Statistical analysis of systematic differences in the calculated pollutant concentrations of the models ECMWF/CAMS (regional reanalysis) and Polyphemus/ DLR

This thesis is submitted in fulfillment of the degree of

Master of Science in Geoinformatics

By

Sathish Kumar Vaithiyanadhan

(Matriculation number: 504941)

Supervised by

Prof. Dr. Edzer Pebesma

Spatio-temporal modelling lab

Institute for Geoinformatics

Münster, Germany

Dr.rer.nat. Frank Baier

Team Atmosphere

Deutsches Zentrum für Luft- und Raumfahrt e.V. (DLR)

German Remote Sensing Data Center,

Oberpfaffenhofen, Weßling

March 2022

Acknowledgement

I wish to acknowledge my primary supervisor from Institute for Geoinformatics, Münster Prof. Dr Edzer Pebesma for supervising my thesis and guiding me throughout this project. I would like to thank my external supervisors M.Sc. Lorenza Gilardi and Dr.rer.nat. Frank Baier for granting me an opportunity to write my master's thesis at "Deutsches Zentrum für Luft- und Raumfahrt e.V. (DLR)". Lorenza continuously encouraged and shared her ideas and was always willing to assist in any way she could throughout the research project. Her ideas and availability for me throughout the thesis were some of the main reasons for the successful completion of the thesis. Special thanks to Mr Ehsan Khorsandi, for his unwavering support throughout the project and also functioned as a technical data source for my thesis. I am gratefully in debt to his valuable contribution to the thesis. Sincerely, I would like to say a big thank you to the wonderful colleague members at the DLR for their energy, understanding, and help throughout my project. To conclude, I cannot forget to thank my family and friends for all the unconditional support in this time which was challenging in every sense.

Abstract

In the last two decades, air pollution was viewed as a very serious issue due to the development of infrastructure all over the world. Environmental stressors such as air temperature, radiation, humidity, wind, noise, pollens, and air pollutants (e.g., O3, NO2, PM10, PM2.5) can affect human health in a variety of ways. With the Copernicus Atmospheric Monitoring Service (CAMS) and the air quality in-situ measurements from the European Environmental Agency, a wealth of data of unprecedented quality and spatiotemporal resolution are available. These data are supplemented by available spatiotemporal high-resolution numerical models like chemical-transport models for the comprehensive description of the environmental conditions. Their advantages are constant coverage and high spatial and temporal resolution. However, it is very important to assess the model performances and comparability with in-situ or satellite observations. The main focus of this paper is to perform a comparison of the outputs of the Copernicus Atmosphere Monitoring Service (CAMS) – Europe Air Quality Reanalysis data and the chemical transport model POLYPHEMUS/DLR, with in-situ measurements (station data). The scope is to assess the discrepancies concerning the different chemical species and to provide statistical indicators like Mean Bias, FGE, RMSE, and Trend Analysis and correction weights describing the different characteristics of the models. Also, a Machine Learning approach was applied as an exploratory task, with the goal to predict concentrations at in-situ stations and to identify the influence of each parameters considered by the Polyphemus model. From the results, it was found that Polyphemus/ DLR model overestimates NO2, PM2.5, and PM10 and underestimates the O3, concentrations in urban and rural areas over the time window considered [June 2016 to Dec 2018]. CAMS outputs especially for PM10 and PM2.5 deviates from station observations though the outputs are corrected using EEA air quality station datasets. Overall, the parameters like surface temperature, boundary layer height and season were found to play a major role in both urban and rural regions. There are also significant changes in the influence of some parameters depending on location. This comparison study will help to understand the model performances (overestimation and underestimation) for each of the pollutants and help to select modelled data for health and air pollution-related research in the future.

Keywords: CAMS, Polyphemus/DLR, In-situ Measurements, Statistical Indicators, Central Europe.

Table of Contents

List of Figures

List of Tables

List of Equations

Abbreviations and Acronyms

CAMS	Copernicus Atmospheric Monitoring Service
EEA	European Environmental Agency
EU	European Union
DLR	Deutsches Zentrum für Luft- und Raumfahrt e.V.
FGE	Fractional Gross Error
RMSE	Root Mean Squared Error
NO_2	Nitrogen Dioxide
O_3	Ozone
$PM_{2.5}$	Particulate Matter whose diameter is smaller than 2.5μm
PM_{10}	Particulate Matter whose diameter is smaller than 10μm
NetCDF	Network Common Data Form
CDO	Climate Data Operators
WHO	World Health Organization
NCD	Noncommunicable Diseases
AQI	Air Quality Index
MACC	Monitoring Atmospheric Composition and Climate
AC	Atmospheric Composition
TNO	Nederlandse Organisatie voor Toegepast Natuurwetenschappelijk Onderzoek
MAE	Mean Absolute Error

1 Introduction

Clean air is one of the prerequisites for the human and ecosystem health. WHO estimated that every year millions of deaths and other health outcomes in cities and their suburbs can be attributable to air pollution (WHO air quality guidelines, 2021). To better understand/quantify the impacts of air pollution, it is necessary to work with advanced technologies to perform qualitative analysis and to monitor the spatial and temporal association between air pollution and the health impairments (D. Superczynski et. al. 2011). The resulting information can be of support for the governments and the public agencies that work to improve the environmental and human health. By comparison with 15 years ago, we now have evidences that air pollution affects human health, ecosystem, and climate by different sources like transportation, industries, etc. According to the latest WHO report (WHO global air quality guidelines 2021), pollutants such as Ozone, Nitrogen Dioxide, and Particulate Matter are the major pollutants impacting on human health all over the world (WHO global air quality guidelines 2021). A proportion of on-communicable diseases (NCD) like Stroke, Chronic Obstructive Pulmonary Disease (COPD), ischemic heart disease, neurological, and other respiratory diseases associated with increased mortality were, in the past decade, attributed to the presence of air pollution (Kampa et. al. 2008). Despite of the evidences collected so far and the numerous studies conducted on the topic, further research is needed to better understand and quantify the impact of air pollution of health. Remote Sensing Data, in-situ station data and numerical models can be used to estimate an Air Quality Index (AQI) (Hashim et. al. 2010). Each data source, however, presents specific limitations in the spatial-temporal resolution, in the accuracy and in the temporal availability. When comparing health data with air pollution data to investigate their causal relationship, it is often necessary to consider heterogeneous air pollution's datasets. Therefore, methods to integrate and validate the different data sources are necessary (Vautard et al., 2012).

Several nations and European Union authorities appraising the pollutant concentrations near the surface level by using efficient climate models. These models help the governments and the public to take immediate actions in case of severe air pollution (Menut and Bessagnet, 2010). These models use meteorological parameters as their input (Baklanov et al., 2014). This is because, the pollutants near the surface level are highly impacted by emission factors and by meteorological parameters like precipitation, surface temperature, boundary layer height, etc. (Delle Monache et al., 2006).

To assess the impact of atmospheric stressors on health and to be able to develop adaptation measures and recommendations, it is necessary to quantify the variability and distribution of pollutants at high spatial and temporal resolution. Chemical transport models can be used in the assessment of air pollution concentration. Their advantages

are the constant coverage and, for some datasets, a high spatial and temporal resolution. In the context of the “Umweltstressoren and Gesundheit” project conducted at the DFD institute of the German Aerospace Agency (DLR), two datasets are being used to estimate the effect of different air pollutants on human health. It is, therefore, important to assess the performances of the two models and to be able to quantify and predict systematic output differences in the concentration of the considered pollutants’ species.

The focus of this Master's thesis is the analyses of the systematic differences between the Chemical Transport models using the standard Statistical indicators with the reference to the European Environmental Agency (EEA) Air Quality Station data. The models considered are the CAMS (Copernicus Atmosphere Monitoring Service) - European regional Reanalysis and the model Polyphemus/DLR.

To analyse the systematic differences between the CAMS and Polyphemus with the reference to in-situ measurements, several statistical indicators like Arithmetic Mean, Standard Deviation, Mean Bias, Root Mean Square Error (RMSE), Fractional Gross Error (FGE), and Trend analysis were calculated. Also, the potential drivers of model accuracy with respect to observations were identified among Polyphemus input parameters exploiting a Machine Learning (ML) algorithm.

1.1 Scope of the study

The main objective of this research is to perform a comparison between the outputs of the Copernicus Atmosphere Monitoring Service (CAMS) – Europe Air Quality Reanalysis data and of the chemical transport model POLYPHEMUS/DLR. In-situ measurements data from EEA for the corresponding period are taken as ground truth. The outputs of the two models are compared to station data to analyse the discrepancies concerning the different chemical species and to provide statistical indicators describing the behaviours of the two models.

As a secondary objective, in the context of model comparison, an approach exploiting ML algorithm was inserted as an exploratory task. This had the scope to identify the importance of some of the input parameters of the Polyphemus model that were used as features to train an algorithm predicting the pollutants concentrations recorded by in-situ stations.

The thesis focusses on the following research questions,

1. How the overestimation/underestimation of the pollutants’ concentrations affects the performance of the models?
2. How extensive is the deviation between the two models with respect to station data?

3. Are there spatial patterns and temporal trends observable in the deviation? (Due to seasonal effects or different background conditions)
4. Can hypotheses be formulated to identify the potential drivers causing Polyphemus deviations with respect to CAMS and station data in urban and rural areas?

1.2 Structure of the report

This chapter introduces the thesis topic, the aim, and the research questions focused on the study. Chapter 2 provides literature references regarding Polyphemus /DLR and CAMS Regional Reanalysis /ECMWF and references to statistical indicators adopted in the study. Information on the area of interest, on the data formats, and on required data formatting for both the models are discussed in chapter 3. Chapter 4 provides descriptions of the statistical methodologies used in the study for the comparison between the models and with in-situ measurements. Chapter 5 presents a discussion about the Random Forest algorithm used to derive the feature importance of the Polyphemus inputs parameters in the prediction of in-situ stations data. Results and discussion are provided in Chapter 6. Chapter 7 delivers the conclusion of this study and together with suggestions for future works.

2 Literature Review

2.1 Model - model - station comparison

The accuracy of the model and accurate representation of pollutants in the surface level by Chemical Transport models are investigated by actual in-situ station observations. There are several established methods to perform a comparison between the model's outputs and in-situ observations. To verify the performance of Chemical Transport models (CTM), one of the efficient ways is using different statistical indicators and analysing the variabilities spatially and temporally (CTM) (Haofei Yu et. al, 2018).

Some of the standard statistical indicators used for model-data comparison with real-world standard observations are Fractional Gross Error (FGE), Root Mean Square Error (RMSE), Mean Bias, Correlation Coefficient, etc., (Wagner, et. al. 2020, Marécal et. al. 2015, CAMS Verification plots: documentation, 2020). Taking these statistical indicators together helps to verify the model performances from different perspectives. Comparing the model's outputs, especially in regions where there is a large number of stations operating continuously like France, Germany, Belgium, etc., with respect to station observations using multiple statistical operators, helps estimating the models' accuracy in different geographical and demographical contexts. Most of the European countries have numerous active stations. (Derwent, et. al. 2010). Statistical indicators like correlation coefficient, FGE, RMSE can be applied to model outputs with respect to in-situ observation. Each statistical indicator has an associated ranking criteria. FGE, for example, ranges between 0 and 2: the more the score is close to zero, the better is the model performance. The overestimation and underestimation of the pollutants' concentrations in the models with respect to station data are investigated deriving the mean bias for the time series of the models with respect to the observations (Tom Grylls, et. al, 2019)

To compare the model outputs (CAMS and Polyphemus) with station observations (EEA air quality station observations), some preliminary data preparations are required:

1. All the datasets considered (CAMS, Polyphemus and station observations) must be interpolated to the same reference location (Regridding).
2. The most common strategy to simplify the analysis is to perform a comparison between spatial and temporal aggregates of the two datasets considered.
3. To overcome the issues with the datasets like verifying station data operation for continuous data availability or with few missing timesteps, a subset of the station data must be performed in the time window considered (E. Solazzo, S. Galmarini, 2015).

Certain assessing criteria should be followed to verify the unstable operation of the stations to overcome the missing concentration values from certain stations for some days or for a longer time (E. Solazzo, S. Galmarini, 2015).

Some of the most efficient methods to compare models and station data are: derive systematic overestimations and underestimations of the model outputs with respect to the station observations considered, derive temporal trends (T. Grylls, et. al, 2019), analyse the deviations between the datasets under various circumstances like different seasons, urban and rural context, diurnal variation, etc., (E. Solazzo, S. Galmarini, 2015).

For air quality and climate-related studies, RMSE is a standard statistical indicator to analyse the performance, especially in the case of model evaluation. Domains like geoscience, atmospheric physics, geology, etc., make extensive use of RMSE for model validation (Savage et al., 2013; Chai et al., 2013). In RMSE, in most of the case of the numerical models, the analysis pursues normal distribution. Along with the verification of the performance, RMSE helps to assess the distribution of the errors. The outliers of the models are also described in RMSE as the outliers are well interpreted as long as they can be described by a normal distribution. For this reason, it is good practice to exclude errors and accounting for the bias prior to the calculation of the RMSE. This is especially true if the model results are extremely biased.

Mean bias is one of the easiest statistical analyses to understand the bias in the model with respect to the in-situ measurements. The mean bias analysis presents only positive and negative values to allow the quantification of possible overestimations and underestimations of the model with respect to station data/other models. Mean bias works well for the model with continuous operations. A certain bias is always expected in model outputs, but it should be closer to zero for better performance of the model (R.J. Stone, 1993).

Fractional gross error (FGE) is similar to RMSE which addresses the positive and negative deviations in the model with respect to observation data. The error metrics here are normalized with respect to the in-situ observations used. It is based on the normalised adaption of mean error in the calculation. FGE helps to calculate the overall error in the model. The main advantage of using FGE for model verification is that FGE does not magnify the outliers as it normalizes the mean errors based on observation? values. FGE is one of the most efficient statistical indicators for model evaluation and comparison as it calculates the fractional difference between the two datasets considered (A. Benedetti et al. 2019). The FGE is especially suited in the event of unevenness of the models. The FGE especially targets the asymmetrical variations of the models as it includes the temporal average (overall time period considered) of the

model and observations. This feature makes FGE slightly different from other time-honoured model evaluation methods with in-situ observations. Furthermore, FGE can be used to compare multiple models at the same time. (Yu et al.,2006).

The correlation coefficient is one of the most common statistical tools that help to analyse the closeness between the datasets considered. Pearson correlation coefficient is particularly suitable to analyse the normally distributed continuous data and a Spearman rank correlation for data with outliers and nonnormally distributed data. The covariance and its absolute magnitude between the model and the station datasets can be illustrated by the Pearson correlation coefficient method. It is calculated as a metric of a linear relationship of the models considered. There are certain assumptions that the datasets considered for correlation should match: the datasets should be randomly sampled and if both the in-situ measurements and the model's datasets (CAMS and Polyphemus) are generally normally distributed, continuous datasets then it would be better to apply linear correlation analysis. Linear regression is one of the most common types of continuous data analysis (Patrick Schober et. al. 2018). Correlation in the datasets can be positive, negative, or null and it is based on the trend direction and on the degree of closeness between the variables considered. This operation provides a basic understanding of the datasets.

Analysing spatial and temporal trends together in an environmental domain over a prolonged period can help us understand how a change in the environment might affect some parameters of interest, for example, understanding the longtime changes in the pollutants' concentrations in the model. Trend analyses can be used in almost every domain to check long-term changes. For example, analysing the career development in a start-up business, changes in land use/land cover, atmospheric pollutants, climate change, and much more (Gaüzère, P, et.al. 2015, Živadinovi´c, 2010). In the case of gridded data, it makes sense to analyse the pixel-wise linear trend to describe the trend over every pixel. The trend considering latitude and longitude gives a great vision about the variability based on locations for a long time. The trend analysis based on different time periods, different spatial variabilities like landscapes, and different land uses gives a vision for a better decision about the longtime changes in the regions i. Statistically, the existence of a trend is confirmed when the p-value of the slope coefficient is 0.05 or smaller (Brigitte and Kerrie, 2019). One of the most common trends tests that fit for both spatial and temporal analysis is Mann-Kendall's trend test.

2.2 Random Forest - Predictive analysis

The Chemical Transport models are influenced by various parameters as potential drivers. Each simulation of the models is based on those parameters considered. The

random forest (L. Breiman, 2001) and its corresponding Mean Decrease Gini (IncNodePurity) based on the Gini impurity index (S. Nembrini et. al., 2018) are the most efficient ML techniques to predict the model simulation and to identify the importance of each parameter in the model simulations. Performing statistical analysis on predicted concentrations with the actual concentrations helps to validate the simulations from Random Forest regression (J. Jake Nichol et. al, 2020).

With reference to all the above-mentioned literature sources, the knowledge gap can be found in applying various statistical indicators for the comparison of the ECMWF / CAMS (Regional Reanalysis) and Polyphemus/DLR models with the in-situ station concentration datasets from EEA. As an exploratory task, a Random Forest algorithm was trained with Polyphemus input parameters to predict in situ stations observation. Afterwards, a feature importance analysis was performed to identify potential drivers of deviation from observations in the Polyphemus model.

3 Study area and Data sources

3.1 Study area

The area considered in this Master Thesis is the one covered Polyphemus domain two (see chapter 3.2.1.1. for a detailed description) with a geographical extent of -1.0625W, 16.9375W, 43.96875N, 51.96875N (Central European domain). Some of the major cities in Europe like London, Paris, Munich, Milan, Frankfurt, Köln, etc., are covered. These cities are considered as some of the major polluted cities in Europe. This area of interest also covers different landscapes like the alps, major water bodies, and multiple different land-use compartments. The extent from Polyphemus domain two is also of interest in the context of the ongoing DLR project “Umweltstressoren and Gesundheit” project from DLR.

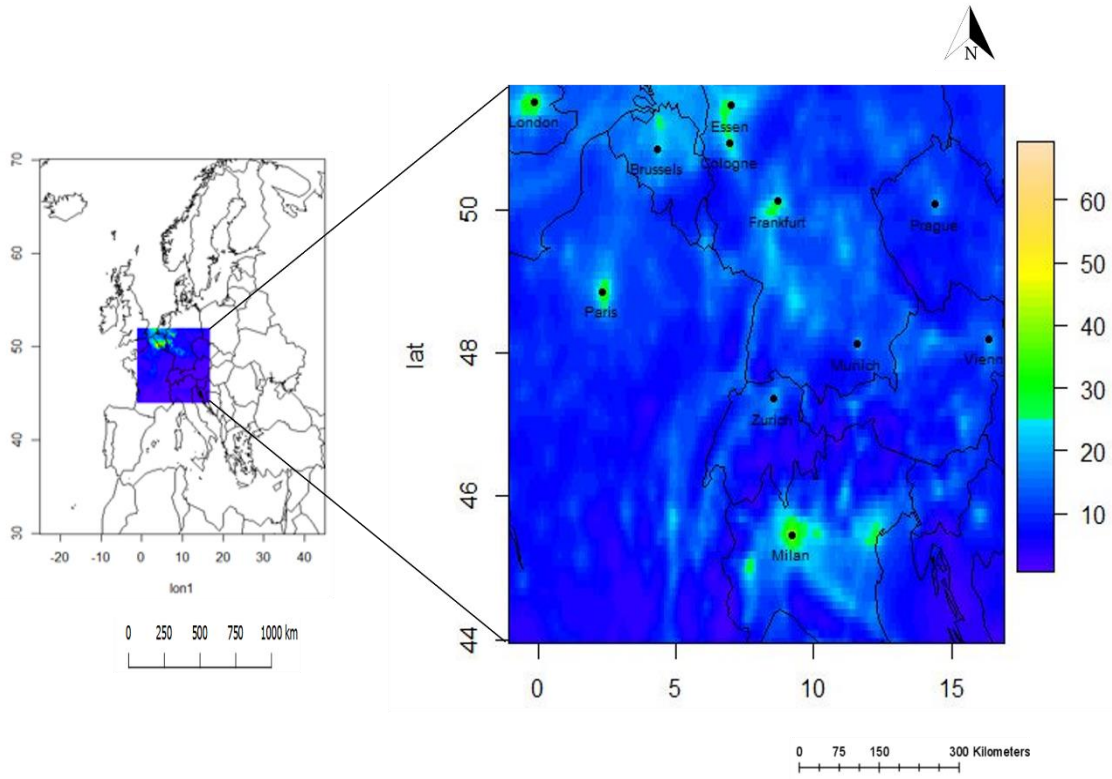


Figure 1: Study area Central European region (Extent of Polyphemus domain two)

The pollutants investigated in this work are NO_2, O_3, PM_{10}, and $PM_{2.5}$. The atmospheric concentrations at the surface level are considered. The time window of the project is from June 2016 to December 2018. This is the time window for which both Polyphemus and CAMS reanalysis were available at the time the study was conducted.

3.2 Data sources

3.2.1 Chemical transport models

Chemical transport models (CTM) are particularly suitable to analyse air pollution near at surface level, as they offer continuous spatial and temporal coverage. However model outputs might differ from reality due to the simplified process, parametrization, supposition of the models (Jacob and Winner, 2009). There are numerous studies that focus on comparing the models with station observations and satellite data, also considering the meteorological parameters that are related to air quality (Smyth et al., 2006).

3.2.1.1 Polyphemus/DLR

Polyphemus/DLR is a Polyphemus model platform developed by the German Aerospace Center (DLR). The model contains several Gaussian, Eulerian, and Lagrangian models as well as chemistry, transport, and aerosol modules. The model is available for three different areas of interest with increasing spatial resolution. Domain one (D1) with European extent, Domain two (D2) with central European extent, and Domain three (D3) with the extent of Southern Germany (Khorsandi et. al. 2018). Polyphemus data are produced for fourteen vertical (altitude) levels, starting from the surface. The project uses surface level concentrations from Polyphemus D2 with the geographical extent of -1.0625W, 16.9375W, 43.96875N, 51.96875N with a horizontal resolution of 0.125°N x 0.0625°W (Chan et al. 2021, Liu et al. 2021).

The model is composed of four autonomous levels: data management, physical parametrization, numerical solvers and the functions (high-level methods) (V. Mallet et al., 2007). Polyphemus is a free running CTM constrained by meteorological data, emissions and chemical/physical processes for tropospheric constituents. The model performs correction and interpolation of meteorological data, land use data, emission inventory data from Nederlandse Organisatie voor Toegepast Natuurwetenschappelijk Onderzoek (TNO), etc. (Khorsandi et. al. 2018, V. Mallet et al., 2007).

The four major numerical models in Polyphemus are a Gaussian plume model, a Gaussian puff model, Polair3D, and Castor. The performance of the model is influenced by some of the potential drivers like boundary layer height, wind, surface temperature, pressure, seasons, etc. (V. Mallet et al., 2007). The model has four data assimilation algorithms, optimal interpolation, ensemble Kalman filter, reduced-rank square root Kalman filter, and 4D-Var. Polyphemus performs Monte Carlo simulations to achieve ensemble forecasts (Evensen, 1994). Polyphemus / DLR intensifies the NO_2 concentrations at the surface level, especially in highly populated regions, industrialized zones, etc. (Song Liu, et. al. 2021, V. Mallet et al., 2007).

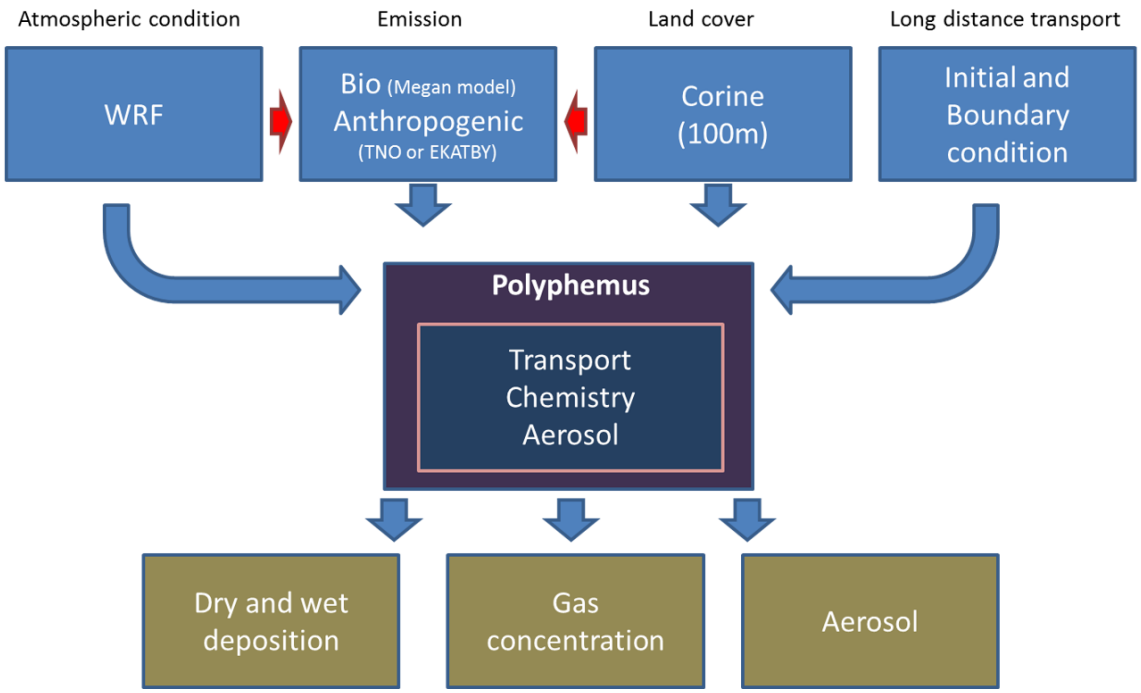


Figure 2: Polyphemus / DLR model structure (source: Khorsandi et. al. 2018)

3.2.1.2 CAMS reanalysis

CAMS reanalysis is the European regional reanalysis dataset of atmospheric composition (AC). The dataset has a horizontal resolution of 0.1° N x 0.1°W and a temporal resolution of one hour. It is an ensemble of nine different regional atmospheric air quality models from Europe and it is partly constrained by data assimilation with EEA station datasets (Marécal et. al. 2015). The models included in CAMS are CHIMERE, EMEP, EURAD-IM, LOTOS-EUROS, MATCH, MOCAGE, and SILAM. Since Oct. 2019 DEHM and GEM-AQ are also included. The ensemble outcomes are more accurate than the single model's ones. The method used in CAMS to produce an ensemble is the median value approach. The method uses the median of the considered models for every grid cell. All the models are re-gridded on a common horizontal resolution of 0.1° N x 0.1°W that approximately covers 10X10 km. The model predicts and monitors the background air pollution levels (CAMS documentation,2022, 20). The geographical extent of CAMS output is 25°W-45°E, 30°N-72°N (European extent). The datasets are available in seven vertical levels starting from surface level. The CAMS Reanalysis data combines the CAMS analysis and forecast data with the EEA station in-situ observations. CAMS reanalysis datasets have a considerably higher level of quality control. (Marécal et. al. 2015). The reanalysis datasets are available for free to download from the datastore link mentioned CAMS reanalysis datastore.

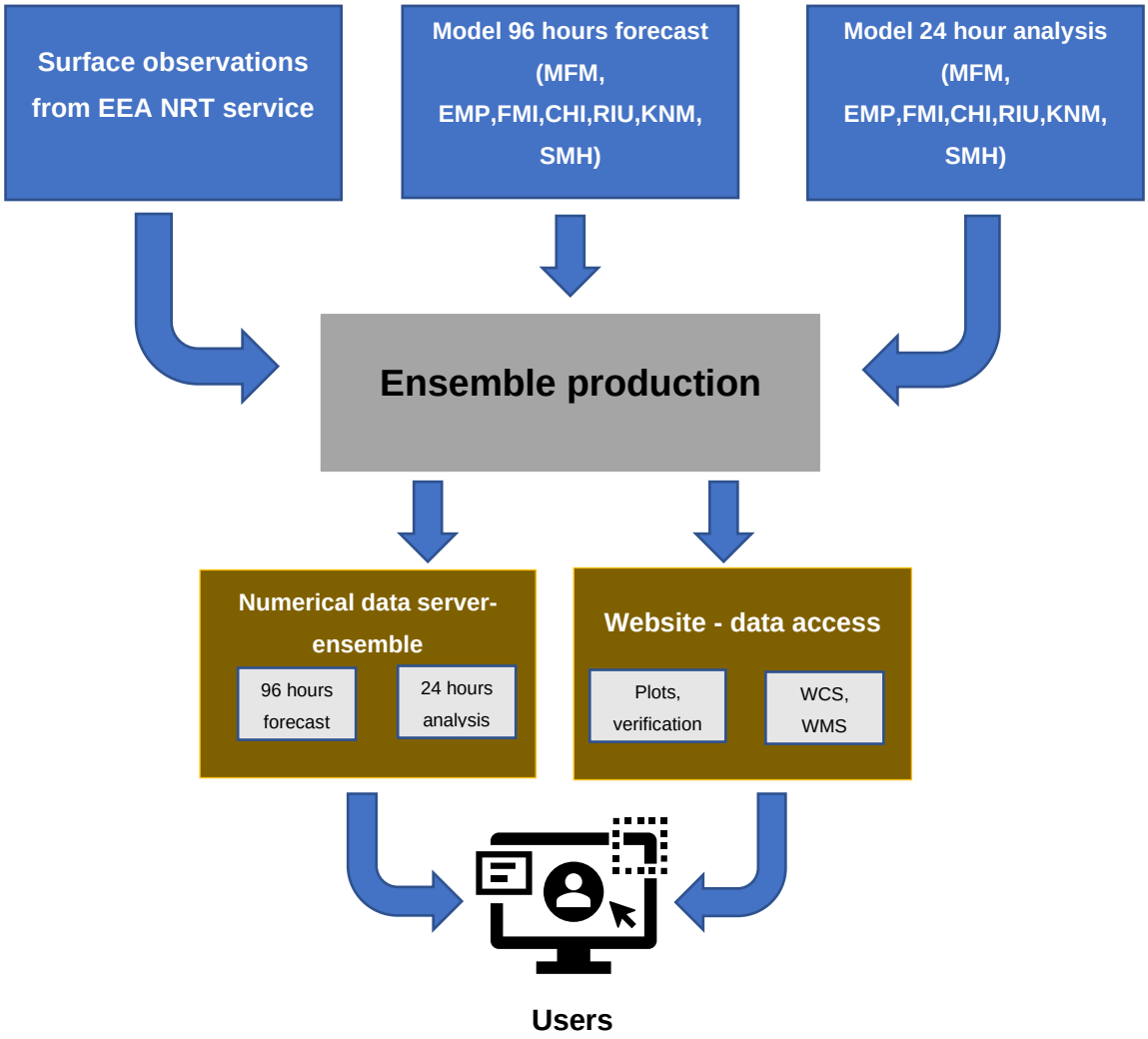


Figure 3: CAMS regional reanalysis model structure (Source: ECMWF CAMS Documentation)

3.2.2 In-situ Measurements (Station datasets)

The air quality station dataset from European Environmental Agency (EEA) is used as the truth dataset to perform a comparison with CAMS and Polyphemus datasets. The stations and concentration measurements of pollutants from EEA are defined by the EU and WHO regulations from stable station locations. Only the validated stations from the EU are taken as valid measurements for analysis. These regulations include 37 European countries except Albania, Kosovo, and Liechtenstein.

There are three different types of air quality measuring stations in EEA regulations based on the origin of emissions. They are,

- Traffic stations – The stations that are close to the city road junctions.
- Industrial stations – The stations that are close to the industrial areas.

- Background stations – The stations where the populations and vegetations play a major role in emissions. (Air quality in Europe report, 2020)

The station areas are also divided into urban, suburban, and rural areas based on their location and surrounding buildings (EEA Air Quality in Europe report, 2020). More than 75% of the data from the stations must be monitored continuously for the pollutants like NO_2, Particulate Matter (PM), O_3. So, the availability of actual station measurements for major pollutants is accessible all the time. The station data available in the area of interest were placed gird with a horizontal resolution of 0.1° N x 0.1°W. If more the one station was present in a grid cell, the mean of the stations' output was assigned to the cell. Only the stations from the category background stations were considered in this work (Air quality in Europe report, 2020, EEA air quality data download portal).

Field	Type	Description
Countrycode	String	Country iso code
Namespace	String	Unique namespace as provided by the country
AirQualityNetwork	String	Network identifier
AirQualityStation	String	Localid of the station
AirQualityStationEoICode	String	Unique station identifier as used in the past AirBase system
Samplingpoint	String	Localid of the samplingpoint
Samplingpoint	String	Localid of the samplingpoint
SamplingProcess	String	Localid of the samplingprocess
Sample	String	Localid of the sample (also known as the feature of interest)
AirPollutant	String	Short name of pollutant. Full list: http://dd.eionet.europa.eu/vocabulary/aq/pollutant/view
AirPollutantCode	String	Reference (URL) to the definition of the pollutant in data dictonary
AveragingTime	String	Defines the time for which the measure have been taken (hour, day, etc)
Concentration	Value	The measured value/concentration
UnitOfMeasurement	String	Defines the unit of the concentration
DateTimeBegin	Datetime	Defines the start time (yyyy-mm-dd hh:mm:ss Z) of the measurement (includes timezone)
DateTimeEnd	Datetime	Defines the end time (yyyy-mm-dd hh:mm:ss Z)of the measurement (includes timezone)
Validity	Integer	The validity flag for the measurement. See http://dd.eionet.europa.eu/vocabulary/aq/observationvalidity/view
Verification	Integer	The verification flag for the measurement. See http://dd.eionet.europa.eu/vocabulary/aq/observationverification/view

Figure 4: EEA station datasets description (Source: EEA air quality data download portal)

This project uses the surface level data (level 1) from both the models and the EEA station observations, the grid size from CAMS reanalysis 0.1° N x 0.1°W was used as a common grid. All the data format used here is NetCDF.

3.3 Data preparation

Comparison studies involve data formatting as a major task for meaningful analysis. Some of the major data formatting needed to be included timestep correction, merging of multiple NetCDF files, temporal aggregation, and re-gridding. These data formatting methods helps making the datasets from different source comparable. The necessary outcomes of the initial data consolidation are: same number of timesteps in all the

datasets, all the datasets being in the same grid size, and projection and creation of common temporal aggregations for the datasets considered.

3.3.1 Panoply

Panoply is a NetCDF visualization software by the NASA Goddard Institute for Space Studies and is freely available for download at the Panoply download link.

Visualizing the model and the station datasets with their properties can be easily done in Panoply. This visualization gives a better understanding of the differences between the datasets like geographical extents, grid sizes, projections, latitudes, longitudes, timesteps, levels, etc., (Panoply documentation, v 1.5.1).

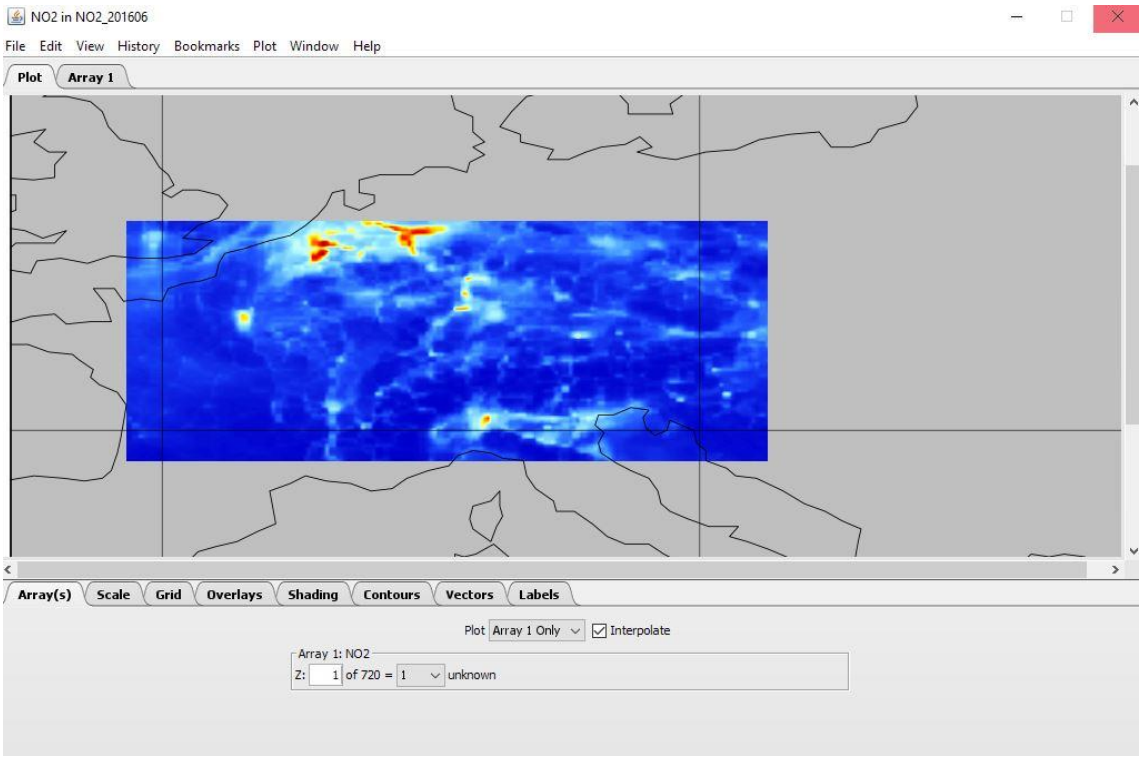


Figure 5: Panoply interface visualizing Polyphemus NO_2 in June 2016, 00:00:00

Panoply can also be used to gain a first impression about the data structure and content (See Table 1 and figure 5). Some of the options in Panoply include latitude-longitude maps (includes georeferenced zonal average plots, contour maps with the combination of dimensions), latitude-vertical grid plots (plots latitude vs vertical profile like levels), longitude-vertical grid plots (plots longitude vs levels), time-latitude/longitude (gridded plot shows a Hovmöller diagram). The main advantage of using Panoply is to understand and verify data as maps, plots, and arrays (Panoply documentation, v 3.1.1).

