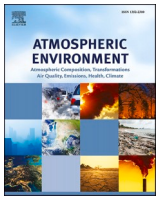
Atmospheric Environment 310 (2023) 119987

Contents lists available at ScienceDirect

Atmospheric Environment

journal homepage: www.elsevier.com/locate/atmosenv



Air pollution prediction using machine learning techniques – An approach to replace existing monitoring stations with virtual monitoring stations

A. Samad [a,*], S. Garuda [b], U. Vogt [a], B. Yang [b]

[a] Institute of Combustion and Power Plant Technology (IFK), Department of Flue Gas Cleaning and Air Quality Control, University of Stuttgart, Germany
[b] Institute of Signal Processing and System Theory (ISS), University of Stuttgart, Germany

HIGHLIGHTS

- Machine Learning models are suitable for pollutant concentration prediction.
- Pollutant concentrations from nearby monitoring stations proved the most effective input parameter.
- The developed methodology is applicable to estimate pollutant concentrations at other locations.
- Virtual monitoring stations can substitute existing monitoring stations.

ARTICLE INFO

Keywords:
Machine learning
Prediction modelling
Air pollution prediction
Multiple linear regression
Random forest
XGboost
Air quality

ABSTRACT

Air pollution in the modern world is a matter of grave concern. Due to rapid expansion in commercial social, and economic aspects, the pollutant concentrations in different parts of the world continue to increase and disrupt human life. Thus, monitoring the pollutant levels is of primary importance to keep the pollutant concentrations under control. Regular monitoring enables the authorities to take appropriate measures in case of high pollution. However, monitoring the pollutant concentrations is not straightforward as it requires installing monitoring stations to collect the relevant pollutant data, which comes with high installation and maintenance costs. In this research, an attempt has been made to simulate the concentrations of $PM_{2.5}$, PM_{10}, and NO_2 at two sites in Stuttgart (Marienplatz and Am Neckartor) using Machine Learning methods. These pollutants are measured with the help of monitoring stations at these locations. Five Machine Learning methods, namely ridge regressor, support vector regressor, random forest, extra trees regressor, and xtreme gradient boosting, were adopted for this study. Meteorological parameters, traffic data, and pollutant information from nearby monitoring stations for the period from January 01, 2018 to 31.03.2022 were considered as inputs to model the pollutants. From the results, it was concluded that the pollutant information from the nearby stations has a significant effect in predicting the pollutant concentrations. Further, it was investigated if a similar methodology can be applied at other locations to estimate pollutant concentrations. This procedure was tested on the data of the monitoring station Karlsruhe-Nordwest which is located in another German city named Karlsruhe. The results demonstrated that this method is applicable in other areas as well.

1. Introduction

Air pollution is the most considerable environmental health risk in all of Europe (European Environment Agency (EEA), 2022). It accounts for mainly cardiovascular and respiratory diseases, causing loss of healthy years of life and premature deaths. In 2019 alone, air pollution has taken a toll on nearly 307,000 lives in Europe (Khomenko et al., 2021). Any substance that changes the natural composition of the air is considered a pollutant (Baumbach, 1996). Apart from living organisms, the pollutants would affect the properties, such as corroding the buildings/structures. In urban cities, the emissions from the combustion of fossil fuels for various transport modes, industries, and household activities account for the main percentage of emissions emitted into the atmosphere (Mosley, 2014). Insufficient air quality monitoring is always

* Corresponding author. Institute of Combustion and Power Plant Technology (IFK), Department of Flue Gas Cleaning and Air Quality Control, University of Stuttgart Pfaffenwaldring 23, 70569, Stuttgart, Germany.
E-mail address: abdul.samad@ifk.uni-stuttgart.de (A. Samad).

https://doi.org/10.1016/j.atmosenv.2023.119987
Received 3 April 2023; Received in revised form 21 July 2023; Accepted 27 July 2023
Available online 28 July 2023

a matter of concern (Duyzer et al., 2015). To tackle pollution, European Union (EU) came up with the approach in 2008 to measure the air quality in areas where people are affected adversely. In case of exceedances of the legal limit values, an air pollution control plan (Luftreinhalteplan) needs to be established to reduce air pollution to stick to the limit values (EU, 2015a). One way to monitor air quality is with the help of the air quality monitoring network, also referred to as a monitoring station. The ambient air quality directive lays down objectives for ambient air quality and methods and criteria for assessing air quality in the member states (EU, 2015b). Achieving adequate coverage with an air quality monitoring network includes factors such as population density, location, cost, and maintenance life-cycle of measuring devices. Increasing the number of monitoring stations is not feasible as per the limited public administration budgets (Spangl et al., 2007a). Machine Learning (ML) can provide a solution to this problem. By utilizing ML techniques robust models can be developed and specific relationships between data collected from monitoring stations and pollutant concentrations at other spatial locations can be proposed.

