states that reached the level of Unhealthy for sensitive groups during the 26 and 27 June 2020, which is in correlation with the backscatter vertical profile measured by

CALIPSO on June 27, 2020, showing dust cloud over Florida state with dust top layer altitude at 4 km.

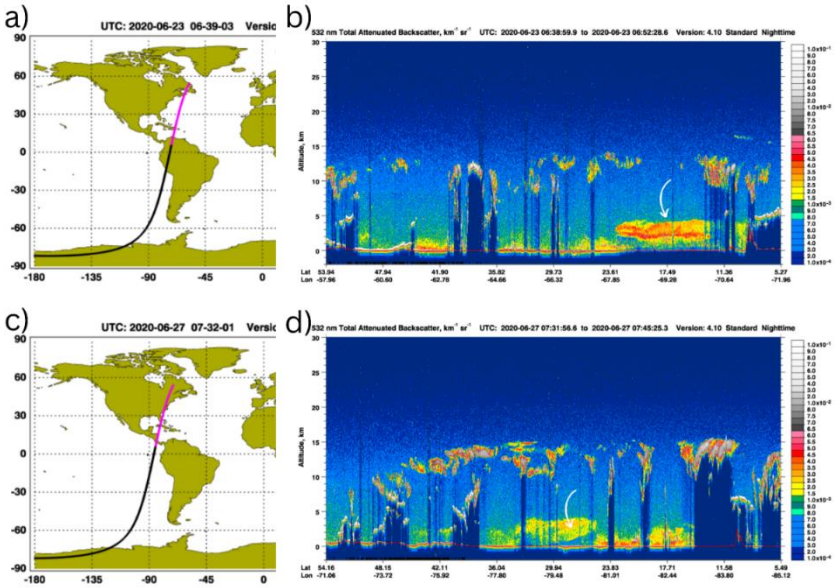

Figure 3.12. CALIPSO 532nm Total Attenuated Backscatter Version 4.10 images, the dust appears in yellow and red in the images. a) Orbit map on 23, June 2020 with area covered in d image coloured in pink b) 532nm Total Attenuated Backscatter on 23, June 2020 with white arrows pointing at the dust top layer over the Caribbean Sea c) Orbit map on 27, June 2020 with area covered in image coloured in pink d) 532nm Total Attenuated Backscatter on 27, June 2020 with white arrows pointing at the dust top layer over Florida state

In order to quantify the dust event, Figure 3.13 and 3.14 show AOD values retrieved from MERRA-2 re-analysis data for Bir Anzarane, Morocco and Nouakchott, Mauritania, for the month of June 2020. According to both, the June 2020 dust event was historical by June standards. For Bir Anzarane, Morocco, the highest AOD value was 3.522 in June 2020, a 188% increase from the highest value registered from 2010 to 2019 (1.87 in June 2017). For Nouakchott, Mauritania, the highest AOD value recorded in June between 2010 and 2019 was 2.78 in June 2010, while in June 2020, the highest AOD

was 5.87, 211% higher. Even though such high AOD levels are exceptional, but not uncommon; during

the record-breaking March 2018 dust outbreak, Solomos *et al.* (2018) and Kaskaoutis *et al.* (2019) observed AOD values over 6.

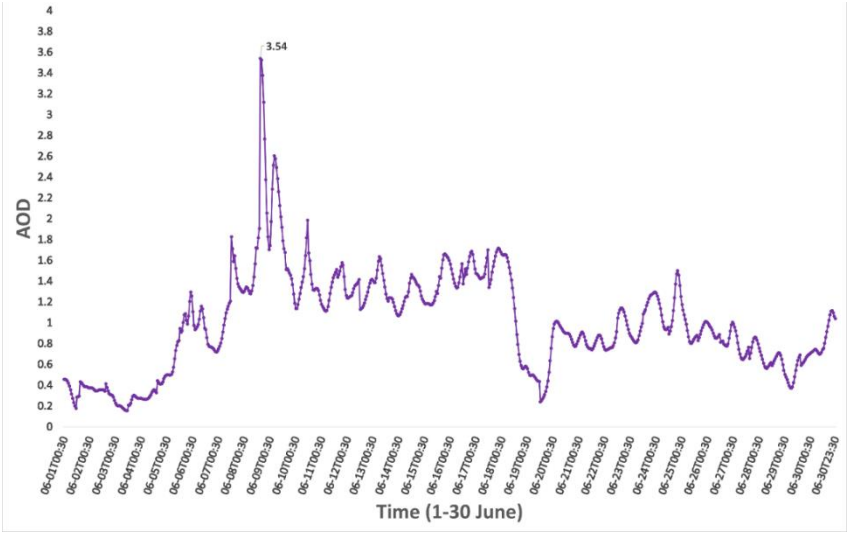

Figure 3.13. Bir Anzarane Morocco AOD values in June 2020

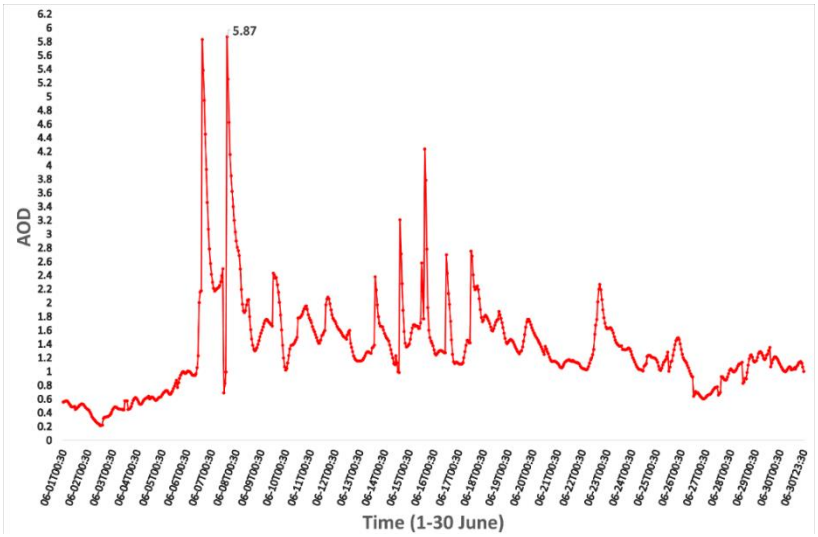

Figure 3.14. Nouakchott, Mauritania AOD values in June 2020

3.2.2 The identification and evaluation of the Saharan dust storm events in Budapest, Hungary between 2018 and 2022

Based on the daily 700 hPa geopotential height, wind maps, and SDE-specific dust transport paths, SDEs were divided into three primary synoptic meteorological groups by Varga, (2020). The various categories were distinguished by certain deterministic atmospheric patterns: Type-1 SDEs were linked to deep atmospheric depressions over Western Europe and north-western Africa. While, dust transport during Type-2 episodes was brought on by Central Mediterranean cyclones, while Type-3 events were defined based on the infrequent dust transport that occurred when dust-loaded air masses approached the Carpathian Basin from the north-western directions (Figure 3.15). From 2018 to 2022, 11 Saharan Dust events (SDEs) were identified in Hungary (Focus on the capital Budapest).

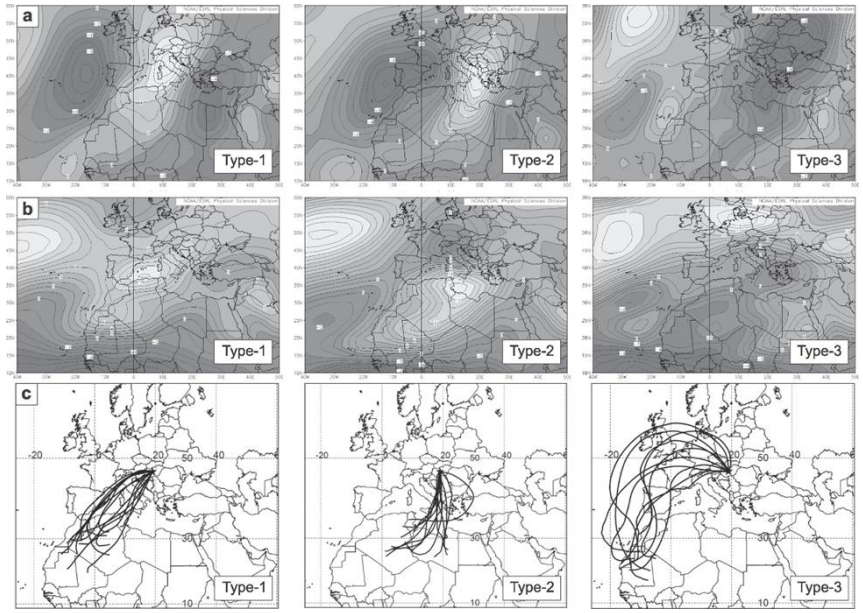

Figure 3.15. Wind flow patterns (mean meridional (a) and zonal (b) wind components at 700 hPa) and (c) specific dust transport routes at 3000 m above surface level by different Saharan dust event types (Varga, 2020).

SDE1: January 7–9, 2018

SDE1 was a type 2 event, Figure 3.16.A show the transport of the dust mass to Hungary at its peak form. The maximum Dust mass was 675.5 mg/m^2, and PM10 daily concentration increased by factor of 2.5, (from 22µg/m^3 in the 4th of January to 55 µg/m^3 in the 7th of January).

SDE2: February 7–8, 2018

SDE2 was also a type 2 as seen in Figure 3.16.B. The maximum Dust mass was 834.7 mg/m^2, and PM10 daily concentration increased by factor of 2, (from 31 µg/m^3 in the 4th of February to 66 µg/m^3 in the 9th of February).

SDE3: October 28–31/01–02 November, 2018

SDE3 was also a type 2 as seen in Figure 3.16.C, it was a two wave SDE, the first wave started to hit on the 28th of October and the second wave on the 1st of November. The maximum Dust mass was 505.5 mg/m^2 on the first wave and 367.4 mg/m^2 on the second wave, and PM10 daily concentration increased by factor of 3, (from 20 µg/m^3 in the 26th of October to 61 µg/m^3 in the 02nd of November).