File "NO2_201606.nc"

File type: NetCDF-3/CDM

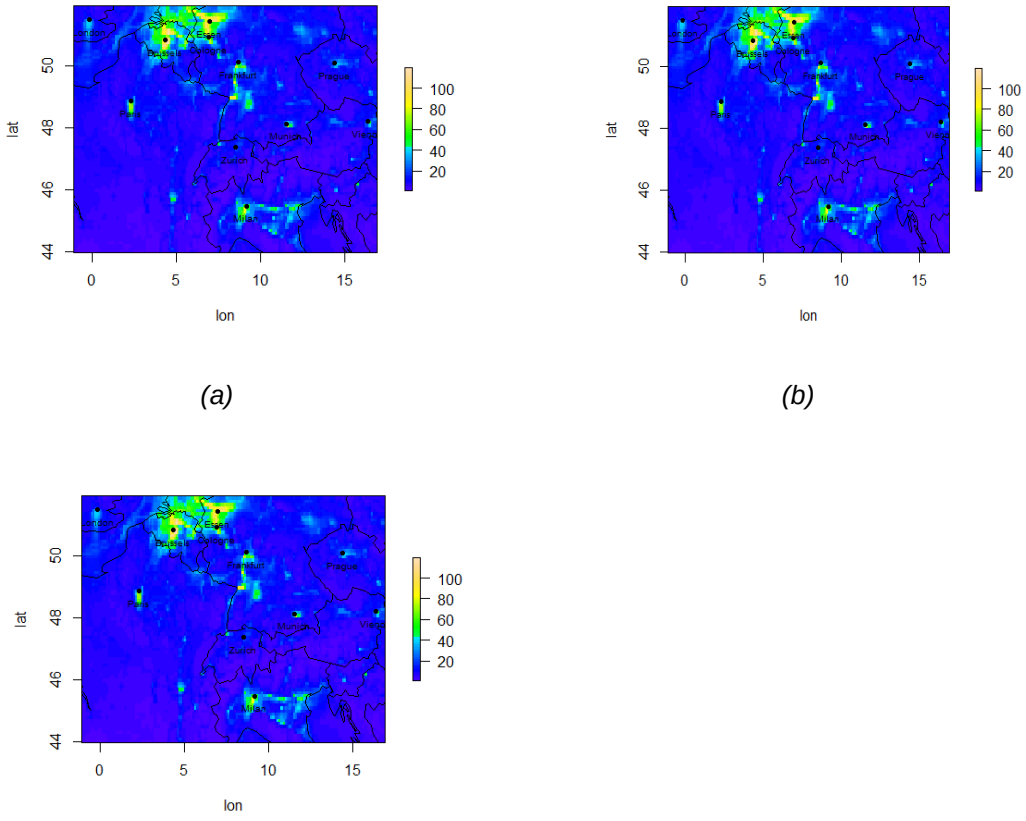
```
netcdf
file:/D:/Master%20Thesis/POLYPHEMUS/NO2%20Monthly%20data%20Regridded
/NO2_201606.nc {
  dimensions:
    longitude = 180;
    latitude = 80;
    z = UNLIMITED;   // (720 currently)
  variables:
    int crs;
      :proj4 = "+proj=longlat +datum=WGS84 +no_defs";

    double longitude(longitude=180);
      :units = "degrees_east";
      :long_name = "longitude";

    double latitude(latitude=80);
      :units = "degrees_north";
      :long_name = "latitude";

    int z(z=720);
      :units = "unknown";
      :long_name = "z";

    float NO2(z=720, latitude=80, longitude=180);
      :_FillValue = -3.4E38f; // float
      :grid_mapping = "crs";
      :proj4 = "+proj=longlat +datum=WGS84 +no_defs";
      :min = 0.3994140625, 0.18756103515625; // double
      :max = 119.8173828125, 122.01513671875; // double

  // global attributes:
  :Conventions = "CF-1.4";
  :created_by = "R, packages ncdf4 and raster (version 3.4-13)";
  :date = "2021-07-22 23:37:00";
}
```

Table 1: Example of visualization of attributes of the Polyphemus NO_2 data in Panoply.

3.3.2 Re-gridding

Re-gridding is a process of interpolating or re-projecting the dataset from its actual grid resolution to the intended grid resolution with the help of source-destination data. Re-gridding was performed to make sure all the datasets are in the same grids and to allow the comparison between them and with station data (Blower, J. and Clegg, A. 2011). This process was applied to the raster datasets to specify the required coordinate reference system. Re-gridding the earth sciences datasets are more complex to relate with real data because of their multidimensional attributes like time, horizontal and

vertical dimensions. Some of the major difficulties for re-gridding are irregular dimensions, multidimensional and high-dimensional datasets, large-sized data, discrete in the case of continuous spatiotemporal data, etc., (Meng Lu et. al, 2018). Some of the re-gridding techniques are interpolation, grid overlapping, area-weighted technique, and others. Out of which, using interpolation techniques is most common for re-gridding multidimensional datasets (Qingkun et. al. 2013, Blower, J. and Clegg, A. 2011). The comparison testing between two common interpolation methods, the Nearest Neighbour interpolation, and Bilinear Interpolation with respect to the real Polyphemus dataset (in figure 6) was done (Meng Lu et. al, 2018).

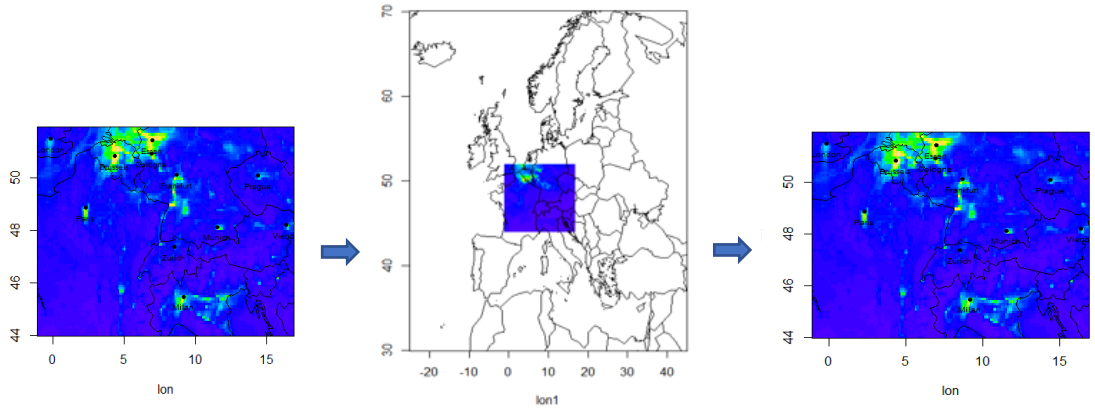


(a) *(b)*

(c)

Figure 6: Comparing the performance of two re-gridding methods applied to Polyphemus data. (a) Polyphemus data in its original grid. (b) Polyphemus data re-gridded using Bilinear interpolation. (c) Polyphemus data re-gridded using Nearest Neighbour interpolation.

From the result, the Nearest Neighbour Interpolation method performed more relevant to the original Polyphemus data in the grids of CAMS, and the same method was used for re-gridding. The actual resolution of the datasets CAMS Reanalysis and Polyphemus is 0.1° N x 0.1°E and 0.125°N x 0.0625°E, respectively.



(a) (b) (c)

Figure 7: Re-gridded Polyphemus data using Nearest Neighbor interpolation method. (a) Polyphemus data visualization for one timestep in its original grid. (b) Re-gridded Polyphemus data in the grid size and geographical extent of CAMS reanalysis, (c) Cropped re-gridded Polyphemus data to its original geographical extent.

3.3.3 Extracting station pixels in the model datasets

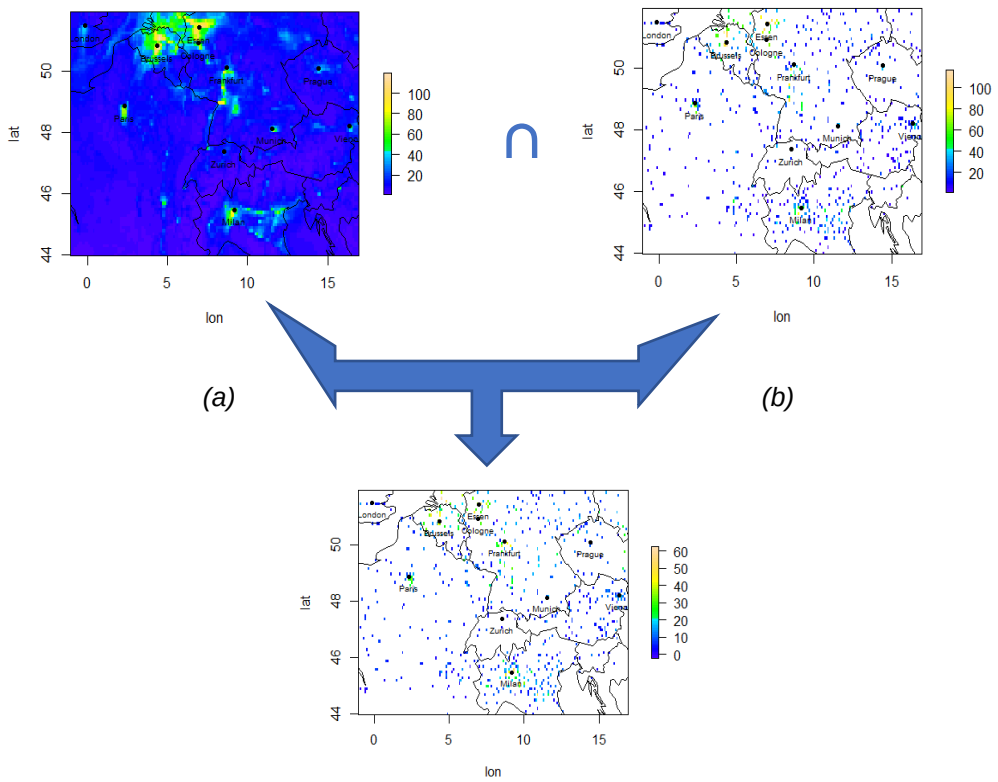


Figure 8: Intersecting Polyphemus and station data to extract the station pixels in the model datasets. (a) Original Polyphemus (model) data NO_2 of timestep 1, (b) Original station data NO_2 of timestep 1, (c) Extracted the same number of station pixels in the model data using intersection (A∩B).

In order to compare the model datasets with the station data, it is very important to extract the same pixels (station locations) of the station datasets in the model data. The challenge here is the pixels of stations varies in every timesteps and must extract the same count of stations in model datasets for every timesteps. For e.g., timestep 2 in NO_2 station data has 130 stations active and timestep 3 in the same data has only 98 stations active.

3.3.4 Timestep correction

Timestep correction in NetCDF involves several subtasks like concatenation of multiple NetCDF files, the removal of specific timesteps, the adjuting in the desired temporal aggregate etc. This is due to some missing temporal values in Polyphemus outputs, due to short temporary outages. All these NetCDF data processing were done in Climate Data Operators (CDO). CDO is a command-line suite to manipulate and analyse NetCDF like Merging NetCDF, removing timesteps, extracting monthly and daily datasets, etc., (See Appendix 7 for detailed processes involved in timestep correction).

At the end of data formatting, all the datasets (CAMS, Polyphemus, and station) are in the same grid size, an equal number of timesteps, latitudes, longitudes, levels, etc., To validate the data formatting, e.g., all the datasets for January are in the same size.

4 Methodology

4.1 Timeseries

In this work, the time series of the model at specific locations, and station data were aggregated at a monthly level. Subsequently, the difference between day and night concentration variations, urban, suburban and rural concentration variations were analyzed.

4.1.1 Trend analysis

In comparison studies, trend analysis is often used in order to observe how different the models and in-situ data develop over time. For example, increasing or decreasing pollutant concentration in the intended area over the period of interest. In statistics, there are many methods to calculate trends. Some of them are Linear Regression, Mann Kendall Test, Spearman's Rank Correlation Test, Theil-Sen Trend Lines. Mann Kendall Trend Test is the method chosen for this work as it is one of the most popular trend tests used. Mann Kendall Test is a distribution-free trend test (Zhang et. al., 2012).

4.1.1.1 Temporal trend analysis

Temporal trend analysis is a trend test that calculates the slope coefficient of the line fitting the data over a period of time considered. This analysis is applied on time series, where preliminary monthly, seasonal or yearly aggregations can be performed. The test used for this work calculates the Kendall's score, the significance value, the variance of Kendall's slope and Kendall's tau statistics. The combination of temporal trends and their spatial visualization is an efficient way to get a clear insight about where the trend increases or decreases in the spatial domain considered (Zhang et. al., 2012; Brigitte and Kerrie, 2019).

4.1.1.2 Spatial visualization of temporal trend

In this case, the trend test is applied pixelwise. As a result, a map plot is produced, that provides the slope coefficient resulting for each pixel. The scope is to have a visual representation of trend differences in several locations. In spatial trend analysis, the trend of each location (pixels) and how it varies over time can be clearly visualized. The deviations in trend varies based on the geographical locations like the alps, low-lying areas, cities, industrial areas, etc., which can be interpreted easily from the spatial trend maps. This type of map can be of support for the policymakers (Zhang et. al., 2012; Brigitte and Kerrie, 2019).

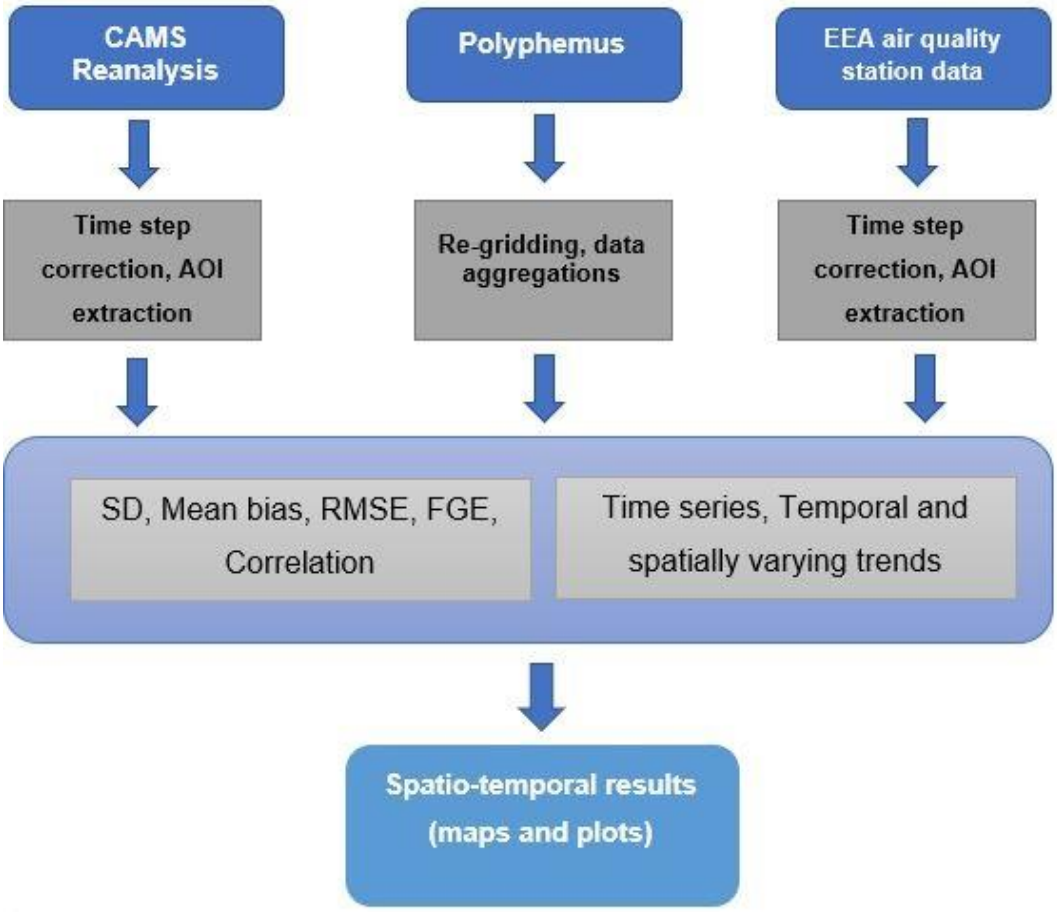


Figure 9: Workflow of Model-model-station comparison

4.2 Statistical Indicators

From the stated literature sources in chapter 2, here is the list of statistical indicators proposed in this work,

1. Arithmetic mean
2. Standard Deviation
3. RMSE
4. FGE
5. Mean Bias
6. Correlation coefficient

Using multiple statistical indicators for model comparison gives a clear view of the performance and also gives a clever idea about how each statistical indicator works in climate model datasets (Marécal et. al. 2015).

4.2.1 Arithmetic mean

The monthly mean of model timesteps for the time series was calculated using the standard arithmetic mean formula (equation:1). Timeseries analysis can be used in

various domains to track the changes over a prolonged time. Here, in our context, the scope of this analysis is to visualize the longtime performance (proposed time window) of the models and station datasets in measuring the pollutant concentrations to ensure the reliability of the model datasets. The analysis involves estimating the trend, deviation, and other parameters (R.L.R. Salcedo et. al., 1998).

$$\bar{x} = \frac{\sum x}{N} \quad \text{.....(1)}$$

Where, $\bar{x}$ = arithmetic mean,

x = concentration values,

N = total numbers. (Wan et al, 2014).

4.2.2 Sample Standard Deviation

To give information about the level of dispersion of data around the mean, we also need standard deviation in statistical analysis. Standard deviation is a measure that explains the actual distribution of the data from its mean and it was calculated using the formula in equation 2. Considering only the mean of the station locations as samples from the standard randomization method, we can calculate the distribution of the data (standard deviation) from the samples. The unit of standard deviation is the same as the unit of the arithmetic mean (Lee et al., 2015).

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{N-1}} \quad \text{.....(2)}$$

Where, s = sample standard deviation,

$\bar{x}$ = arithmetic mean,

x_i = concentration values,

N = total numbers. (Lee et al., 2015).

4.2.3 Mean Bias

In the case of comparing model and real-world observations (station) datasets, mean bias can be used to evaluate the deviation between them. From the results of mean bias, an insight can be obtained about how the model datasets are biased (positively or negatively biased) from the station observations (ground truth). Mean bias can be calculated using the formula in equation 3. The closer is the bias to zero, the better the accuracy of the model. Mean bias can be represented in the unit $\mu g.m^{-3}$ (R.J. Stone, 1993).

$$MB = \frac{1}{N} \sum_i (f_i - o_i) \text{.....(3)}$$

Where, f_i = model concentration value,

o_i = station concentration value. (V. Marécal et al, 2015).

4.2.4 Root Mean Square Error (RMSE)

The RMSE is one of the most used statistical methods to perform comparisons between two datasets. It is the measure of the standard deviation of differences between the model and the station observations. The closer is RMSE to zero, the better the accuracy of the model. RMSE can be represented in the unit $\mu g.m^{-3}$ (T. Chai and R. R. Draxler, 2014).

$$RMSE = \sqrt{\frac{1}{N} \sum_i (f_i - o_i)^2} \text{.....(4)}$$

Where, f_i = model concentration value,

o_i = station concentration value. (V. Marécal et al, 2015)

4.2.5 Fractional Gross Error (FGE)

FGE is defined as a normalized version of mean errors. It measures the overall errors in the model concentrations with respect to the station data. FGE uses absolute values instead of squared values as in RMSE. The calculations of FGE results in a value between 0 to 2. The closer is FGE to zero, the better the accuracy of the model. FGE can be represented in the unit $\mu g.m^{-3}$ (CAMS Verification plots: documentation, 2020).

$$FGE = \frac{2}{N} \sum_i \left| \frac{f_i - o_i}{f_i + o_i} \right| \text{ ....(5)}$$

Where, f_i = model concentration value,

o_i = station concentration value. (V. Marécal et al, 2015)

4.2.6 Correlation Coefficient

Correlation is a measure of the linear relationship between the datasets considered. In our case, it's a range of similarities found between the model and station observations. The standard range of correlation is between -1 and 1. The closer is the correlation to 1, the better the accuracy of the model (CAMS Verification plots: documentation, 2020).

The method proposed in this work makes use of Pearson's correlation coefficient and the formula to calculate Pearson's correlation coefficient can be found in equation 6.

$$R = \frac{\frac{1}{N}\sum_i (f_i - \bar{f})(o_i - \bar{o})}{\sigma_f \sigma_o} \quad \text{.............................(6)}$$

Where, R = correlation coefficient,

f_i = mean of model concentration values,

o_i = mean of station concentration value,

f_σ = standard deviation of model concentration values,

o_σ = standard deviation of station concentration value. (V. Marécal et al, 2015)

5 Predictive analysis - Random Forest

An exploratory task in the thesis consists in understanding the influence of some input parameters of the Polyphemus model on the resulting pollutants concentration and their deviation from station observations. This is achieved by training a Random Forest (RF) algorithm feeding it with Polyphemus inputs as training features and pollutants concentrations records from stations as the target. Here, the ML techniques like Random Forest (RF) regression and Gini importance are used. For prediction analysis, RF is one of the capable ensembles learning methods suitable for both ML classification and regression. The main advantage of using RF is that it uses an ensemble of decision trees, that can also work on nonlinear relationships between features. Furthermore, other advantages of the simple decision trees like easy utilization, high accuracy, no scaling, working on data overfitting, etc., also apply (J. Jake Nichol et. al, 2020). In this study, the dataset was split into 75% for training and 25% for testing. The hourly data of both features and target from the respective locations were used. After tuning the parameters of the algorithm to define the maximum number of nodes and the number of trees trying multiple combinations, the model performance was at its best with the combinations of 3000 as a maximum number of nodes and 3000 as a maximum number of trees. The accuracy of the predicted concentrations was assessed using the standard statistical indicators like Correlation, RMSE, FGE, Mean bias, and Mean Absolute Error (MAE) as proposed in chapter 2. The RF regression predicts concentrations based on equation 7.

$$RF(N) = \frac{1}{N}\sum_{n=1}^{N} T_n(x) \quad \text{.....(7)}$$

Where, x = training sample,

For N trees, $T_1, \ldots, T_N$ (J. Jake Nichol et. al, 2020).

This process was applied to urban and rural areas of interest. In both cases, the importance of each Polyphemus input parameter in predicting station data was derived. From the results of the previous statistical analysis, out of all the regions considered, Paris and its metropolitan region and some suburban and rural regions of Southern France were those that showed more deviations from the in-situ measurements. For this reason, these two regions were taken as areas of interest (see figure 11).

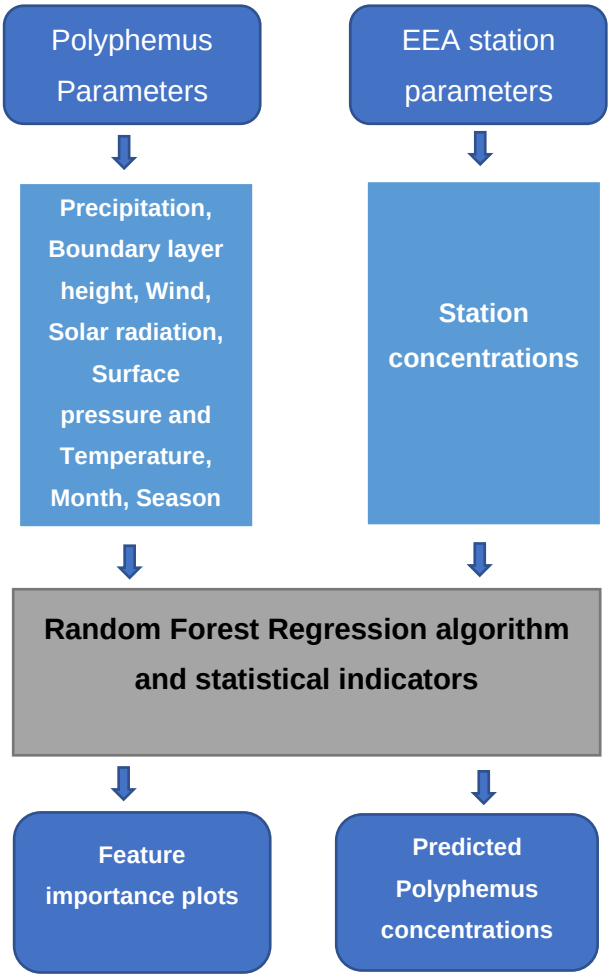
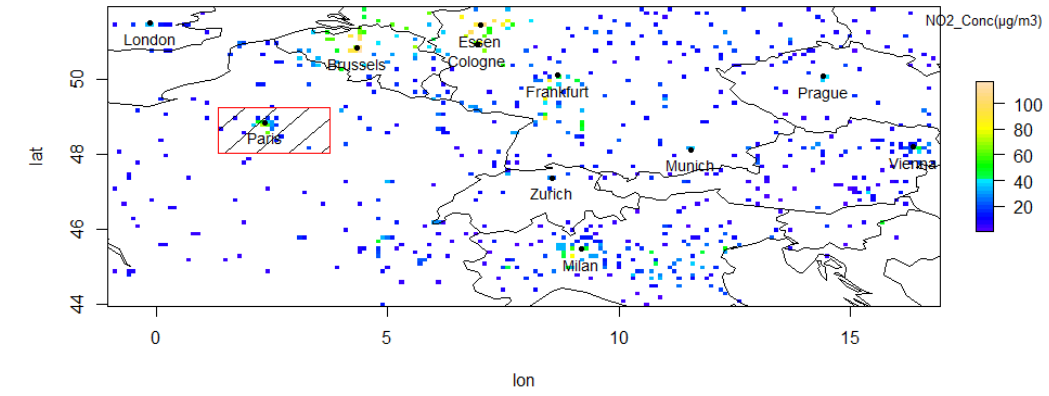


Figure 10: Random Forest Regression algorithm workflow.

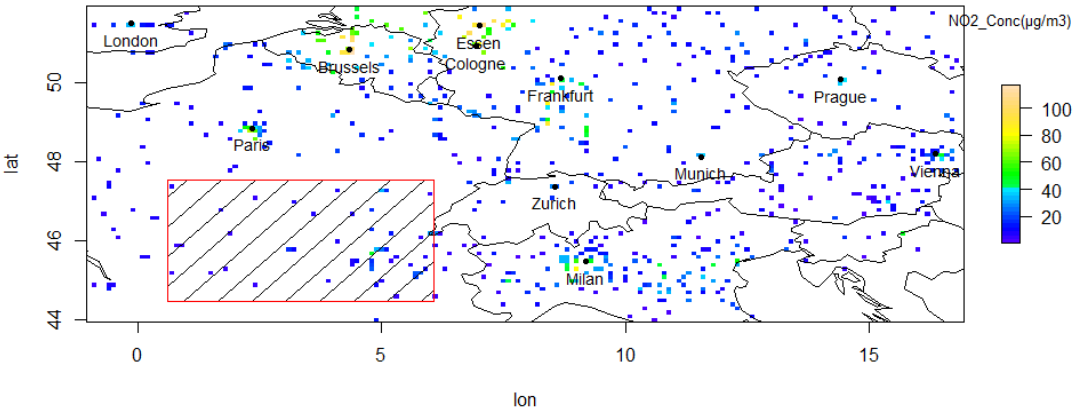
5.1 Potential drivers

While CTM simulations are based on well-known mathematical functions and parameters, there is an ongoing discussion on what is the influence of each parameter in the model outputs. The insight about finding the potential drivers that influence the performance of the model helps the atmospheric scientists to improve the model simulations to make them more comparable with the real-world measurements and to consider other potential parameters for the better performance of the model. It also helps to identify the parameters that worsen model results (J. Jake Nichol et. al, 2020). The ultimate tool to identify the potential drivers is to improve the prediction accuracy of the model. In this work, to analyze the feature importance, the Gini importance method was used (IncNodePurity). It is a Mean Decrease Gini based on the Gini impurity index. Gini importance was preferred in this work because of the presence of non-linearities in the data used. Gini impurity index will be measured by the residual sum of squares of overall

trees. This value for each parameter is the predictive strength of that parameter in the simulation (J. Jake Nichol et. al, 2020).



(a)



(b)

Figure 11: (a) The region of Paris metropolitan and its suburban considered. (b) The suburban and rural regions of Southern France considered.

Some of the parameters from Polyphemus considered for the analysis and their sources are in table 2. The parameter, Polyphemus concentration was not considered to train the model. It was used only to compare the predicted concentration values along with actual station values.

Parameters	Datasource	Input/Output
Date	Date series from June 2016 to Dec 2018.	Input
Precipitation	Polyphemus parameter	Input
Boundary layer height	Polyphemus parameter	Input
Surface wind velocity	Polyphemus parameter	Input
Solar radiations	Polyphemus parameter	Input
Polyphemus concentrations	Polyphemus parameter	Input
Surface pressure	Polyphemus parameter	Input
Surface temperature	Polyphemus parameter	Input
Month	Month series from June 2016 to Dec 2018.	Input
Season	Seasonal series from June 2016 to Dec 2018.	Input
Station concentrations	EEA air quality datasets	Output

Table 2: List of input and output parameters used in the RF Regression model.

6 Results and Discussions

This chapter presents the main results of the analysis of spatial and temporal variability in models and station datasets using different statistical indicators for all the pollutants considered. The main goal is a better understanding of the potential drivers of Polyphemus model results for urban and rural areas.

6.1 Model-model-station comparison

6.1.1 NO_2

The spatial and temporal time-series of NO_2 based on model and station data for the time window June 2016 to Dec 2018 give a clear view of how concentrations from different data sources can vary for the same location and timestep. Here for NO_2, The models and the station datasets for NO_2 are more comparable in the time series analysis. Though mean values deviate, model outputs and station data follow the same pattern. A seasonal variation can be observed for NO_2 where the concentrations decrease in summer and increase in winter (fig. 12).

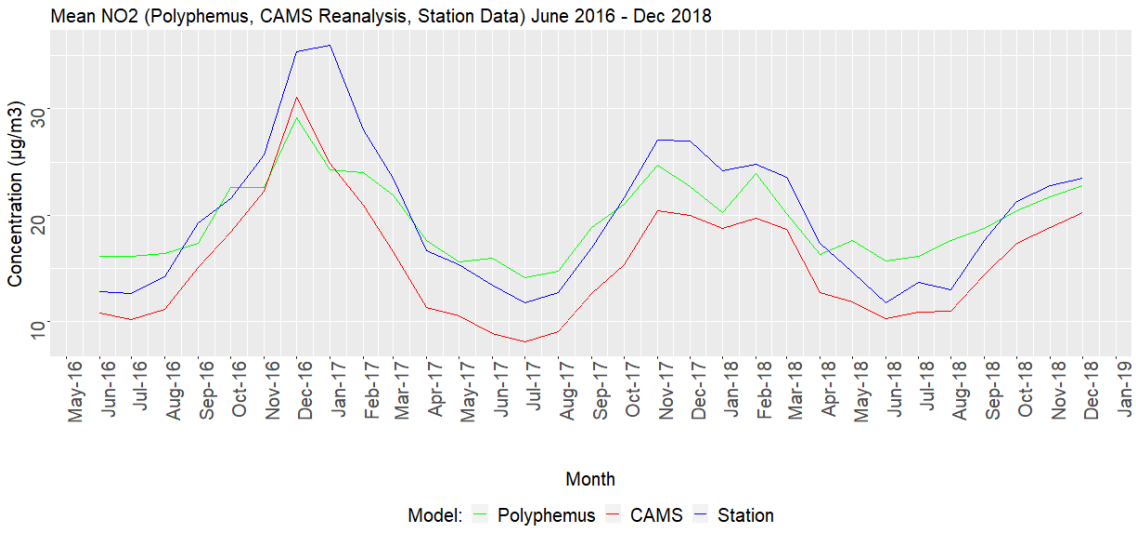


Figure 12:Time-series of NO_2 Model-model-station comparison from June 2016 to Dec 2018.

In order to have a spatial visualization of output differences between models and station data, a plot is produced where values from the three data sources corresponding to station locations are represented in a map. The main advantage of this type of visualization is that it is possible to identify the deviations in each pixel of the datasets. From fig. 13, NO_2 mean for the month of December 2016 (winter month) in model-model-station comparison, it is clear that the concentrations in metropolitan cities like Paris, Milan, etc., show maximum values and the model outputs slightly differs from the station data. But in the suburban regions, Polyphemus models show a slight underestimation in concentrations.

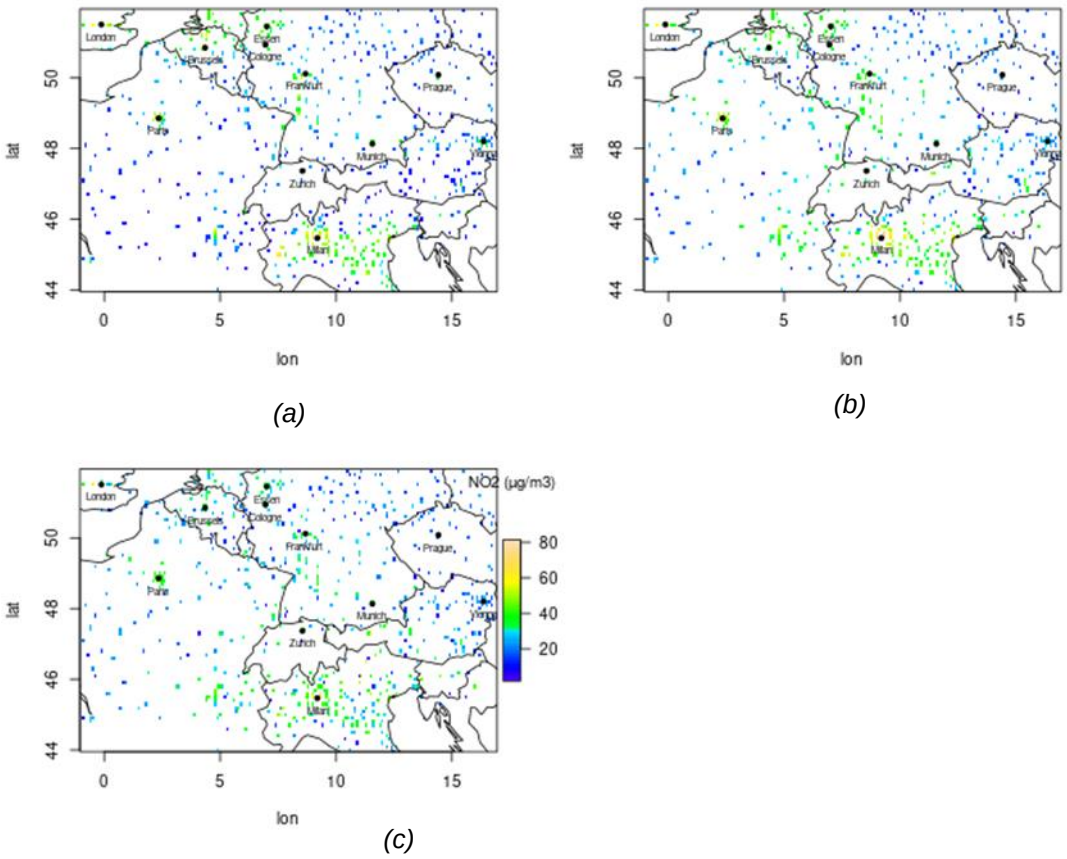
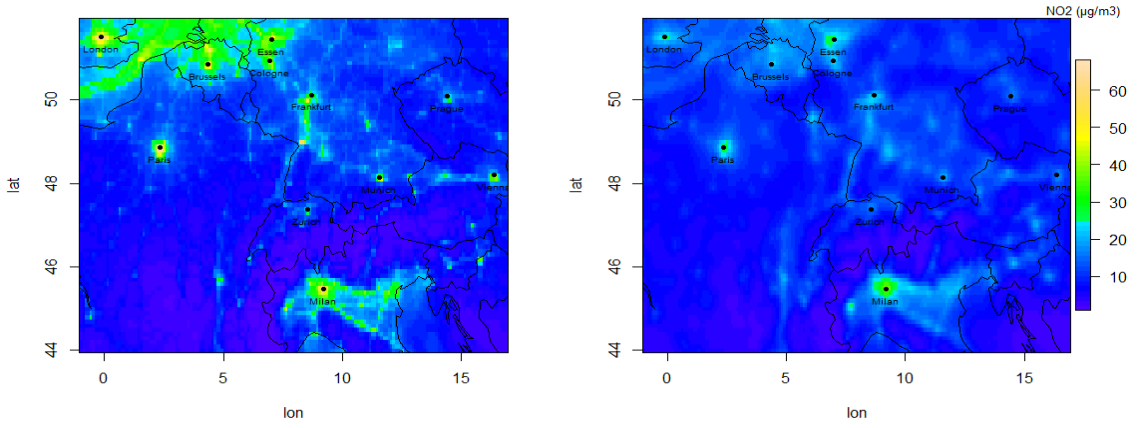
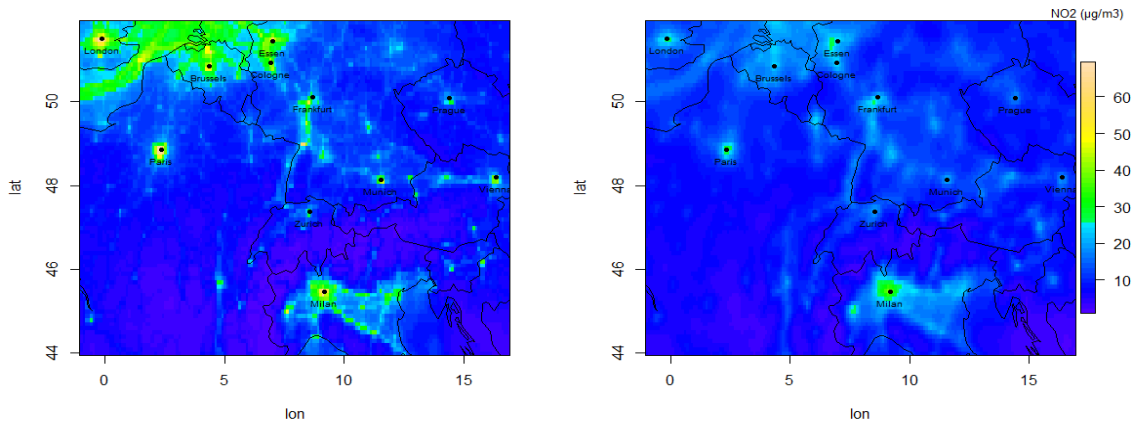


Figure 13: NO_2 monthly mean of models and station data for the month of December 2016 in the pixels of station locations (a): Polyphemus, (b): CAMS, (c): station datasets.