1.1. Machine learning and its components

Artificial intelligence is the highest paradigm where an element can sense, reason, make decisions, and adapt from its mistakes. ML is a subset of artificial intelligence which enables the element to learn without explicitly being coded by a set of rules (Xu et al., 2021). Features or independent variables are the inputs provided to the algorithm enabling it to capture the relationship with the variable of interest. The output to be estimated is called a target or dependent variable. At times features can be the input data itself, and sometimes new features could be created to provide new hints for the algorithm to learn. This concept of creating new features with the help of existing features is called feature engineering (Dong and Liu, 2018). ML model is a method applied to the input data, that tries to detect its relationship with the target value. This entire process is referred to as Training. When the model is applied to new unseen inputs, to evaluate the relationship learned is referred to as Testing (Goodfellow et al., 2016). The bias and variance hold fundamental importance in evaluating the performance of the ML model. Bias is the degree of effectiveness with which the model learns from the input data (training data). High bias means the model was unable to capture the relationship between the training data and the target value. This phenomenon is referred to as underfitting (Goodfellow et al., 2016). When the ML model is established and new unseen data (test data) is supplied to make predictions, the extent to which the predictions correspond to the test data refers to variance. The model is said to be overfitting if it learns the training data to the extent that it negatively affects the model performance on new data (Goodfellow et al., 2016). Thus, the goal of a model is to have low bias and low variance. The hyperparameters are the external configuration of the model that can be tuned to optimize the ML model algorithm, which minimizes the loss when applied to a particular data (Goodfellow et al., 2016).

1.2. Machine learning models

Estimating the pollutant concentration can be carried out with the help of traditional models, such as chemical transport and dispersion models. However, these models depend on several physical and chemical formulas, which makes it a challenging task (Vlasenko et al., 2021). These models involve complex flow control equations. Although, with advancements in data-driven methods, processing these has become easier, still working with these is a challenging task. Applying the ML models has also produced reliable estimates lately and is thus used widely. The advantages of ML models are the ease of computing and inexpensiveness compared to the traditional methods (Xing et al., 2020). Estimating pollutant concentration is an active area of research because one can try to reduce the dependency on networks or sensors used, making an approximate estimation. Following are the ML models that were applied to predict pollutant concentrations.

1.2.1. Ridge regression

Ridge regression is a parameter estimation method. In linear regression, the parameters are the weights learned by the model and the output is a linear combination of inputs. In the case of non-linearity among the inputs, linear regression tries to fit a line. Though it has a low bias, it might suffer from high variance. However, a ridge regressor tries to add a bias into the coefficients to have a minimizing effect on variance. The advantage of this method is that if there is noise in the original data, the bias forces the loss to be small by minimizing the weights (McDonald, 2009).

1.2.2. Support vector regression

The concept of support vectors was originally introduced by Vapnik et al. (1996) to solve pattern recognition problems on multidimensional data. Using support vector regression, it is possible to reduce error or loss within a defined range. Unlike ridge regression, outputs are often approximated with a line. This method offers the flexibility that it fits not only on a line but also on a curve depending on the type of function used.

1.2.3. Ensemble methods

The ensemble is a particular way to combine different models strategically to solve a particular problem (Zhang and Ma, 2012). The main goal of the ensemble is to eliminate the weakness of individual models by integrating them. There are many forms of ensembles, such as Bagging, Boosting and Stacking (Zhang and Ma, 2012). In Boosting, a sequence of models is built so that the residuals of one model are provided as targets for the next model, and so on, which in turn reduces the bias (Zhang and Ma, 2012). In Stacking, the outputs from individual models are provided to another estimator, called the meta-model, to obtain the final output (Wolpert, 1992). By far, the ensemble methods have shown promising results compared to any other ML models (Zhang and Ma, 2012). In a decision tree, branching at any node depends on a feature that should minimize the loss function.

1.2.3.1. Bagging. Bagging is the short form for Bootstrapped Aggregation. In Bagging, individual models are trained separately and the output of all models is averaged to reduce variance (Zhang and Ma, 2012). The base models are decision trees that easily tend to overfit. To eliminate overfitting, the concept of Bagging is introduced. There are various forms of Bagging, such as random forest and extra trees regressor. In the end, the output of different base learners is combined by taking the average (He et al., 2021).

1.2.3.2. Random forest and extra trees regressor. Random forest is a variant of bagging that builds a multitude of decision trees to obtain the output (Liaw and Wiener, 2001). Sampling features are termed column sampling and data points as row sampling. Trees are built with row and column samples. The advantage of building the model in such a way is that it is robust in estimating new data points. Extra trees regressor performs similarly to random forest, with one more level of randomization (Geurts et al., 2006). This method helps to achieve slightly better performance compared to the random forest, also helping in reducing the training time. extra trees > random forest

1.2.3.3. Boosting. In Boosting, unlike Bagging, the fundamental difference is that the models are built successively (Zhang and Ma, 2012). The main idea is to build a base model and estimate the residual error (the difference between actual and estimated target values). The next model is built based on residuals of the previous stage as the target. This process is continued iteratively until the lowest residual error is reached. Thus, the final model combines all models in each iteration (Han et al.,

2019). XGBoost is a variant of boosting algorithm which includes both row and column sampling.

1.3. Performance evaluation metrics

Different metrics are available to evaluate the model performance. This research work primarily focused on predicting the pollutant concentration, which is a real-valued number. Hence, the regression error metrics were considered for this work as suggested in the study by Botchkarev, Alexei (Botchkarev, 2018). For all the metric demonstrations, N denotes the number of data points, y_i represents the actual value of each data point, and $\hat{y}_i$ represents the predicted value of the data point.