SDE4: April 23-27, 2019

Even SDE4 was a type 2 as shown in Figure 3.16.D. This event was also a 2 waves SDE, the first wave started to hit on the 23rd of April and the second wave on the 26th of April. The maximum Dust mass was 993.8 mg/m^2 on the first wave and 952.9 mg/m^2 on the second wave, and PM10 daily concentration increased by factor of 1.7, (from 28 µg/m^3 in the 20th of April to 48 µg/m^3 in the 26th of April).

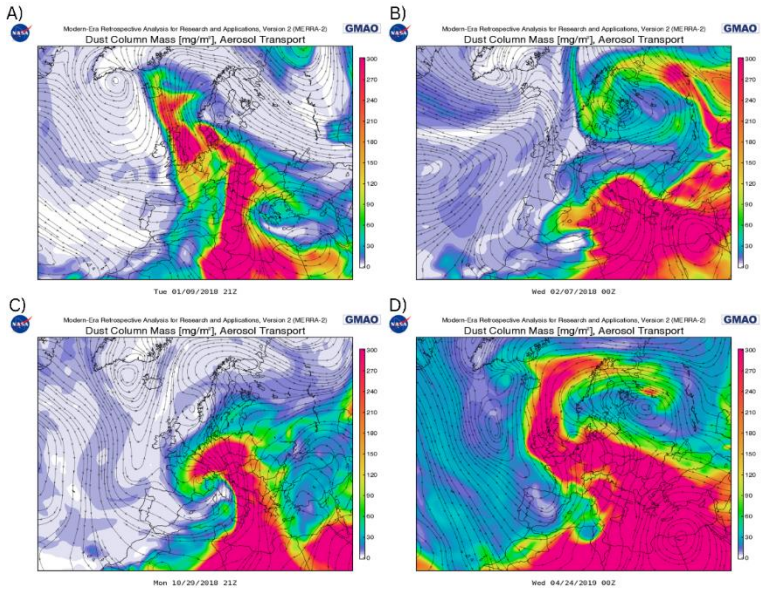

Figure 3.16. Dust Column Mass representation at its peak for A) SDE1, B) SDE2, C) SDE3, and D) SDE4.

SDE5: May 13-20, 2020

SDE5 was a type 1 as illustrated in Figure 3.17.A. This event was also a 2 waves SDE, the first wave started to hit on the 14th of April and the second wave on the 16th of April. The maximum Dust mass was 654.5 mg/m^2 on the first wave and 612.1 mg/m^2 on the second wave, and PM10 daily concentration increased by factor of 1.5, (from 18 µg/m^3 in the 12th of May to 28 µg/m^3 in the 20th of May).

SDE6: February 06-08, 2021

SDE6 is a type 1 (Figure 3.17.B). The maximum Dust mass was 989.9 mg/m^2, and PM10 daily concentration increased by factor of 2, (from 14 µg/m^3 in the 4th of February to 31 µg/m^3 in the 6th of February).

SDE7: February 23-26, 2021

SDE7 was a type 3 (Figure 3.17.C). The maximum Dust mass was 691.9 mg/m^2, and PM10 daily concentration increased by factor of 3, (from 29 µg/m^3 in the 21st of February to 92 µg/m^3 in the 26th of February).

SDE8: June 22-25, 2021

SDE8 was a type 1 (Figure 3.17.D). The maximum Dust mass was 719.8 mg/m^2, and PM10 daily concentration increased by factor of 2, (from 10 µg/m^3 in the 20th of February to 23 µg/m^3 in the 25th of February).

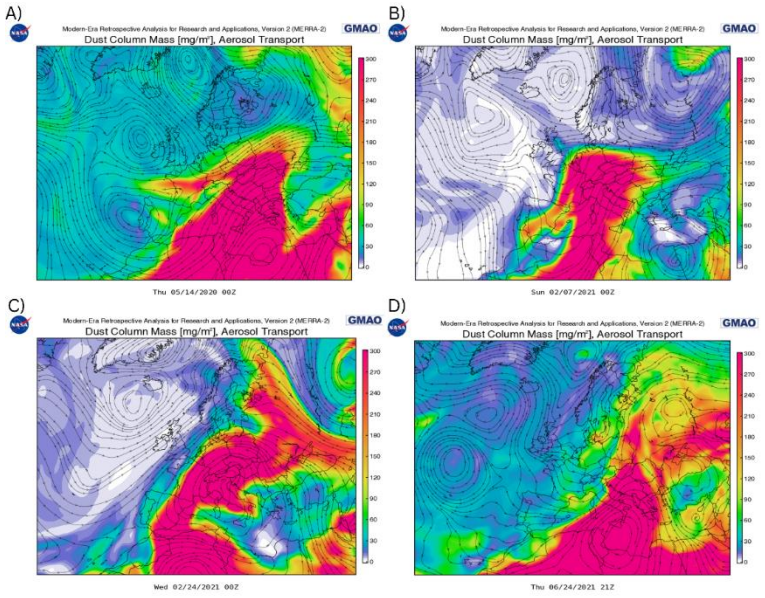

Figure 3.17. Dust Column Mass representation at its peak for A) SDE5, B) SDE6, C) SDE7, and D) SDE8.

SDE9: March 17-20, 2022

SDE9 was a type 3 (Figure 3.18.A). The maximum Dust mass was 719.8 mg/m^2, and PM10 daily concentration increased by factor of 2, (from 10 µg/m^3 in the 20th of February to 23 µg/m^3 in the 25th of February).

SDE10: March 29-31, 2022

SDE10 was a type 3 (Figure 3.18.B). The maximum Dust mass was 679.9 mg/m^2, and PM10 daily concentration increased by factor of 1.4, (from 29 µg/m^3 in the 26th of March to 43 µg/m^3 in the 30th of March).

SDE11: April 22-24, 2022

SDE11 was a type 2 (Figure 3.18.C). The maximum Dust mass was 699.7 mg/m^2, and PM10 daily concentration increased by factor of 2.4, (from 12 µg/m^3 in the 19th of April to 29 µg/m^3 in the 21st of April).

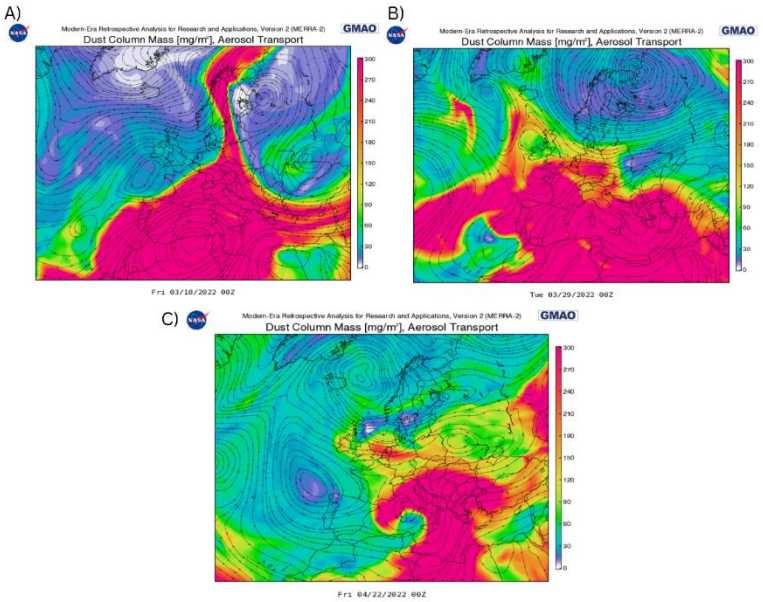

Figure 3.18. Dust Column Mass representation at its peak for A) SDE9, B) SDE10, C) SDE11

During the period of 2018 to 2022, type 2 SDEs were dominant (5 times – Figure 3.19.I), while type 1 and 3 both occurred 3 times each. In addition, February, March and April months are the months where the most of SDEs happened (7 times – Figure 3.19.II), and SDEs occurring in those months are more likely to be severe events, since the maximum dust mass registered is in

that period (SDE4), and also associated with an increase of PM10 daily average concentration by a factor of 2 or more.

With increasing distance from the source, dust's grain size decreases. When transported over long distances, coarse particles typically do not exceed 20μm because of their higher settling velocity (Does *et al.*, 2016). Mahowald *et al.* (2014) hypothesized that because coarser particles tend to settle out more readily, dust in the high atmosphere is finer grained than dust that has been deposited. Moreover, a high Saharan Dust Mass during SDE could lead to high increase in PM10 concentrations, but that depends on the dust particles size and deposition velocity, which mean that a relationship between dust mass during SDE and PM10 concentration is not always a direct positive relationship. Varga, (2020), highlight contravention of the numerical simulations that estimate the mineral grains sizes during SDE, and that the bulk of global and regional dust models only use a few size-bins with a rather restricted size range, hence mineral grains larger than 20 μm are typically not taken into consideration in the numerical simulations, and the direct measurements of individual particles illustrate that the mineral grain size during a SDE in the Carpathian region is about 40 μm.