The yearly NO_2 mean (figure 14) between the model datasets (both Polyphemus and CAMS) gives an insight into the yearly deviations of model results. For model-model comparisons, from the mean of the years 2017 and 2018, main deviations dominate in the northwest regions, high concentrations in Paris, and also the regions in northern Italy. In general, Polyphemus shows significantly higher NO_2 values.





(b)

Figure 14: (a) Yearly NO_2 mean for 2017 in the extent of Polyphemus domain two (left: Polyphemus mean, right: CAMS mean). (b) Yearly NO_2 mean for 2018 in the extent of Polyphemus domain two (left: Polyphemus mean, right: CAMS mean).

The standard deviation for model-model-station comparison (fig. 15) describes the distribution of the modelled NO_2 values around the mean with respect to the station observations. The distribution of CAMS follows the same pattern as station data but the Polyphemus differs from the pattern of the observations.

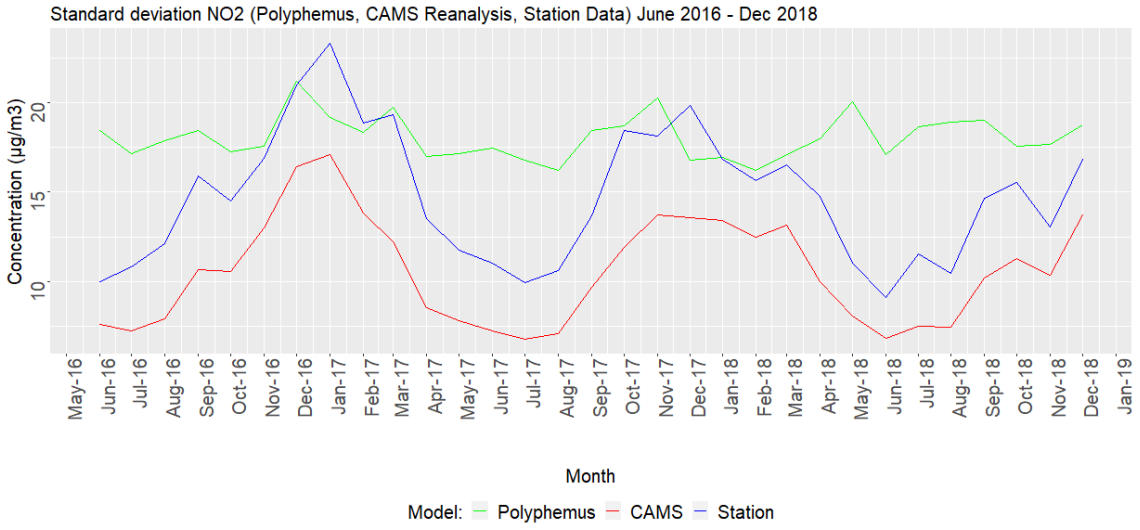
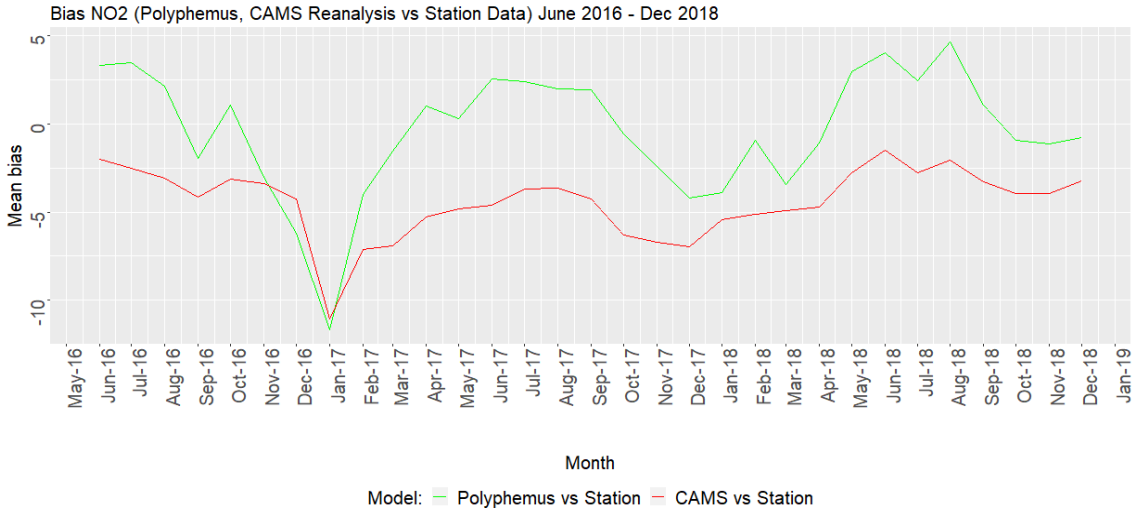


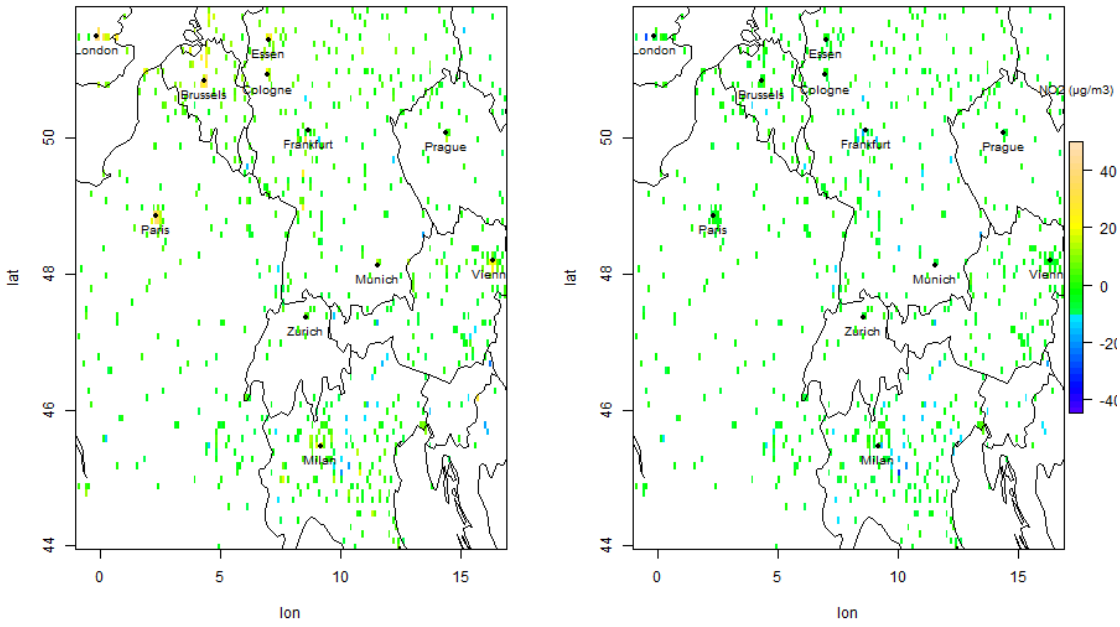
Figure 15: NO_2 Standard deviation between model-model-station from Jun 2016 to Dec 2018.

Mean bias calculated for NO_2 shows that there are seasonal changes in the performance of both models when compared with station data. The mean bias calculated for Polyphemus results is generally closer to zero than the one calculated for CAMS. CAMS datasets are negatively biased during the entire time window considered (fig. 16a). The spatial mean of the mean bias results also gives strong evidence that there are strong spatial seasonal changes. Models' outputs results are positively biased. In winter a slight

negative bias is observable for Polyphemus while the CAMS model results are highly positively biased in summer (fig.16b). Spatial and temporal mean bias for NO_2 show varying results with station observations, but Polyphemus outperforms during summer in some station locations.



(a)

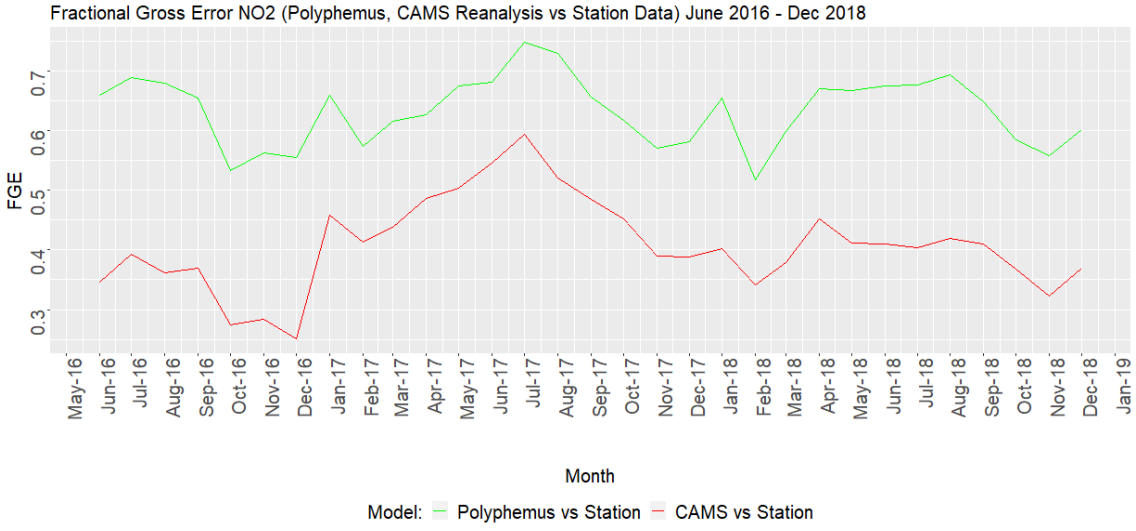


(b)

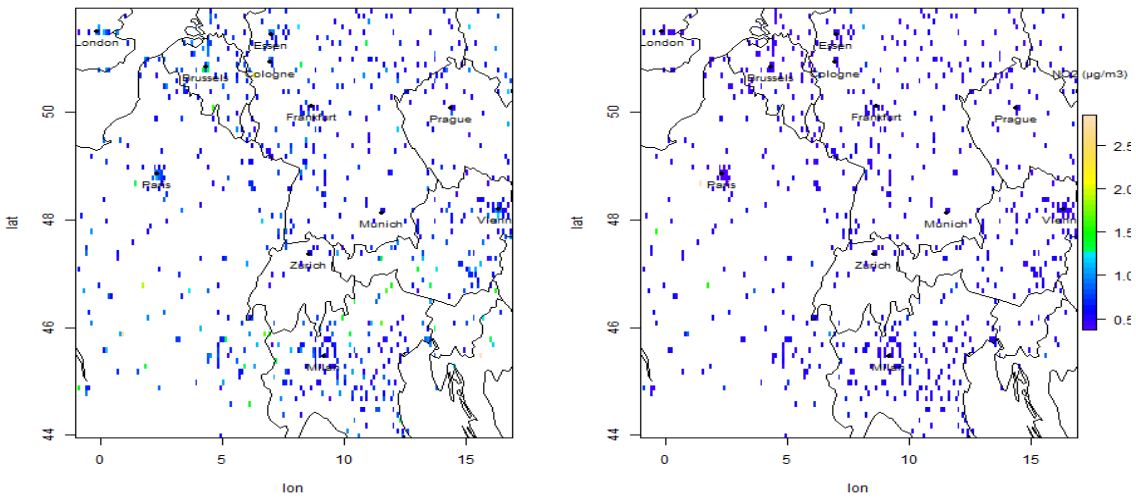
Figure 16: (a) Temporal NO_2 mean bias model-model-station from June 2016 to Dec 2018. (b) Spatial NO_2 mean bias model-model-station for June 2017 (left: Polyphemus mean bias, right: CAMS mean bias).

The results from FGE for NO_2 are presented in fig. 17(a) and provide diverging results in comparison with the previously calculated mean bias. FGE calculated for CAMS outperforms Polyphemus as the FGE values of CAMS are closer to zero than those of

Polyphemus. As both the FGE values calculated for the two models remain lower than one in the time window considered, for NO_2 we can consider that both the model predictions are comparable to station observations and both models show the same FGE pattern. This result is compatible with the fact that observational data are included in the model ensemble used to produce CAMS reanalysis. There is a significant difference between the performance of the models in summer, especially in the urban regions. For winter, deviations between the models in the Alps regions over Northern Italy and Southern Switzerland are easily noticeable.



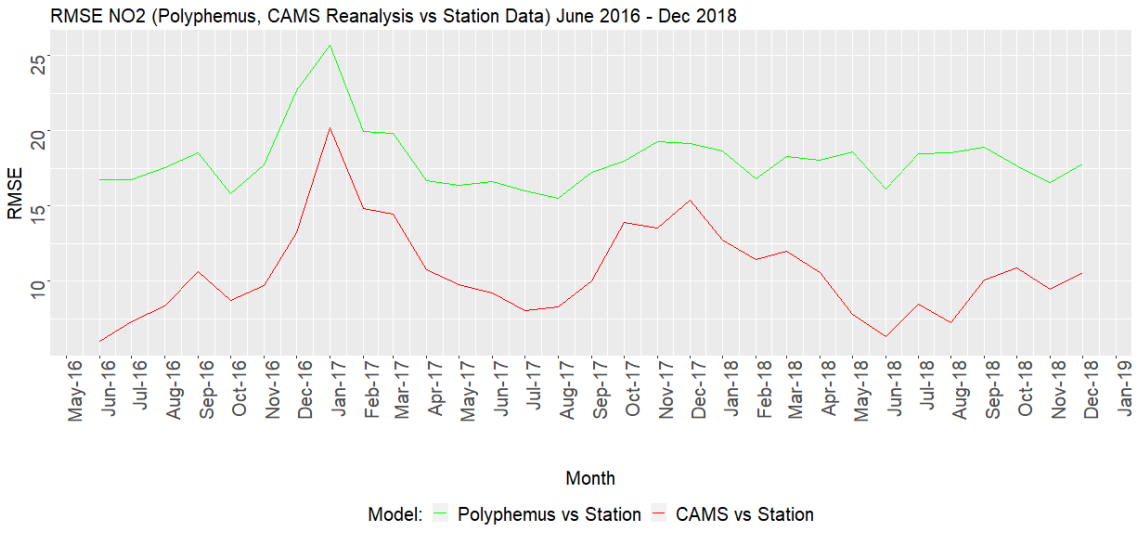
(a)



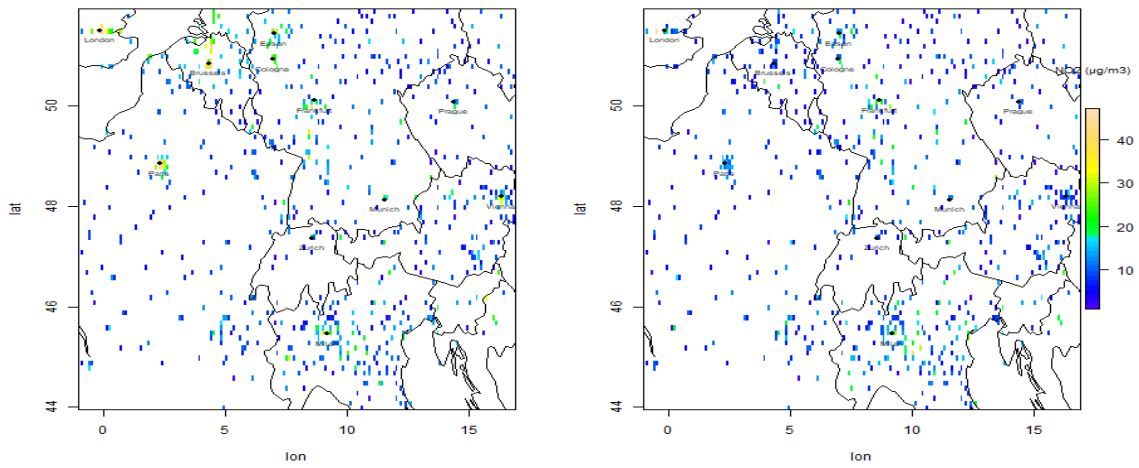
(b)

Figure 17: (a) Temporal NO_2 FGE model-model-station from June 2016 to Dec 2018. (b) Spatial NO_2 FGE model-model-station for June 2017 (left: Polyphemus mean FGE, right: CAMS mean FGE).

For NO_2, the results of RMSE adds strength to the results of FGE. Though none of the model outputs gets close to zero, the RMSE calculated for CAMS shows that the CAMS model has better performances with respect to Polyphemus. RMSE analyses CAMS model performs better in summer than winter (fig. 18a). Pixel-wise, other than the Alps regions and some metropolitan cities, all the regions are comparable with the station observations in winter. But there are deviations between the models in the urban regions observed in both summer and winter (fig. 18b).



(a)

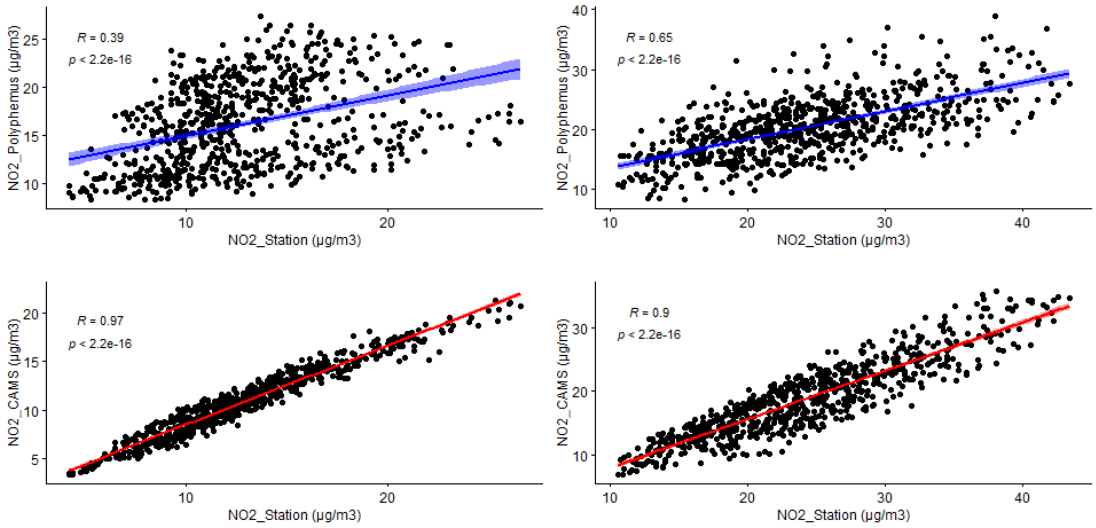


(b)

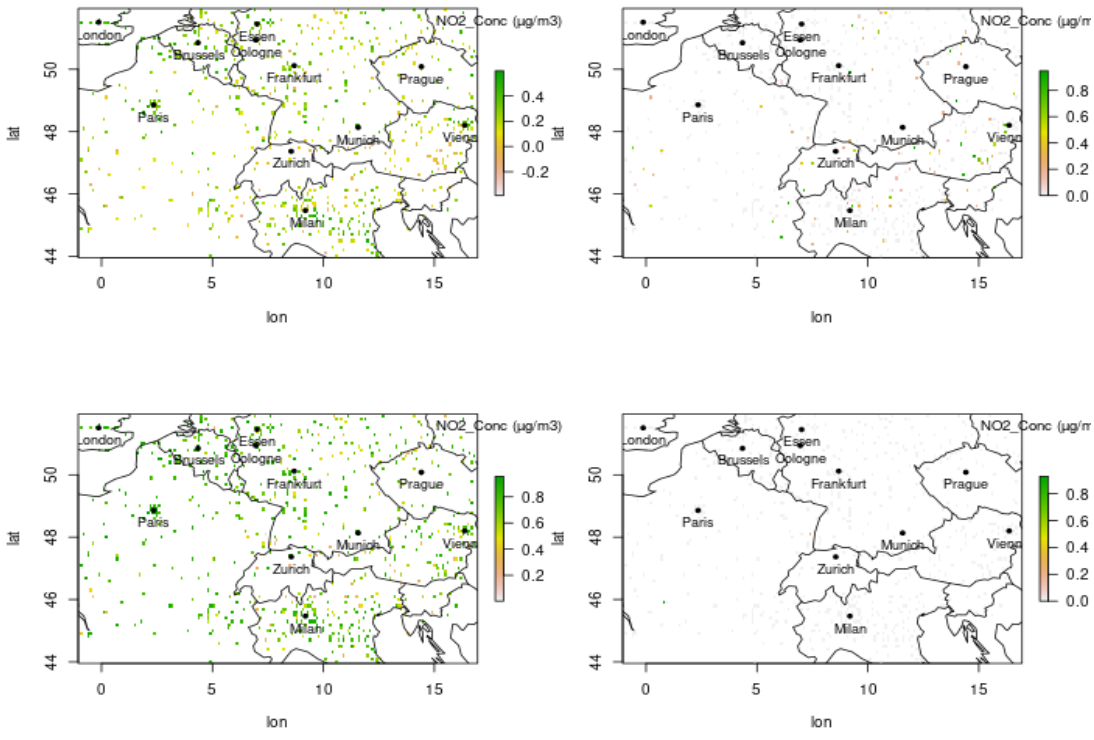
Figure 18: (a) Temporal NO_2 RMSE model-model-station from June 2016 to Dec 2018. (b) Spatial NO_2 RMSE model-model-station for June 2017 (left: Polyphemus mean RMSE, right: CAMS mean RMSE).

The results of correlation analysis for NO_2 give insight that CAMS outputs are strongly correlated with the station data with an average correlation coefficient of 0.9 for both summer and winter seasons. The correlation analysis carried out for Polyphemus and

station observations shows that the correlation coefficient varies between summer and winter. In winter the average correlation coefficient is 0.7 while in summer is 0.4. Correlation in Polyphemus follows a seasonal pattern with considerable correlation with station data in winter but the correlation in CAMS remains constant throughout the year (fig. 19a).



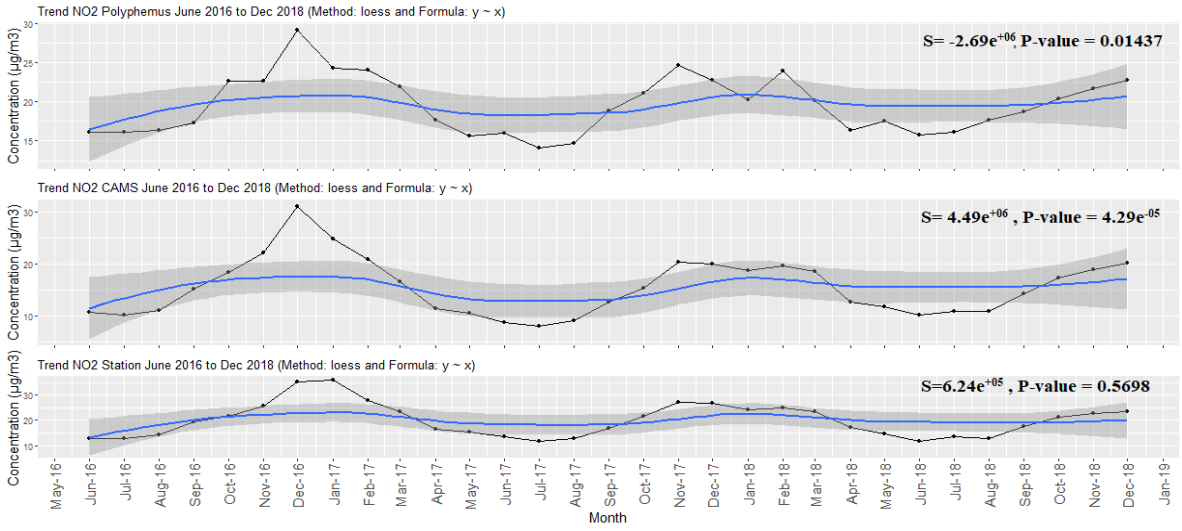
(a)



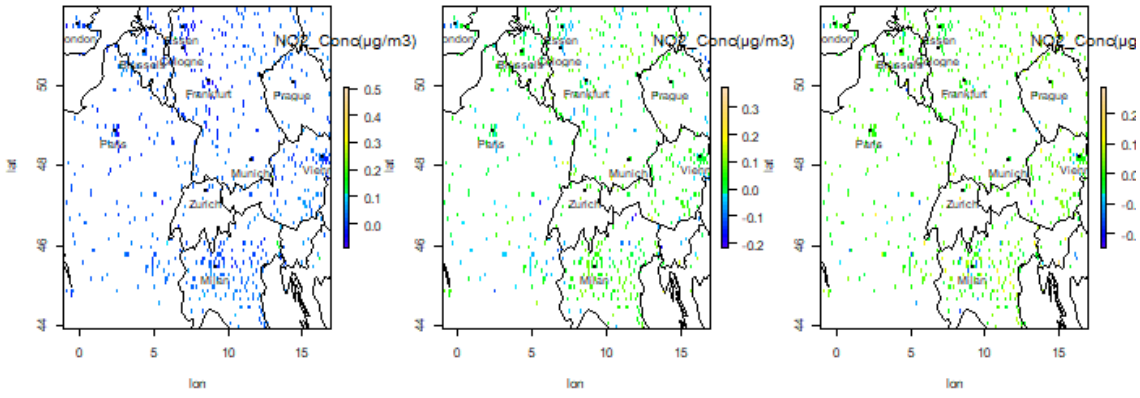
(b)

Figure 19: (a) NO_2 Temporal Correlation model-model-station on June 2016 (upper left: Polyphemus vs station, lower left: CAMS vs station) and Jan 2018 (upper right: Polyphemus vs station, lower right: CAMS vs station). (b) NO_2 Spatial correlation model-model-station on June 2016 with P-value (upper left: Polyphemus vs station, upper right: its P-value, lower left: CAMS vs station, lower right: its P-value).

With respect to spatial variation, there is a higher correlation in the urban regions and remains positive. For Polyphemus, the correlation varies with close to null and negative values for several suburban and rural regions and these changes in suburban and rural areas continue in both summer and winter. Pixel-wise, several significant locations are changing based on the seasons. For example, in summer, some of the western regions of Polyphemus and the Alps regions over northern Italy have several significant stations in summer and they become less active in winter. The correlation results for Polyphemus are favourite in summer with several locations showing excellent correlation all over the year for CAMS. (*see Appendix 1 for more NO_2 statistical spatial results*).



(a)



(b)

Figure 20: (a) NO_2 Temporal Trend model-model-station from Jun 2016 to Dec 2018. (b) NO_2 Spatially varying trend model-model-station from Jun 2016 to Dec 2018 (left: Polyphemus trend, middle: CAMS trend, right: station trend).

The temporal and spatially varying trend for NO_2 in CAMS and Polyphemus during the time window gives an insight into the longtime performance of the models. Temporally, there is a strong seasonal trend between the model and the station observations for NO_2

with regular fluctuations in summer and winter. There is a positive seasonal trend in both the models as analysed from Mann Kendall's trend test over the complete time window (fig.20a). Spatial Mann Kendall's trend test show varying results based on geographical locations. For CAMS, there is a negative trend in most of the urban regions and positive trends in the suburban regions over the complete period of interest and the negative trend is observed from the Polyphemus data. There is a significant difference in the results from spatial and temporal trends, each conveys trends from a different perspective *(see appendix 5 Table A1 for Mann Kendall's trend statistics results for NO_2).*

6.1.2 O_3

For O_3, time series from both the model results were found that the surface O_3 level increases when the surface NO_2 decreases, especially in winter (Bauwens et al.,2020). The results from both models are comparable to the station observations. From figure 19, it is observable that the CAMS follows the same pattern of station observations considered and the Polyphemus deviates slightly from the station observations. There is also a strong seasonal variation recorded with increasing concentrations in summer and decreases in winter (fig. 21).

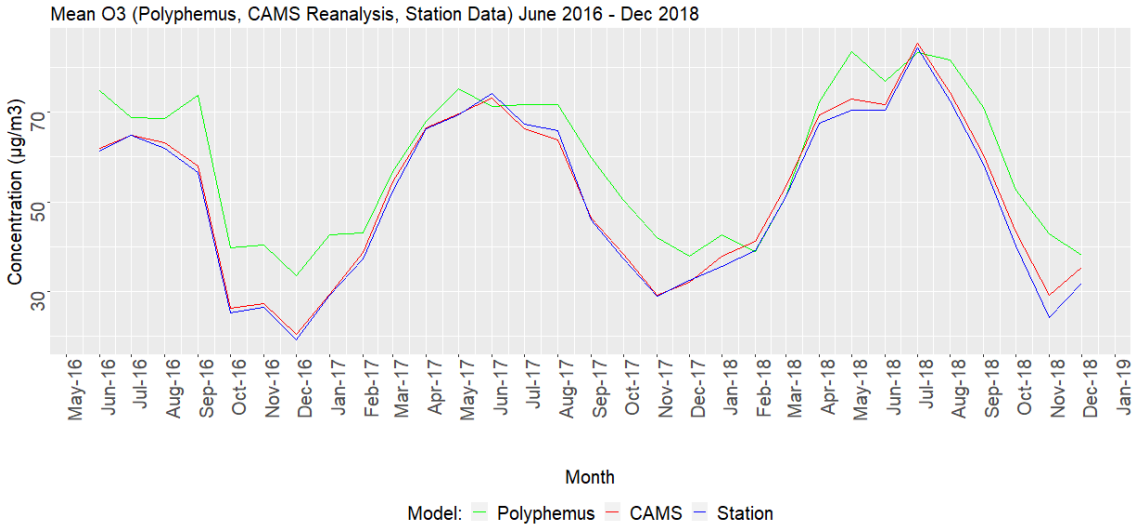
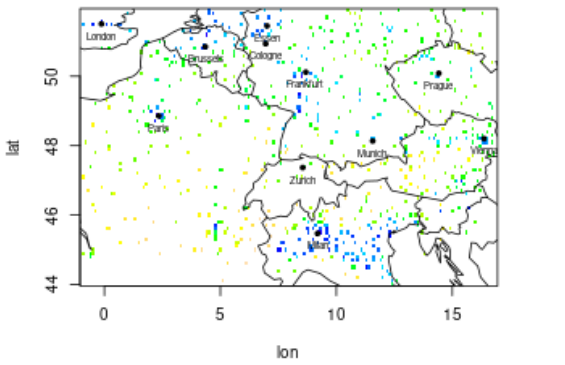


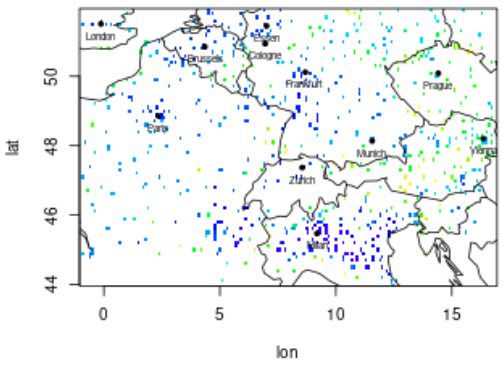
Figure 21: Time-series of O_3 Model-model-station comparison from June 2016 to Dec 2018.

There is a considerable difference in the model results based on geographical locations and seasonal variations. High concentrations in the southern and western regions of the area of interest were observed during summer from Polyphemus outputs and also considerable deviations in most of the suburban and rural regions during winter seasons. Both the model results are comparable in the urban regions during winter and high deviations in summer in both urban and rural regions. The CAMS results are more similar to the station observations throughout the time period. There are no significant differences between CAMS and station observations. They follow the same pattern in

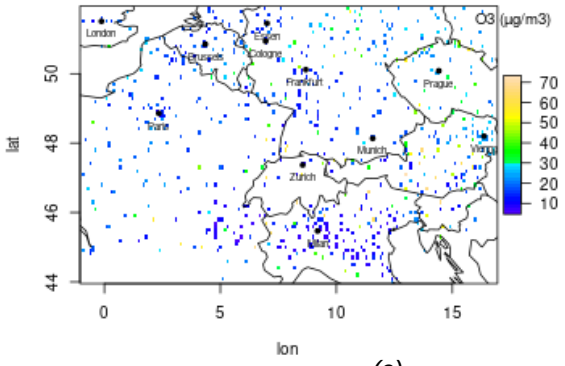
all the seasons (fig. 22). Overall from the time period considered, the model results in urban areas like Paris, Milan, Münich, Frankfurt, etc., are comparable to the station observations. For O_3, higher deviations are observed in the rural regions of the domain.



(a)

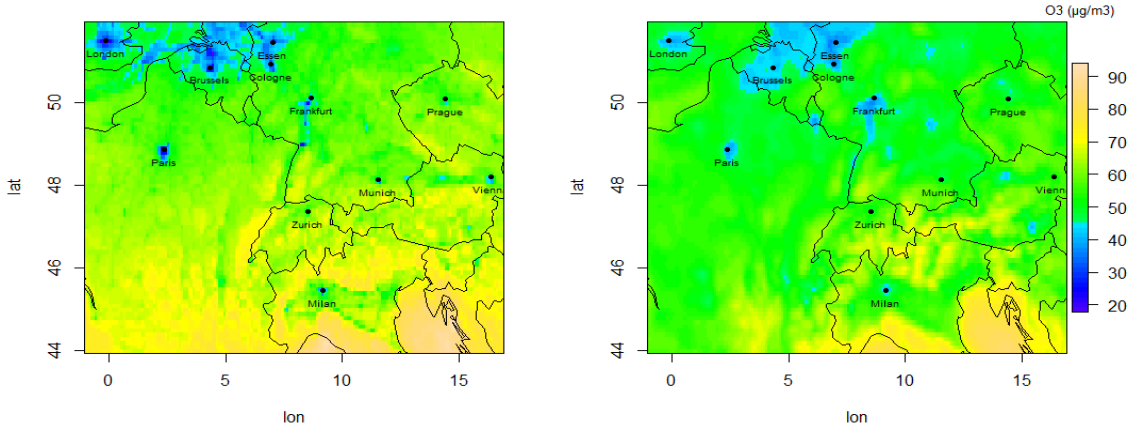


(b)

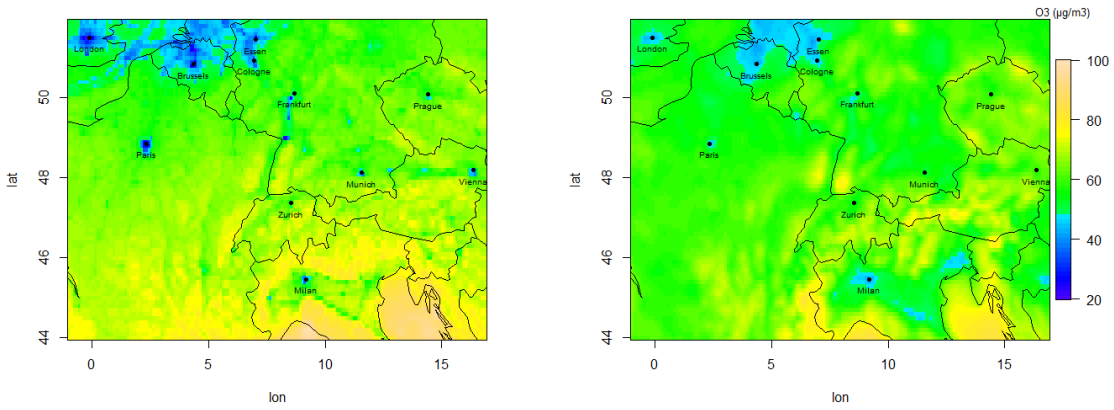


(c)

Figure 22: O_3 monthly mean of models and station data for the month of December 2016 in the pixels of station locations (a): Polyphemus, (b): CAMS, (c): station datasets.