- R Square (R^2): R^2 demonstrates how much variability in the dependent variable can be explained by the model. It is represented as shown in Equation A1 in Appendix. As N is the number of the total data points, and $\overline{y_i}$ is the average of all total data points (Botchkarev, 2018). It is a good metric to check the fitting of the model with data points. The R^2 value near 1 represents the best possible accuracy.
- Mean Absolute Error (MAE): It is the magnitude of the difference between the actual and predicted outcomes. For N data points, MAE is defined as shown in Equation A2 in Appendix (Botchkarev, 2018). This performance metric is robust to outliers.
- Root Mean Squared Error (RMSE): It is expressed as the square root of the mean square error. The advantage of RMSE is the differentiability and can also be used as a loss function. RMSE can never be negative and is defined in Equation A3 in Appendix (Botchkarev, 2018).
- Mean Absolute Percentage Error (MAPE) defines the percentage deviation of the predicted value from the actual value. For N data points, it is expressed as shown in Equation A4 in Appendix (Botchkarev, 2018).

1.4. Literature review

In a research study, the mentioned regression model was implemented to estimate PM_{10} concentration (target) in Chonburi, Thailand with the help of independent input data, which are meteorological and pollutants. The meteorological inputs included air pressure, precipitation, air temperature, relative humidity, and wind speed. The pollutants included carbon monoxide (CO), nitrogen monoxide (NO), nitrogen dioxide (NO_2), sulfur dioxide (SO_2), Black Carbon (BC), methane (CH_4), Non-Methane Hydro Carbon (NMHC) and ozone (O_3) from 2006 to 2008 (Saithanu and Mekparyup, 2014). Another study by Rybarczyk and Zalakeviciute estimated $PM_{2.5}$ concentration with regression models based on time. Primarily the regression models were built by segregating the day into three-time segments such as 6 a.m.–10 a.m., 10 a.m. to 2 p.m. and 2 p.m.–7 p.m. for the capital city of Quito, Ecuador. In this study, initially, the models were built for each of the time periods based on ease of data availability. Traffic was considered for the three time periods and was segregated into high, medium and low. The model showed an R^2 score of 0.27 with these settings. By adding meteorological data including solar radiation, air temperature, air pressure, precipitation, relative humidity, and wind speed as features, an improvement in R^2 score to 0.38. Finally, the trace gas concentrations SO_2, NO_2, O_3 and CO were considered as well, which improved the R^2 score to 0.8. The limitation of this study was the extra cost associated with measuring the trace gas concentrations (Rybarczyk and Zalakeviciute, 2017).

Support vector regression was used to estimate CO concentration for the region of New South Wales in Australia, dividing the entire region into 100 grids. The authors estimated CO concentration in all 100 grids using four different sets of features that included CO concentrations from four monitoring stations within the study area, latitude and longitude of the grids, hour, day of the week and season. The authors did not include meteorological parameters for this estimation. In this manner, the same spatial pollutant monitoring networks were employed in modelling (Hu et al., 2016). Support vector regression was also applied to predict O_3 concentration in Delhi, India. Different kernels related to support vector regression, such as linear, polynomial, and radial basis functions were employed. The work stated the best possible feature set to forecast O_3 concentration with five input parameters, namely the ozone for the previous two days and meteorological inputs of air temperature, relative humidity, and sunshine hours. Finally, a comparison between performance metrics of linear regression and multiple layer perceptron along with support vector regression was performed concluding that support vector regression was able to capture non-linear trends effectively with radial basis function kernel used when compared to linear regression and multiple layer perceptron (Chelani, 2009). In another study, similar experiments were carried out to forecast the air quality index in Tehran utilizing the pollutant information of various monitoring stations located across Tehran, from the past two days to forecast the hourly air quality index for the next 24 h. The study was conducted for period of 2008–2013. This study explored different kernel functions and a forecast of a pollutant map with different AQI for different locations in Tehran was obtained (Ghaemi et al., 2018).

The ensemble methods have been employed in the context of air pollution estimation because of their wide popularity and applicability. In most of the studies using this method, meteorological parameters such as air temperature, air pressure, relative humidity, and wind speed were considered as input parameters. These variables vary depending on location and play a crucial role in rapidly varying pollutant concentrations. $PM_{2.5}$ forecasting was performed on a single monitoring station, by including the mentioned meteorological parameters in Delhi, India. Overall eleven models were utilized and the R^2 scores were compared. However, in this study, the outputs from two different models were also combined to see the improved performance. It was concluded that combining two algorithms can produce slightly enhanced performance overall when compared to a standalone algorithm (Kumar et al., 2020a). In another study, a total of 23 features along with $PM_{2.5}$ concentration from 37 monitoring stations were considered. The R^2 score obtained with various ensemble ML models and artificial neural networks was compared with and without the inclusion of aerosol optical depth. This study concluded that because of missing values in $PM_{2.5}$ and aerosol optical depth data obtained from satellite, the performance capabilities of the artificial neural network were reduced (Zamani Joharestani et al., 2019). A random forest model for ozone estimation was built at Research Academy for Environmental Sciences in Beijing, China (Zhan et al., 2022). A linear hybrid machine learning model was applied for $PM_{2.5}$ concentration estimation in China (Song et al., 2021). In London, $PM_{2.5}$ has been estimated using the widely available PM_{10} and NO_X emissions with the help of regression modelling as well as the machine learning method (Random Forest) and a combination of both (Analitis et al., 2020). In another study, different ML models were developed to estimate $PM_{2.5}$ and NO_X for three different monitoring networks using local pollution estimates, meteorological data, and emissions from vehicles. The main objective of this study was to check which of the ML model provide the best possible performance and which were the influential variables. Six ML models were investigated to estimate the prediction capability of $PM_{2.5}$ and NO_2 (Li et al., 2020). Daily CO concentration was estimated in Taiwan for the study period from 2000 to 2018 from which the last two years were used for evaluation. Three models using a deep neural network, random forest and XGBoost were used. The authors concluded that XGBoost had the highest R^2 score of 0.85, followed by random forest and neural network with 0.84 and 0.81 respectively. In comparison, a simple regression model yielded an R^2 score of 0.69 (Wong et al., 2021). A machine learning method to estimate $PM_{2.5}$ concentrations was applied across China with remote sensing, meteorological parameters and land use information (Chen et al., 2018). In a study in Munich Germany, the XGBoost model was built using meteorological parameters, precursors and simulations of