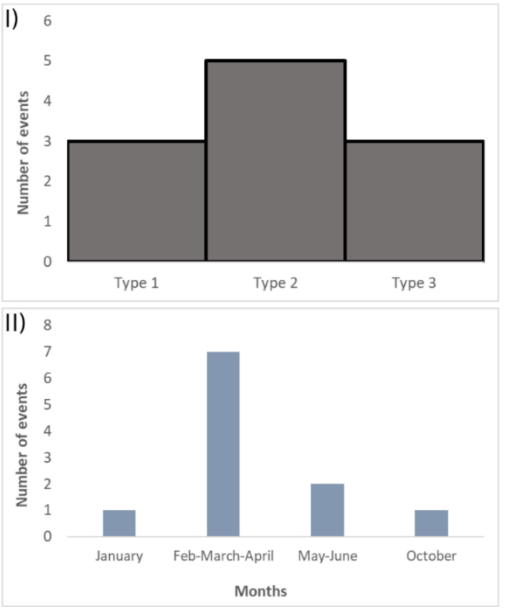

Figure 3.19. Frequency distribution of SDEs by number of events and I) Types, and II) Months of occurrence

3.2.3 Case study of the Saharan dust effects on PM10 and PM2.5 concentrations in Budapest in March 2022

3.2.3.1 PM10 and PM2.5 concentrations during the Saharan dust events

March 2022 dust storm events were type 3 events as described in previous chapter (3.4). The daily PM10 and PM2.5 concentrations increased in each Saharan dust event at a different percentage rate. During the first SDE (SDE9) in March 2022 (17th -20th), the daily PM10 and PM2.5 concentration jumped from 34 µg/m^3 and 18 µg/m^3 in 16th to 46 µg/m^3 and 28 µg/m^3 in the 18th and then start to decrease to reach 25 µg/m^3 and 15 µg/m^3 in the 20th. For the second SDE (SDE10) in March 2022 (28th – 31st), the daily PM10 and PM2.5 concentration changed from 29 µg/m^3 and 13 µg/m^3 on the 27th to 43 µg/m^3 and 18 µg/m^3 on the 30th after which begin to decline to attain 16 µg/m^3 and 14 µg/m^3 in the 1st of April 2022. Hourly PM10 and PM2.5 concentrations (Figure 3.20) provide details on how the hourly concentration changed with the SDE.

SDE9 was more intense than SDE10, as the effects were seen on the level of PM10 and PM2.5. The peak hourly concentration for PM10 was 86 µg/m^3 and 57 µg/m^3 for SDE9 and SDE10 respectively, while for PM2.5 it reached 51 µg/m^3 and 27 µg/m^3 as hourly concentration for SDE9 and SDE10 respectively.

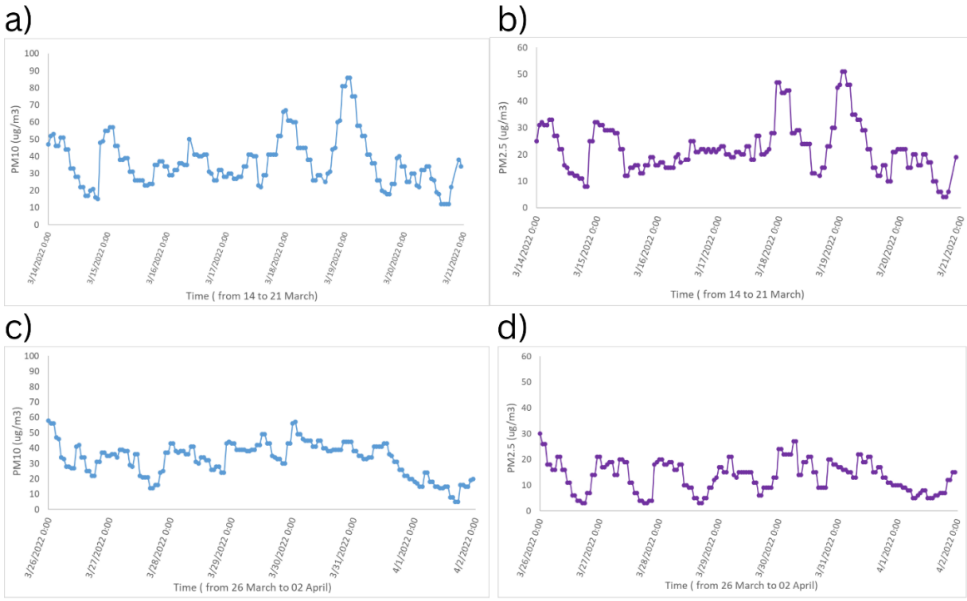

Figure 3.20. Hourly concentration of a) PM10 during SDE9, b) PM2.5 during SDE9, c) PM10 during SDE10, and d) PM2.5 during SDE10

About weather conditions, from 8 to 12 March 2022, the daily maximum temperature was between 7 and 10˚C, the wind direction from the North and Northeast direction, and the maximum wind speed was between 3.5 and 5 m/s, and no precipitations were registered. There was a slight increase in the maximum temperature (between 8 and 15˚C) in the following days, while from the 15th of March wind pattern full of Saharan dust arrived by western winds and in-ground surface the wind speed didn't exceed 3.5 m/s with no precipitations during SDE9.

The temperature continued to increase after SDE9, ranging from 16 and 21˚C as the daily maximum temperature, and start to decrease from the 31st of march and returns to the same levels that were at the beginning of March 2022, from the 2nd of April 2022 (between 5 and 9˚C). On the other hand, Saharan dust clouds were transported by West-Southwest, Southwest, and South-Southwestern winds, and maximum wind speed on the ground surface ranged from 3 and 5 m/s during the SDE10, and start to increase from the 1st of April 2022 to exceed 6 m/s as maximum wind speed, and no precipitations occurred on those days.

With increasing distance from the source, dust's grain size decreases. When transported over long distances, coarse particles typically do not exceed 20 µm because of their higher settling velocity. (Mahowald *et al.*, 2014; Does *et al.*, 2016) thought that dust in the upper atmosphere is finer grained than dust that has been deposited because coarser particles drop off more rapidly. Seasonally, summertime is when Saharan dust is coarser than wintertime. The high dust load in both SDEs triggered the increase of hourly PM10 and PM2.5 concentration.

The Sahara is the main source of dust in the Northern Hemisphere, and it is obvious that it has an impact on many different continents, from the fertilization of South America to the air quality in Europe. The Saharan dust storms affects the PM concentrations depending on the intensity of the storm. During March 2022 SDEs, PM10 and PM2.5 concentrations in an urban background air quality station in Budapest increased by 12 µg/m^3 and 10 µg/m^3 respectively during SDE9, and 14 µg/m^3 and 5 µg/m^3 during SDE10 as daily average concentrations. in Both SDEs the effect on PM10 was almost the same, while SDE9 raised the PM2.5 concentrations more than SDE10.

The effects of March 2022 Saharan dust events were similar to the one that was identified in 2016, on October 15th, which washed out a large amount of Saharan dust in the central European region, where it has several impacts, from flight security to air quality and impacts on solar radiation (Rostási *et al.*,

2022). However, March 2022 Saharan dust events had a more significant impact on air quality all over Europe, affecting a wider geographical area of western, central, and northern Europe, from Spain up to Scandinavia that led to an increase in PM concentrations (Liaskoni *et al.*, 2023; Uzunpinar *et al.*, 2023), and was associated with dust-infused cirrus clouds that persisted for nearly a week, affecting weather patterns and cloud cover over the region.

With the changing of the world climate, the intensity and the number of the Saharan dust storms episodes increase, and many models are still improving to provide more accurate forecast and to analyse the dust effects on different meteorological and air quality parameters.

3.3 Estimation and evaluation of PM concentrations

3.3.1 Evaluation of PM surface concentrations simulated by Version 5.12.4 of NASA's MERRA-2 Aerosol Reanalysis over Hungary in the period between 2019 and 2021

3.3.1.1 First Approach

In the machine learning algorithm, we used a split of 0.8x0.2. 80% of the data were used to train the model and 20% for the validation of the predicted values given by the trained model. To arrive to the results presented, we tried the model for many times, and each time the parameters of the machine learning algorithms (mainly the number of trees) were changed until arriving to the maximum results that can be achieved, where beyond that point the results whether they stopped improving or the performances start to decline.

For all the location chosen for this study, the use of equations 1 and 2 to estimate PM10 and PM2.5 result in an R^2 less than 0.1, and R^2 improved when coupling 5 species concentration used in equations 1 and 2 with meteorological data and AOD (Figure 3.21).

For Budapest Gilice tér station, in case of estimating PM10, MLR and OLS had low and a similar R^2 (0.22 and 0.21), RMSE (15.4 and 15.5 µg/m^3) and MAE (11.4 and 11.5 µg/m^3) values. For RF, R^2 of 0.75 achieved when N (the number of trees) was equal to 300, and RMSE and MAE were 9.1 µg/m^3 and 6.4 µg/m^3 respectively. While, for XT R^2 was equal to 0.78 when N=300, and RMSE and MAE were 8.1 µg/m^3 and 5.7 µg/m^3 respectively. Additionally, in case of estimating PM2.5, MLR and OLS also had low and a similar R^2 (0.27), RMSE (9.1 µg/m^3) and MAE (6.7 µg/m^3) values. For RF maximum value of R^2 (0.75) obtained when N was equal to 300, RMSE and MAE were 5.3 µg/m^3 and 3.5 µg/m^3 respectively. While, for XT, R^2 was equal to 0.8 when N=300, RMSE and MAE were 4.7 µg/m^3 and 3.1 µg/m^3 respectively.

The same was for Kecskemét, for PM10, low R^2 of 0.17 and 0.16 were obtained in MLR and OLS model respectively, and RMSE were 15.5 µg/m^3 and 15.6 µg/m^3, and MAE was identical for both models (11.5 µg/m^3). For RF, value of 0.66 for R^2 when N=270, and RMSE and MAE were 9.8 µg/m^3 and 6.8 µg/m^3 respectively. While, for XT R^2 was equal to 0.75 when N=300, and RMSE and MAE were 8.5 µg/m^3 and 6 µg/m^3 respectively. Moreover, in case of estimating PM2.5, MLR and OLS again had low R^2 (0.23 and 0.22 respectively), RMSE (11.3 µg/m^3 for both models) and MAE (7.5 µg/m^3 for both models). For RF maximum value of R^2 (0.69) obtained when N was equal to 300, RMSE and MAE were 7.1 µg/m^3 and 4.7 µg/m^3 respectively. While, for XT, R^2 was equal to 0.76 when N=300, RMSE and MAE were 6.2 µg/m^3 and 3.9 µg/m^3 respectively.