(a)



(b)

Figure 23: (a) Yearly O_3 mean for 2017 in the extent of Polyphemus domain two (left: Polyphemus mean, right: CAMS mean). (b) Yearly O_3 mean for 2018 in the extent of Polyphemus domain two (left: Polyphemus mean, right: CAMS mean).

The comparison between the models for O_3 has variant yearly results. Overall, the concentrations in urban areas are more comparable than the rural areas. But, there are also deviations in the regions of southern France where the Polyphemus concentrations are high. The distribution of concentrations observed in the Polyphemus for O_3 is completely vice versa to model results of NO_2 (fig. 23 a&b).

The distribution of the model datasets over time for O_3 is represented in figure 22. The standard deviations between the model-model-station comparison follow the same pattern for all the datasets considered. This resembles the coordination between the models and station datasets for the pollutant O_3. Also, the distribution follows the seasonal fluctuations in summer and winter which makes the model predictions best for O_3 (fig. 24).

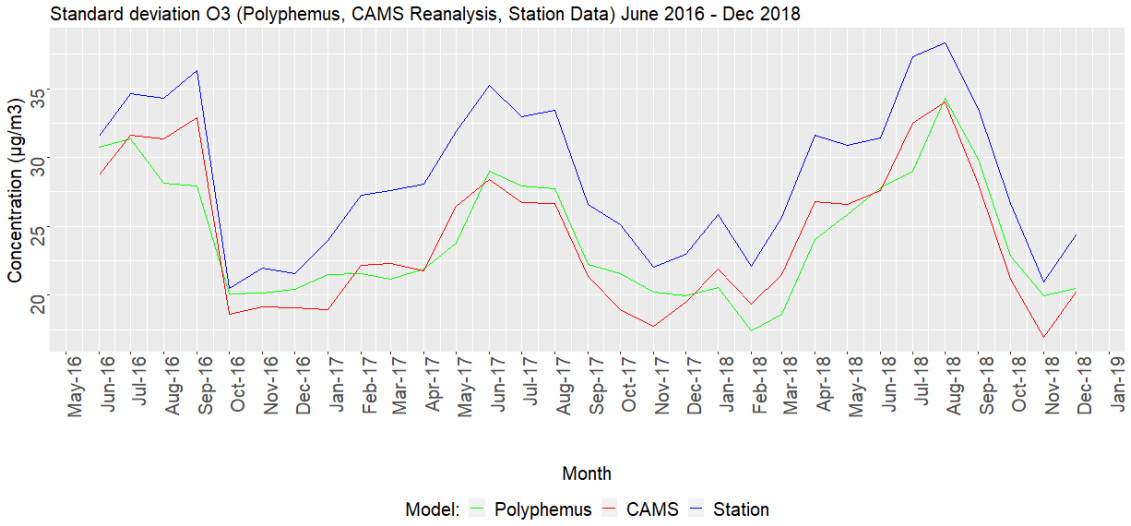
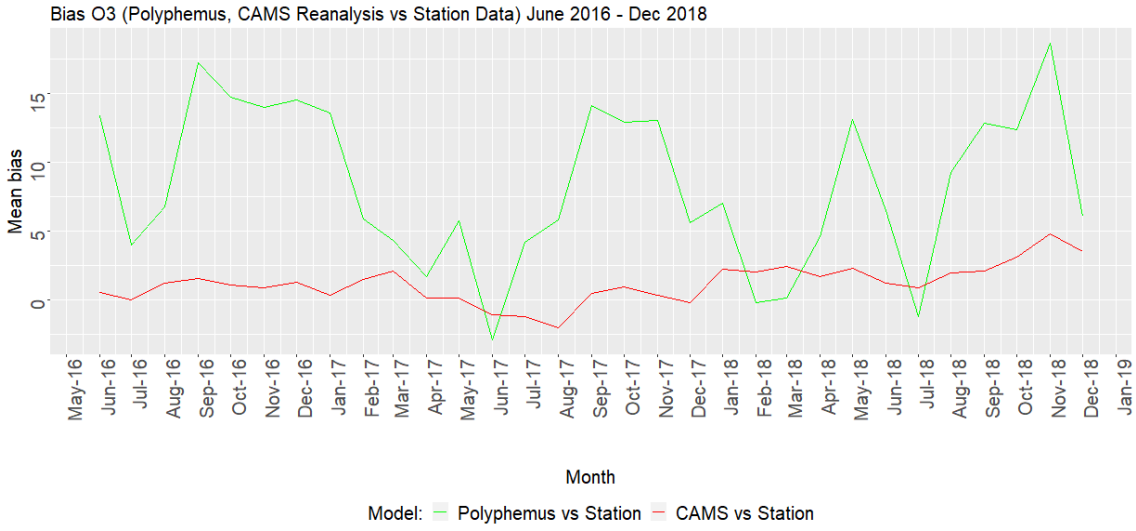
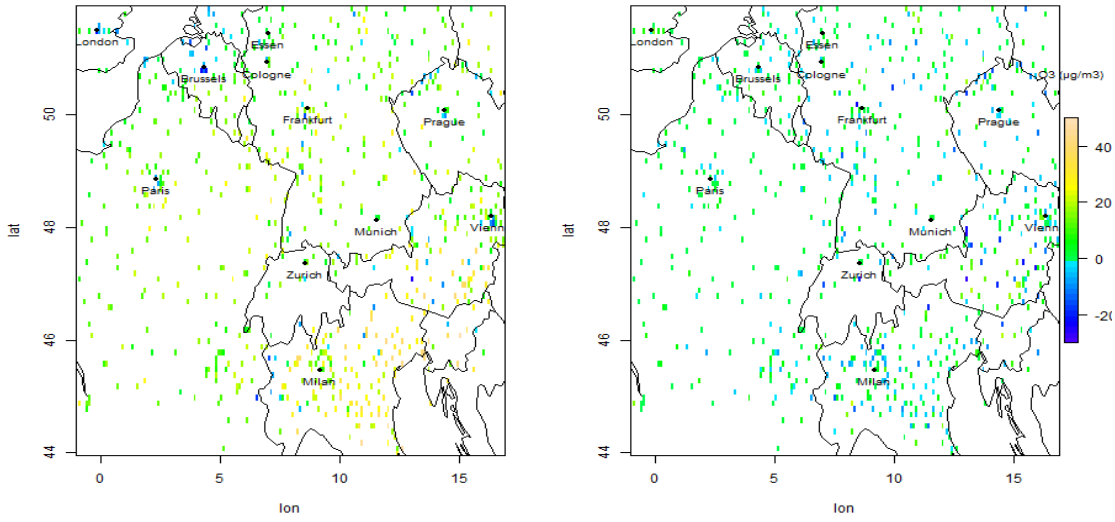


Figure 24: O_3 Standard deviation between model-model-station from Jun 2016 to Dec 2018.

From figure 25a, it can be noticed that the mean bias for O_3 in CAMS results are closer to zero which makes the CAMS outputs more comparable to in-situ observations than the Polyphemus outputs. There are no seasonal observable patterns in the models with respect to the station concentrations. Polyphemus concentrations are biased towards positive. Considering station pixels, there are notable changes in the mean bias that vary based on locations and are also positively biased with respect to station data throughout the time window. Pixel-wise, there are significant seasonal differences in the model's outputs in urban and rural areas. The alps regions in northern Italy and some rural areas of southern France show a high positive bias (fig. 25b). The mean bias of O_3 has comparatively differing results from the mean bias of NO_2. (*see Appendix 2 for more O_3 statistical spatial results*).



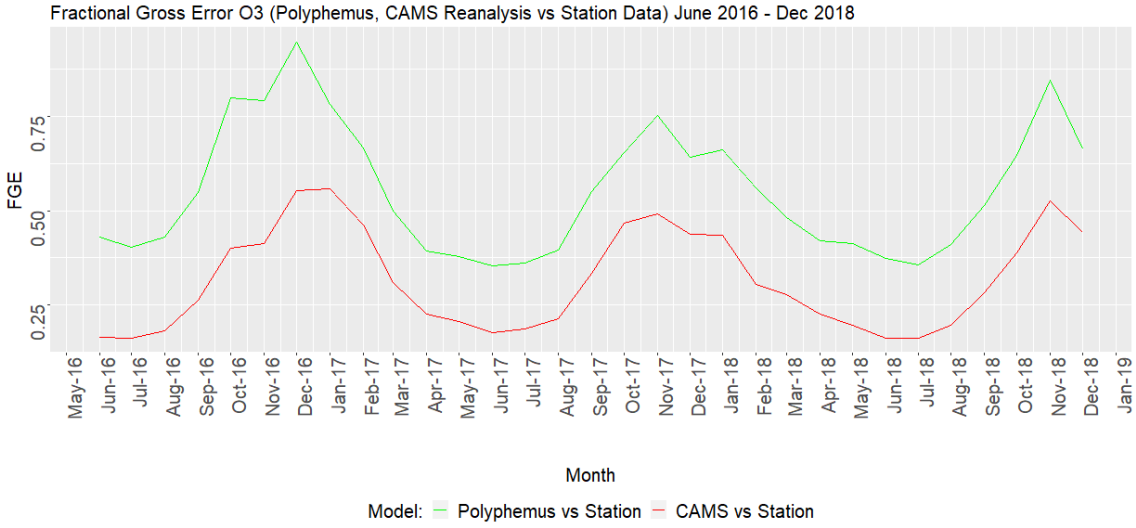
(a)



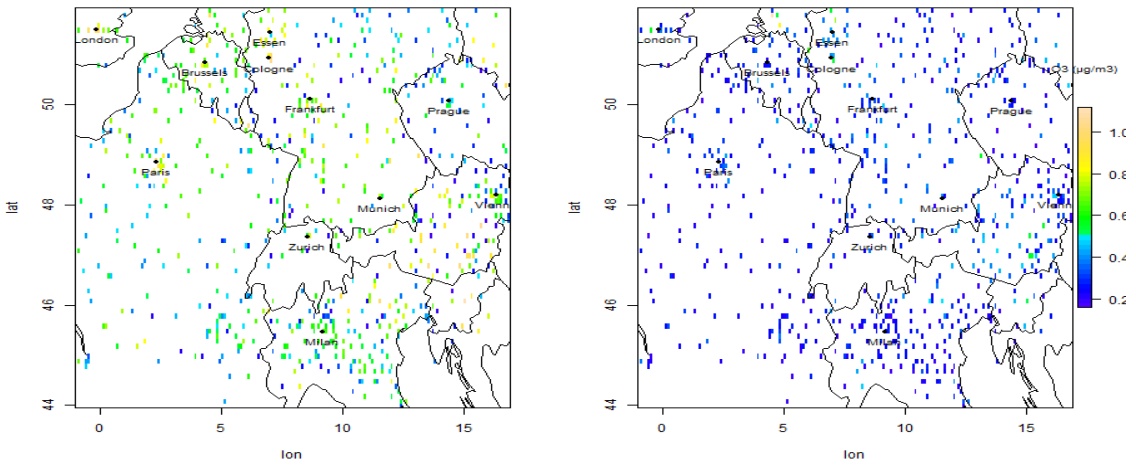
(b)

Figure 25: (a) Temporal O_3 mean bias model-model-station from June 2016 to Dec 2018. (b) Spatial O_3 mean bias model-model-station for June 2016 (left: Polyphemus mean bias, right: CAMS mean bias).

The FGE comparing Polyphemus, CAMS with respect to the station data for O_3 gives the easily understandable results in figure 26a. As the FGE results from both models are not higher than 1, the models have good predictions compared to real-world measurements. Though the performances of both the models are comparable with the station observations, FGE for CAMS is closer to zero than Polyphemus. There is a strong spatial and temporal seasonal pattern between the models and the station datasets. Both the spatial and temporal results justify the same FGE over the time window with mild variation over the alps regions in Northern Italy (fig. 26b).



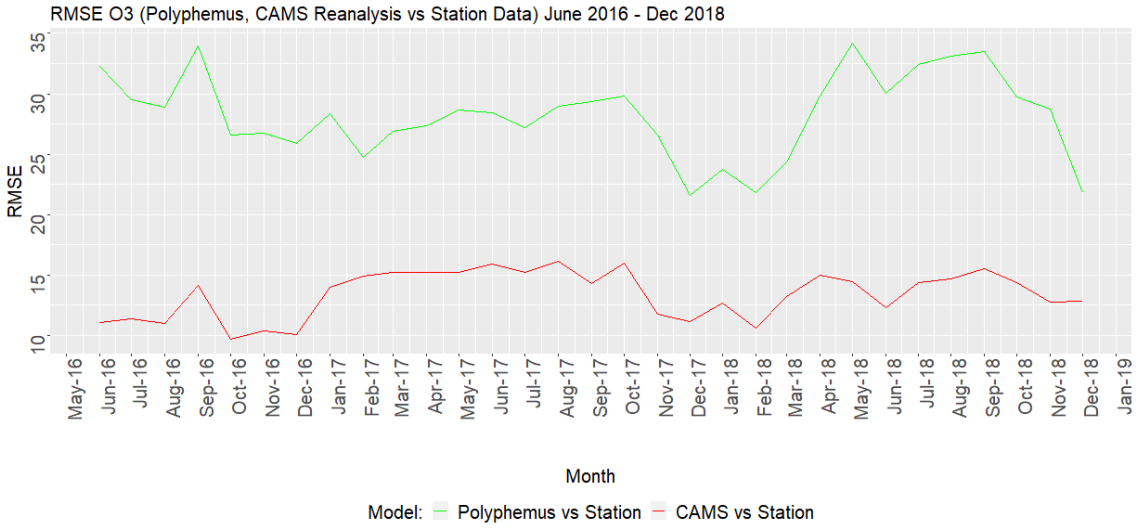
(a)



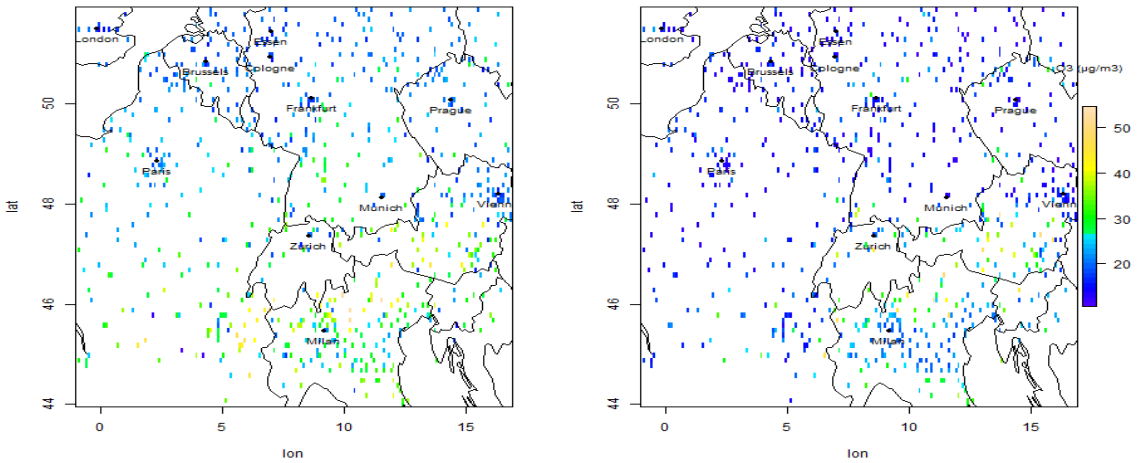
(b)

Figure 26: (a) Temporal O_3 FGE model-model-station from June 2016 to Dec 2018. (b) Spatial O_3 FGE model-model-station for Sept 2016 (left: Polyphemus FGE, right: CAMS FGE).

The results of RMSE and FGE for O_3 show that CAMS performs better with respect to the real-world measurements. RMSE of CAMS is close to zero than Polyphemus zero resulting the healthier predictions that are similar to station observations. There are no significant patterns followed in the results of RMSE for O_3 (fig. 27a). The deviation between the models in the urban areas for O_3 is high and persists throughout the period of interest. Most of the cities in the area of interest show deviations in Polyphemus with respect to the station datasets. There are some mild deviations observed in the rural regions and CAMS performs better throughout the domain. The deviations in the alps regions are higher during the winter seasons.



(a)

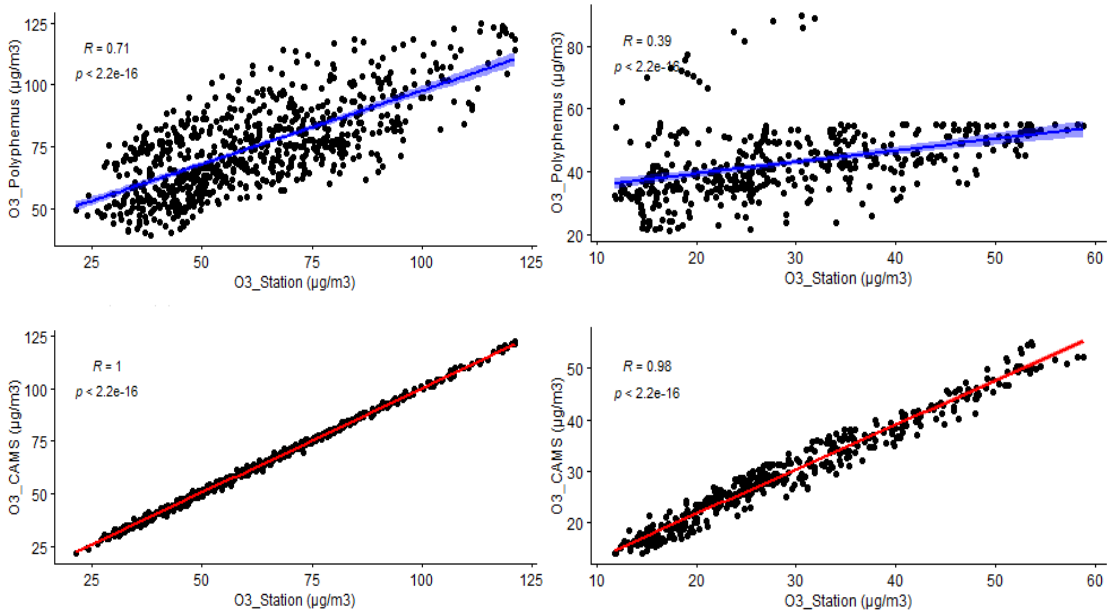


(b)

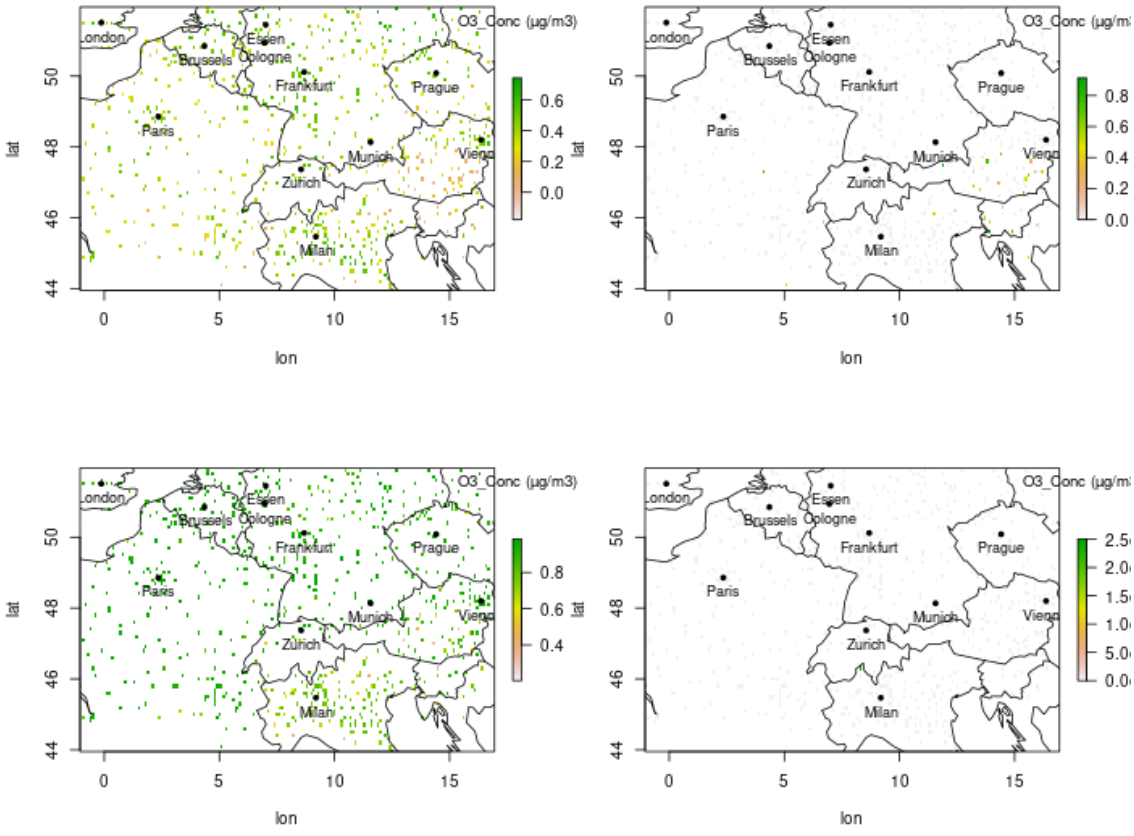
Figure 27: (a) Temporal O_3 RMSE model-model-station from June 2016 to Dec 2018. (b) Spatial O_3 RMSE model-model-station for Nov 2017 (left: Polyphemus RMSE, right: CAMS RMSE).

The spatial and temporal correlation coefficient for CAMS shows that CAMS datasets for O_3 are correlated with station data at an average of 0.98. This result can also be

evidenced by other statistical indicators. Some of the summer and autumn months show an excellent correlation near 1 for the CAMS model.



(a)

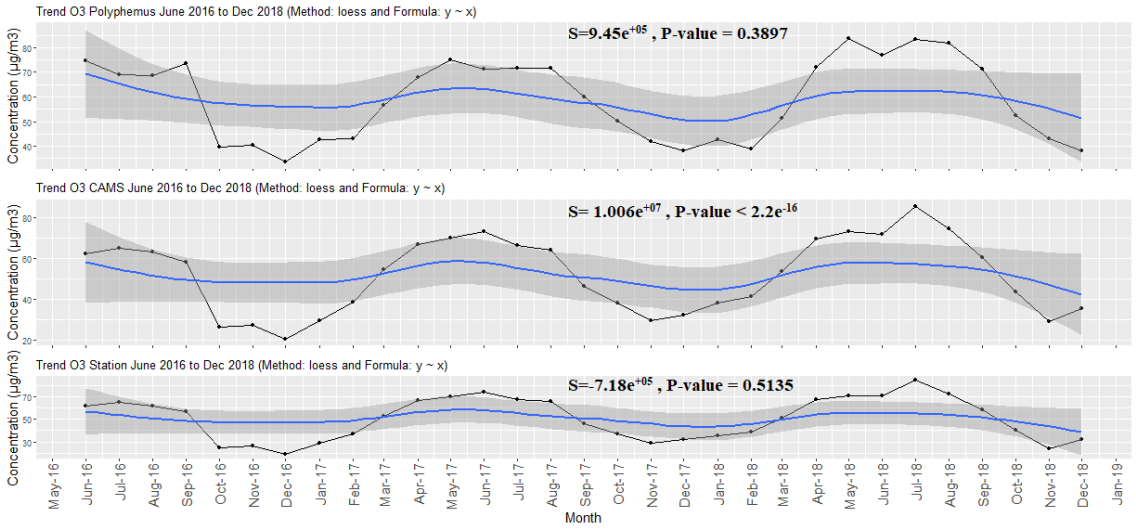


(b)

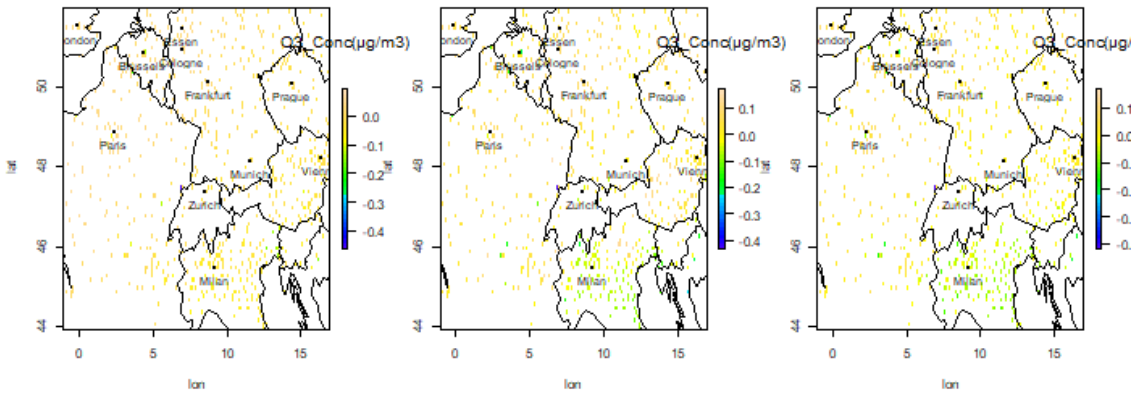
Figure 28: (a) NO_2 Temporal Correlation model-model-station on June 2016 (upper left: Polyphemus vs station, lower left: CAMS vs station) and Jan 2017 (upper right: Polyphemus vs station, lower right: CAMS vs station). (b) NO_2 Spatial correlation model-model-station on May

2018 with P-value (upper left: Polyphemus vs station, upper right: its P-value, lower left: CAMS vs station, lower right: its P-value).

As it was found for NO_2, Polyphemus O_3 shows seasonal deviations. On average, it is 0.7 during summer and 0.4 in winter. The correlation coefficient for Polyphemus in winter are not that comparable to station observations. The varying correlation results are due to the observed outliers in the Polyphemus outputs. Without these, results would much improve. The outliers in the datasets are visible in figure 28(b).



(a)



(b)

Figure 29: (a) O_3 Temporal Trend model-model-station from Jun 2016 to Dec 2018. (b) O_3 Spatially varying trend model-model-station from Jun 2016 to Dec 2018 (left: Polyphemus trend, middle: CAMS trend, right: station trend).

There are considerable differences observed between the models during both summer and winter in spatial correlation analysis. The deviations in the western and the southern regions of the area of interest are easily noticeable from figure 28b. The significant station pixels in the model datasets were found in the eastern and southern regions of the area

of interest. The correlation analysis performs better in summer with several significant locations for Polyphemus and remains excellent correlation all over the year for CAMS. It was observed that there is a strong seasonal trend over the time window with regular fluctuations in summer and winter for O_3. It was analysed that there is a negative seasonal trend in both the models from Mann Kendall's trend test with a positive significance linear trend over the complete time window (fig.29a).

Spatial Mann Kendall's trend test results in varying trends in the alps and southern regions of the area of interest for both the models. There is a positive trend for O_3 in most of the major cities and rural areas in the north and central regions in the area considered. There is a deviation in the regions of northern Italy which has a negative trend in CAMS that resembles the trend in real-world measurement and it is a positive trend in Polyphemus (fig.29b). *(see appendix 5 Table A2 for Mann Kendall's trend statistics results for O_3).*

6.1.3 PM_{10}

For PM_{10}, the model's outputs and the station observations are comparable. Though there are no seasonal patterns followed by Polyphemus, the model datasets follow the actual pattern of the real-world observations. There are no considerable deviations between the datasets in the station pixels over the time window (fig. 30). The deviations in January 2017, also evident in other statistical analyses for PM_{10} are due to data gaps.

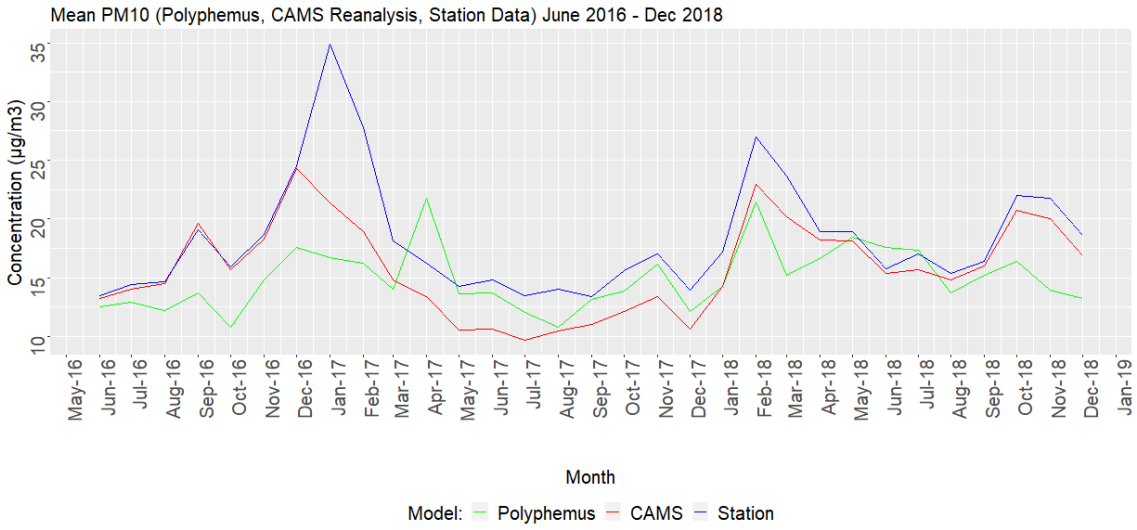


Figure 30: Time-series of PM_{10} Model-model-station comparison from June 2016 to Dec 2018.

In the pixel-wise analysis for PM_{10}, there are considerable deviations between the models' outputs beyond seasons. It was observed in Polyphemus that the concentrations of PM_{10} have nearly been underestimated in most of the rural regions and severe outlier in Paris and its surrounding station pixels with respect to the station observations

throughout the time window. There are seasonal patterns observed for PM_{10} in the pixel-wise analysis. The underestimation of the concentrations by Polyphemus is continuous in all the seasons. CAMS outputs are comparable with station datasets (fig. 31). Between the models, in Polyphemus, there is an extreme outlier observed in some of the urban regions in France and Belgium. These outliers exist in some cities during our complete time period considered (June 2016 to Dec 2018).

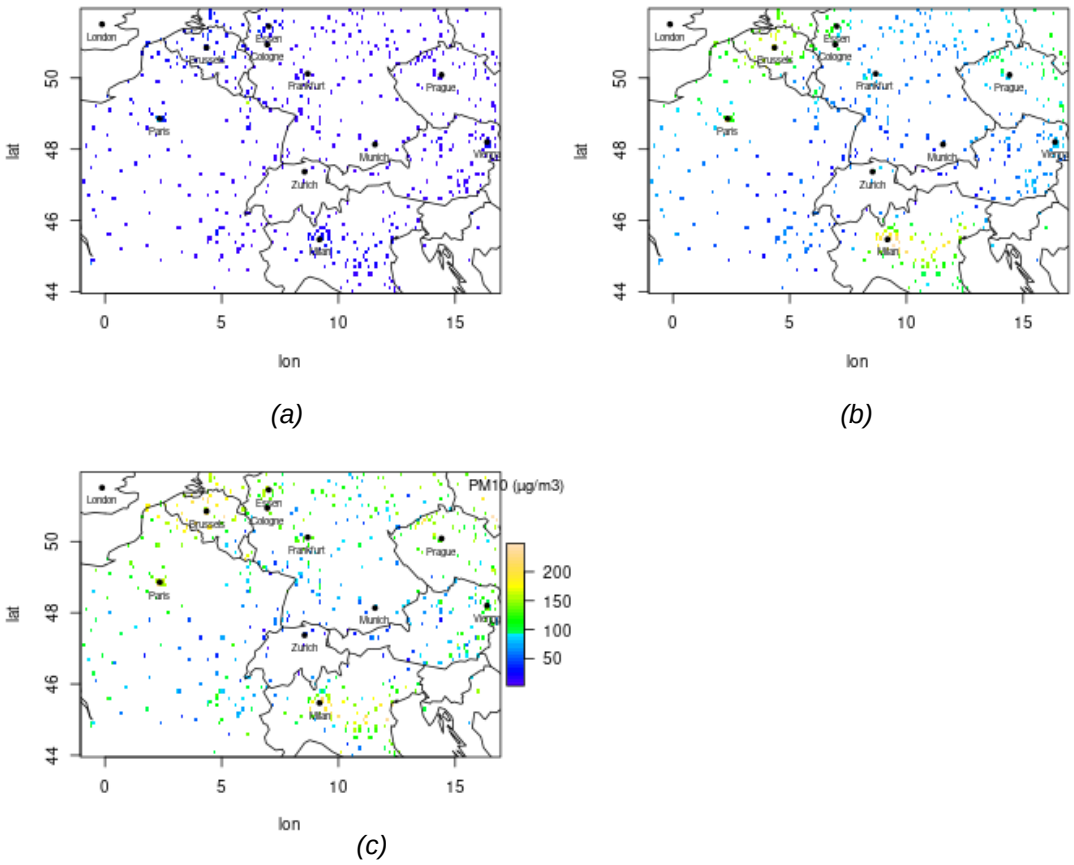


Figure 31: PM_{10} monthly mean of models and station data for the month of June 2016 in the pixels of station locations (a): Polyphemus, (b): CAMS, (c): station datasets.

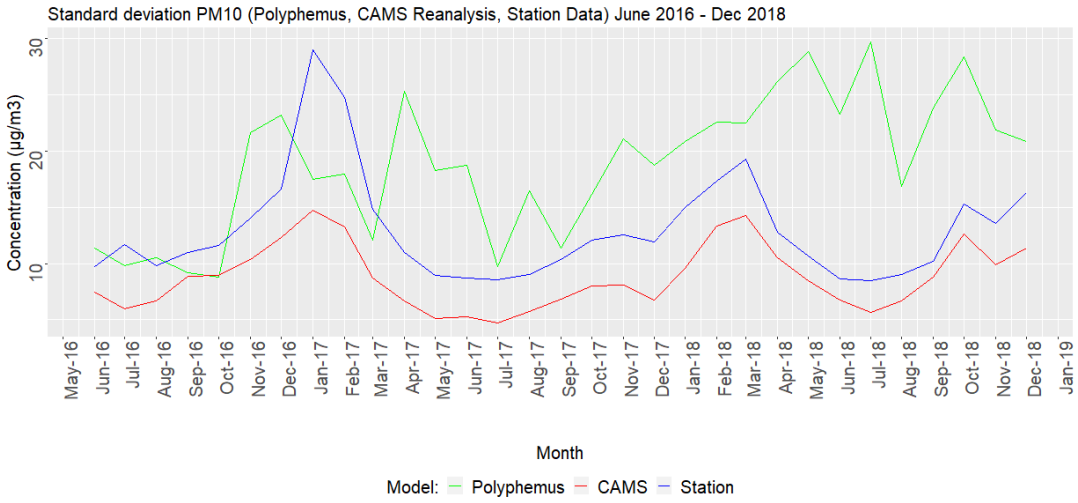
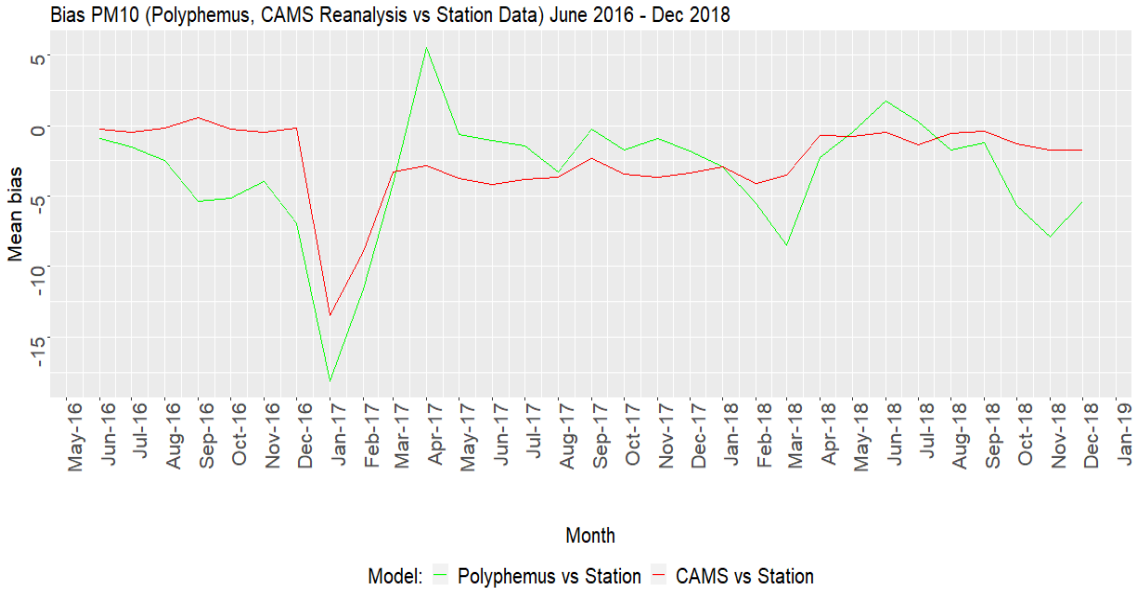


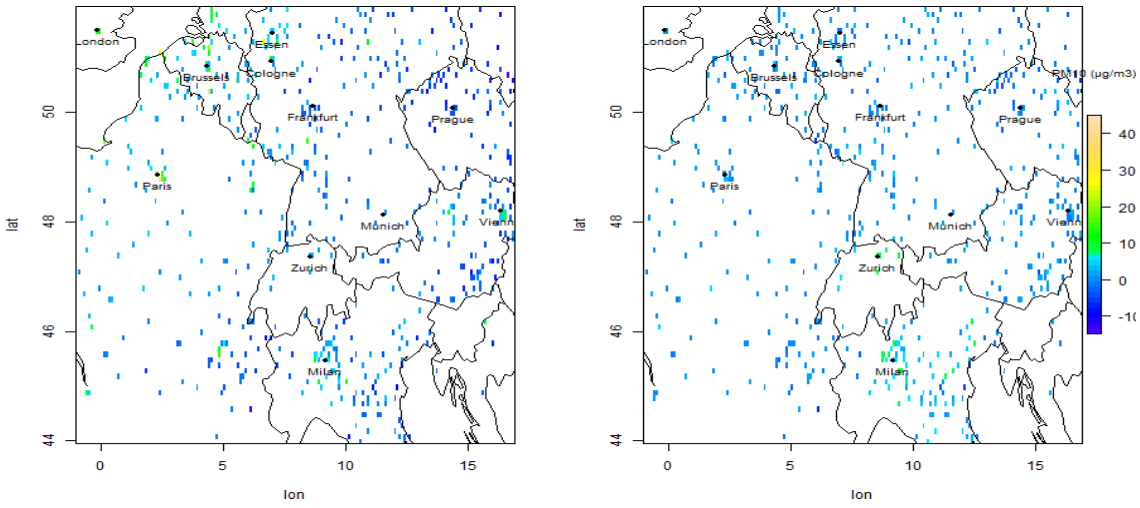
Figure 32: PM_{10} Standard deviation between model-model-station from Jun 2016 to Dec 2018.