ozone concentration obtained from the CAMS2 dataset to estimate ozone concentration. The objective of this study was to investigate the significance of precursor information in modelling surface ozone using ML. The meteorological parameters such as air temperature, relative humidity, boundary layer height, wind speed and wind direction as well as in-situ ozone precursors (NO, NO_2 and CO), and satellite ozone precursors (column NO_2 and HCHO) along with CTM simulations (CAMS model surface O_3) were used as input parameters. Additionally, day of the week and season was also considered (Balamurugan et al., 2022). Satellite-Based estimates of daily NO_2 exposure in China were tested using a hybrid random forest and spatiotemporal kriging model (Zhan et al., 2018).

A study done by Isam Drewil and Jabbar Al-Bahadili in 2022 (Isam Drewil and Jabbar Al-Bahadili, 2022) proposed a model that combines the Genetic Algorithm (GA) with Long Short Term Memory (LSTM) to optimize hyperparameters and predict pollution levels for the next day, focusing on four key pollutants: PM_{10}, $PM_{2.5}$, CO, and NO_X. One of the primary challenges associated with LSTM is the selection of appropriate parameters, such as window size and the number of units in LSTM. The application of the metaheuristic GA offers a successful solution to this issue, allowing for more flexible performance in predicting pollution levels (Isam Drewil and Jabbar Al-Bahadili, 2022). Another study performed by Du et al. (2020) covered the effectiveness of four advanced machine learning methods for spatial data handling: SVM, semi-supervised and active learning, ensemble learning, and deep learning. These methods have been applied to address classification, regression, and inversion problems, showcasing their ability to improve performance in spatial data analysis. However, it should be noted that the scope of machine learning and spatial data handling is broad, and this review only covers a subset of methods (Du et al., 2020). Another method to investigate the spatial and temporal variations of atmospheric pollutants is the use of satellite-based measurement data with ground-based measurement results. Such a research was performed to investigate seven cities located near the South Gobi deserts (Filonchyk et al., 2020). The analysis covered the period from January 1, 2016 to December 31, 2018. The main pollutants examined were particulate matters ($PM_{2.5}$ and PM_{10}) and gaseous pollutants (SO_2, NO_2, CO) (Filonchyk et al., 2020). Kumar et al. focused on evaluating different interpolation techniques for air quality mapping in Mumbai, India (Kumar et al., 2020b). The authors compared the effectiveness of several interpolation methods, including inverse distance weight (IDW), Kriging (spherical and Gaussian), and spline techniques using data collected from air quality monitoring stations in the city. In terms of statistical assessments, the IDW method indicated a better fit between predicted and observed values. These findings suggested that the IDW approach performs favorably among the interpolation techniques tested in this study (Kumar et al., 2020b).

1.5. Objectives

Forecasting and estimation of pollutants with the help of ML models are becoming an active area of research. The detailed literature review paved the path to knowing which inputs would be influential in estimating the pollutant concentrations. There are very few studies available in which it is tried to implement the ML model built for one location to another. In light of the background information provided, this research addresses the knowledge gap concerning the estimation of pollutant concentrations. The dependency of pollution concentration variation on the location along with pollution sources particular to the locality that lead to pollution makes it a challenging task to apply such techniques. In general, the traffic trend can relate to the pollutant concentration variation during peak and off-peak times. The motivation behind this study lies in the need for accurate estimates of pollutant concentrations, especially in areas without monitoring stations.

The aim of this research is to model the pollutant concentrations using ML models at selected locations of Marienplatz and Am Neckartor in Stuttgart, Germany to provide an opportunity to replace the existing monitoring stations with a virtual monitoring station. Another objective was to check the applicability of the developed methodology in other locations apart from Stuttgart. To achieve these objectives multiple ML models were tested with meteorological parameters, traffic data in the form of a number of vehicles passing per minute and pollutant concentrations from other monitoring stations as input variables.

The findings of this study may hold relevance for policymakers and researchers, enabling evidence-based decision-making and targeted pollution control measures, for instance to know at which location it is important to have the monitoring station and, in its absence, can a model be used in order to continue the estimation. By fulfilling these objectives, this research sets the stage for future investigations and the development of effective environmental management strategies, where monitoring is the starting point.

2. Methodology

2.1. Conceptualization

As mentioned before, the main goal of this research was to predict the pollutant concentrations at a certain location. The pollutants of interest in this study were $PM_{2.5}$, PM_{10}, and NO_2. To investigate the applicability of the models, two locations in the city of Stuttgart were chosen. The data for this research was collected from different sources. After data acquisition, an initial analysis of the pollutants $PM_{2.5}$, PM_{10}, and NO_2 concerning outliers and missing values was conducted. As a continuation step, the relevant features and their relation to pollutant concentration were investigated in detail. Further, the splitting of data into train and test was performed. Then the training of ML models was carried out with the choice of hyperparameter setting. In the end, it was tested if the applied method can be used for other locations.