Furthermore, for Kazincbarcika in case of PM10, R^2 of 0.24 was obtained in MLR and OLS models, and RMSE was 18.4 µg/m^3 and 18.5 µg/m^3, and MAE 13.3 µg/m^3 and13.4 µg/m^3 respectively. For RF, value of 0.7 for R^2 when N=300, and RMSE and MAE were 11.4 µg/m^3 and 7.6 µg/m^3 respectively. While, for XT, R^2 was equal to 0.77 when N=300, and RMSE and MAE were 10 µg/m^3 and 6.6 µg/m^3 respectively. Additionally, in case of PM2.5, MLR and OLS again had low R^2 (0.34 and 0.33 respectively), RMSE (15.3 µg/m^3 for both models) and MAE (11.2 µg/m^3 for both models). For RF maximum value of R^2 (0.75) obtained when N was equal to 265, RMSE and MAE were 9.4 µg/m^3 and 6.3 µg/m^3 respectively. While, for XT, R^2 was equal to 0.8 when N=300, RMSE and MAE were 8.3 µg/m^3 and 5.5 µg/m^3 respectively.

a)

PM10	R2	RMSE	MAE
Eq1	>0.1	22.7	15.7
MLR	0.22	15.4	11.4
RF	0.72	9.1	6.4
XT	0.78	8.2	5.7
OLS	0.21	15.5	11.5

b)

PM2.5	R2	RMSE	MAE
Eq2	>0.1	11.7	7.9
MLR	0.27	9.1	6.7
RF	0.75	5.3	3.5
XT	0.8	4.7	3.1
OLS	0.27	9.1	6.7

c)

PM10	R2	RMSE	MAE
Eq1	>0.1	24.9	15.1
MLR	0.17	15.5	11.5
RF	0.66	9.8	6.8
XT	0.75	8.5	6
OLS	0.16	15.6	11.5

d)

PM2.5	R2	RMSE	MAE
Eq2	>0.1	14.2	8.9
MLR	0.23	11.3	7.5
RF	0.69	7.1	4.4
XT	0.76	6.2	3.9
OLS	0.22	11.3	7.5

e)

PM10	R2	RMSE	MAE
Eq1	>0.1	26	17
MLR	0.24	18.4	13.3
RF	0.7	11.4	7.6
XT	0.77	10	6.6
OLS	0.24	18.5	13.4

f)

PM2.5	R2	RMSE	MAE
Eq2	>0.1	22.1	14.2
MLR	0.34	15.3	11.2
RF	0.75	9.4	6.3
XT	0.8	8.3	5.5
OLS	0.33	15.3	11.2

Figure 3.21. Tables of performance statistic parameters for Budapest Gilice tér a) PM10, b) PM2.5, and Kecskemét c) PM10, d) PM2.5, and Kazincbarcika e) PM10 and f) PM2.5

The MERRAero dataset is useful tool to estimate PM10 and PM2.5 concentrations. The results show that the congruence in hourly PM10 and PM2.5 values between the observation and the calculated values based on equations 1 and 2 was inconsequential in all locations of the study. However, estimated PM10 and PM2.5 got better when coupling the estimations with meteorological data and component concentrations used in equations 1 and 2. In the three locations chosen for this study, MLR and OLS had poor R^2 (between 0.16 and 0.34), while the best R^2 was always achieved in case of XT model. The high RMSE and MAE results in case of Kazincbarcika compared to Budapest and Kecskemét, is due to the fact that concentrations of PM10

and PM2.5 registered in Kazincbarcika are higher than in Budapest and Kecskemét.

The use of sophisticated machine learning algorithms like RF and XT, gave better estimations of PM10 and PM2.5, in comparison to linear regression machine learning (MLR and OLS), and that is because of the complicated non-linear relationship between PM10 and PM2.5 to other variables like meteorological data.

3.3.1.2 Second Approach

Figure 2.7, describe the method used in this section. The estimation in this approach was done only using 4 machine learning algorithms (MLR, OLS, RF, and XT) to estimate PM2.5 concentrations based on 10 variables (AOD, O_3, NO_2, SO_2, T, Ws_{10}, Ws_{50}, RH, P, and PBLH). Figure 3.22 summarize the results of the MLR, OLS, RF, and XT models.

Multiple linear regression and Ordinary least square regression:

Except for Kazincbarcika where R^2 was in good range (0.65 and 0.64) for MLR and OLS, the value obtained in case of Budapest Gilice tér and Kecskemét were low values (between 0.29 and 0.32).

Random Forest:

For RF, the peak performance was achieved when N=1700 for the three locations. The results show that R^2 value using the RF regression machine learning algorithm was 0.69, 0.71, 0.83 for Budapest Gilice tér, Kecskemét, and Kazincbarcika respectively, overall RMSE was 5.9, 6.9 and 7.9 µg/m^3 and MAE was 4, 4.5 and 5.1 µg/m^3 respectively.

Extra Tree regression:

For XT, the peak performance was achieved when N =1000 for Kecskemét, and Kazincbarcika and 1100 for Budapest Gilice tér. The results show that R^2 value between the estimated and observed PM2.5 using the XT regression machine learning algorithm was 0.73, 0.75, and 0.84 for Budapest Gilice tér, Kecskemét, and Kazincbarcika respectively, overall RMSE was 5.5, 6.4, and 7.6 µg/m^3, and MAE was 3.7, 4.2, and 4.8 µg/m^3.

I)	R2	RMSE	MAE
MLR	0.3	8.9	6.5
RF	0.69	5.9	4
XT	0.73	5.5	3.7
OLS	0.29	9	6.5

II)	R2	RMSE	MAE
MLR	0.32	11.2	7.5
RF	0.71	6.9	4.5
XT	0.75	6.4	4.2
OLS	0.32	11.2	7.5

III)	R2	RMSE	MAE
MLR	0.65	11.4	8.1
RF	0.83	7.8	5.1
XT	0.84	7.6	4.8
OLS	0.64	11.5	8.2

Figure 3.22. Tables of performance statistic parameters for I) Budapest Gilice tér, II) Kecskemét, and III) Kazincbarcika

In RF algorithm, each tree in the ensemble is constructed from a sample selected with substitute from the training set. Additionally, while partitioning each node throughout tree construction, the optimum split is determined by selecting either all input features or a random subset of size. The goal of these two randomness sources is to reduce the variance of the forest estimator. Individual decision trees, in fact, have a large variation and tend to overfit. Forests with injected randomness provide decision trees with partially dissociated prediction errors. Some inaccuracies can be eliminated by taking an average of such projections. RFs minimize variance by merging various trees, sometimes at the expense of a modest bias increase. In reality, the variance decrease is frequently large, resulting in a superior overall model. The way splits are produced in XT algorithm goes even further. A random subset of candidate features is employed, much as in RF, but instead of looking for the most discriminative thresholds, thresholds are produced at random for each candidate feature, and the best of these randomly-generated thresholds is chosen as the splitting criterion. This generally allows for a little reduction in model variance at the price of a slight increase in bias.

The MERRAero dataset is a valuable tool for estimating PM10 and PM2.5 concentrations. The results show that the congruence in hourly PM10 and PM2.5 values between the observation and the calculated values based on equations 1 and 2 was inconsequential in all study locations. However, estimated PM10 and PM2.5 improved when coupling the estimations with meteorological data and component concentrations used in equations 1 and 2. In the three sites chosen for this study, MLR and OLS had poor R^2 (between 0.16 and 0.34), while the best R2 was always achieved in the case of the XT model. The high RMSE and MAE results in the case of Kazincbarcika compared to Budapest and Kecskemét can be noticed due to the concentrations of PM10 and PM2.5 registered in Kazincbarcika, which are higher than in Budapest and Kecskemét.

In addition to the absence of nitrate particle concentrations, Provençal *et al.* (2017) explains the incongruence between observed and simulated PM2.5, which is probably due to a combination of [SO_4], [OC] and [BC] differences. Additionally, Buchard *et al.* (2016) noticed a disparity in carbonaceous particle concentrations in suburban areas of the United States. Many additional research has proposed adding nitrate concentrations to improve MERRA-2 PM2.5 estimates (He *et al.*, 2019; Ma, Xu and Qu, 2020), while poor MERRA-2 PM estimations were claimed to be caused mostly by the use of the Goddard Earth Observing System, version-5 (GEOS-5) model's bottom-up emission database and meteorological issues in GOES-5 simulations (Song *et al.*, 2018). According to Ali *et al.* (2022), multiple statistical models can be used to estimate PM2.5 using MERRA-2 aerosol reanalysis data, with the random forest model having the highest accuracy. Their results indicate that the random forest model is an appropriate choice for calculating PM2.5 concentrations in China.

It was demonstrated that machine learning is a valuable method for predicting PM2.5 by using algorithms to estimate PM2.5 based on MERRA-2 AOD, Meteorological, NO_2, O_3, and SO_2 data in 3 years (2019 to 2021) for Budapest. A comparison between 4 machine learning approaches revealed that the Extra-Tree regression model outperformed other models like RF, MLR and OLS. For Budapest the results of XT model for estimation of PM2.5 give an R^2 of 0.73, RMSE of 5.5 $\mu g/m^3$, and MAE of 3.7 $\mu g/m^3$. For Kecskemét an R^2 of 0.75, RMSE of 6.4 $\mu g/m^3$, and MAE of 4.2 $\mu g/m^3$. And for Kazincbarcika an R^2 of 0.84, RMSE of 7.6 $\mu g/m^3$, and MAE of 4.8 $\mu g/m^3$. The use of sophisticated machine learning algorithms like RF and XT gave better estimations of PM10 and PM2.5, compared to linear regression machine learning (MLR and OLS); that is because of the complicated non-linear

relationship between PM10 and PM2.5 to other variables like meteorological data.

3.3.2 Calibration of CAMS PM2.5 data over Hungary using machine learning

Python 3.9.17 was used to write a code that performs data preprocessing, model training, prediction, evaluation, and visualization, for the data using the LightGBM regression model.

Promising results were obtained from the calibration of CAMS PM2.5 data using the LightGBM algorithm. The correlations before and after training the model were analysed, revealing noticeable improvements in prediction accuracy (Figures 3.23 and 3.24). Before training, the correlations between the observed and CAMS PM2.5 data varied across the stations, ranging from 0.07 to 0.20. However, after training, the correlations significantly increased, ranging from 0.78 to 0.88. These enhanced correlations demonstrate the efficacy of the LightGBM algorithm in capturing the relationships between the input features and PM2.5 levels, leading to improved accuracy in predicting air quality.