The standard deviation of PM_{10} reflects the time-series of PM_{10} that the CAMS data are continuous and equally distributed around the mean with respect to station datasets. The CAMS standard deviation is low and follows the pattern of station data. The pattern of Polyphemus standard deviation deviates from real-world observations (fig. 32).

With respect to the station observations, the mean bias results show that there is no significant systematic error in both the model's results to compare station observations. Though the models' outputs are negatively biased in some months, it is also important that the biases are close to zero. But the monthly pixelwise mean bias analysis results show there is a seasonal pattern in the performance of Polyphemus and CAMS. The Polyphemus outputs are negatively biased in summer and bias close to zero shows good predictions in later winter. The alps and western regions of the area of interest are low biased towards station concentrations. The CAMS model outputs are positively biased and more comparable towards station pixels in summer and highly biased outwards in cities and other regions during winter (fig. 33b). The temporal and spatial mean bias for PM_{10} conveys some varying results with respect to time and geographical locations. (*see Appendix 3 for more PM_{10} statistical spatial results*).



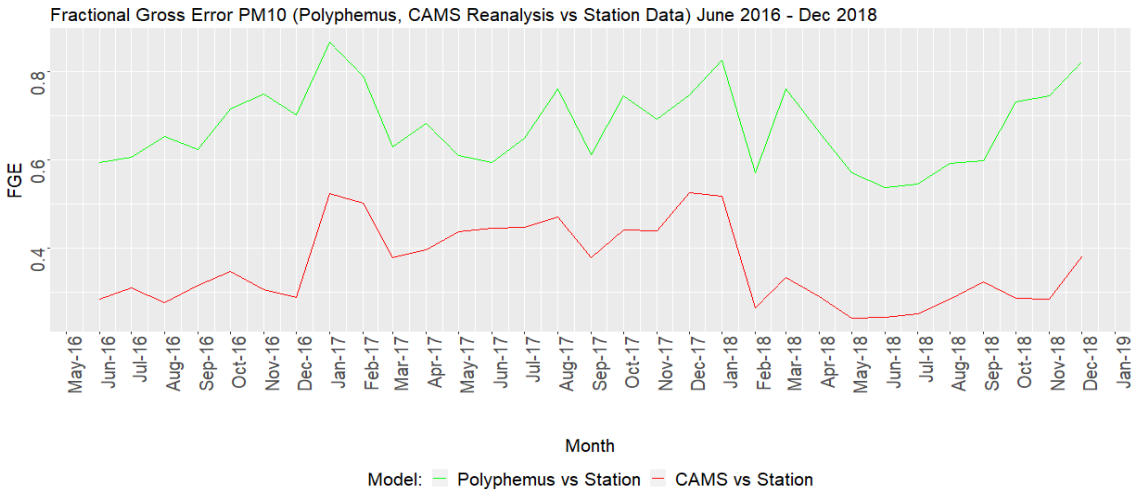
(a)



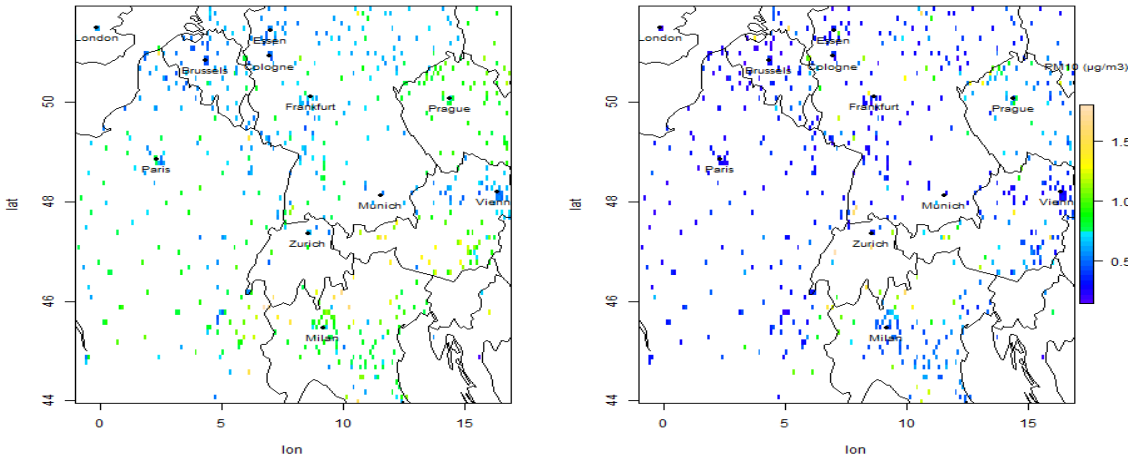
(b)

Figure 33: (a) Temporal PM_{10} mean bias model-model-station from June 2016 to Dec 2018. (b) Spatial PM_{10} mean bias model-model-station for June 2016 (left: Polyphemus mean bias, right: CAMS mean bias).

The results of FGE and mean bias for PM_{10} is also varying from the previous results of NO_2 and O_3. It is visible from the FGE for PM_{10} that CAMS performs better than Polyphemus with respect to station data. As the FGE for both the models are less than one, it makes both the model results are comparable with the station datasets. FGE for PM_{10} follows varying patterns over time (fig. 34a). The FGE spatial analysis also adds evidence to temporal analysis that FGE for both the models are less than one and CAMS performs better. The performance of both the models are more comparable in spring, summer, and autumn than winter. There are some changes observed in the model's outputs in later 2018 that model predictions are more comparable with the station observations.



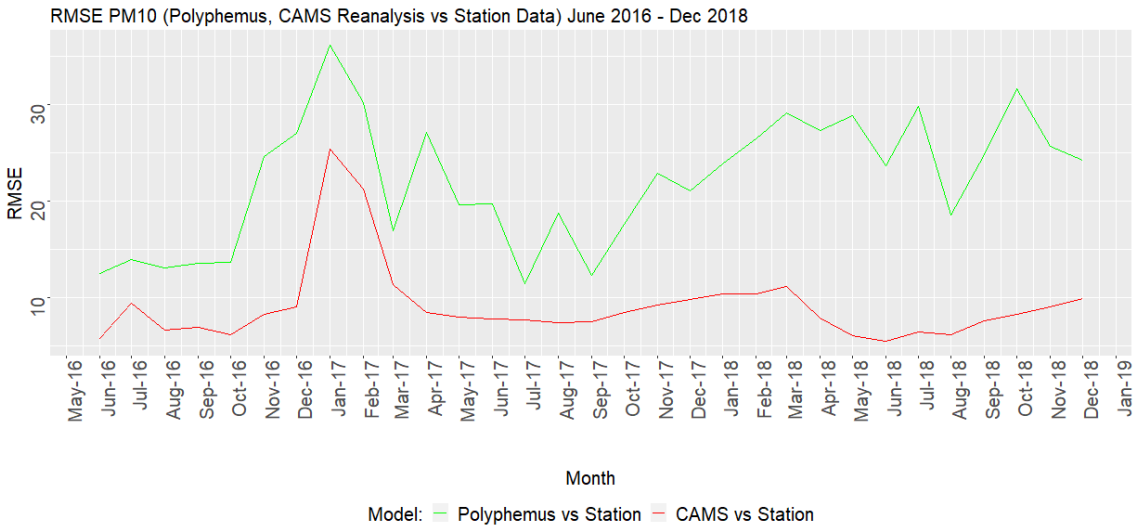
(a)



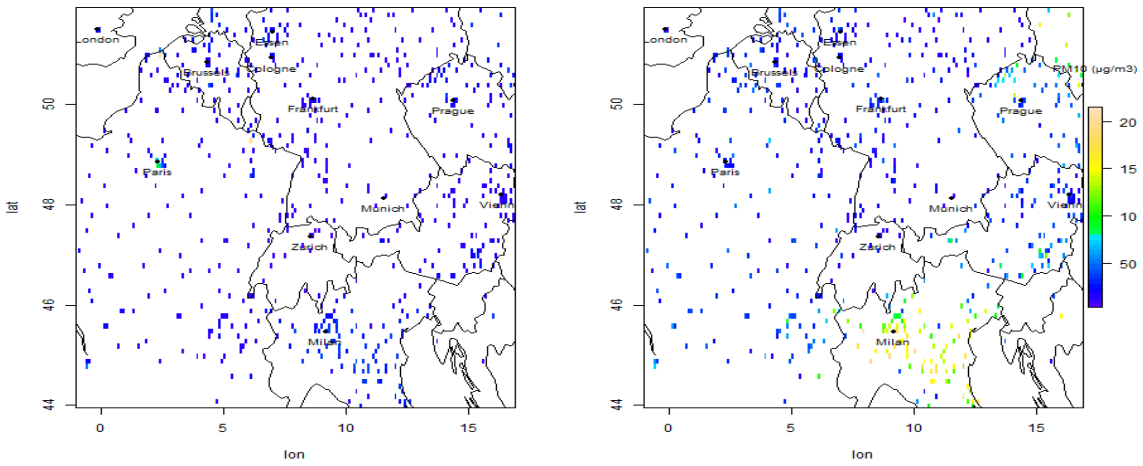
(b)

Figure 34: (a) Temporal PM_{10} FGE model-model-station from June 2016 to Dec 2018. (b) Spatial PM_{10} FGE model-model-station for Dec 2016 (left: Polyphemus FGE, right: CAMS FGE).

The RMSE for CAMS, Polyphemus with respect to the station datasets show that there is no seasonal pattern over time. As FGE, RMSE for model-model-station comparison also evidences that CAMS outperforms Polyphemus (fig. 35a). There are considerable deviations observed in the northern Italian regions during winter. Especially these deviations in Italian regions are spotted in CAMS. RMSE with respect to station observations analysed that the models' outputs are better in summer than winter (fig. 35b).



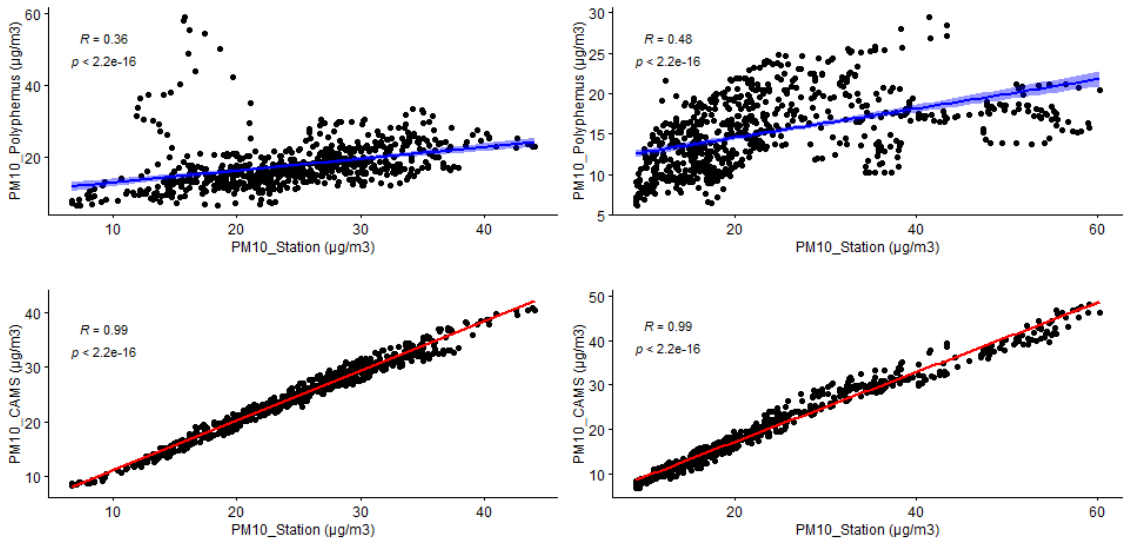
(a)



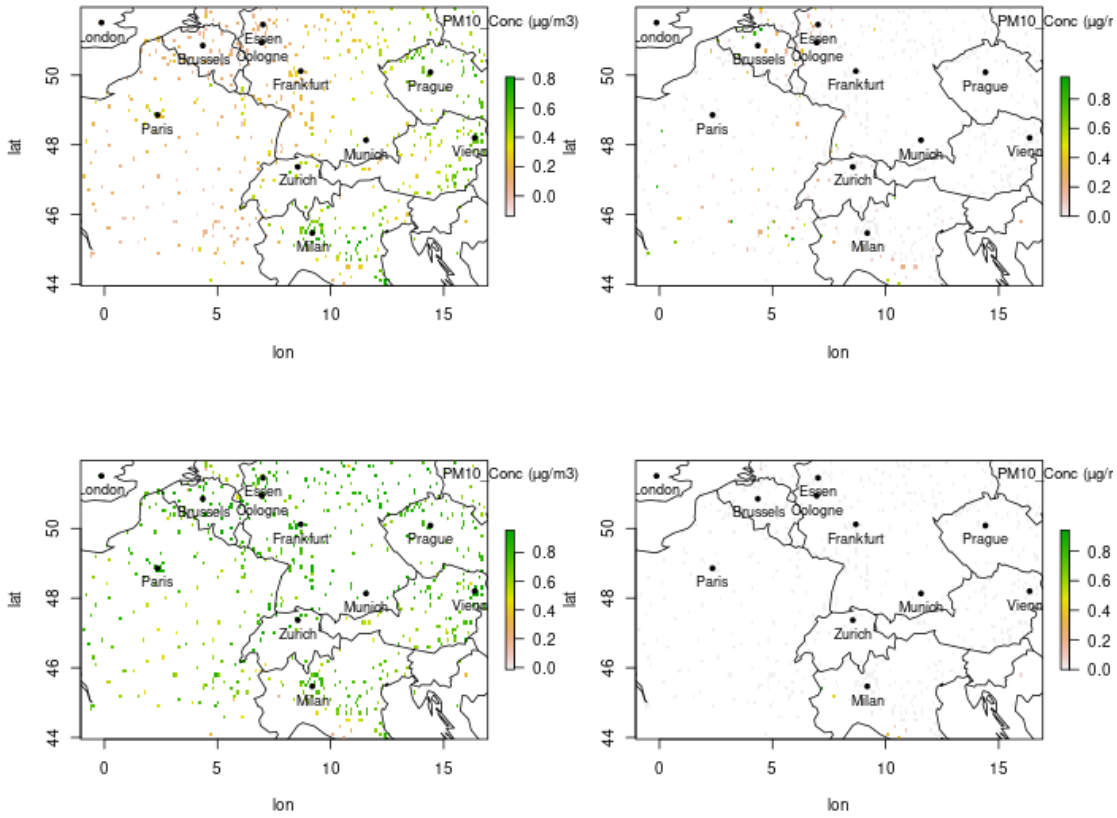
(b)

Figure 35: (a) Temporal PM_{10} RMSE model-model-station from June 2016 to Dec 2018. (b) Spatial PM_{10} RMSE model-model-station for Nov 2017 (left: Polyphemus RMSE, right: CAMS RMSE).

Out of all the statistical indicators analysing PM_{10} datasets, the results from correlation give the exact idea about how relevant the models predicted with the real world. On average, CAMS correlated 0.95 and Polyphemus correlated 0.34 with respect to station observations. The outliers in the PM_{10} Polyphemus outputs indicate some systematic model error. Here it is probable due to a specific event as the points describe a trajectory.



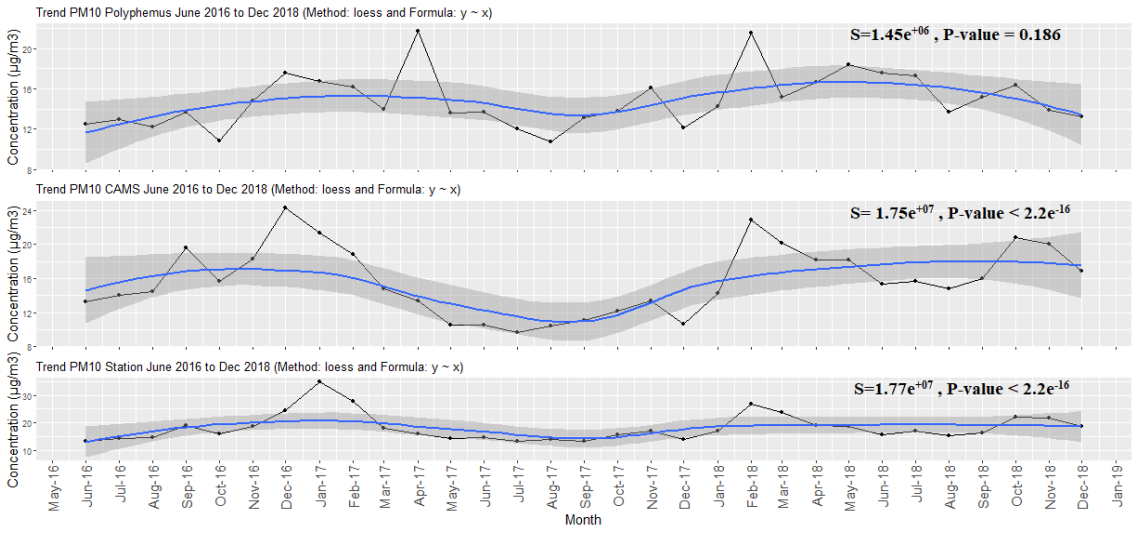
(a)



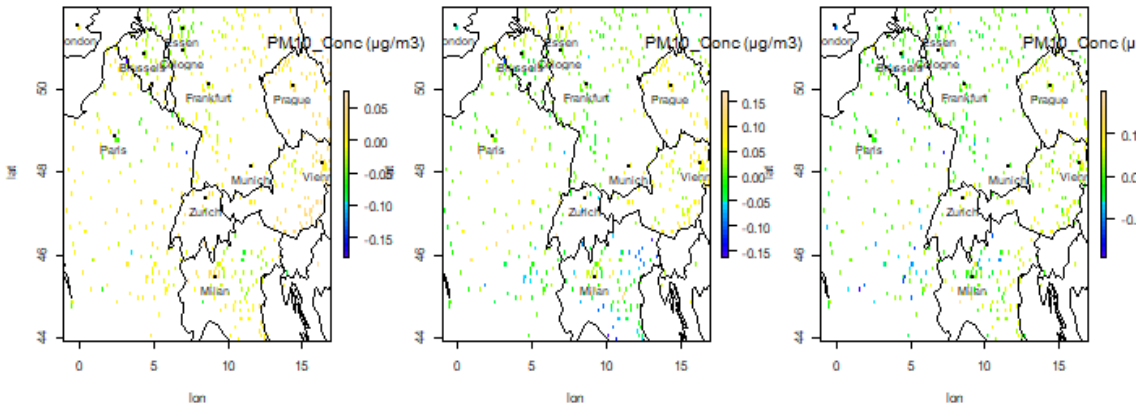
(b)

Figure 36: (a) PM_{10} Temporal Correlation model-model-station on Dec 2016 (upper left: Polyphemus vs station, lower left: CAMS vs station) and Mar 2018 (upper right: Polyphemus vs station, lower right: CAMS vs station). (b) PM_{10} Spatial correlation model-model-station on Oct 2017 with P-value (upper left: Polyphemus vs station, upper right: its P-value, lower left: CAMS vs station, lower right: its P-value).

It was found for Polyphemus that correlations reach as low as 0.1 and 0.3 for some months. There is no seasonal pattern found in these results. But the results from CAMS maintain their high correlations with station observations (fig. 36a). Spatial correlation also shows that the Polyphemus does not compare as well with station datasets. Some of the stations in Germany and Switzerland show a good correlation in summer also some eastern regions in winter (fig. 36b).



(a)



(b)

Figure 37: (a) PM_{10} Temporal Trend model-model-station from Jun 2016 to Dec 2018. (b) PM_{10} Spatially varying trend model-model-station from Jun 2016 to Dec 2018 (left: Polyphemus trend, middle: CAMS trend, right: station trend).

With a strong seasonal pattern, the model-model-station comparison for PM_{10} shows a strong positive temporal trend in all the datasets considered. The Linear trend model also shows significance results. The spatial Mann Kendall's trend test results from figure 37b show a strong positive trend for Polyphemus outputs. Though the CAMS and station datasets show positive trends for most areas, there are also some stations in northern Italy, West Germany and Belgium where negative trends are observed over the time window considered. *(see Appendix 5 Table A4 for Mann Kendall's trend statistics results for PM_{10}).*

6.1.4 $PM_{2.5}$

Like PM_{10}, models outputs for $PM_{2.5}$ are also comparable with station observations. The patterns of both models and station $PM_{2.5}$ datasets were observed as similar as PM_{10}. CAMS generally compares well with real-world observations. There are clear seasonal patterns found with increasing concentrations in winter. Both the model results were observed within the mean of the station observations. The mean concentrations of Polyphemus look flat but the CAMS and station concentrations have a mild seasonal impact that there are low concentrations observed in summer and a bit rising in winter (fig. 38).

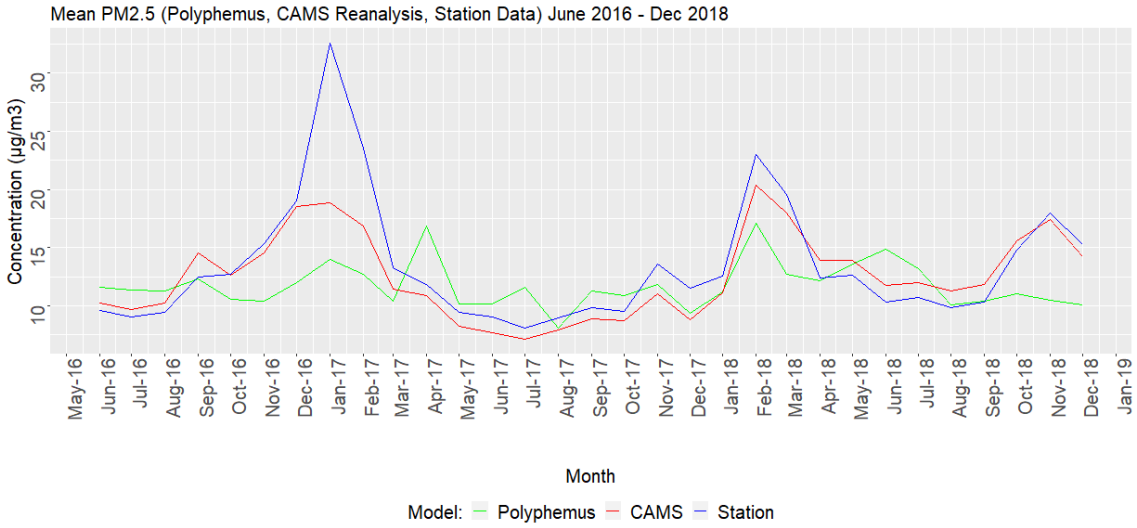


Figure 38: Time-series of $PM_{2.5}$ Model-model-station comparison from June 2016 to Dec 2018.

The spatial time series analysis for $PM_{2.5}$ also has comparable results between the models and the station datasets (*see appendix 5, Table A3 for Mann Kendall's trend statistics for $PM_{2.5}$*). There is a mild deviation found in the Polyphemus model results during the winter season in the locations like northern Italy, some stations in Vienna, Prague, France including Paris, etc., During summer, the station in Belgium and the areas covered in Germany show considerable deviations. These deviations in $PM_{2.5}$ vary based on geographical locations and continue to exist over years considered. For Polyphemus, the concentrations for most of the station pixels in the suburban and rural areas are underestimated and outliers were found in the cities like Paris, Milan, etc., which affects the performance of the models to a large extent. These outliers are also visible in the model-model comparison for $PM_{2.5}$. There are also mild deviations found in the CAMS results in some of the station pixels in Belgium and that exists over time. CAMS also has accurate predictions in the southern regions of the domain including Milan and its suburban, Vienna and the surrounding rural regions, etc., (fig. 39).

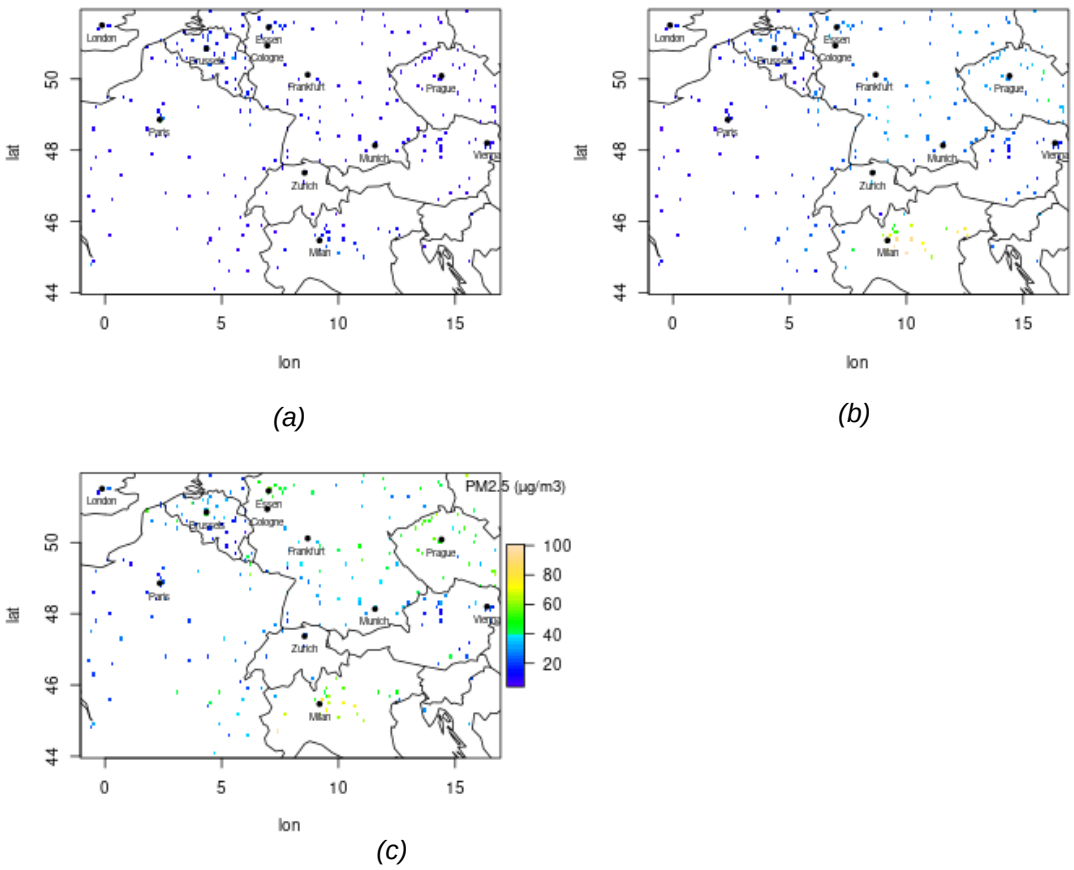


Figure 39: $PM_{2.5}$ monthly mean of models and station data for the month of July 2016 in the pixels of station locations (a): Polyphemus, (b): CAMS, (c): station datasets.

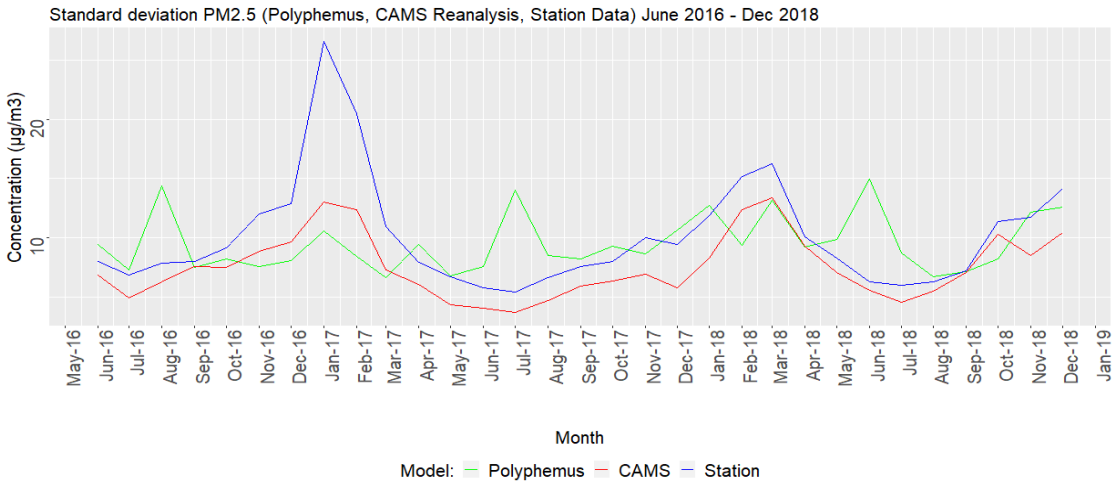
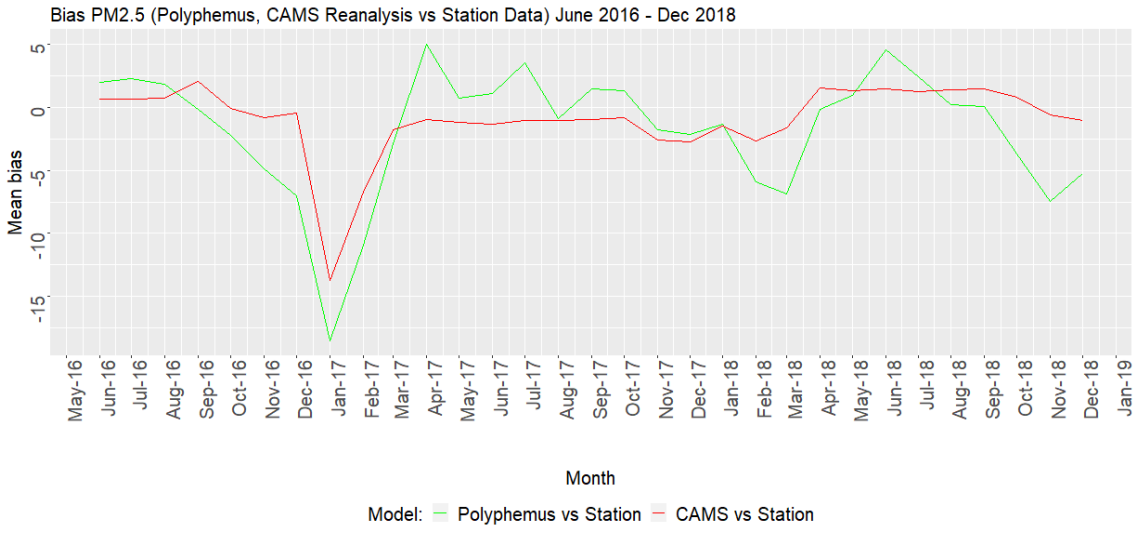
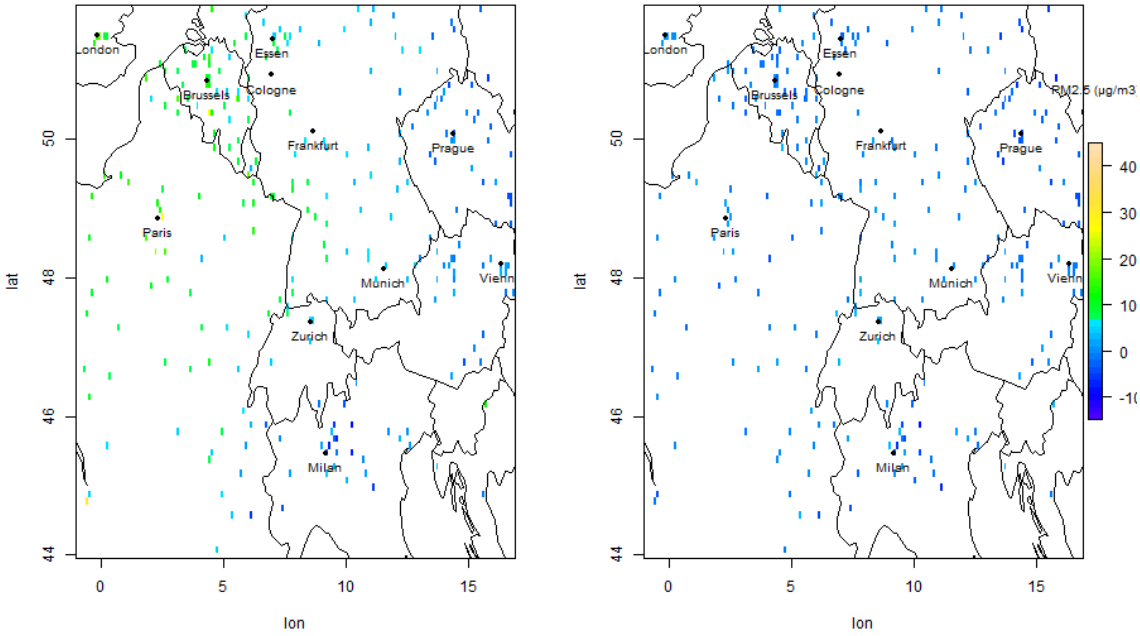


Figure 40: $PM_{2.5}$ Standard deviation between model-model-station from Jun 2016 to Dec 2018.

The clustered distribution of standard deviation of model and station concentrations evidences the dispersion of models' outputs around the mean (fig. 40). This standard deviation result justifies the mean time series of $PM_{2.5}$ and follows the same pattern. The $PM_{2.5}$ concentrations from models and stations are continuous and well distributed.



(a)

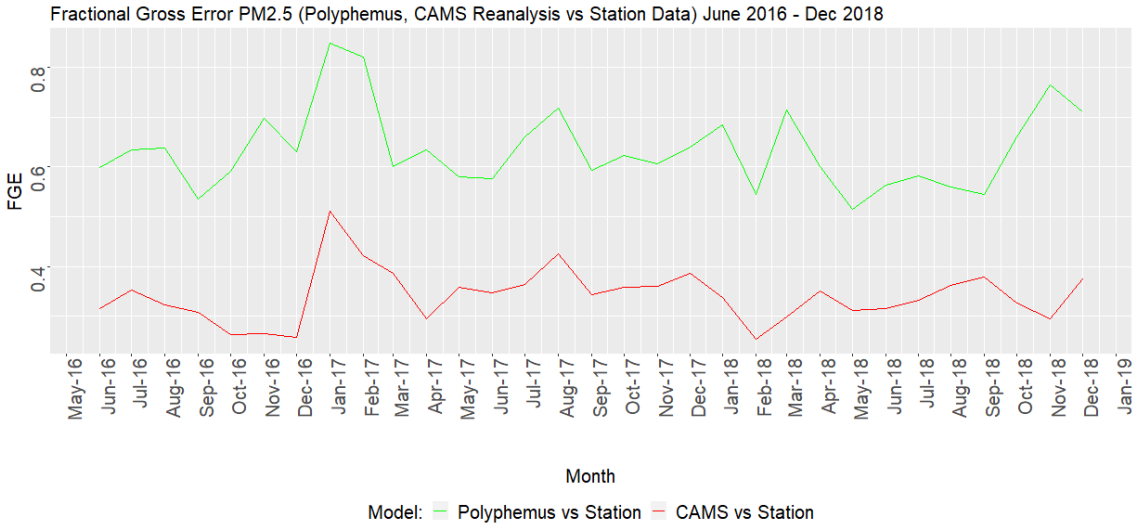


(b)

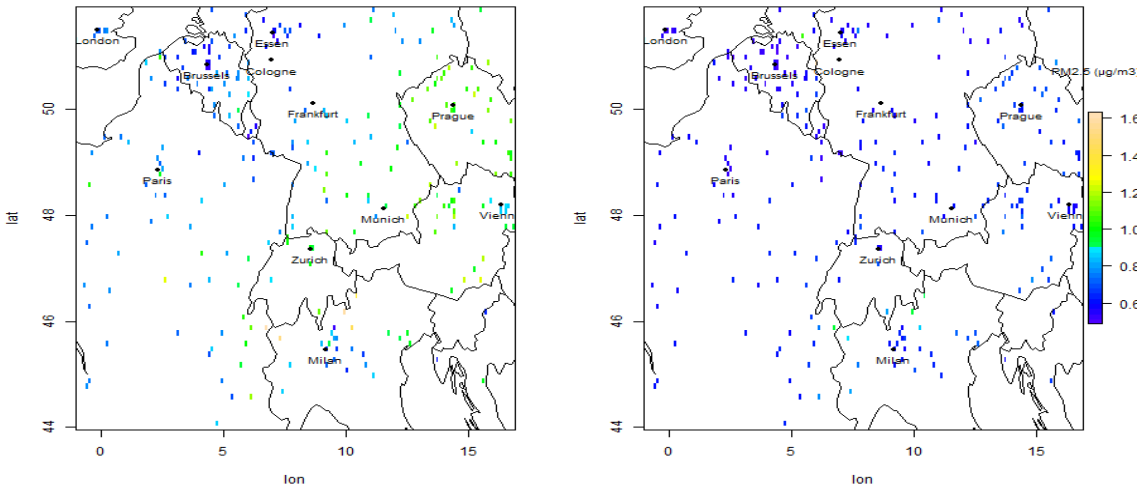
Figure 41: (a) Temporal $PM_{2.5}$ mean bias model-model-station from June 2016 to Dec 2018. (b) Spatial $PM_{2.5}$ mean bias model-model-station for April 2017 (left: Polyphemus mean bias, right: CAMS mean bias).