2.2. Study area

The pollutants of interest $PM_{2.5}$, PM_{10} and NO_2 were to be modeled at two locations in Stuttgart, namely Am Neckartor and Marienplatz which have different characteristics. In Fig. 1, the locations Am Neckartor and Marienplatz are shown with blue and black location pins respectively. Image 1 in this figure represents the aerial view of Am Neckartor location and the small red circle indicates the position of the monitoring station. The monitoring station at Am Neckartor can be seen in Image 2. Similarly, Images 3 and 4 in this figure depict the aerial view of Marienplatz and the monitoring station at Marienplatz respectively. The distance between the two stations is around 3 km.

To perform any ML modelling, data acquisition is the first step. Data were gathered from different sources. The pollutant and meteorological data at Marienplatz were obtained from the Department of Flue Gas Cleaning and Air Quality Control at the Institute of Combustion and Power Plant Technology, University of Stuttgart (Samad and Vogt, 2020). The remaining pollutant and meteorological data were gathered from the Baden-Württemberg State Institute for the Environment (Landesanstalt für Umwelt Baden-Württemberg – LUBW). The LUBW measures the pollutants by establishing a fixed air quality monitoring network at different locations in the city. The traffic data were obtained by the integrated traffic control center in Stuttgart (IVLZ).

2.3. Machine learning workflow

2.3.1. Data processing

The period for the entire study was from January 01, 2018 till 31.03.2022. After data accumulation, the next step was data preprocessing which enabled to identify data quality. In this step, missing value analysis and outlier removal were carried out. Missing values are the number of hourly observations that are not available for reasons such as maintenance procedures or malfunctions in the monitoring

Fig. 1. Am Neckartor and Marienplatz monitoring stations, Stuttgart.

network. The missing values of the pollutants at the monitoring station Am Neckartor were below 5% of the total data and at Marienplatz, this count was below 20%. These values were not imputed since it does not include any bias in the model (Demertzis et al., 2015). For removing the outliers, interquartile range method was used with varying values of fences as suggested in a study done by Hubert et al. (Hubert and Vandervieren, 2008).

2.3.2. *Independent variables/features*

The selection of variables that were considered for this research study was based on the literature review and the data availability. The goal was to select the most informative and influential variables that have an impact on the outcome of interest while excluding irrelevant or redundant variables. By considering this, domain knowledge was vital in identifying the variables as the pollutant concentration of one component can vary depending on temporal factors, meteorological factors, traffic situation, topography and concentration of other pollutants. To visualize this, the concentration variation of the pollutants concerning time, meteorological parameters, traffic and other pollutant concentrations is shown in the form of plots in Figs. 2, 3 and 5.

An amalgamation of features was derived from the data collected to estimate the hourly pollutant values. The features are broadly classified into temporal, meteorological, traffic, and pollutants from other spatially distributed monitoring networks. The temporal features include hourly, daily, weekly and monthly values as the target output variates with these features. Automobiles play an important role in air quality and account for pollution (Long and Carlsten, 2022; Sun and Zhu, 2019). The traffic data obtained from the integrated traffic control

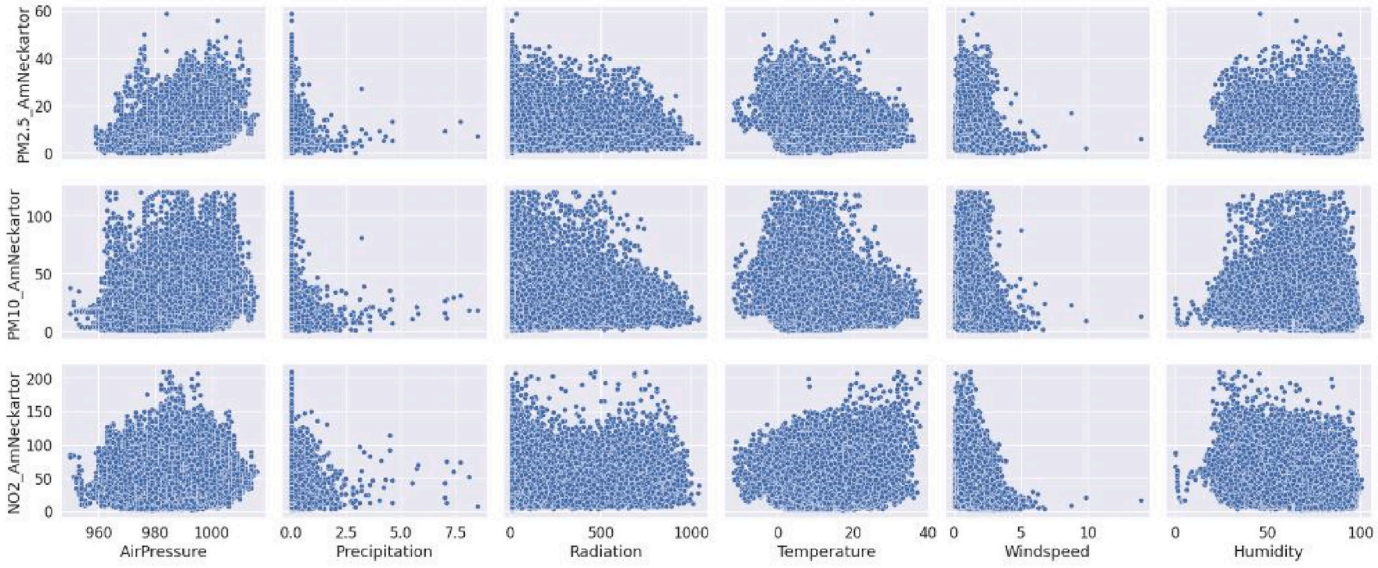


Fig. 2. Scatter plot of meteorological parameters and pollutants at monitoring station Am Neckartor.