The evaluation metrics, such as the R^2 scores and root mean squared error (RMSE), were utilized to assess the model's performance. The R^2 scores, which measure the model's ability to explain the variance in observed PM2.5 values, ranged from 0.61 to 0.77. This indicates that the model accounted for 61% to 77.4% of the variance, indicating a good fit to the data. Furthermore, the RMSE values, representing the average magnitude of the differences between predicted and observed values, ranged from 5.31 to 9.92 $\mu g/m^3$. Lower RMSE values indicate higher precision and accuracy in the model's predictions.

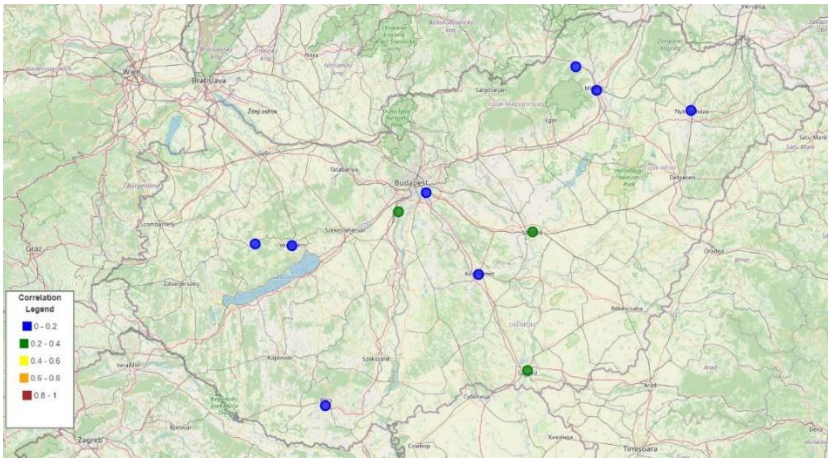

Figure 3.23. Correlation map between CAMS and In-situ PM2.5 before calibration

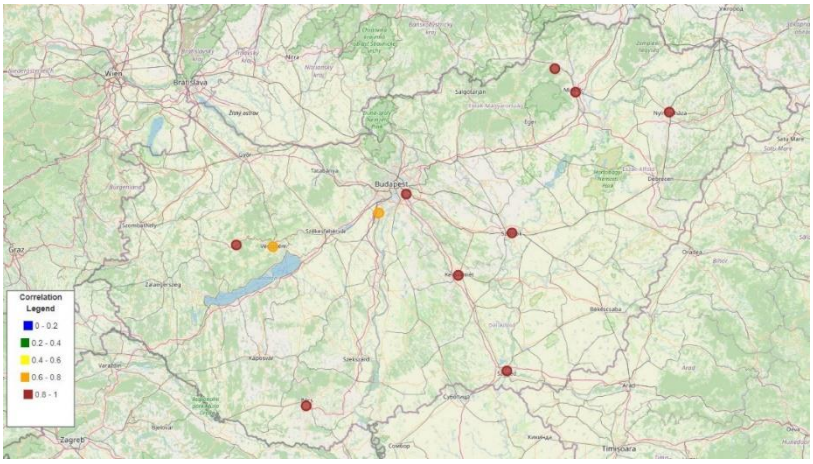

Figure 3.24. Correlation map between the predicted and In-situ PM2.5 after calibration

Furthermore, the scatter plots (Figures 3.25 and 3.26) also shed light on the model's predictive power by showcasing the proximity of the predicted PM2.5 values to the observed values. The close alignment between the predicted and in situ PM2.5 data points in the scatter plots signifies the model's ability to capture the underlying patterns and relationships. The proximity between these points reinforces the improved correlations observed after training, substantiating the effectiveness of the LightGBM algorithm in calibrating CAMS PM2.5 data. The plots demonstrate enhanced correlations and close alignment between predicted and in situ PM2.5 data points, highlighting the algorithm's ability to accurately predict PM2.5 levels.

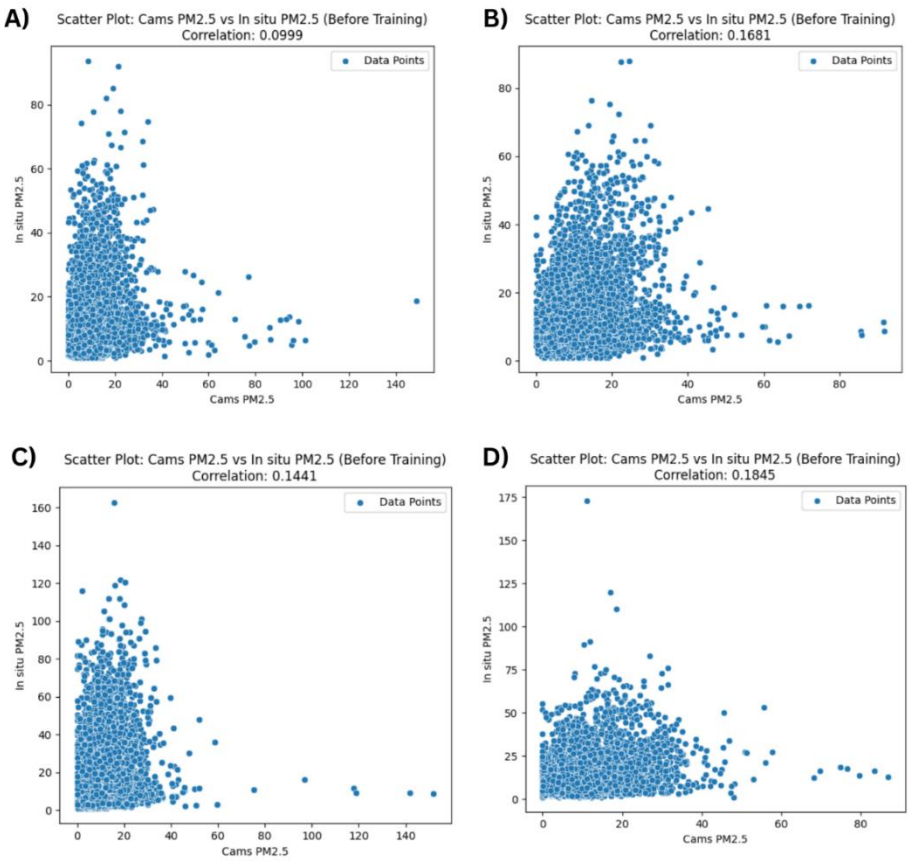

Figure 3.25. Scatter plots of raw CAMS PM2.5 data for A) Ajka station, B) Budapest Gilice ter station, C) Kazincbarcika station, D) Kecskemet station

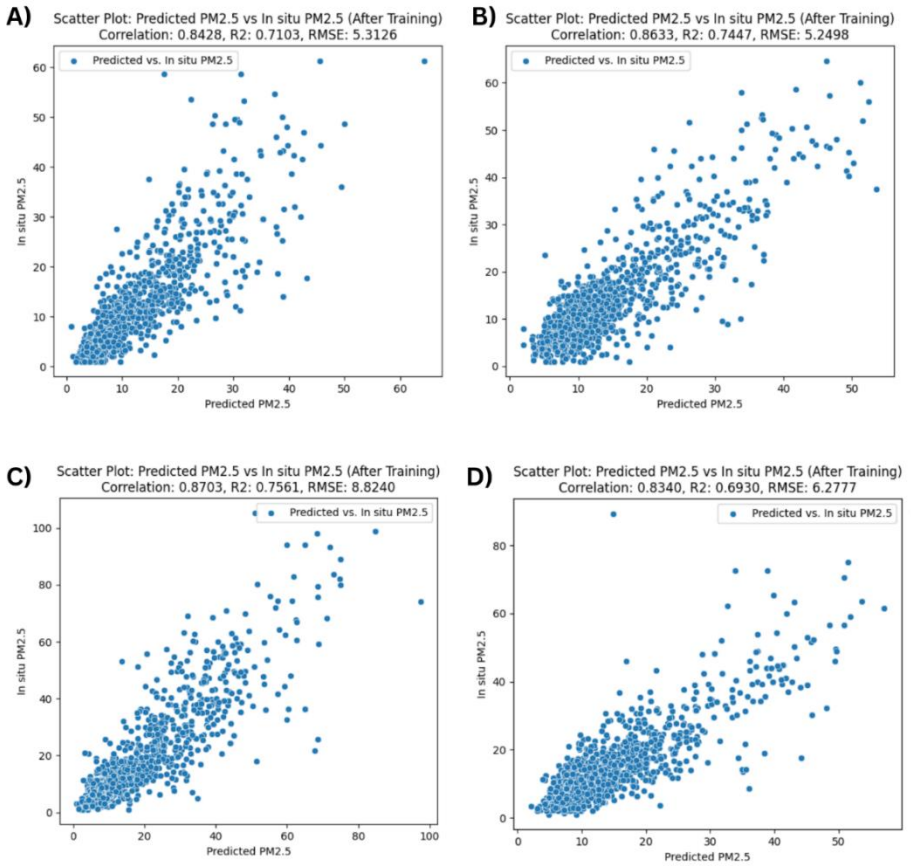

Figure 3.26. Scatter plots of predicted and In-situ PM2.5 for A) Ajka station, B) Budapest Gilice ter station, C) Kazincbarcika station, D) Kecskemet station

The results of this study calibration of CAMS PM2.5 data using the LightGBM algorithm align with other studies' conclusions, emphasizing the necessity of increasing the accuracy and dependability of CAMS datasets. In agreement with Ali *et al.* (2022), who reported CAMS overestimation in several places, our findings suggest that CAMS PM2.5 data had greater correlation values after calibration. Before training, the correlation values ranged from 0.0719 to 0.2072, demonstrating a moderate relationship between CAMS PM2.5 and in-situ PM2.5 readings. However, following training, the correlation values improved significantly, ranging from 0.7869 to 0.8820, indicating a greater link between the calibrated PM2.5 levels and in-situ observations. (Gueymard and Yang, 2020) also emphasized the limits of raw CAMS PM2.5 estimates, such as coarse spatial resolution and modelling biases. Our findings back with their conclusions, as results of the current paper

showed low correlation values between raw CAMS and measurements PM2.5 concentrations. However, after applying the LightGBM calibration technique, the correlations increased significantly, and RMSE values were low, indicating that the calibrated PM2.5 estimates were more accurate.