The mean bias analysis between the models and the station datasets shows that bias from both the models lies near zero and continues to exist the same throughout the time window and makes the datasets comparable with real-world data (fig. 41a). There is a considerable deviation in the CAMS outputs with respect to the station pixels during summer. The bias measured in the pixelwise analysis for Polyphemus has some good results in summer and varies a mild in winter. Though CAMS follows the actual pattern of the station observations, the spatial bias analysis shows that there are also some

significant deviations in CAMS for $PM_{2.5}$ with respect to real-world measurements (fig. 41b) (*see Appendix 4 for more $PM_{2.5}$ statistical spatial results*).



(a)

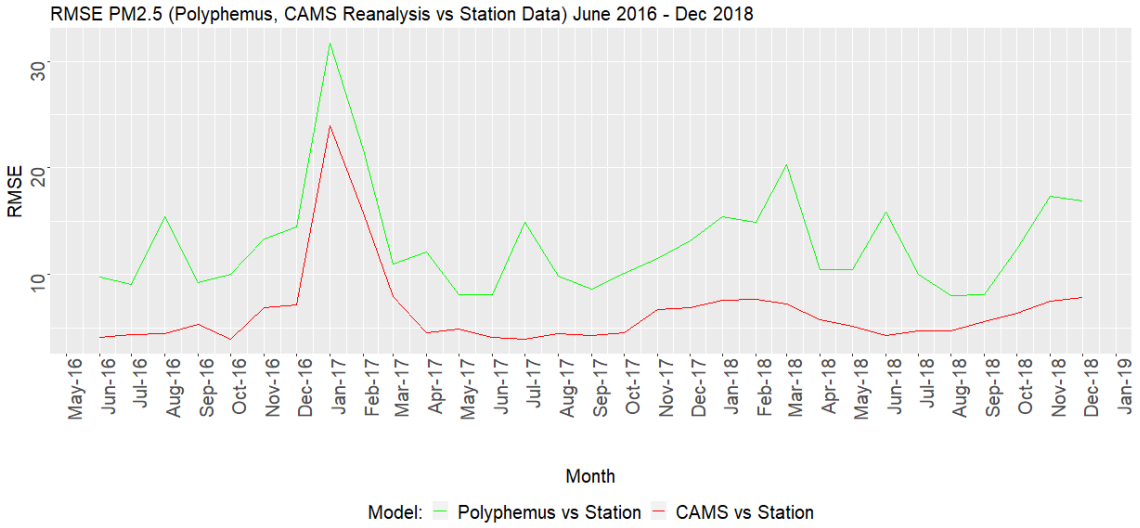


(b)

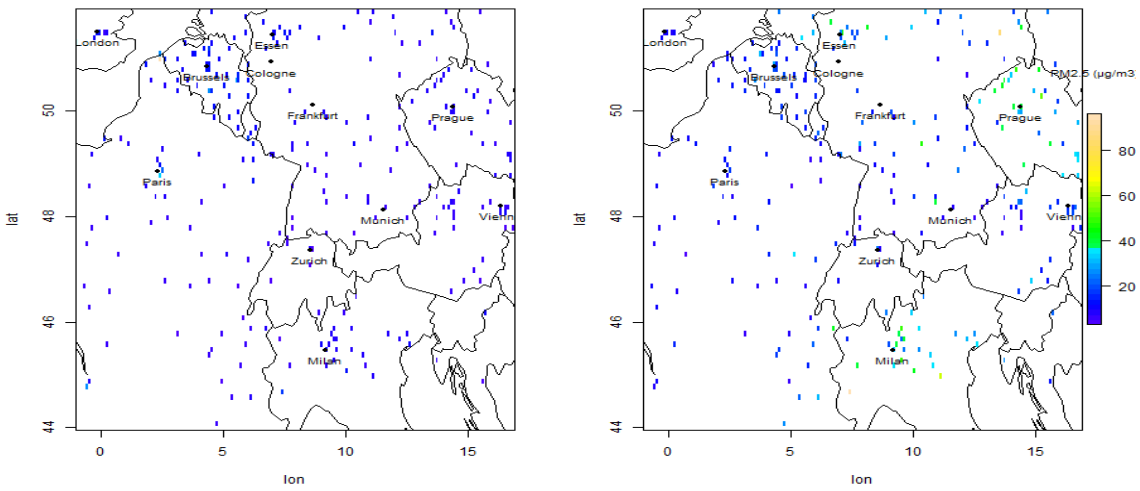
Figure 42: (a) Temporal $PM_{2.5}$ FGE model-model-station from June 2016 to Dec 2018. (b) Spatial $PM_{2.5}$ FGE model-model-station for Jan 2017 (left: Polyphemus FGE, right: CAMS FGE).

The FGE for $PM_{2.5}$ results that the FGE for both the models is less than one which makes the model's outputs more comparable with the real-world measurements. Like for other pollutants from previous results, the CAMS shows better performance for $PM_{2.5}$ compared to Polyphemus with respect to station data. There is no seasonal pattern followed and the distribution of errors from both the models are not continuous (fig. 42a). The spatial FGE analysis has varying results that there are no considerable deviations between the models for $PM_{2.5}$ in summer and mild variations in the urban regions of

Polyphemus during winter. The concentrations of Polyphemus for $PM_{2.5}$ in winter 2018 were quite different and more comparable to the station observations (fig. 42b).



(a)

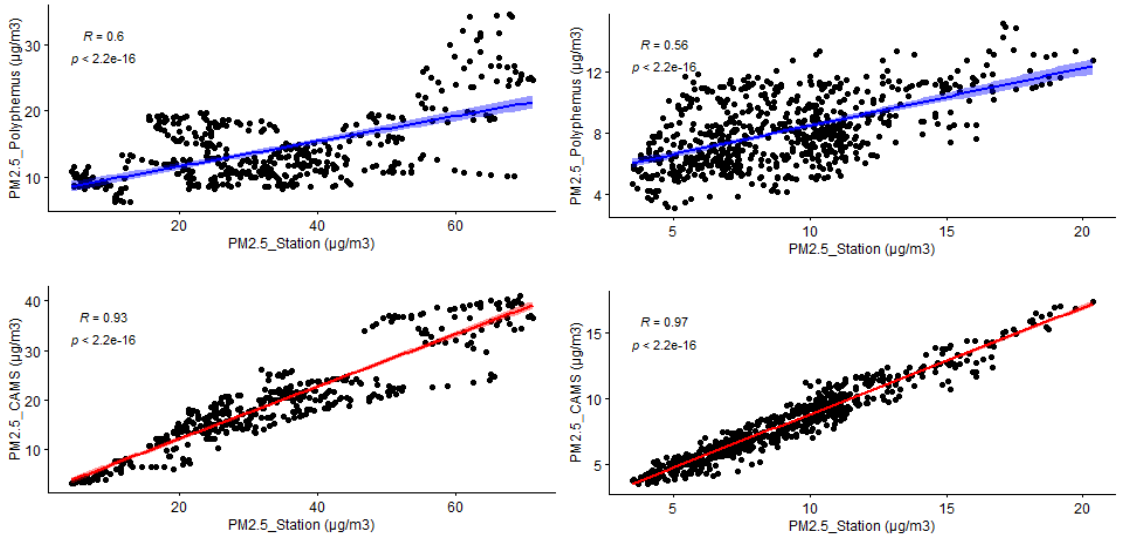


(b)

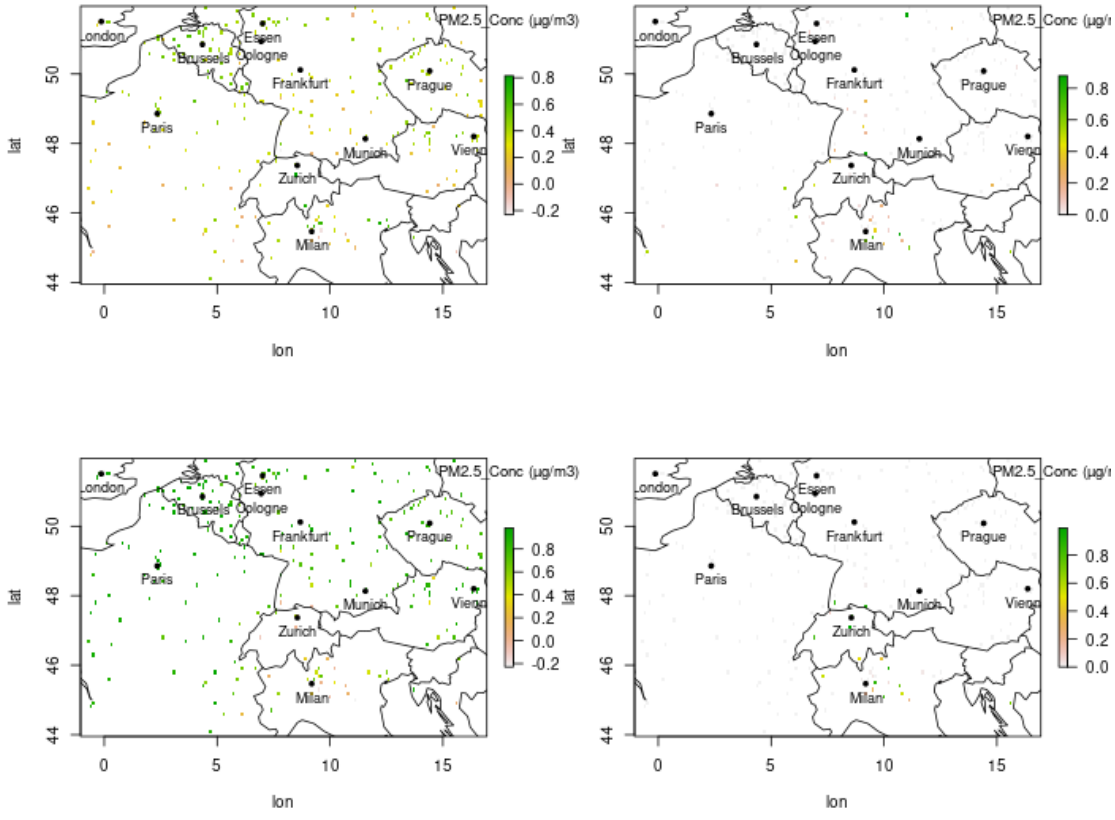
Figure 43: (a) Temporal $PM_{2.5}$ RMSE model-model-station from June 2016 to Dec 2018. (b) Spatial $PM_{2.5}$ RMSE model-model-station for July 2017 (left: Polyphemus RMSE, right: CAMS RMSE).

Also for RMSE, like other pollutants, the results are quite similar to FGE for $PM_{2.5}$. With mild changes in both FGE and RMSE, both convey the same results while comparing models concentrations. The extreme deviation in January 2017 is due to missing datasets. The RMSE in CAMS is almost continuous and seasonal. It is very close to 1 which makes the model more comparable with the real-world measurements. The RMSE for $PM_{2.5}$ from Polyphemus outputs are discrete and no pattern is followed. Though CAMS outperforms Polyphemus, the RMSE of Polyphemus is also relatable with the station observations (fig. 43a). There is a deviation found in the areas like Milan and Prague that affects the performance of the model the most. These deviations are observed in both the models and are a bit severe in CAMS. A mild deviation from CAMS

in these regions during summer was observed. In spatial RMSE, comparing every station pixels, Polyphemus performances are a bit ahead of CAMS under different geographical conditions (fig. 43b).



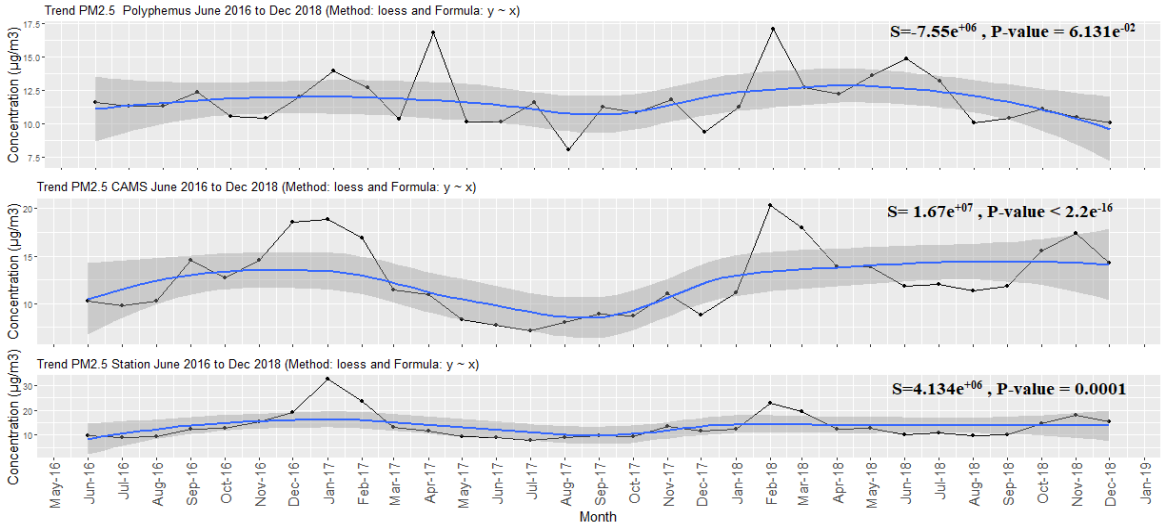
(a)



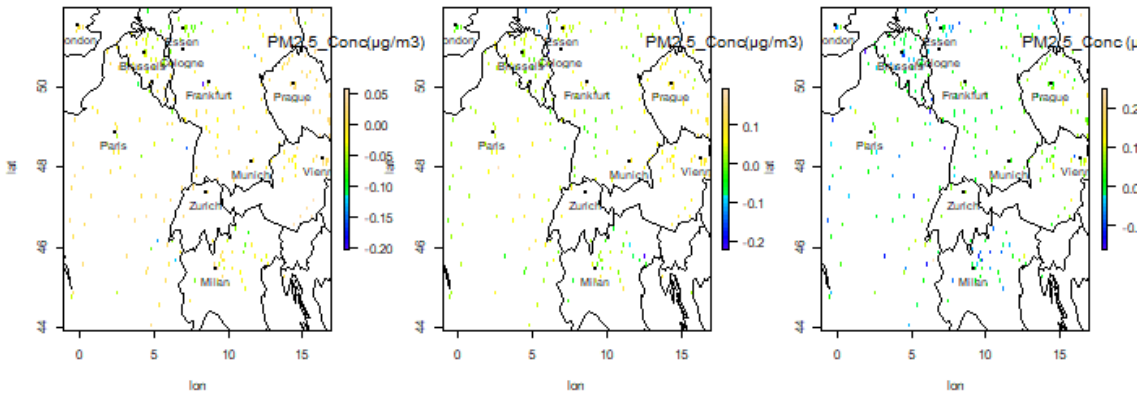
(b)

Figure 44: (a) $PM_{2.5}$ Temporal Correlation model-model-station on Jan 2017 (upper left: Polyphemus vs station, lower left: CAMS vs station) and Aug 20172018 (upper right: Polyphemus vs station, lower right: CAMS vs station). (b) $PM_{2.5}$ Spatial correlation model-model-station on Dec 2016 with P-value (upper left: Polyphemus vs station, upper right: its P-value, lower left: CAMS vs station, lower right: its P-value).

The spatial and temporal correlation analysis for $PM_{2.5}$ results on average of 0.94 and 0.5 correlation for CAMS and Polyphemus respectively. There are no seasonal deviations found in the correlation analysis that the correlations coefficients are discrete throughout the time window (fig. 44a). In Polyphemus, the constant deviations are found in the station pixels from some stations in southern France, western Germany, and Belgium (especially in summer). This relates to varying seasonal temporal trends for $PM_{2.5}$ in the area of interest. Both the models and the station datasets follow the same pattern throughout the period considered and a positive trend was observed in both temporal and spatially varying trend analysis for both the models and the station datasets. A strong positive trend was observed in the model results with mild deviations from the real-world observations (fig. 45) (*see appendix 5 Table A3 for Mann Kendall's trend statistics results for $PM_{2.5}$*).



(a)

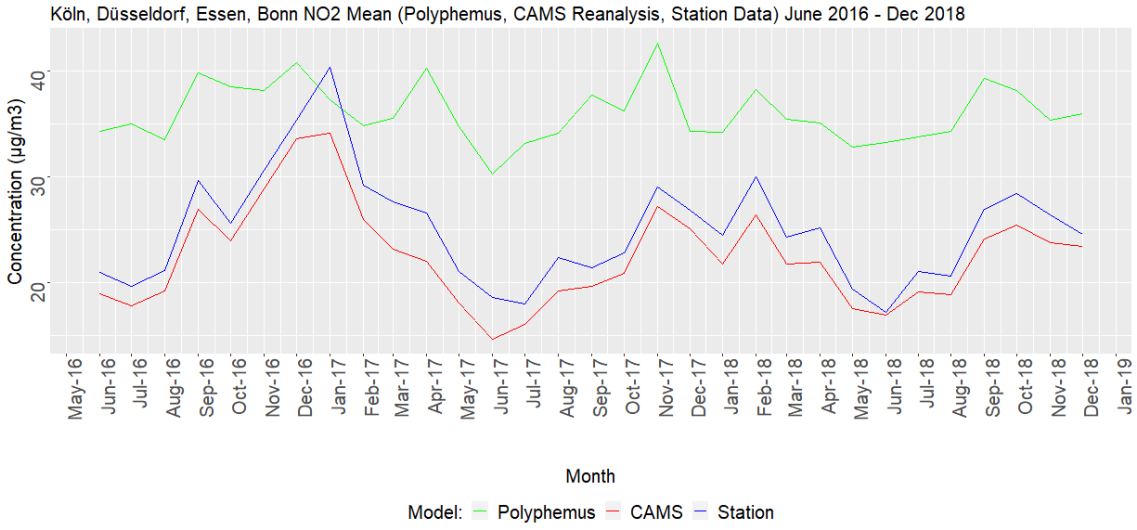


(b)

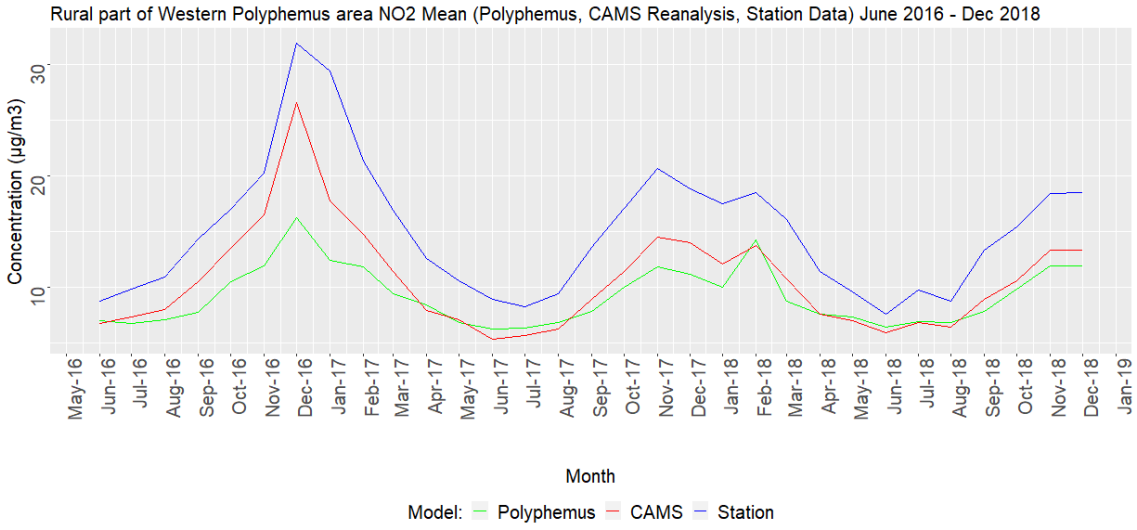
Figure 45: (a) $PM_{2.5}$ Temporal Trend model-model-station from Jun 2016 to Dec 2018. (b) $PM_{2.5}$ Spatially varying trend model-model-station from Jun 2016 to Dec 2018 (left: Polyphemus trend, middle: CAMS trend, right: station trend).

6.2 Performance of the models in urban and rural areas

This time-series analysis is to understand the performance and deviations of CAMS and Polyphemus in the urban and the rural regions. The main scope is to analyse the underestimations and overestimations in the model's performance for all four pollutants. For this analysis, some of the major cities in the area of interest like Paris, Köln, Düsseldorf, Essen, Bonn, Milan, Prague are considered for analysing the urban changes. The suburban and rural regions from the western (Southern France) and eastern (Austria and Czech Republic) Polyphemus regions were taken for analysing the rural changes. The datasets were formatted on monthly basis.



(a)

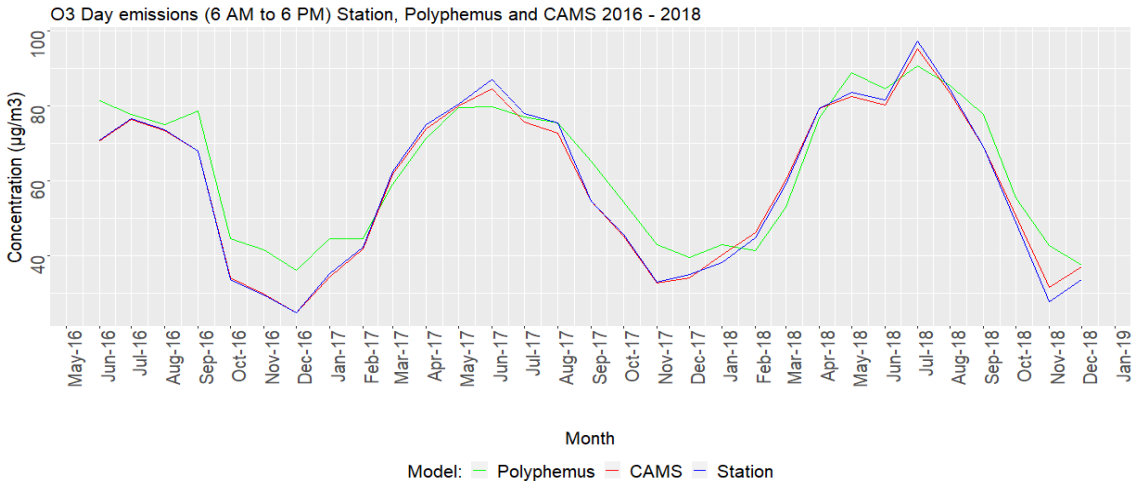


(b)

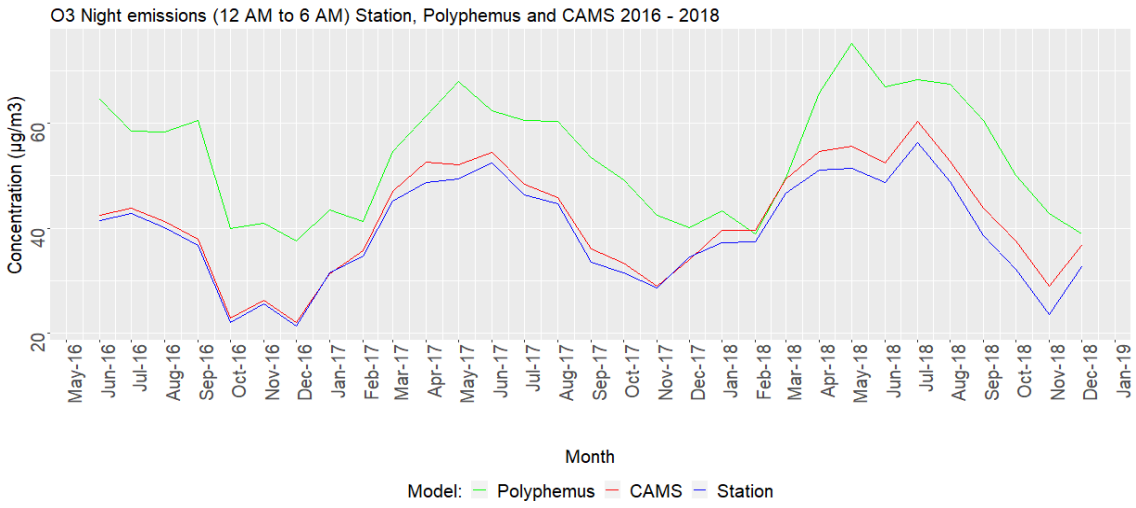
Figure 46: (a) NO_2 Time series of Polyphemus, CAMS, and Station data in Urban regions(Köln, Düsseldorf, Essen, and Bonn). (b) NO_2 Time series of Polyphemus, CAMS, and Station data in rural regions (eastern Polyphemus regions).

It was found that the Polyphemus model overestimates the NO_2, $PM_{2.5}$ and PM_{10} concentrations, especially in the urban areas (metro cities) over its domain, and underestimates the concentrations in very rural areas. The CAMS reanalysis datasets follow the pattern of station datasets for all the pollutants. There is an exceptional case for the city of Milan that the performance of Polyphemus for PM_{10} and $PM_{2.5}$ are quite different. The concentrations of PM_{10} and $PM_{2.5}$ from both models are comparable with the station observations. Strong deviations in Polyphemus outputs were observed for the pollutants like NO_2 and O_3. For Paris and Cities in western Germany, the deviations from all the pollutants in Polyphemus with respect to station datasets are highly noticeable (figure 46 a and b).

6.3 Performance of the models in day and night times



(a)



(b)

Figure 47: (a) O_3 Time series of Polyphemus, CAMS, and Station data during the day (b) O_3 Time series of Polyphemus, CAMS, and Station data during night.

As a part of the overestimation and underestimation analysis, the performance of the Polyphemus and CAMS during day and night times was analysed. The daytime considered is from 6 AM to 6 PM and the night time is from 12 AM to 6 AM. The complete area from Polyphemus domain two was taken for this analysis and the datasets were formatted on monthly basis.

It was analysed that there were no considerable deviations in the models' outputs during the day and night times. Other than O_3, all the other pollutants have similar performances during the day and night times and follow the same pattern. There is a strong deviation in O_3 that the Polyphemus datasets show a strong and very significant overestimation during the nighttime with respect to the station observations (fig. 47b).

6.4 Predictive analysis

The potential drivers of the Polyphemus model for the given parameters (figure 10 & table 2) for the regions indicated in figure 11 were analysed. The analysis was performed for all 4 pollutants in urban (Paris) and rural (Southern France) areas. The datasets are formatted on an hourly basis for all the parameters considered. Based on the Gini impurity index (x-axis in figure 48 and 49), Parameters like the boundary layer height, wind, and season vary the most in the urban and rural regions and the parameters like surface temperature and boundary layer height play a major role in both the regions for all the pollutants.

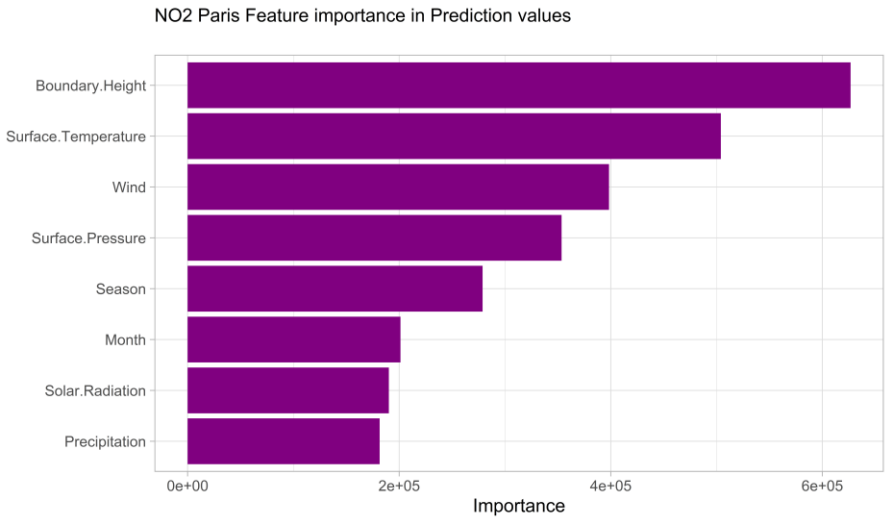


Figure 48: NO_2 Potential drivers from Polyphemus in the urban region (Paris)

For NO_2, boundary layer height and surface temperature parameters in urban areas and surface temperature and season parameters in rural areas are considered as the most important parameters used by the RF algorithm to predict the in situ stations

observations. Solar radiation and precipitation are the least utilized parameters for the predictions.

For O_3, there are no considerable differences in the influence of parameters in the urban and rural regions. The parameters like surface temperature and boundary layer height are the most important parameters used for prediction Surface pressure, month, and precipitation are the least utilized parameters to predict O_3 concentrations.

For PM_{10}, the parameters that varies a lot in urban and rural regions are boundary layer height, wind, season, and month. In the urban region, boundary layer height and surface temperature play a major role. Whereas in rural areas, the boundary layer height was not highly utilized for PM_{10} concentration prediction by RF algorithm. The parameters like surface temperature, month, and season play a major role in rural areas.

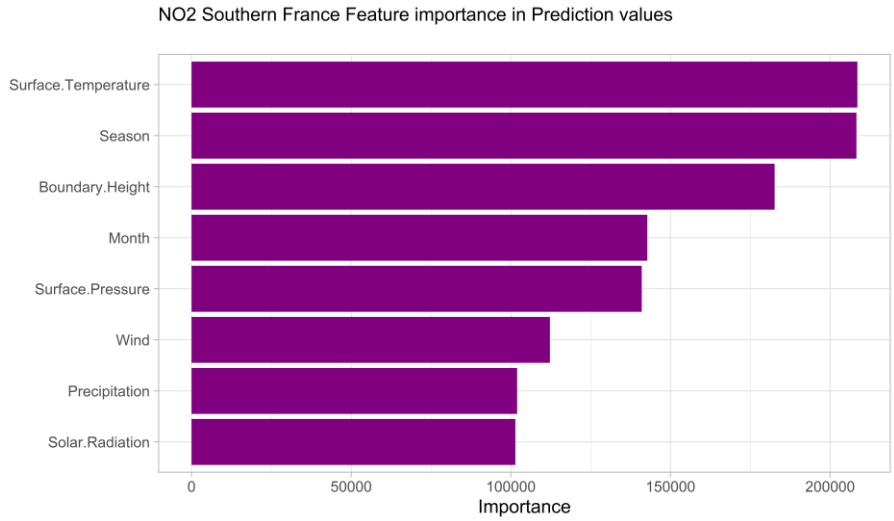


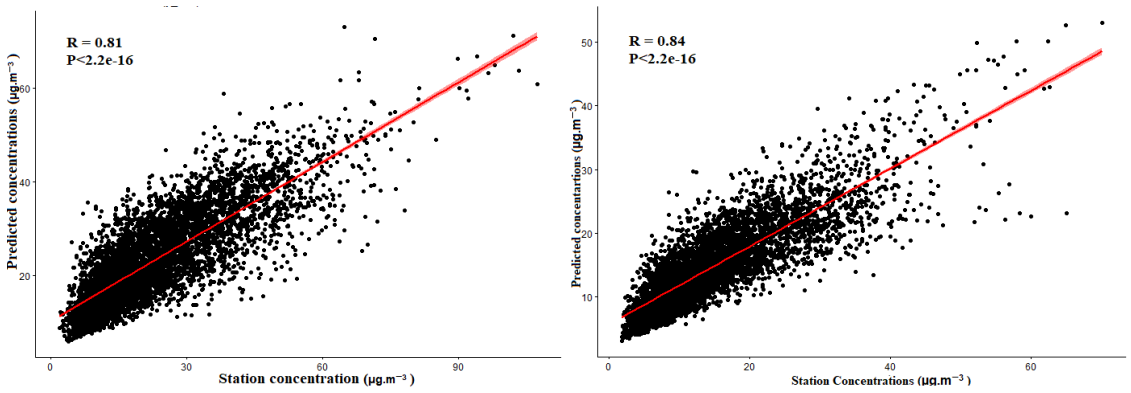
Figure 49: NO_2 Potential drivers from Polyphemus in the rural region (southern France).

There are considerable similarities between PM_{10} and $PM_{2.5}$ in the case of potential drivers for Polyphemus in urban and rural areas. Some of the varying parameters are boundary layer height, precipitation, season, and month. Like other pollutants, the surface temperature is the most utilized parameter for $PM_{2.5}$ in both regions followed by surface pressure. Solar radiation is the least utilized parameter in both urban and rural regions followed by season parameter in urban and boundary layer height parameter in rural areas. (See Appendix 6 for O_3, PM_{10}, and $PM_{2.5}$ potential drivers in Polyphemus results).

6.4.1 Accuracy assessment

The accuracies of the predicted concentrations are assessed with respect to station concentrations using the proposed statistical indicators. Comparing both urban and rural

areas for all the four pollutants, the predicted concentrations are better correlated to station observations than the actual Polyphemus concentrations (compare figure 50 with figures 19 for NO_2). The statistical indicators that accessed the accuracies of the predicted concentrations indicate that rural areas considered have a better prediction for all the pollutants. These high correlated results are due to the station concentrations as target parameters. The same case as CAMS reanalysis datasets assimilated with the station datasets.



(a) *(b)*

Figure 50: (a) Correlation between Station and the predicted NO_2 concentrations in the urban region – Paris. (b) Correlation between Station and the predicted concentrations NO_2 in the rural region – Southern France.

Urban: Paris Metropolitan

	NO_2	O_3	$PM_{2.5}$	PM_{10}
%Var explained	65.29	82.42	69.34	72.82
Correlation	0.80	0.90	0.85	0.86
RMSE	8.32	11.66	5.67	4.25
Mean Bias	-0.10	-0.14	0.09	-0.00042
FGE	0.28	0.23	0.24	0.30
MAE	6.15	8.90	4.12	2.88

Table 3: Statistical validation of the predicted concentrations for Paris and it's suburban with respect to station observations.

Rural: Southern France

	NO_2	O_3	$PM_{2.5}$	PM_{10}
%Var explained	70.27	86.94	75.9	79.91
Correlation	0.84	0.93	0.87	0.90
RMSE	5.04	9.61	3.72	2.92
Mean Bias	0.04	-0.04	-0.02	-0.011
FGE	0.25	0.15	0.18	0.21
MAE	3.61	7.28	2.70	1.99

Table 4: Statistical validation of the predicted concentrations for Southern France and its suburbs with respect to station observations.

High correlation significance was observed with respect to station concentrations in both the urban and rural regions. Also, other statistical indicators like FGE, RMSE, mean bias, mean absolute error (MAE) show that the predicted concentrations from all the pollutants are well comparable to station observations. (See tables 3 and 4 for accuracy assessment statistics results).

7 Conclusion and future works

The main objective of this research is to compare the differences between the outputs of the Copernicus Atmosphere Monitoring Service (CAMS) – Europe Air Quality Reanalysis data and of the Polyphemus/DLR with respect to EEA air quality station observations using statistical indicators. The spatial and temporal visualizations of all the statistical analyses give insight into model adequacy and uncertainty with respect to the station datasets considered. It was clear from the beginning that CAMS reanalysis compares better to the EEA station measurements for all the pollutants considered. This is mainly because the CAMS reanalysis are the datasets with data assimilation of station data. Identifying the overestimations/ underestimations of the concentration values involves time series analysis, mean bias, performance in the urban and rural areas, for day and night times. It was found that the concentrations of Polyphemus tend to be are overestimated in cities like Paris, Köln etc., especially for NO_2, $PM_{2.5}$ & PM_{10}, and sightly underestimated in rural areas with respect to EEA Station concentration. For O_3, there are no considerable differences in the model's performances in urban and rural areas. The time series to analyse the day and night differences in concentrations showed a mild overestimation of O_3 in Polyphemus during the nighttime. There are no significant differences in the model's outputs for other pollutants between the day and night times. Analysing the deviations in the models with respect to the station data involves also statistical analyses like RMSE, FGE, correlation coefficient. From both the spatial and temporal statistical analysis, the CAMS reanalysis performs better compared to the EEA station observations. Though CAMS outperforms Polyphemus, the predictions of PM_{10} and $PM_{2.5}$ from both models are quite good compared with station observations showing no significant deviations between the models for PM_{10} and $PM_{2.5}$.