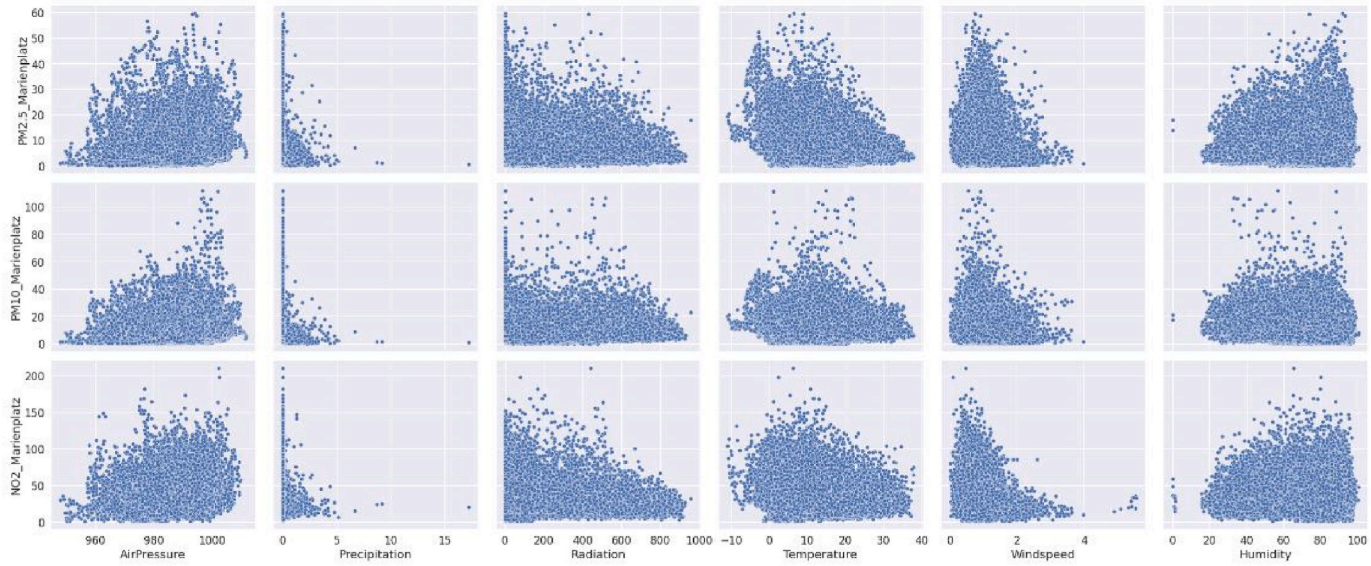


Fig. 3. Scatter plot of meteorological parameters and pollutants at monitoring station Marienplatz.

center (IVLZ) Stuttgart provided the traffic information, i.e. minute average of the number of vehicles passing by the monitoring stations. As mentioned in the literature, meteorological parameters play a vital role in the variation of pollutant concentrations. Occasionally these meteorological conditions change rapidly, resulting in the transportation of pollutants. High wind speed causes the dispersion of pollutants, transporting them from a few meters to kilometers (Latini et al., 2002). Precipitation causes pollutants to settle down, also at times the pollutants are trapped inside the snow (Tian et al., 2021). Both scenarios lower the pollutant concentrations. Apart from the climatic conditions, the topography of the city can account for increased concentrations as in the case of Stuttgart (2008).

The considered meteorological features were air temperature, relative humidity, air pressure, wind speed, global radiation and precipitation. The pair plots for the pollutants $PM_{2.5}$, PM_{10} and NO_2 and their relationship with the meteorological parameters are shown in Figs. 2 and 3 for monitoring station Am Neckartor and Marienplatz respectively. It can be seen that the pollutant concentrations show a negative trend with a high amount of precipitation and windspeed and vice versa. The pollutants can alter the amount of light that can reach the earth's surface (Khodakarami and Ghobadi, 2016). Thus, lowering the radiation levels may result from high pollutant levels (Khodakarami and Ghobadi, 2016). This trend was particularly seen in the case of $PM_{2.5}$ and reduced further for PM_{10} and NO_2. Another notable observation was the impact of humidity for all three pollutants where low pollutant concentrations were observed with low humidity.

Every city has a different topology attributed to itself. Thus, monitoring with the help of a single monitoring network is merely possible (Verghese and Nema, 2022). Monitoring stations can be broadly categorized into hot spot, background, and commercial stations depending on the monitoring site. A hot spot station is one situated right next to the source, like traffic. In contrast, the background station is where pollution levels are not directly affected by emission sources and are represented by land cover and population (Spangl et al., 2007b). At a commercial station, pollutant levels are accounted for in both scenarios, i.e., hotspot and background. In Fig. 4 the air pollution monitoring network in Stuttgart is shown. The monitoring stations marked by blue pins are the ones operated by LUBW and the monitoring station marked by black pin is operated by IFK, University of Stuttgart. The name of the monitoring station, category and measured pollutants are listed in Table 1.