Furthermore, the present results are compatible with the findings of (Jin *et al.*, 2022), who stressed the possibility of distinct calibration schemes to improve CAMS products. After calibration, the rise in coefficient of determination (R^2) values demonstrates the reduction in modelling biases and the improved performance of the calibrated PM2.5 data.

Overall, the calibration results using LightGBM algorithm are consistent with prior studies, suggesting that calibration approaches can significantly increase the accuracy and reliability of CAMS PM2.5 estimates. The calibrated PM2.5 data better matches with ground-based observations by correcting overestimation and lowering modelling biases, giving more reliable information for air quality assessments and decision-making processes. These findings emphasize the importance of calibration in increasing the utility and reliability of CAMS PM2.5 data for environmental monitoring and public health activities.

4 Conclusions and recommendations

The study of PM pollutants is a broad field with many branches, from studying the chemical composition of the PM particles to improving measurements and developing computer-based models to forecast PM pollution of any kind. The research conducted in this PhD thesis has contributed valuable insights into the impact of obstacles, hills, and Saharan dust storms on PM concentrations and the use of satellite-based models and machine learning algorithms to estimate PM concentrations and calibrate CAMS PM2.5 data. Based on the findings of this research, the following conclusions and recommendations can be made:

The study of the effects of simple obstacles on the PM10 concentration provides valuable insights into the effects of obstacle height, distance from the source, and wind speed on PM10 concentration in different sensor locations. The study's results indicate that obstacles can significantly impact PM10 concentration levels, with higher obstacle heights and greater distances from the source leading to decreased PM10 concentration levels. Additionally, the study found that wind speed plays a crucial role in PM10 concentration levels, with higher wind speeds leading to higher PM10 concentration levels. The study's findings are consistent with previous research on the effects of urban architecture and wind speed on PM concentrations. For example, a study on the influence of wind speed on airflow and fine particle transport within different building layouts of an industrial city found that height variation and layout of urban architecture affect the local concentration distribution of PM (Mei *et al.*, 2018). Similarly, another study on the effects of windbreaks on particle concentrations from agricultural fields under various wind conditions showed that vegetation barriers can alter particle transport by affecting airflow (Chang *et al.*, 2019). In conclusion, the results of the research paragraph provide valuable insights into the effects of obstacle height, distance from the source, and wind speed on PM10 concentration levels. By considering the impact of obstacles and wind speed on PM10 concentration levels, policymakers and urban planners can develop effective strategies to minimise the impact of PM10 on public health in urban environments.

The study the impact of hill elevation on the dispersion of PM plumes showed that the correlation between PM10 concentration and wind speed at different sensor locations was affected. The results show that hills can significantly affect the dispersion pathway of PM particles, with different slopes creating other flows of PM dispersion. The study found that the correlation between PM10 concentration and wind speed decreases as the height of the hill increases, indicating changes in the wind flow created by different elevations

of the hill and that, in general, topography can significantly impact the level of PM concentrations. In a study done by Wen *et al.* (2022), they discuss the quantitative disentanglement of topography's geographical impacts on PM2.5 pollution in China. They emphasise that mountains significantly impact the spatial heterogeneity of PM2.5 pollution levels. The study found that high-altitude mountains and plateaus experience lower levels of PM2.5 pollution, while plains and surrounding platforms and hills suffer from severe pollution. Also, the mountain's blocking effects begin to play an efficient role when their altitudes reach a specific value; however, the exact altitude values vary by different mountains, with a value of 163 m for all typical mountains with absolute PM2.5 concentration differences between their two sides greater than 10 μg/m3. Even though the experiments included relatively low height, it showed that height can alter the PM concentrations, even in short range. In conclusion, the research results provide valuable insights into the effects of hill elevation on the dispersion of PM plumes and the correlation between PM10 concentration and wind speed at different sensor locations.

The study of Saharan dust storms in Hungary revealed that they increased PM10 and PM2.5 concentration levels, and the seasonality and frequency are changing. Varga (2020) indicates numerous intense events happened after 2014 when an unusually significant amount of mineral dust was washed out. All occurred between the end of October and February, and the increase in frequency and intensity of wintertime dust depositional events in the Carpathian Basin (Hungary mainly) is attributed to climatic conditions. Our study for the period between 2018 and 2022 showed changes in the frequency of the intense Saharan dust storms in Hungary (more likely to happen between February and April) and also the intensity (Recent March 2022 as an example), and dominant Type-2 events that are connected to Central Mediterranean cyclones which is responsible for dust transport.

Machine learning is a powerful tool that can be used to estimate PM concentrations from MERRA-2 data. In the study, we estimate PM concentrations using two approaches, mainly coupling MERRA-2 AOD and five PM components with meteorological data and MERRA-2 AOD and in-situ measurements of primary air pollutants (SO2, NO2, O3) with meteorological data. The best results were given using the Extra-tree machine learning algorithm in the two approaches for the three stations chosen for Hungary (Budapest, Kazincbarcika, and Kecskemet) with an R2 between 0.73 and 0.84. Our models performed well for the selected stations compared to other studies that used MERRA-2 data to estimate PM concentrations using machine learning. Dhandapani, Iqbal and Kumar (2023) Apply machine

learning models on MERRA2 data to predict surface PM2.5 concentrations over India. Overall, the study evaluates the utility of Machine Learning (ML) models, focusing on XGBoost (XGB), Random Forest (RF), and LightGBM (LGBM) individual models, as well as a stacking technique. The authors compared the performance of these models and found that the stacking technique (R2= 0.77) outperformed unique models (R^2=0.73), showing the best hourly prediction in the eastern (R^2 = 0.80) and northern regions (R^2 = 0.63). In another study by Sayeed *et al.* (2022), they evaluated the performance of the machine learning model in estimating PM2.5 concentration, which outperformed the MERRA-2 empirical estimation of PM2.5 and exhibited a small and uniform bias throughout the day and in all seasons and proved to be better at estimating PM2.5 than the MERRA-2 practical calculations. In conclusion, Coupling MERRA-2 and meteorological data with other meaningful parameters and employing machine learning to predict PM concentrations can yield promising results, as demonstrated in our case for the three stations in Hungary.

For the calibration of CAMS PM2.5 data using machine learning in Hungary, our model improved the degree of accuracy of CAMS PM2.5 data from low R (>0.25) to higher R (<0.79), demonstrating the effectiveness of calibration schemes using LightGBM machine learning model in reducing modelling biases and enhancing the performance of CAMS PM2.5 data in Hungary. Jin *et al.* (2022) proposes a calibration method to improve the accuracy of CAMS PM2.5 data, using the Extremely Tree machine learning model, resulting in significant accuracy improvement, with R reaching up to 0.81 and RMSE decreasing by about 60% for the original CAMS PM2.5 For China, US and Africa. Overall, CAMS reanalysis datasets require significant improvement for use in local and regional air quality monitoring, and our study showed a significantly better correlation between the Calibrated PM2.5 and in-situ measurements of PM2.5 over Hungary, suggesting an improved accuracy of Calibrated CAMS PM2.5.

Concerning the recommendations, further research on the impact of obstacles on PM concentrations: It is recommended that the impact of different types of obstacles on PM concentrations be understood and that more accurate models to predict PM concentrations in the presence of obstacles be developed.

Use of satellite-based models to estimate PM pollutants: It is recommended to use satellite-based models to estimate PM pollutants in Hungary and to compare the results with in situ measurements to validate the accuracy of the models.

Further research on the impact of Saharan dust storms on PM concentrations: It is recommended that the impact of different types of dust storms on PM concentrations be understood and that more accurate models to predict PM concentrations during dust storms be developed.

Use of machine learning algorithms to estimate PM concentrations: It is recommended to use machine learning algorithms to estimate PM concentrations in other locations in Hungary and to expand the study by using more extensive data.

In conclusion, the research conducted in this PhD thesis has contributed valuable insights into the impact of various factors on PM concentrations. It has provided recommendations for further research to improve the accuracy of PM concentration predictions. This research's findings can help protect public health and the environment by providing accurate and reliable PM concentration data.

5 New scientific results

1. *As a result of small-scale experiments, I have shown that the PM10 concentration from a point source depends on the wind speed (Ws=0-2.9 m/s) and the height and position of a simple "obstacle" placed between the source and the measurement point - with a critical ratio of obstacle distance from the source over obstacle height ranging (OD/OH) from 0 to 2.3. A linear equation established the relationship between the mentioned parameters to calculate PM10 concentration, reaching an R^2 of 0.79. I declare that, the obstacle's position emerges as an essential determinant shaping the estimated PM10 concentration, underscoring its profound significance in our findings.*

Research conducted a multiple regression analysis to predict $PM10_A$ concentration using "OH" (Obstacle height), "OD" (Distance of the obstacle from the source), "$PM10_C$" (PM10 concentration), and "Ws" (Wind speed) as independent variables. The regression model equation was:

$$PM10_A = 143.07 - 71.86 * OH - 171.42 * OD + 1.23 * PM10_C + 12.34 * Ws$$

Results showed a moderately significant positive connection between $PM10_A$ and the independent variables (R=0.89). The model explained 79% of $PM10_A$ variation (R^2=0.79). Obstacle height, distance from the source, and wind speed had significant positive effects on $PM10_A$.
The experiment elicits know how a simple obstacle in the form of a solid barrier can reduce the PM concentrations. And using the equation, it is possible to strategically position the obstacle to obstruct the PM particle movements and opt for the optimal possible height to effectively isolate the area from direct PM transportation to minimize the PM concentration levels to acceptable levels.