The temporal and spatially varying trend analysis for all the pollutants strong seasonal variations. There is a significant positive long-term trend in both model results over the time window analyzed. Pixelwise, there is a considerably varying trend observed between the models and the station data based on locations. The spatially varying positive trends in between the datasets were analysed for NO_2 and for other pollutants. There are relatable positive trends in cities and industrial areas, neutral and negative trends in rural areas for models, and station data.

Identifying the potential drivers of the Polyphemus model involves Random Forest Regression ML algorithm and proposed statistical analysis to access the accuracy of the predicted concentrations. Some of the potential drivers considered for Polyphemus that cause deviations in urban and rural areas are:

- For NO_2 & PM_{10}, parameters like boundary layer height, seasons, and wind.

- For O_3, parameters like surface pressure, seasons, and month.
- For $PM_{2.5}$, parameters like precipitation, boundary layer height, seasons, and month.

The parameters like surface temperature and season play a major role in both the urban and rural areas for all pollutants. The interesting fact here is the strong influence of the boundary layer height in urban regions for all pollutants. Whereas, it's completely vice versa in rural areas for PM_{10} and $PM_{2.5}$. the influence of parameters like precipitation and wind in the model is varying. As can be expected, the RF algorithm performs better than the model for all the pollutants in both the urban and rural areas with an average correlation of 0.8 for all the pollutants. All statistical indicators accessed compare well to station concentrations.

Some of the suggested work for the future are:

- Using satellite data (Sentinel-5P) along with station measurements as truth data for model comparison.
- Verify statistical variations between models, station, and satellite datasets.
- The feature analysis should be extended with more detailed data in the future.

8 References

A. Benedetti et al.: The value of satellite observations for Asian dust analysis and short-range prediction of Asian dust, 2019, Published by Copernicus Publications on behalf of the European Geosciences Union. https://doi.org/10.5194/acp-19-987-2019

Air pollution fact sheet 2013 Germany, European Environment Agency, 2013. https://www.eea.europa.eu/themes/air/air-pollution-country-fact-sheets/germany-air-pollutant-emissions-country-factsheet

Air quality in Europe — 2020 report, European Environment Agency, 2020. doi:10.2800/786656, https://op.europa.eu/en/publication-detail/-/publication/447035cd-344e-11eb-b27b-01aa75ed71a1

B. Mohammed Hashim & M. Abdullah Sultan, Using remote sensing data and GIS to evaluate air pollution and their relationship with land cover and land use in Baghdad City, 2010, Iranian Journal of Earth Sciences 2 (2010) / 120-124.

Baklanov, A., et. al: Online coupled regional meteorology chemistry models in Europe: current status and prospects, Atmos. Chem. Phys., 14, 317–398, doi:10.5194/acp-14-317-2014, 2014.

Bauwens et al. (2020). Impact of coronavirus outbreak on NO2 pollution assessed using TROPOMI and OMI observations. Geophysical Research Letters. DOI: https://doi.org/10.1029/2020GL087978

Blower, J. and Clegg, A. (2011) Fast regridding of large, complex geospatial datasets. In: Com.Geo 2011: The 2nd International Conference on Computing for Geospatial Research & Applications, 23-25 May 2011, Washington D.C., pp. 1-6. (10.1145/1999320.1999350) Available at http://centaur.reading.ac.uk/19928/

Brigitte Colin and Kerrie Mengersen, Estimating Spatial and Temporal Trends in Environmental Indices Based on Satellite Data: A Two-Step Approach, Received: 9 November 2018; Accepted: 11 January 2019; Published: 17 January 2019, https://doi.org/10.3390/s19020361

CAMS Regional: European air quality analysis and forecast data documentation, 2022, https://confluence.ecmwf.int/display/CKB/CAMS+Regional%3A+European+air+quality+analysis+and+forecast+data+documentation

CAMS Verification plots: documentation, 2020, http://macc-raq-int.meteo.fr/doc/USER_GUIDE_VERIFICATION_STATISTICS.pdf

CAMS50_2018SC2_D2.0.2-U2_Models_documentation_202003_v2, Regional Production, Updated documentation covering all Regional operational systems and the ENSEMBLE , Issued by: METEO-FRANCE / G. Collin, Date: 05/03/2020.

Chai, T., et. al. Evaluation of the United States National Air Quality Forecast Capability experimental real-time predictions in 2010 using Air Quality System ozone and NO2 measurements, Geosci. Model Dev., 6, 1831– 1850, doi:10.5194/gmd-6-1831-2013, 2013.

Chan, K.L.; Khorsandi, E.; Liu, S.; Baier, F.; Valks, P. Estimation of Surface NO2 Concentrations over Germany from TROPOMI Satellite Observations Using a Machine Learning Method. Remote Sens. 2021, 13, 969. https://doi.org/10.3390/ rs13050969

Delle Monache, L., Deng, X., Zhou, Y., and Stull, R.: Ozone ensemble forecasts: 1. A new ensemble design, J. Geophys. Res., 111, D05307, doi:10.1029/2005JD006310, 2006.

Derwent, et. al. Evaluating the Performance of Air Quality Models, 2010, http://www.defra.gov.uk/

E. Khorsandi, F. Baier, T. Erbertseder, M. Bittner, "Air quality monitoring and simulation on urban scale over Munich," Proc. SPIE 10793, Remote Sensing Technologies and Applications in Urban Environments III, 1079303 (26 October 2018); doi: 10.1117/12.2503969

E. Solazzo, S. Galmarini, Comparing apples with apples: Using spatially distributed time series of monitoring data for model evaluation, 2015, http://dx.doi.org/10.1016/j.atmosenv.2015.04.037

ECMWF CAMS reanalysis documentation:
https://confluence.ecmwf.int/display/CKB/CAMS%3A+Reanalysis+data+documentation

Evensen, G.: Sequential data assimilation with a nonlinear quasigeostrophic model using Monte Carlo methods to forecast error statistics, J. Geophys. Res., 99, 10 143–10 162, 1994.

Gaüzère, P.; Jiguet, F.; Devictor, V. Rapid adjustment of bird community compositions to local climatic variations and its functional consequences. Glob. Chang. Biol. 2015, 21, 3367–3378. doi:10.1111/gcb.12917.

Haofei Yu, et. al. Cross-comparison and evaluation of air pollution field estimation methods, 2018, https://doi.org/10.1016/j.atmosenv.2018.01.045.

J. Jake Nichol et. al. Machine learning feature analysis illuminates disparity between E3SM climate models and observed climate change, 2020. Journal of Computational and Applied Mathematics, https://doi.org/10.1016/j.cam.2021.113451

Jacob, D. J. and Winner, D. A.: Effect of climate change on air quality, Atmos. Environ., 43, 51–63, https://doi.org/10.1016/j.atmosenv.2008.09.051, 2009.

L. Breiman, Random forests, Mach. Learn. 45 (1) (2001) 5–32, http://dx.doi.org/10.1023/A:1010933404324.

Lee et al. Standard deviation and standard error of the mean, Copyright ⓒ the Korean Society of Anesthesiologists, 2015. Online access in http://ekja.org

Liu et al., An improved tropospheric NO2 column retrieval algorithm for TROPOMI over Europe, 2021, https://doi.org/10.5194/amt-2021-39

Marilena Kampa & Elias Castanas (2008). Human health effects of air pollution, Environmental Pollution. Vol. 151, pp. 362-367.

Mathiesen, P., Kleissl, J., 2011. Evaluation of numerical weather prediction for intra-day solar forecasting in the continental United States. Sol. Energy 85, 967–977. https://doi.org/10.1016/j.solener.2011.02.013

Meng Lu et. al, Multidimensional Arrays for Analysing Geoscientific Data, 2018, ISPRS Int. J. Geo-Inf. 2018, 7, 313; doi:10.3390/ijgi7080313 www.mdpi.com/journal/ijgi.

Menut, L. and Bessagnet, B.: Atmospheric composition forecasting in Europe, Ann. Geophys., 28, 61–74, doi:10.5194/angeo-28-61- 2010, 2010.

Panoply A Tool for Visualizing NetCDF-Formatted Model Output, v-3.1.1, by Robert Schmunk, NASA Author: A.Seidenglanz, Uni Bremen. 2012.

Panoply NetCDF Visualization Software v. 1.5.1, by Robert Schmunk http://www.giss.nasa.gov/tools/panoply/

Patrick Schober et. al., Correlation Coefficients: Appropriate Use and Interpretation, Article in Anesthesia and Analgesia, February 2018, DOI: 10.1213/ANE.0000000000002864.

Qingkun Yu1 , Xiaoning Guan2, Regridding and data interpolation of projection domain and Radon domain for super-resolution tomograpic reconstruction, 978-1-4673-5887-3/13/$31.00 ©2013 IEEE.

R. J. Stone, Improved statistical procedure for the evaluation of solar radiation estimation models, Solar Energy Vol. 5 I, No. 4, pp. 289-291, 1993.

R.L.R. Salcedo et. al. Time series analysis of air pollution data, 1998, Atmospheric environment 33 (1999) 2361 – 2372.

S. Nembrini, I.R. Ko, M.N. Wright, C. Lu, The revival of the Gini importance? 34 (May) (2018) 3711–3718, http://dx.doi.org/10.1093/ bioinformatics/bty373.

Savage, N. H.,et.al. Air quality modelling using the Met Office Unified Model (AQUM OS24-26): model description and initial evaluation, Geosci. Model Dev., 6, 353–372, doi:10.5194/gmd-6-353-2013, 2013.

Smyth, S., Yin, D., Roth, H., Jiang, W., Moran, M. D., and Crevier, L. P.: The impact of GEM and MM5 meteorology on CMAQ air quality modeling results in eastern Canada and the northeastern United States, J. Appl. Meteorol., 45, 1525e1541, https://doi.org/10.1175/JAM2420.1, 2006.

Song Liu, et. al. An improved tropospheric NO2 column retrieval algorithm for TROPOMI over Europe. 2021. https://doi.org/10.5194/amt-2021-39

Stephen D. Superczynski & Sundar A. Christopher (2011). Exploring Land Use and Land Cover Effects on Air Quality in Central Alabama Using GIS and Remote Sensing. Remote Sens. 2011, 3, 2552-2567; doi:10.3390/rs3122552.

T. Chai and R. R. Draxler, Root mean square error (RMSE) or mean absolute error (MAE)? – Arguments against avoiding RMSE in the literature, 2014. doi:10.5194/gmd-7-1247-2014

Tom Grylls, et. al, Evaluation of an operational air quality model using large-eddy simulation, 2019, https://doi.org/10.1016/j.aeaoa.2019.100041

V. Marécal, et. al, “A regional air quality forecasting system over Europe: the MACC-II daily ensemble production”, Geosci. Model Dev. Discuss., 8, 2739–2806, 2015: doi:10.5194/gmdd-8-2739-2015

Vautard, R., et. al. Evaluation of the meteorological forcing used for the Air Quality Model Evaluation International Initiative (AQMEII) air quality simulations, Atmos. Environ., 53, 15–37, 2012.

Vivien Mallet, D. Quélo, B. Sportisse, M. Ahmed de Biasi, I. Korsakissok, et al.. Technical Note: The air quality modeling system Polyphemus. Atmospheric Chemistry and Physics, European Geosciences Union, 2007, 7 (20), pp.5487. ff10.5194/acp-7-5479-2007ff. ffhal-00328547

Wagner, A, et al. 2021. Comprehensive evaluation of the Copernicus Atmosphere Monitoring Service (CAMS) reanalysis against independent observations. Elem Sci Anth, 9: 1. DOI: https://doi.org/10.1525/elementa.2020.00171

WHO global air quality guidelines. Particulate matter (PM2.5 and PM10), ozone, nitrogen dioxide, sulfur dioxide and carbon monoxide. ISBN 978-92-4-003422-8 (electronic version) ISBN 978-92-4-003421-1 (print version), World Health Organization 2021. https://apps.who.int/iris/handle/10665/345329

Xiang Wan et.al., Estimating the sample mean and standard deviation from the sample size, median, range and/or interquartile range, BMC Medical Research Methodology 2014, 14:135 http://www.biomedcentral.com/1471-2288/14/135

Yu, S., Eder, B., Dennis, R., Chu, S.-H., and Schwartz, S. E.: New unbiased symmetric metrics for evaluation of air quality models, Atmos. Sci. Lett., 7, 26–34, 2006.

Zhang et. al., Spatial and temporal analysis of water quality trends in the Min River Basin, 2012, https://ieeexplore.ieee.org/document/6349581

Živadinovi´c, I.; Ilijevi´c, K.; Gržeti´c, I.; Popovi´c, A. Long-term changes in the eco-chemical status of the Danube River in the region of Serbia. J. Serb. Chem. Soc. 2010, 75, 1125–1148. doi:10.2298/JSC091102075Z.

Appendix

Due to page constraints of the thesis, the additional maps and plots of all the statistical analyses and the R scripts performing all the statistical analyses are uploaded in the Github repository named **Master-Thesis---Statistical-Analysis**

(https://github.com/SathishVaithiyanadhan/Master-Thesis---Statistical-Analysis.git)

Appendix 1 – NO_2 Spatial results.

More NO_2 spatial results for all the statistical indicators for model-model-station comparison are presented here.

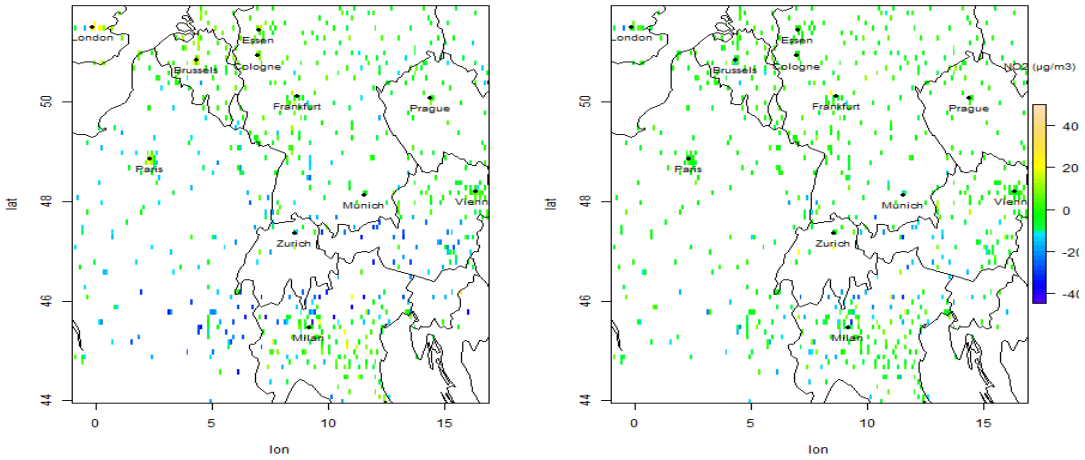


Figure A1: NO_2 Spatial Mean bias in winter Dec 2016 (left: Polyphemus mean bias, right: CAMS mean bias).

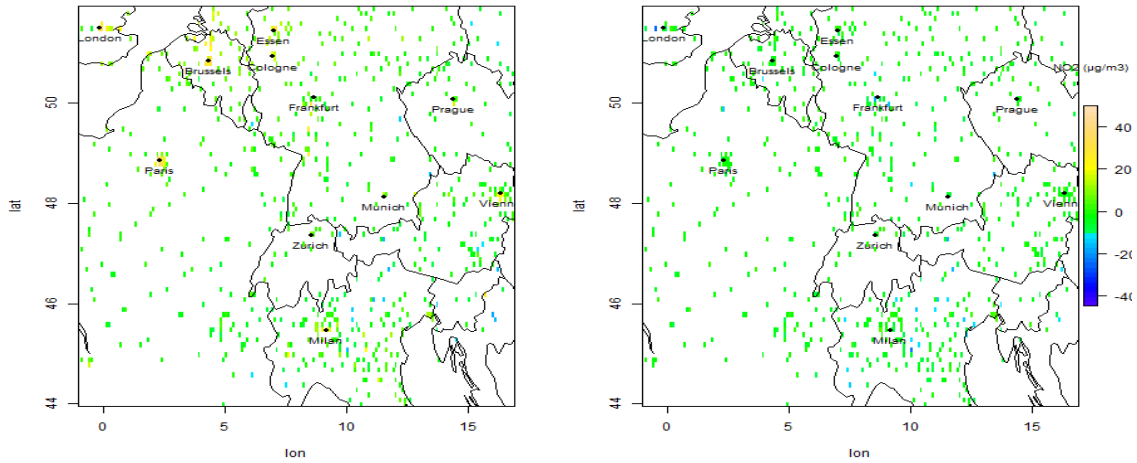


Figure A2: NO_2 Spatial mean bias in summer July 2017 (left: Polyphemus mean bias, right: CAMS mean bias).

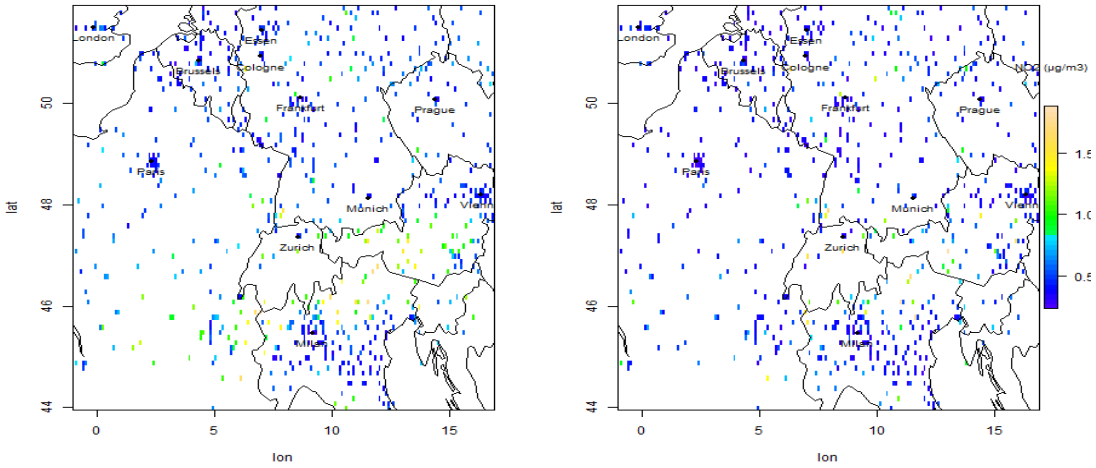


Figure A3: NO_2 Spatial FGE in Winter Dec 2016 (left: Polyphemus FGE, right: CAMS FGE).

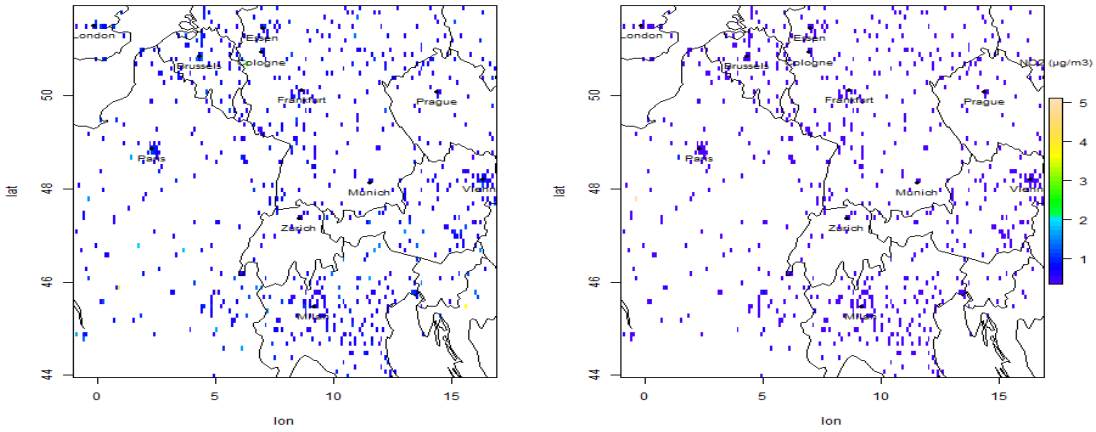


Figure A4: NO_2 Spatial FGE in summer July 2017 (left: Polyphemus FGE, right: CAMS FGE).

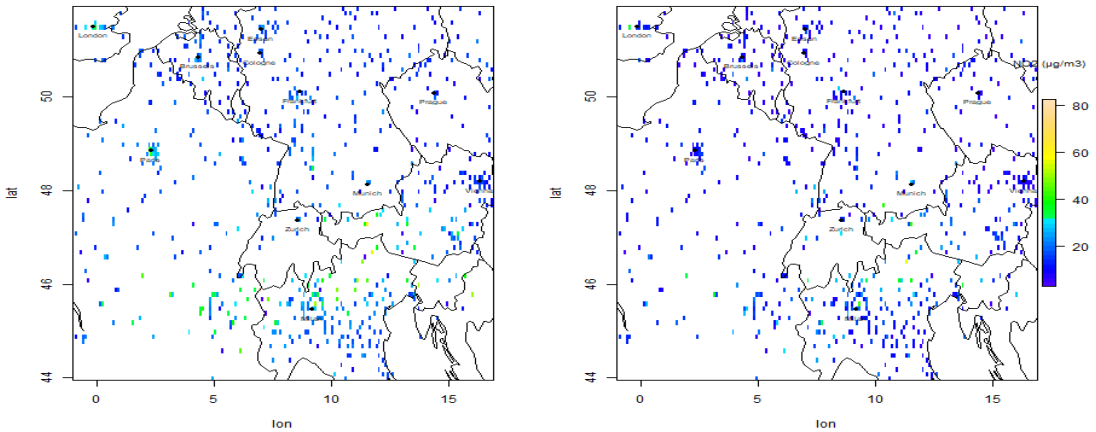


Figure A5: NO_2 Spatial RMSE in Winter Dec 2016 (left: Polyphemus RMSE, right: CAMS RMSE).

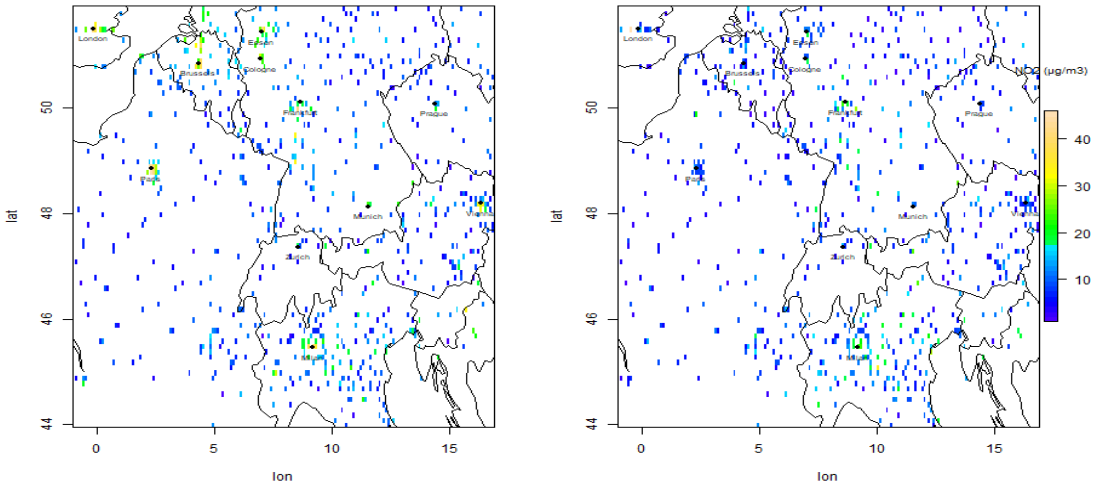


Figure A6: NO_2 Spatial RMSE in summer July 2017 (left: Polyphemus RMSE, right: CAMS RMSE).

Appendix 2 – O_3 Spatial results.

More O_3 spatial results for all the statistical indicators for model-model-station comparison are presented here.

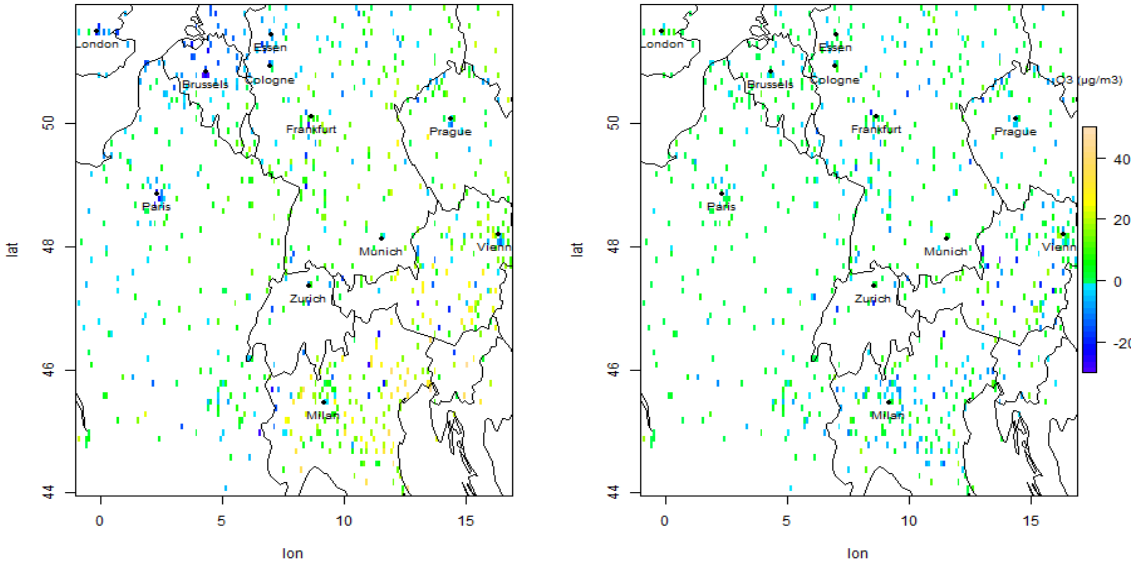


Figure A7: O_3 Spatial mean bias in summer July 2016 (left: Polyphemus mean bias, right: CAMS mean bias).

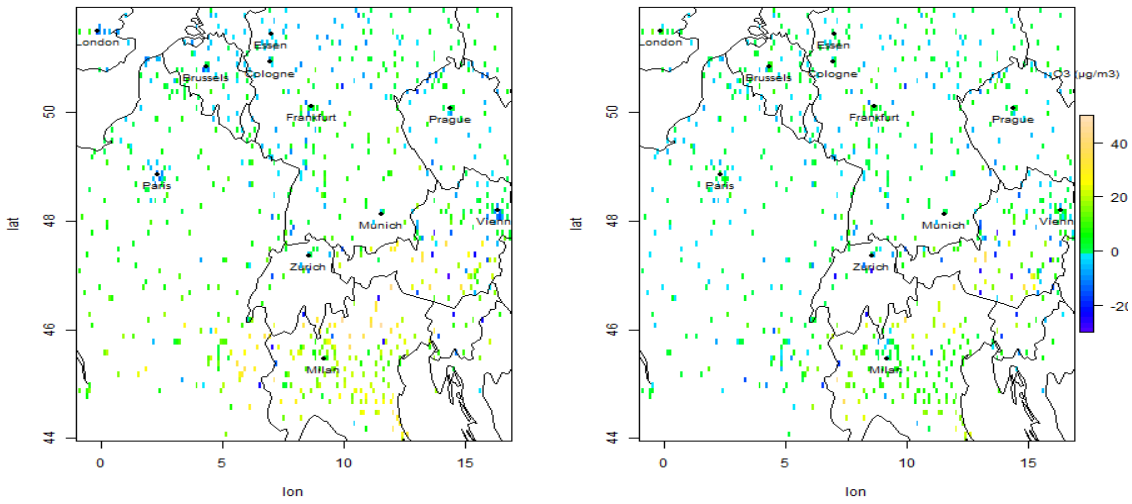


Figure A8: O_3 Spatial Mean bias in winter Jan 2018 (left: Polyphemus mean bias, right: CAMS mean bias).

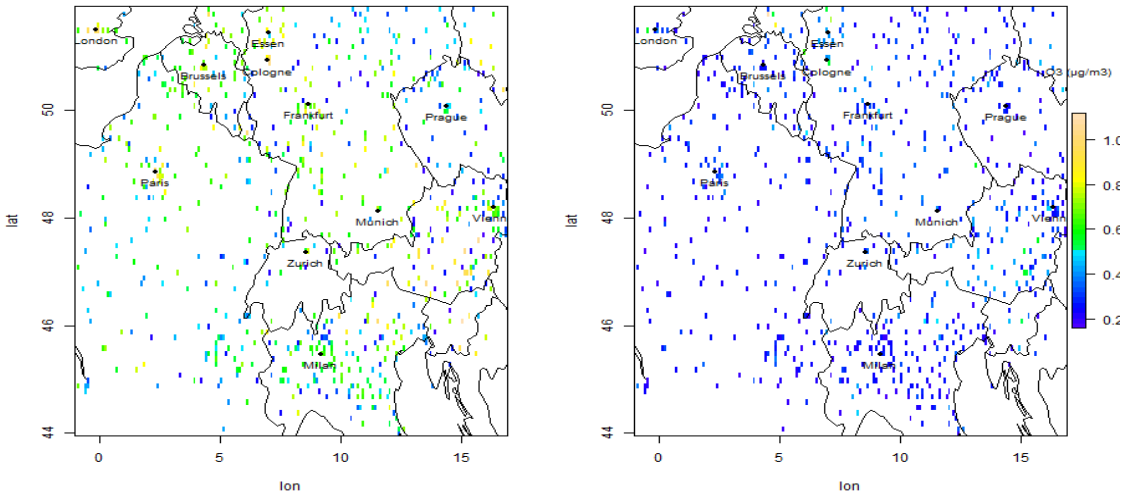


Figure A9: O_3 Spatial FGE in autumn Sept 2016 (left: Polyphemus FGE, right: CAMS FGE).

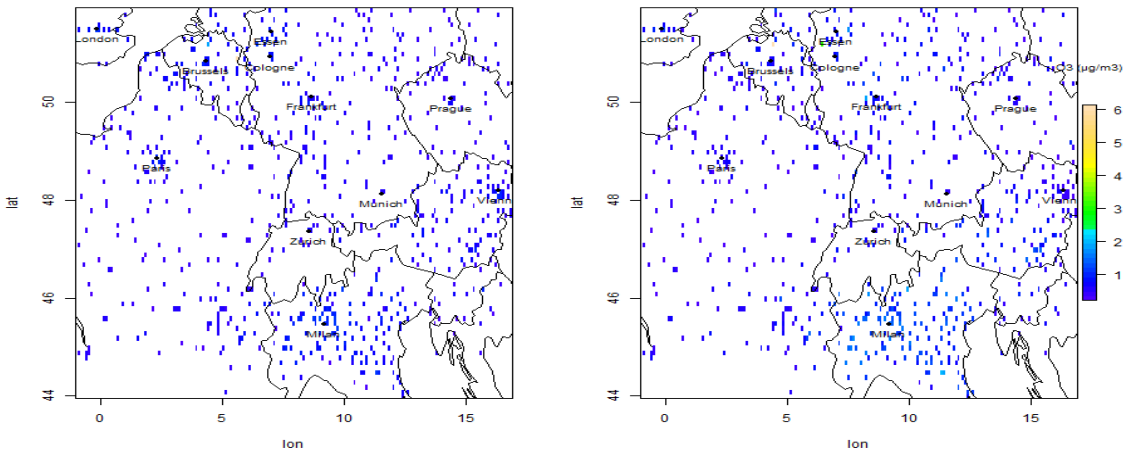


Figure A10: O_3 Spatial FGE in winter Feb 2018 (left: Polyphemus FGE, right: CAMS FGE).

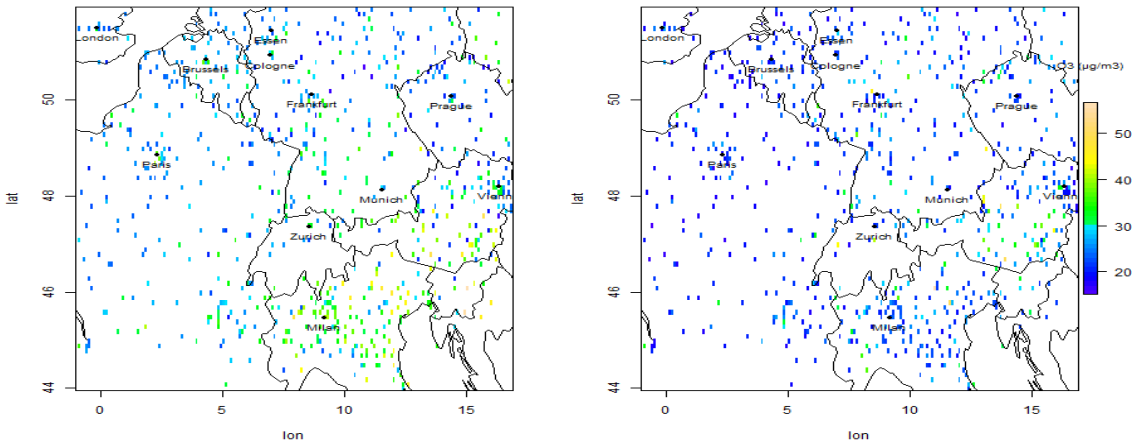


Figure A11: O_3 Spatial RMSE in summer August 2016 (left: Polyphemus RMSE, right: CAMS RMSE).

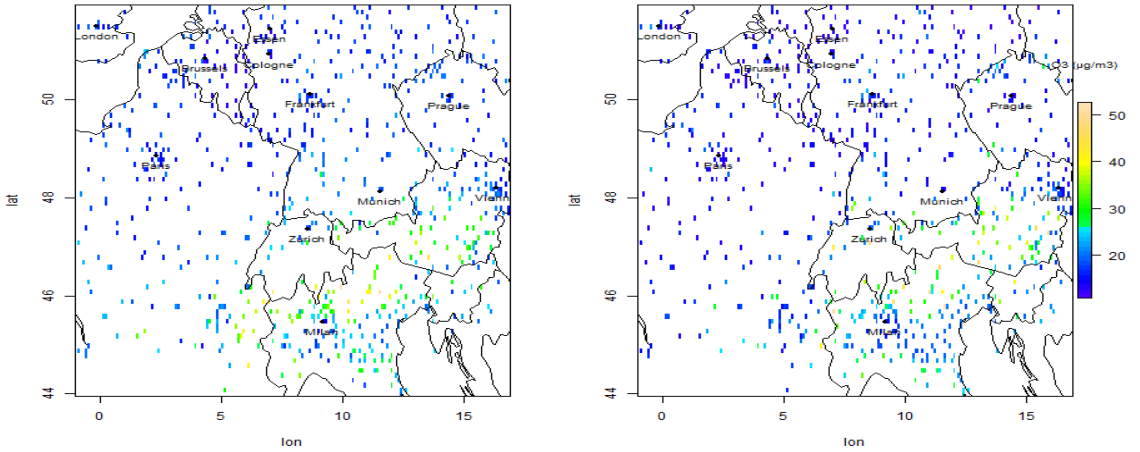


Figure A12: O_3 Spatial RMSE in winter Dec 2017 (left: Polyphemus RMSE, right: CAMS RMSE).

Appendix 3 – PM_{10} Spatial results.

More PM_{10} spatial results for all the statistical indicators for model-model-station comparison are presented here.

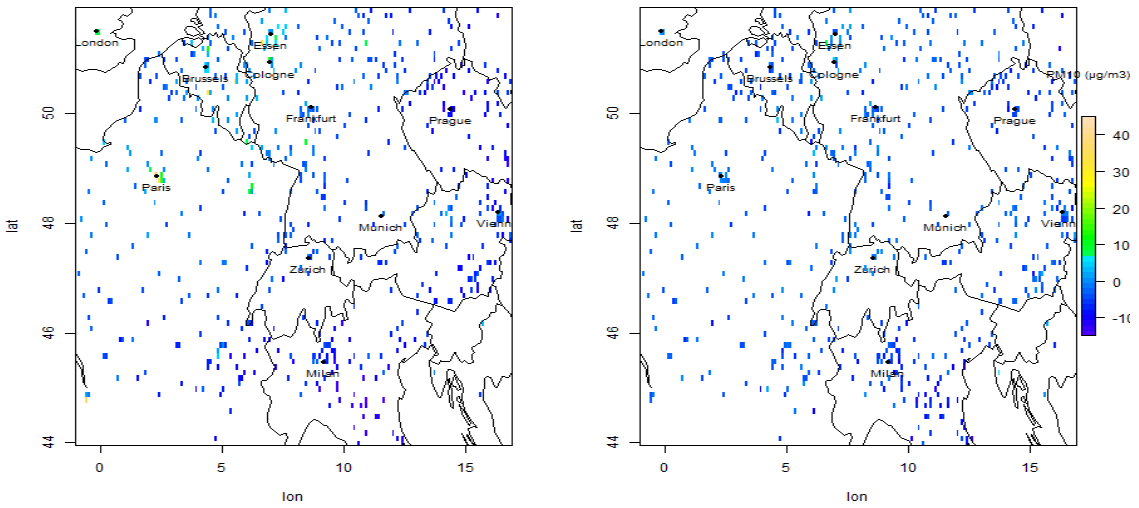


Figure A13: PM_{10} Spatial mean bias in summer Aug 2017 (left: Polyphemus mean bias, right: CAMS mean bias).