Since the pollutants to be modeled are also measured by monitoring

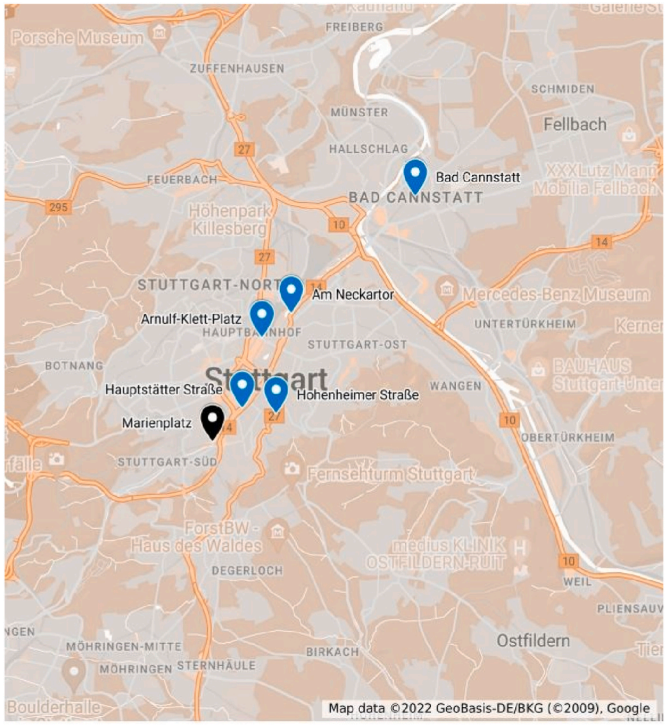


Fig. 4. Air pollutant monitoring network in Stuttgart showing continuous monitoring stations operated by LUBW (blue pin) and IFK, University of Stuttgart (black pin).

stations at different sites, using this spatial information can assist in estimating the pollutants of interest $PM_{2.5}$, PM_{10} and NO_2 at the two locations Marienplatz and Am Neckartor. The pollutant concentration is affected by certain factors that are not considered for ML models to make it less complex such as the location characteristics, topography, varying emission sources, and occasional activities, e.g. construction, festivals, etc. One of the reasons to include pollutant concentration from other stations as an input was to consider such factors indirectly using the pollutant concentration of other stations. The relationship between pollutants measured at different monitoring stations is shown with a Spearman rank correlation matrix (Akoglu, 2018) in Fig. 5. This plot

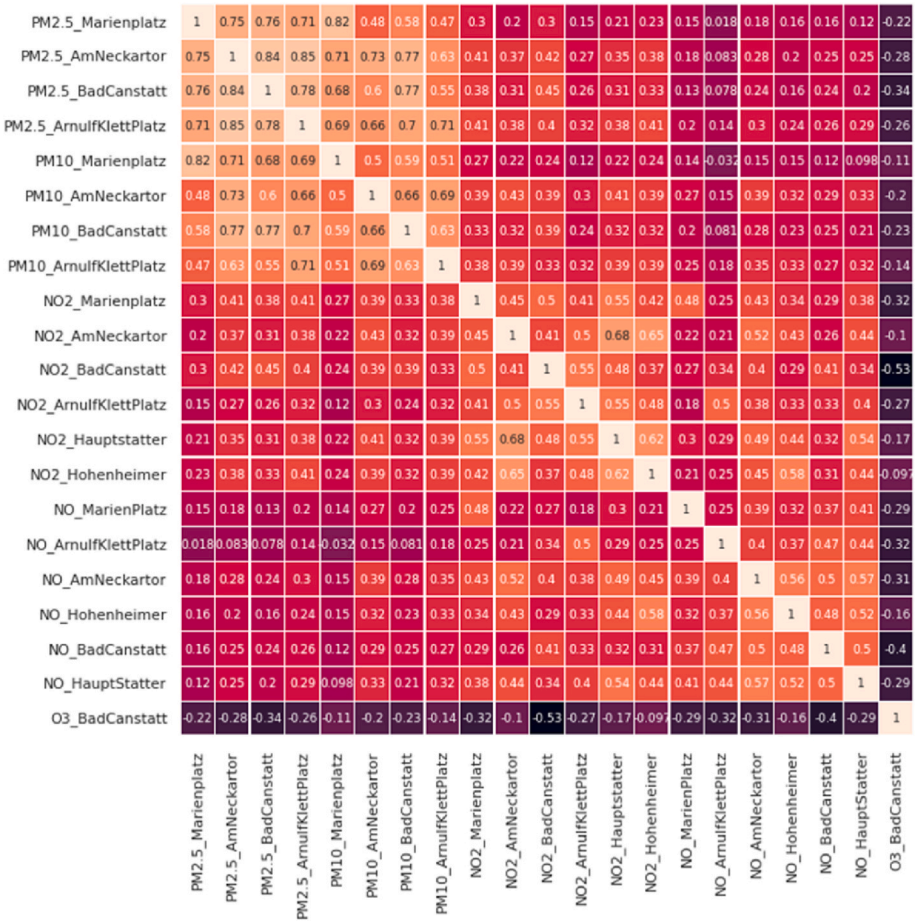


Fig. 5. Spearman rank correlation matrix between pollutants.

Table 1
Monitoring networks along with the list of pollutants being measured.

Name of monitoring station	Type	Measured pollutants
Bad Cannstatt	Background	PM_{10}, $PM_{2.5}$, NO_2, NO and O_3
Am Neckartor	Hot spot, traffic	PM_{10}, $PM_{2.5}$, NO_2 and NO
Arnulf Klett Platz	Commercial	PM_{10}, $PM_{2.5}$, NO_2 and NO
Hohenheimer Straβe	Hot spot, traffic	NO_2 and NO
Hauptstatter Straβe	Hot spot, traffic	NO_2 and NO
Marienplatz	Commercial	PM_{10}, $PM_{2.5}$, NO_2, NO and O_3

measures the relationship between two variables in order of their ranks. Thus, it essentially provides a measure of the monotonic relationship between those two variables.