2. *I conclude that hill height influences PM10-wind speed correlations negatively as a result of a series of outdoor trials investigating the effects of different ground surface elevations (H=0-1 m) on the dispersion of PM10 and PM2.5 pollutants at various wind speeds (Ws=0-6 m/s) in short range, revealing the complicated interplay between topography and air pollution patterns.*

The study examines the correlations between PM10 concentrations recorded by three different sensors (S1, S2, and S3) over three different experimental situations with varied wind speeds on a hill with varying heights. Correlations between PM10 concentrations and wind speed differ between sensors (S1, S2, S3) and experimental cases on a hill with variable elevation.
Sensor 1 (S1), which is near to the source, has a negative correlation of -0.5 in Case 3, but weaker and inverted correlations to -0.18 and 0.2 in Cases 2 and 1 respectively. Sensor 2 (S2) positioned at the edge of the hill's slope, has continuous positive correlations in all cases, however they are weaker in case 2 (0.27) and 3 (0.18), probably due to changes in nearby terrain. Sensor 3 (S3), located atop the hill, retains positive associations, although they decline as elevation increases, yielding values of 0.65 (Case 3), 0.84 (Case 2), and 0.8 (Case 1), suggesting that shifting wind patterns impact PM10 transport.

3. *HYSPLIT dust simulations offer compelling insights into the origin and trajectory of Saharan dust particles. Our analysis reveals that these particles observed in the Gulf of Mexico during the June 2020 Saharan storm unmistakably trace back to the Moroccan and Mauritanian Saharan regions. Dust storms emerged from specific hotspots, such as Tinduf near the Moroccan border and Adrar, Tiris Zemmour, and Tagant in Mauritania. Markedly, the June 2020 Saharan dust storm was associated with the highest June aerosol optical depths recorded, exceeding AOD=3.5 in Bir Anzarane, Morocco, and an astonishing AOD=5.5 in Nouakchott, Mauritania, affecting the PM concentrations to unhealthy levels in several US Gulf States, further substantiating that due to climate change, the Saharan dust storms are getting more intense, especially in the Moroccan and Mauritanian Sahara.*

   Based on HYSPLIT dust simulation and also comparing the simulation results maps with MODIS AOD average maps, the dust storm in the region of Morocco and Mauritania started on the 14th of June 2020, from the region of Tinduf close to the borders of Morocco with Algeria, Adrar Tiris Zemmour and Tagant in Mauritania. Due to the wind field heading towards the Atlantic oceans, the dust was transported across the ocean to the American continents. In addition,

HYSPLIT cluster analysis from many places in Morocco and Mauritania showed a significant percentage of PM10 particles that negatively affected the PM10 and PM2.5 concentrations in the Caribbean Sea and US coastal in the Gulf of Mexico, originated from places like Bir Anzarane, Morocco, Nouakchott, and Tichit Mauritania, and Bordj Badji Mokhtar Algeria. In addition, there was an increase in Bir Anzarane, Morocco, in AOD with a value of 3.522. This marks a surge of 188% compared to the highest recorded value between 2010 and 2019, 1.87 in June 2017. Similarly, in Nouakchott, Mauritania, the peak AOD value from 2010 to 2019 was 2.78 in June 2010. However, there was a rise in June 2020, with an AOD value reaching 5.87, representing an increase of around 211%.

4. *I conclude that events where the Saharan dust transport was brought on by Central Mediterranean cyclones to Hungary, - called type 2 Saharan dust storm events - were dominant in 2018 and 2022 and usually happen in February, March, or April, with a maximum hourly dust mass between 450 and 1000 mg/m2. The 2 Saharan Dust storms in March 2022 raised the concentrations of PM10 and PM2.5 in Budapest by 12 µg/m3 and 10 µg/m3, respectively, during the first Saharan Dust event and by 14 µg/m3 and 5 µg/m3 during the Second Saharan Dust event.*

   Based on my evaluation of the Saharan dust storm events in Hungary, between 2018 and 2022, 11 SDEs were identified in Hungary. Type 2 SDEs were dominated in that period and characterised by high Dust mass, negatively affecting the PM concentrations. And most of the time, the SDEs were likely to occur between February and April. Moreover, March 2022 was a unique month due to two extreme outbreaks of Saharan Dust events (14-19 and 28-31), which were unusual throughout the study period. In general, the Saharan dust events between 2018 and 2022 were associated with an increase of PM10 daily average concentration by a factor of 2 or more, according to PM10 concentration measurements from an urban background air quality station in Budapest.

5. *I find that PM10 and PM2.5 concentrations simulated from MERRAero data, encompassing five PM species (SO_4, OC, BC, DS and SS), AOD, and meteorological parameters (T, Ws_{10}, Ws_{50}, RH, P and PBLH), between 2019 and 2021 accurately estimated using the Extra-Tree regression model for three cities in Hungary (Budapest, Kecskemét and Kazincbarcika), achieving R^2 values between 0.75 and 0.8 for PM10 and PM2.5.*

Based on my evaluation of the estimated PM10 and PM2.5 based on the five PM species simulated by the MERRAero hourly data collection of (SO_4, OC, BC, DS, and SS), in conjunction with aerosol optical depth (AOD) and meteorological parameters (T, Ws_{10}, Ws_{50}, RH, P, and PBLH) for the period spanning 2019 to 2021, using a variety of machine learning algorithms, it is discerned that the Extra-Tree regression model consistently produced the most favourable outcomes. The quantitative results, disaggregated by location, are as follows:

- In Budapest, the determination coefficient (R^2) reached 0.78 and 0.8 for the estimation of PM10 and PM2.5, respectively.
- In Kecskemét, the R^2 values achieved were 0.75 and 0.76 for PM10 and PM2.5 estimation, respectively.
- For Kazincbarcika, the R^2 values obtained for PM10 and PM2.5 were 0.77 and 0.8, respectively.

The significance of these results lies in their potential to enhance air quality monitoring and forecasting in urban areas such as Budapest, Kecskemét, and Kazincbarcika. The Extra-Tree regression model demonstrates robust predictive capabilities, with R^2 values consistently approaching or exceeding 0.75.

6. *Utilizing Machine learning algorithm (Extra-Tree regression model) to estimate PM2.5 concentrations based on MERRA-2 AOD, meteorological data (T, Ws_{10}, Ws_{50}, RH, P and PBLH), and in-situ measurements of NO_2, O_3, and SO_2 over three years (2019 to 2021) in 3 locations in Hungary (Budapest, Kecskemét and Kazincbarcika) underscores the importance of machine learning in PM2.5 prediction attaining an R^2 ranging from 0.73 to 0.83, and RMSE between 5.5 and 7.6 $\mu g/m^3$.*

The utilisation of machine learning algorithms to estimate PM2.5 concentrations based on a comprehensive dataset comprising MERRA-2 AOD, meteorological data, and in-situ measurements of NO2, O3, and SO2 over three years (2019 to 2021) for 3 locations in Hungary underscored the effectiveness of machine learning as a valuable predictive tool but also revealed the superiority of the Extra-Tree regression model over alternative approaches. Specific results for each location are as follows:

- In case of Budapest, I had an R^2 of 0.73, RMSE of 5.5 $\mu g/m^3$, and MAE of 3.7 $\mu g/m^3$
- In case of Kecskemét, I had an R^2 of 0.75, RMSE of 6.4 $\mu g/m^3$, and MAE of 4.2 $\mu g/m^3$.
- In case of Kazincbarcika, I had an R^2 of 0.84, RMSE of 7.6 $\mu g/m^3$, and MAE of 4.8 $\mu g/m^3$.

These findings hold paramount importance as they affirm the applicability of machine learning for precise PM2.5 predictions, offering a robust and versatile methodology for air quality assessment and prediction in these specific geographical areas. Such accurate predictive models are instrumental for public health, urban planning, and environmental management

7. *Using the LightGBM algorithm in calibrating CAMS PM2.5 data for 11 air quality stations in Hungary reveals a remarkable improvement in data accuracy and alignment with in-situ measurements with post-calibration, correlations substantially increased, with values ranging from 0.78 to 0.88, underscoring a solid association between calibrated CAMS data and actual PM2.5 measurements, and a coefficient of determination values ranging from 0.61 to 0.77.*

    - Correlation analysis shows initial alignment between raw CAMS data and in-situ measurements, with correlations before training ranging from 0.071 to 0.207. After training, correlations significantly improve, ranging from 0.787 to 0.882, demonstrating a strong association between calibrated CAMS data and in-situ PM2.5 measurements.
    - The coefficient of determination values ranges from 0.618 to 0.774, indicating a substantial portion of the variance in in-situ PM2.5 measurements is explained by the calibrated CAMS PM2.5.

- Lower root mean square error values reflect reduced discrepancies between the calibrated CAMS PM2.5 and actual measurements, indicating improved accuracy and precision.

The findings underscore the critical role of calibration in improving the accuracy of air quality data (such as CAMS PM raw data). Enhanced correlations, higher coefficient of determination values, and reduced root mean square error values following machine learning calibration are scientifically significant and have direct practical implications.

6 SUMMARY

Studying the PM pollutants is broad field that have many branches, from studying the chemical composition of the PM particles to improving measurements and developing computer-based models to forecast PM pollutions of any kind. In the initial phase of my research, I looked over the literature in a few relevant subfields, which led to Saharan dust storm study both the June 2020 event and to know the effect of climate change on the triggering and transport of Saharan dust to Hungary. In addition to understand the dispersion of PM particles around simple obstacle and in small range elevated hills, as well as understanding the relationship between inside and outside PM concentrations. Moreover, deep search in the literature lights the fact that no one has done an estimation of the PM pollutants in Hungary using the Satellite based models, despite the fact of the increasing interest in this subfield of research and the rise of the number of papers published in order to improve the use of the Satellite datasets to estimate one of the major and dangerous air pollutants.