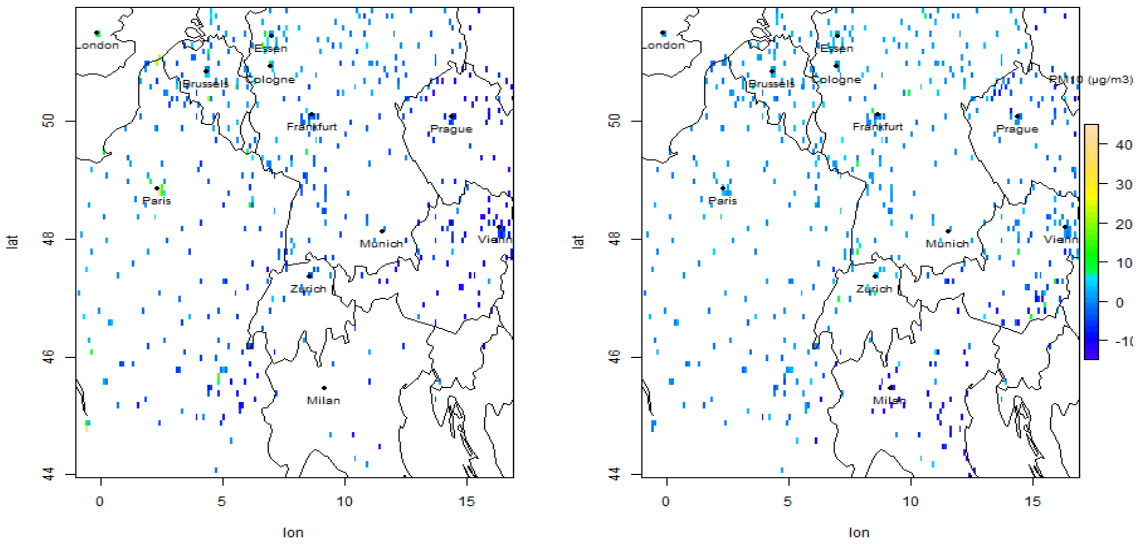


Figure A14: PM_{10} Spatial mean bias in winter Dec 2018 (left: Polyphemus mean bias, right: CAMS mean bias).

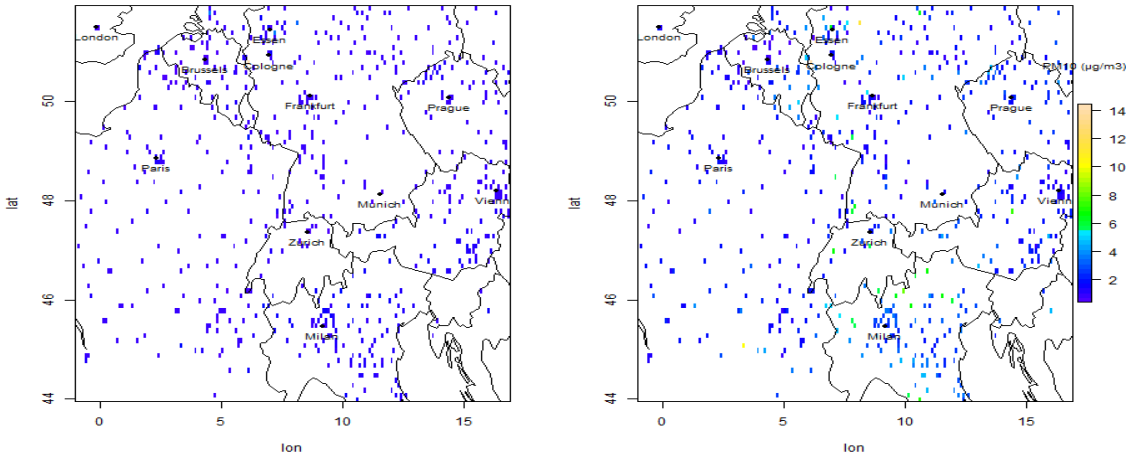


Figure A15: PM_{10} Spatial FGE in winter Dec 2018 (left: Polyphemus FGE, right: CAMS FGE).

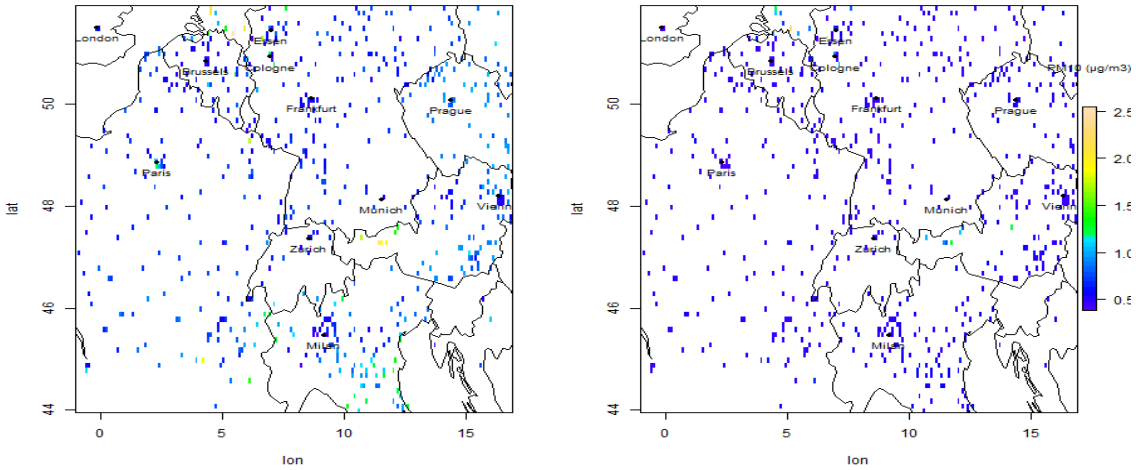


Figure A16: PM_{10} Spatial FGE in summer Aug 2017 (left: Polyphemus FGE, right: CAMS FGE).

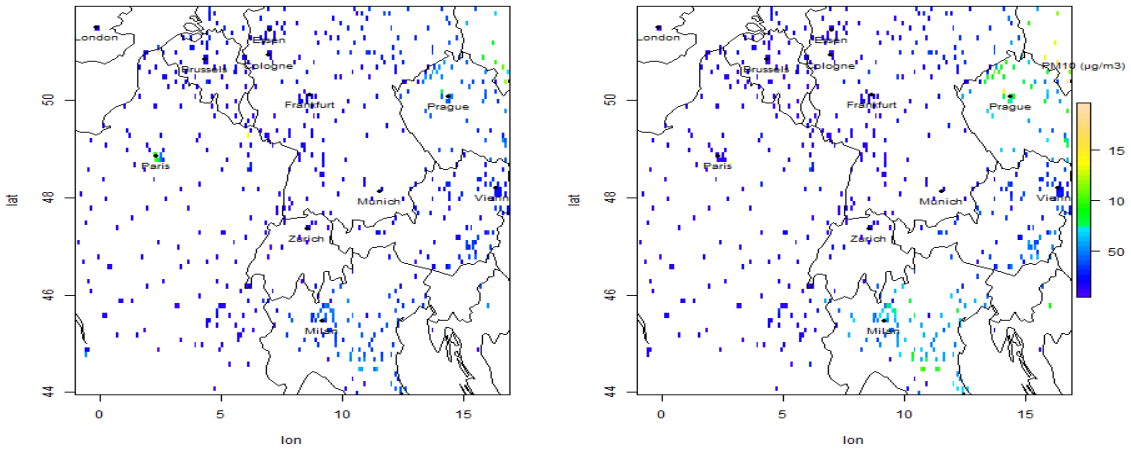


Figure A17: PM_{10} Spatial RMSE in winter Feb 2017 (left: Polyphemus RMSE, right: CAMS RMSE).

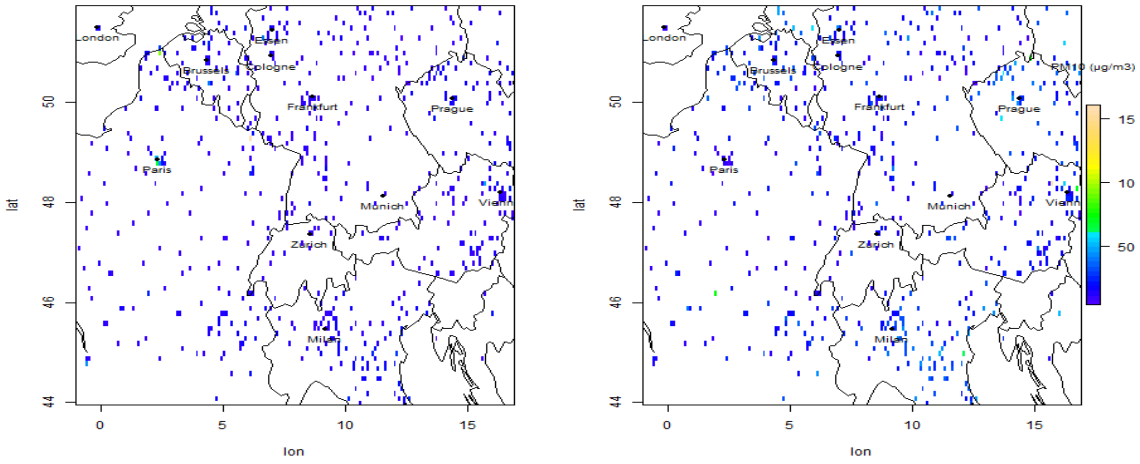


Figure A18: PM_{10} Spatial RMSE in summer Aug 2017 (left: Polyphemus RMSE, right: CAMS RMSE).

Appendix 4 – $PM_{2.5}$ Spatial results.

More $PM_{2.5}$ spatial results for all the statistical indicators for model-model-station comparison are presented here.

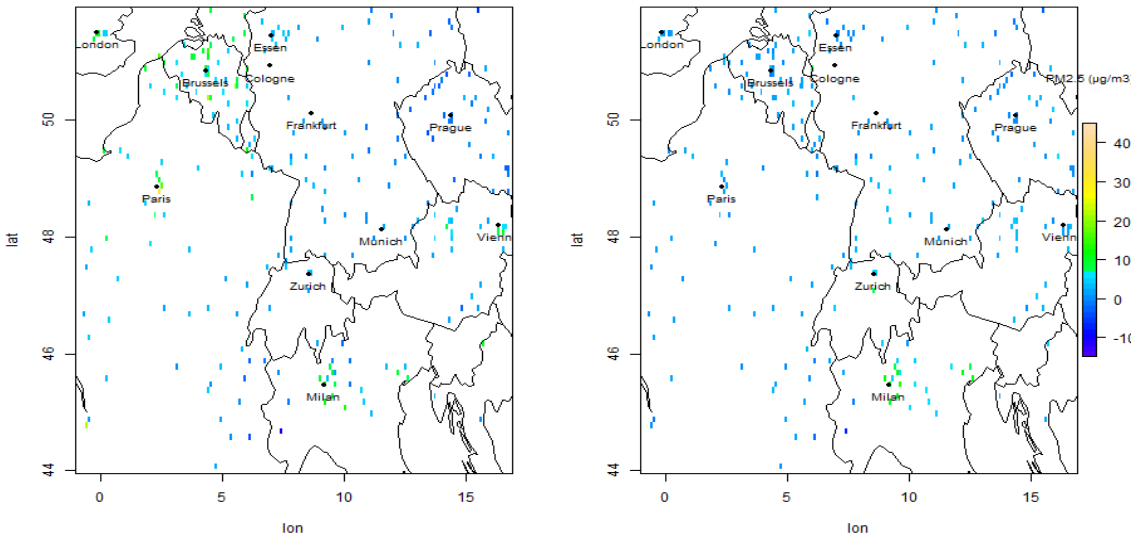


Figure A19: $PM_{2.5}$ Spatial mean bias in summer June 2016 (left: Polyphemus mean bias, right: CAMS mean bias).

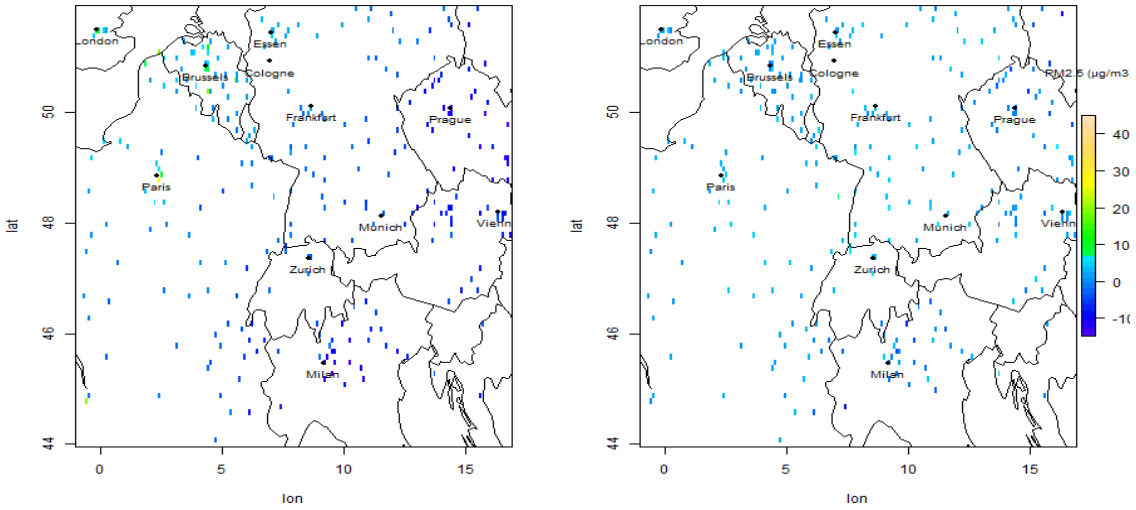


Figure A20: $PM_{2.5}$ Spatial mean bias in autumn Oct 2018 (left: Polyphemus mean bias, right: CAMS mean bias).

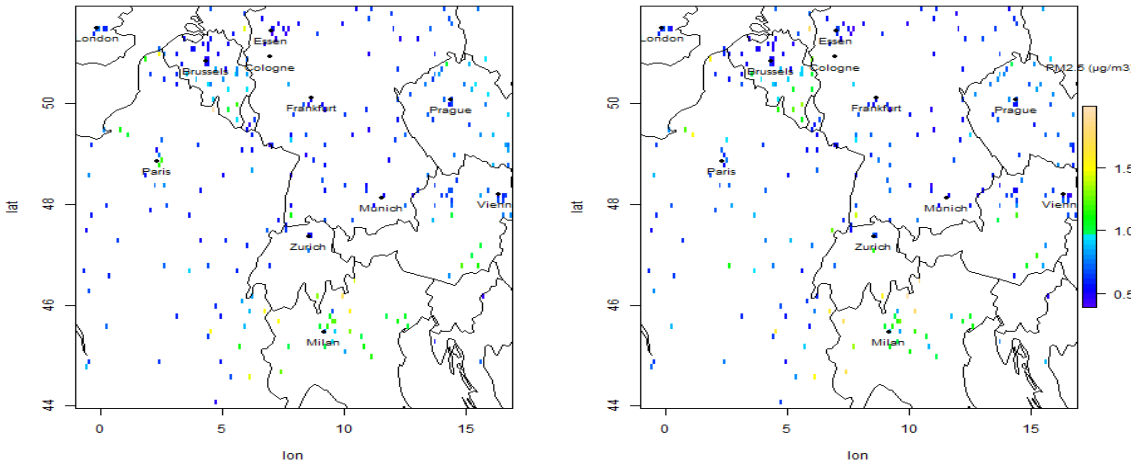


Figure A21: $PM_{2.5}$ Spatial FGE in winter Dec 2017 (left: Polyphemus FGE, right: CAMS FGE).

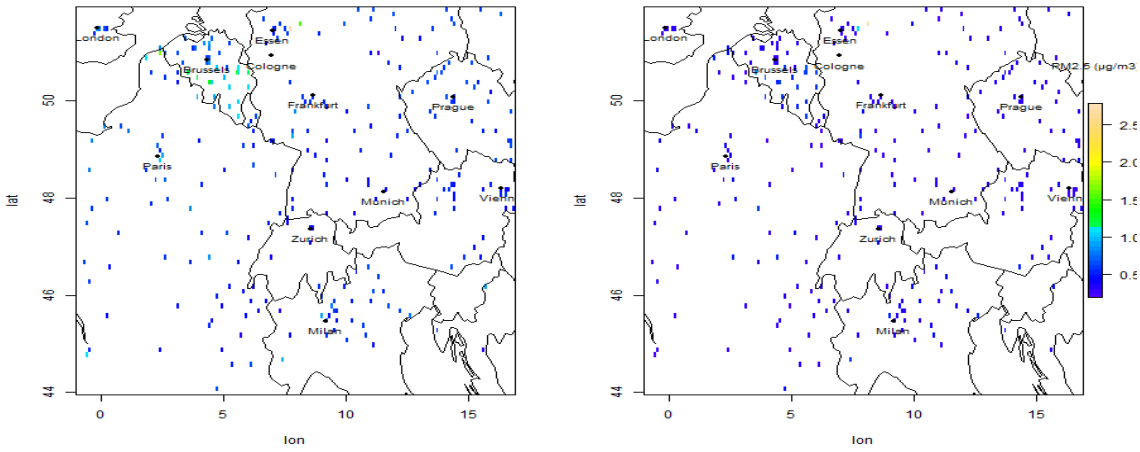


Figure A22: $PM_{2.5}$ Spatial FGE in summer July 2018 (left: Polyphemus FGE, right: CAMS FGE).

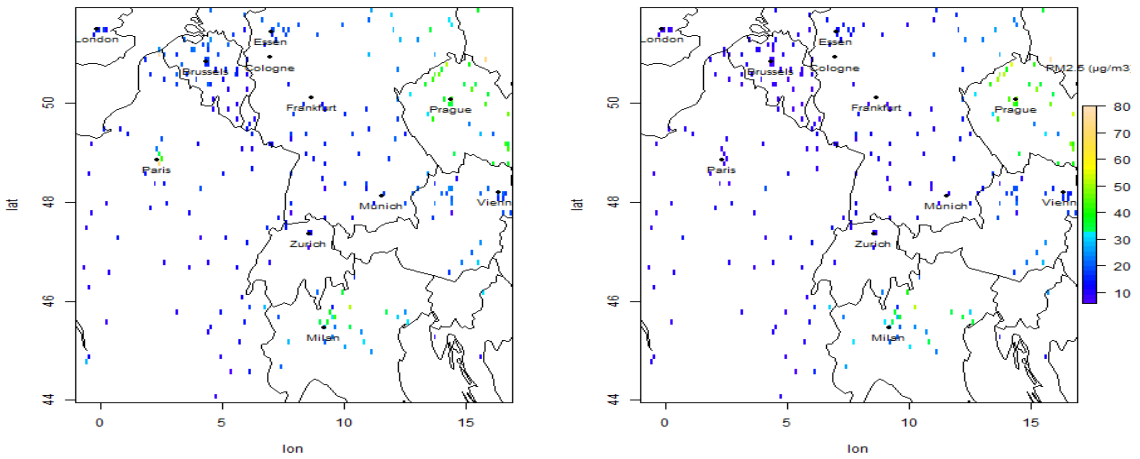


Figure A23: $PM_{2.5}$ Spatial RMSE in winter Feb 2017 (left: Polyphemus RMSE, right: CAMS RMSE).

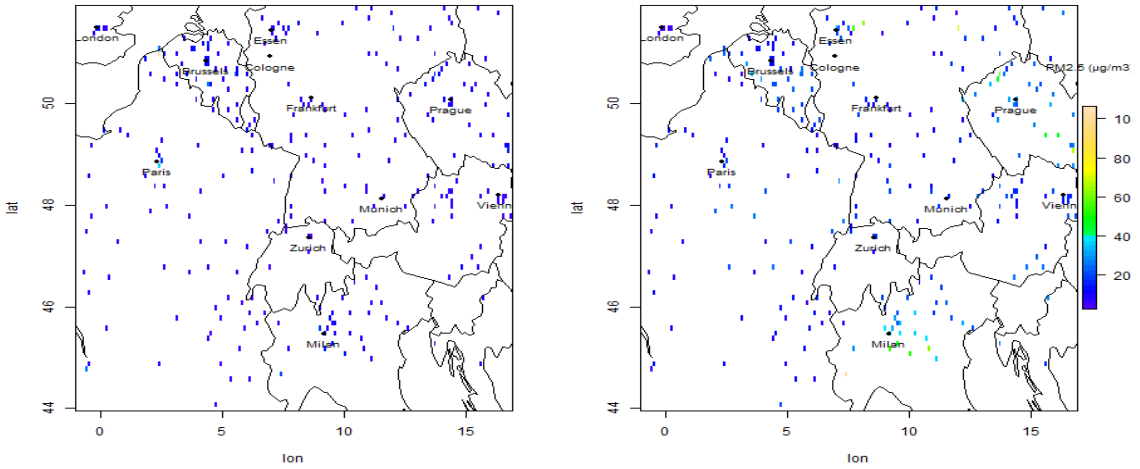


Figure A23: $PM_{2.5}$ Spatial RMSE in summer Aug 2018 (left: Polyphemus RMSE, right: CAMS RMSE).

Appendix 5 – Mann Kendell's Trend Statistics for all the pollutants (Model-model-station comparison).

NO_2	Kendall's score	P-value	Numbers	Variance of Kendall's slope	Kendall's tau statistics
CAMS Reanalysis	$4.496e^{+06}$	$4.29e^{-05}$	22152	$1.207e^{+12}$	$1.832e^{-02}$
Polyphemus	$6.245e^{+05}$	0.5698	22152	$1.207e^{+12}$	$2.547e^{-03}$
EEA station data	$-2.690e^{+05}$	0.01437	22152	$1.207e^{+12}$	$-1.0965e^{-02}$

Table A1: Mann Kendall's Trend statistics for Polyphemus, CAMS, and Station observations NO_2.

O_3	Kendall's score	P-value	Numbers	Variance of Kendall's slope	Kendall's tau statistics
CAMS Reanalysis	$1.006e^{-05}$	$< 2.2e^{-16}$	22152	$1.207e^{+12}$	$4.1019e^{-02}$
Polyphemus	$9.453e^{+05}$	0.3897	22152	$1.207e^{+12}$	$3.8532e^{-03}$
EEA station data	$-7.180e^{+05}$	0.5135	22152	$1.207e^{+12}$	$-2.9267e^{-03}$

Table A2: Mann Kendall's Trend statistics for Polyphemus, CAMS, and Station observations O_3.

$PM_{2.5}$	Kendall's score	P-value	Numbers	Variance of Kendall's slope	Kendall's tau statistics
CAMS Reanalysis	$1.6763e^{+07}$	$< 2.2e^{-16}$	22152	$1.207e^{+12}$	$6.8324e^{-02}$
Polyphemus	$-7.557e^{+06}$	$6.131e^{-12}$	22152	$1.207e^{+12}$	$-3.0804e^{-02}$
EEA station data	$4.1348e^{+06}$	0.00016	22152	$1.207e^{+12}$	$1.6853e^{-02}$

Table A3: Mann Kendall's Trend statistics for Polyphemus, CAMS, and Station observations $PM_{2.5}$.

PM_{10}	Kendall's score	P-value	Numbers	Variance of Kendall's slope	Kendall's tau statistics
CAMS Reanalysis	$1.7537e^{+07}$	$< 2.2e^{-16}$	22152	$1.207e^{+12}$	$7.1482e^{-02}$
Polyphemus	$1.4518e^{+06}$	0.1865	22152	$1.207e^{+12}$	$5.9175e^{-03}$
EEA station data	$1.798e^{+07}$	$< 2.2e^{-16}$	22152	$1.207e^{+12}$	$7.2543e^{-02}$

Table A4: Mann Kendall's Trend statistics for Polyphemus, CAMS, and Station observations PM_{10}.

Appendix 6 – Polyphemus potential drivers

Potential drivers for the pollutants O_3, $PM_{2.5}$, and PM_{10}.

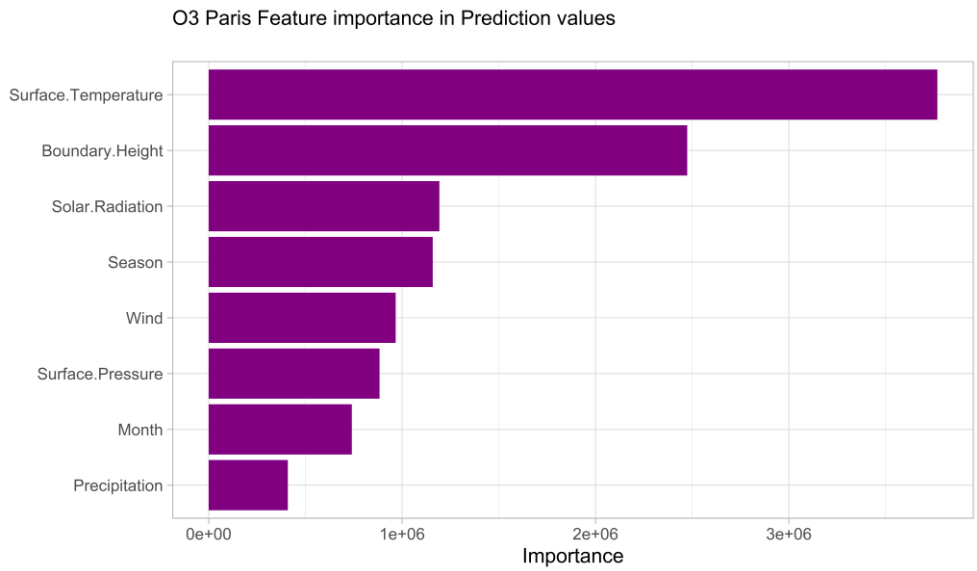


Figure A26: O_3 Polyphemus potential drivers in the urban region (Paris)

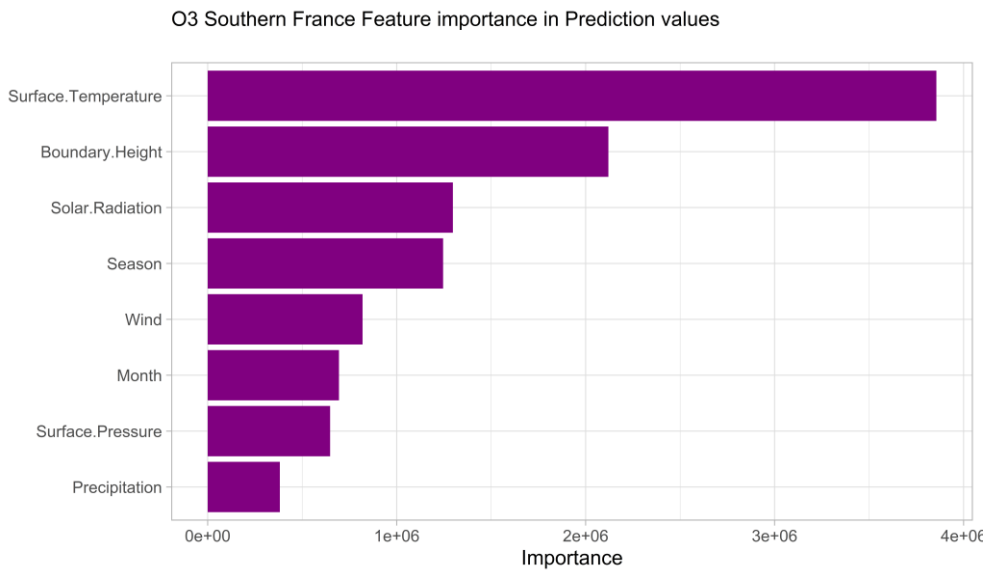


Figure A27: O_3 Polyphemus potential drivers in the rural region (southern France)

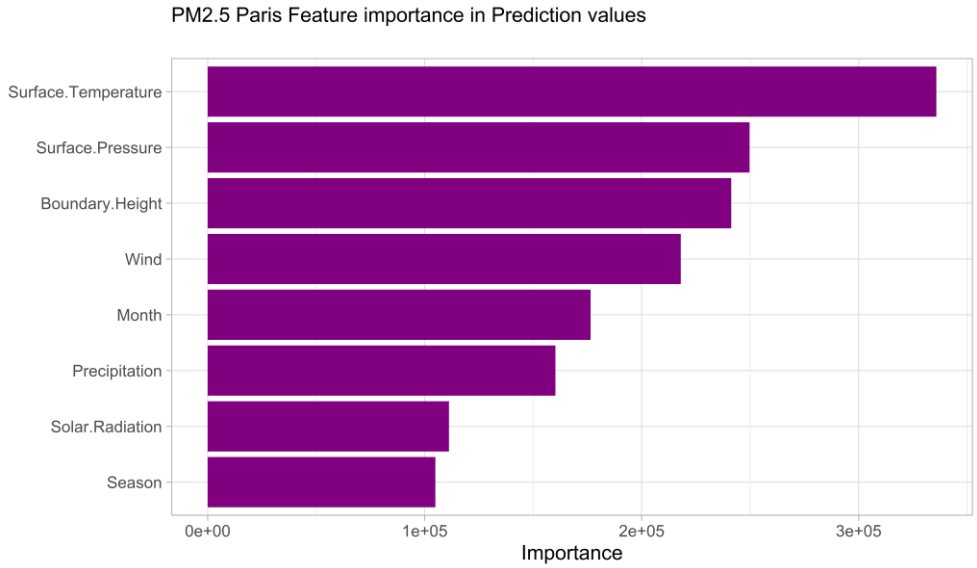


Figure A28: $PM_{2.5}$ Polyphemus potential drivers in the urban region (Paris)

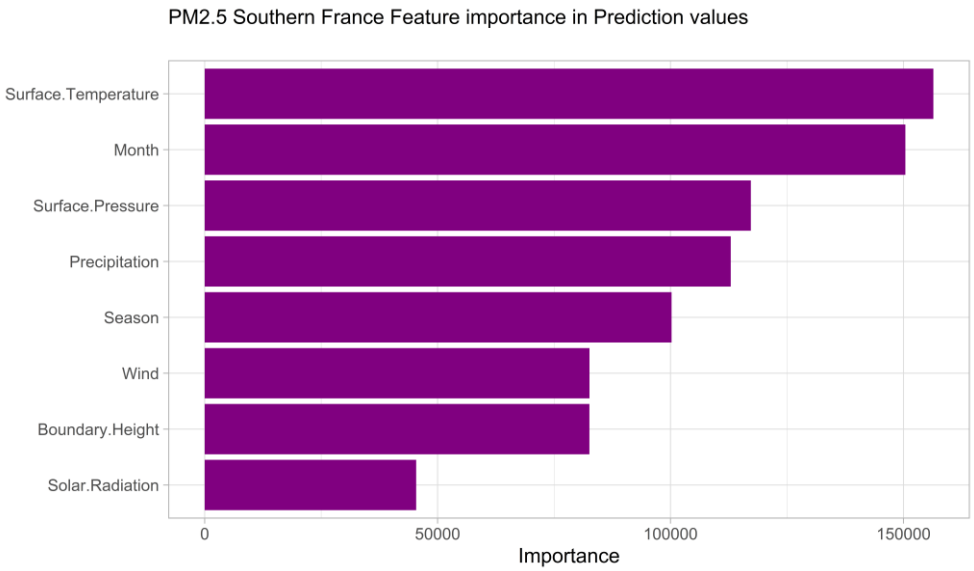


Figure A29: $PM_{2.5}$ Polyphemus potential drivers in the rural region (southern France)

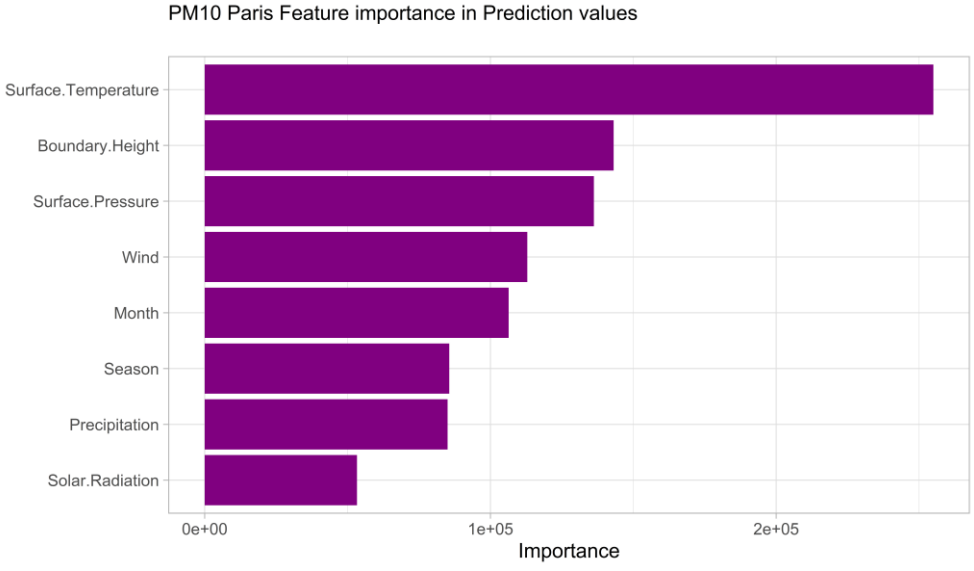


Figure A30: PM_{10} Polyphemus potential drivers in the urban region (Paris)

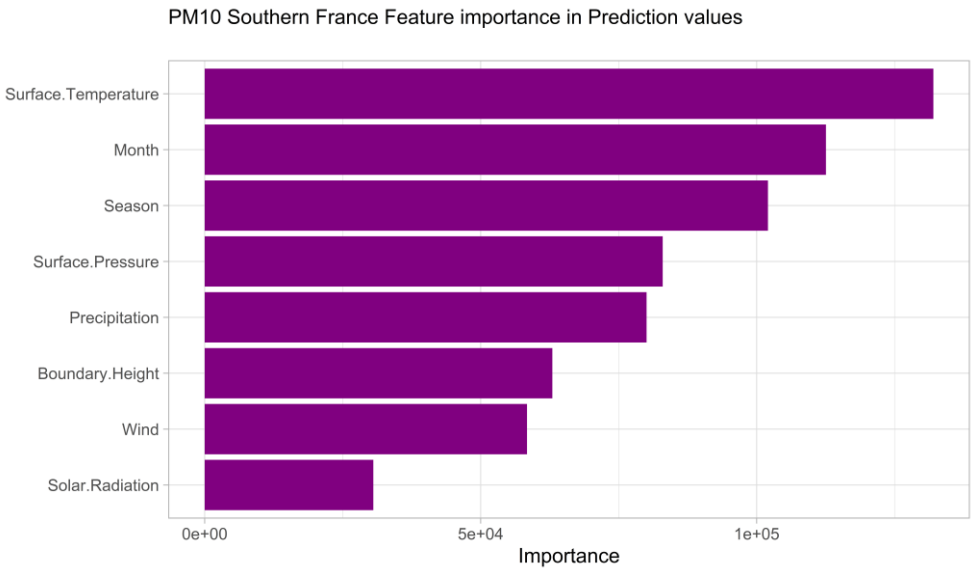


Figure A29: PM_{10} Polyphemus potential drivers in the rural region (southern France)

Appendix 7 Data formatting – Timestep correction

Concatenation of NetCDF

The source Polyphemus datasets are daily NetCDF with hourly timesteps. Concatenating daily data into monthly data helps in processing and visualization of NetCDF in RStudio. In CDO, concatenating NetCDF can be done using the following command.

```
$ cdo cat *.nc out.nc          # To Merge NetCDF files.
```

This command merges all the NetCDF files in the folder together as one NetCDF without changing the actual attributes of the files. All the files should be in the same grid size and projection.

Removing timesteps

The command for deleting the timesteps in CDO is used in several tasks for data correction. Timesteps are removed in some cases like:

- *Removing the first timestep in Polyphemus data:* Polyphemus datasets were created with 25 timesteps in daily data and CAMS data with 24 timesteps. It is mandatory to remove the 1st timestep from Polyphemus data. To remove the 1st timestep from the data for all the Polyphemus data in a loop,

```
$ files=*.nc; for i in $files; do cdo -seltimestep,2/25 ${i} r${i}.nc; done
# for selecting 2to 25 timesteps from the Polyphemus data
```

- *Remove days (Timesteps) from the CAMS Reanalysis data that are not recorded in the Polyphemus data:* Certain days are not recorded in the Polyphemus data. For a better comparison of the models, those days from the CAMS reanalysis data must be removed just to make sure both the model datasets have the same number of timesteps for each month (e.g., 744 timesteps for January, 720 for April, etc.,).

```
$ for i in $(ls); do cdo -delete,timestep=1, 577/648 ${i} r${i}wgs.nc; done
#For deleting the timesteps from the CAMS data.
```

Declaration of Academic Integrity

I hereby confirm that this thesis on **Statistical analysis of systematic differences in the calculated pollutant concentrations of the models ECMWF/CAMS (regional reanalysis) and Polyphemus/ DLR** is solely my own work and that I have used no sources or aids other than the ones stated. All passages in my thesis for which other sources, including electronic media, have been used, be it direct quotes or content references, have been acknowledged as such and the sources did cited.

_____

(date and signature of the student)

I agree to have my thesis checked in order to rule out potential similarities with other works and to have my thesis stored in a database for this purpose.

_____

(date and signature of the student)