The study of small-scale PM dispersion around simple obstacle demonstrated that Obstacle height, Distance of the obstacle from the source, and Wind speed had significant positive effects on PM10 concentration after the obstacle. The analysis revealed a moderately significant positive connection between the dependent variable (PM10 concentration after the obstacle) and the set of independent factors, as indicated by the correlation coefficient (R) of 0.89. Moreover, the independent variables in the model collectively explained approximately 79% of the variation in the dependent variable, as reflected by the coefficient of determination (R^2) of 0.79. Overall, the research provides valuable insights into the impact of obstacle height, distance from the source, and wind speed on PM10 concentration and confirms the transport behaviour of PM particles in both small-scale experiments and larger-scale urban settings.

For the study of the effects of small hills on PM concentrations, the results revealed that at low wind speeds (0 and 0.7 m/s), the average concentrations of PM10 and PM2.5 were similar for all three cases. However, at higher wind speeds (2.4, 3.7, and 5.1 m/s), the average concentrations of PM10 and PM2.5 were significantly higher in the 1m height and 0.8 m height cases compared to the flat ground surface. Furthermore, the study showed that the difference in ground surface elevation between the 1m height and 0.8 m height cases had a notable impact on PM dispersion. The elevated ground surface (hill) altered the dispersion pathways of PM particles, resulting in higher concentrations in certain areas. Sensor 1 recorded higher PM concentrations in cases 2 and 3

compared to case 1, especially at wind speeds below 3 m/s, primarily due to the reflective effect of the hill and low wind speeds. Sensor 2 registered higher PM concentrations before the hill, indicating particle trapping in that area. The study also employed multiple linear regression to estimate PM10 concentration at the top of the hill based on measurements from sensor near the source, and sensor at the bottom of the hill, wind speed, and hill height. The regression analysis showed a strong positive correlation (R=0.9) between the dependent variable (PM10 concentration at the top of the hill) and the combination of independent variables (mentioned above). Approximately 82% of the variance in PM10 concentration at the top of the hill was explained by the independent variables (R^2=0.82). Also, the correlation coefficient between measured PM10 by all three sensors and wind speed demonstrates that hill height is important in shaping correlations between PM10 and wind speed, revealing intricate connections between topography and air pollution pattern.

For the dust storm simulation over the Sahara Desert (Moroccan and Mauritanian regions) using HYSPLIT, the average PM10 concentration between 0 and 100m reached severe levels according to the HYSPLIT dust simulation results. Regions like Dakhla-Oued Ed-Dahab in Morocco, Adrar and Tiris Zemmour in Mauritania had higher PM10 concentrations (higher than 100 µg/m^3) and AOD values (between 0.7 and 1) during the 4 days of the dust storm. Moreover, PM10 particles were transported over the Atlantic Ocean to the Caribbean Sea and the Gulf of Mexico, causing raise in the level of concentrations in those regions. The tropospheric level of the Caribbean Sea and the Gulf of Mexico was loaded by dust particles transported from the study area. Bir Anzarane Morocco, Nouakchott and Tichit Mauritania, and Bordj Badji Mokhtar Algeria all contributed to the high PM10 concentrations observed in the Martinique islands and the southern United States, while the top altitude of the dust layer was between 4 and 4.5 km, according to the backscatter vertical profile measured by CALIPSO. Therefore, PM10 concentration and AOD revealed their peak values during June 2020 dust storm, and this is evidenced by AOD values recorded at Bir Anzarane, Morocco, and Nouakchott, Mauritania, both of which are historical by June norms. There was a rise in AOD, with a value of 3.52 in Bir Anzarane, Morocco. This represents an increase of 188% above the highest recorded figure between 2010 and 2019, which was 1.87 in June 2017. Similarly, from 2010 to 2019, the AOD value in Nouakchott, Mauritania, was 2.78 in June 2010. However, there was a spike in June 2020, with an AOD value reaching as high as 5.87, reflecting a 211% increase.

For the identification and evaluation of the Saharan dust storm events in Budapest, Hungary between 2018 and 2022, type 2 SDEs predominated in that period with 5 occurrences, whereas types 1 and 3 appeared three times each. In addition, February, March, and April have seen the most SDEs (7 times), and SDEs occurring in those months are more likely to be severe events because that period had the highest dust mass recorded (SDE4 - April 23-27, 2019) and an increase in PM10 daily average concentration of at least a factor of 2.

For the case study of the Saharan dust effects on PM10 and PM2.5 concentrations in Budapest in March 2022, the two type 3 dust storms contributed to an increase in PM10 and PM2.5 concentration levels. The PM10 concentrations increased by 12 and 14 $\mu g/m^3$, during first and second Saharan dust events, while for PM2.5 the concentration rise by 10 and 5 $\mu g/m^3$, in first and second Saharan dust events respectively, highlighting the fact that first Saharan dust event had bigger impact on PM2.5 that the second Saharan dust events, in contrast for PM10, the impact of the two Saharan dust events were nearly similar.

For the evaluation of PM surface concentrations simulated by Version 5.12.4 of NASA's MERRA-2 Aerosol Reanalysis over Hungary in the period between 2019 and 2021, the estimation of the PM10 and PM2.5 concentrations done in two approaches. The first approach involved estimating PM10 and PM2.5 using equations 1 and 2, that calculate the PM10 and PM2.5 in function of BC, OC, DS, SO_4 and SS concentrations given by MERRA-2 Aerosol analysis dataset and compare it with real measurements of PM10 and PM2.5, in addition to estimations using machine learning algorithms such as MLR, OLS, RF, and XT, and the data used in machine learning algorithm is coupled with meteorological data (T, P, RH, Ws_{10}, Ws_{50} and PBLH) and AOD. The second approach used the machine learning techniques used in the first approach to estimate PM2.5 and this time based on AOD in conjunction with observations of NO_2, O_3, SO_2, and meteorological data (T, P, RH, Ws_{10}, Ws_{50} and PBLH). And, both first and second approaches were applicated in 3 cities in Hungary, Budapest, Kecskemét, and Kazincbarcika. In case of the first approaches, results showed that using XT model gave the best results for all the three locations of the study, for Budapest I got an R^2 of 0.78 and 0.8 for PM10 and PM2.5 estimations respectively, and for Kecskemét an R^2 of 0.75 and 0.76, in addition for an R^2 of 0.77 and 0.8 for PM10 and PM2.5 estimations respectively for Kazincbarcika, proving the effectiveness of the XT machine learning model in estimating the PM concentrations. Moreover, for the second approach, Estimating the PM2.5 using XT model also gave the

best results. The best R^2 achieved was for Kazincbarcika with value of 0.84, followed by Kecskemét with value of 0.75 and Budapest with value of 0.73.

The use of Satellite based data, coupled with meteorological data can give accurate estimations of PM concentrations, especially PM2.5, where it is highlighted in many research studies that PM2.5 have a complex relationship with AOD and can be used to predict it's concentrations. Furthermore, the use of machine learning or deep learning methods prove to be useful tool in PM estimations in the study that I've done in this thesis, and my study can be expended by using bigger data and to other locations that could cover all the Hungarian territory.

Finally for the calibration of CAMS PM2.5 data, results reveal significant improvements in various metrics. The correlation coefficients before (ranged from 0.07 to 0.20) and after (ranging from 0.78 to 0.88) of the calibration method demonstrate noteworthy enhancements, indicating a stronger alignment between the CAMS PM2.5 data and in situ measurements. Additionally, the coefficient of determination (R^2) (ranged from 0.61 to 0.77) exhibits substantial increases, highlighting the improved predictive power of the calibrated data. The calibration process also leads to reductions root mean squared error (RMSE), indicating decreased variability between predicted and observed PM2.5 values.

These calibration outcomes have implications not only for Hungary but also for other countries grappling with air quality issues. Accurate and reliable CAMS PM2.5 data serves as a vital resource for governments, environmental agencies, and health organizations worldwide. By leveraging calibration techniques like LightGBM, countries can enhance the quality of their air quality datasets, leading to more accurate assessments of pollution levels and better-informed decision-making.

7 Relevant publications related to the thesis

MTMT: https://m2.mtmt.hu/api/author/10072503

Refereed papers in foreign languages:

1. **Qor-el-aine A.**, Béres, A., Géczi, G. (2022): Case Study of the Saharan Dust Effects on PM10 and PM2.5 Concentrations in Budapest in March 2022, JOURNAL OF CENTRAL EUROPEAN GREEN INNOVATION 10: Suppl 1 pp. 67-78., 12 p. https://doi.org/10.33038/jcegi.3500
2. **Qor-el-aine A.**, Géczi, G., Béres, A. (2021): Dust Storm simulation over the Sahara Desert (Moroccan and Mauritanian regions) using HYSPLIT, Atmospheric Science Letters, e1076. https://doi.org/10.1002/asl.1076 ; Impact factor (2022): 2.992; Q1.
3. **Qor-el-aine A**., Béres A., Géczi G. (2021): The concentration level of PM10 in southern Poland (Katowice, Krakow, and Rzeszów) during the year 2018, Science, Technology and Innovation, 14(3-4), 27–34. https://doi.org/10.55225/sti.8
4. **Qor-el-aine A.**, Benécs J., Béres A., Géczi G. (2021): Small scale experiments of PM10 dispersion around obstacles, Hungarian Agricultural Engineering, Vol. 40, pp. 96-101., HU ISSN 2415-9751. https://doi.org/10.17676/HAE.2021.40.96
5. **Qor-el-aine A**., Béres A., Géczi G. (2021): The nitrogen dioxide (NO2 and PM10 pollution level in Debrecen, Miskolc, and Nyíregyháza Hungary in the previous 4 years**,** Hungarian Agricultural Research, Budapest, Hungary, Vol. 30, No. 2, pp. 04-10., HU ISSN 1216-4526.
6. **Qor-el-aine A**., Benécs J., Béres A., Géczi G. (2020): Evaluation of particulate matter low-cost sensors: laboratory case study**,** Mechanical Engineering Letters, Gödöllő, Hungary, Vol. 20, pp. 67-72., HU ISSN 2060-3789.

Refereed papers in Hungarian:

7. Géczi, G., **Qor-el-aine A.**, Béres, A. (2021): Debrecenben megfigyelt magas PM10 koncentráció elemzése, ACTA AGRONOMICA ÓVÁRIENSIS 62: Különszám I. pp. 170-183, 14 p.