From the correlation matrix in Fig. 5, the pollutant at each station is denoted as pollutant_stationname. The correlation ranges from −1 to 1. If the correlation is near to one then the features are positively correlated, where −1 means negatively correlated. It can be seen from the correlation matrix that PM2.5_Marienplatz has a decent correlation of 0.76 and 0.7 with PM2.5_BadCanstatt and PM2.5_ArnulfKlettPlatz respectively. As a common observation, $PM_{2.5}$ between different locations is highly correlated compared to PM_{10} followed by NO_2. Ozone is negatively correlated with $PM_{2.5}$, PM_{10} and NO_2. This matrix serves as substantial support in including pollutants from other monitoring networks in establishing the ML models at Marienplatz and Am Neckartor.

2.3.3. Train and test split

After identifying the features, the next step was to divide the available hourly data into training and test data. The training data was selected from January 01, 2018 to 31.03.2021 and the test period from April 01, 2021 to 31.03.2022. Approximately 75% of the data were used for training and the remaining test data (25%) was for one complete year. The advantage of selecting an entire year for test data is that it covers the seasonal variation that the pollutants can be influenced by during an entire year. The train and test split percentages were based on the literature review in which many authors suggested the division in this range. A 10-fold cross-validation technique was employed, i.e., dividing the train set into ten equal parts and, each time, training nine parts and evaluating the remaining part as the validation set. Thus, in this way, the generalizing ability of the model increases while training and hence when evaluated on the test may yield plausible results. This form of evaluation on one part after training on nine parts is to find out the best set of hyperparameters for each of the models.

2.3.4. Model training and hyperparameter tuning

Five different ML models were trained with the same training data obtained after the split. The training strategy is shown in Fig. 6, in which each row contains various temporal, meteorological, and traffic features and corresponding targets (one pollutant measurement) fed to each ML model. Here the row represents the hourly timestamp represented with the naming convention T1 to Tn. For every timestamp, say Tx, the corresponding model outputs are named O1 to O5 for the ML models M1 to M5. The minimum, average, and maximum values of the obtained outputs for every timestamp were analyzed. Obtaining minimum, average, and maximum value gives a range of pollutant concentrations at each timestamp. Performing the average of results provides more robust outputs as suggested in several studies (Wichard and Ogorzalek, 2004; Maqsood et al., 2004; Talebizadeh and Moridnejad, 2011) that averaging could result in enhanced performance.

The individual ML models along with their hyperparameter settings

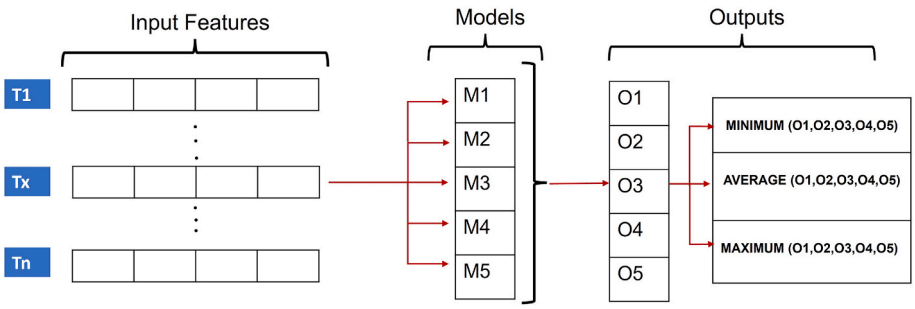


Fig. 6. ML workflow including input features, models and outputs.

are listed in Table 2. The selection of ML models for this study was based on the literature review. From the previous studies predicting air pollutant concentrations, these ML models were proposed. All the used models are available in the Scikit-Learn3 library. The best possible hyperparameter set was obtained after several trials avoiding the case of overfitting. Finally, the build models were evaluated on the test set.

3. Results and discussion

To estimate the hourly pollutant concentration in the study areas, the ML models are built for the following scenarios depending on the data availability and the pollutant concentration variation concerning different features.

- Scenario 1: In this scenario, ML models were built by providing the meteorological and temporal as input features.
- Scenario 2: In scenario 2, to the features of scenario 1, traffic was included as an input feature.
- Scenario 3: For scenario 3, along with the features of scenario 2, the same pollutant concentration from the background monitoring station (Bad Cannstatt) was introduced.
- Scenario 4: For scenario 4, along with the features of scenario 2, pollutants from the other stations mentioned in Table 1 other than monitoring station Marienplatz and monitoring station Am Neckartor were considered.

3.1. Scenario 1

In Scenario 1, temporal features such as month, day and hour were considered. The meteorological features such as air pressure, precipitation, global radiation, temperature, windspeed, and humidity were given as input features. The ML models were applied with a ten-fold cross-validation technique to obtain the optimal set of hyperparameters. After training, the models were evaluated using the test data set to assess the model's performance. The models were optimized by performing hyperparameter tuning. The performance metrics for individual pollutants at both locations were established with the purpose to investigate the model performance.

Table 2
ML models along with Hyperparameters used.

ML models	Hyperparameters and range
Ridge Regressor (RIDGE)	Regularizer λ
Support Vector Regressor (SVR)	Kernel, C, degree, epsilon
Random Forest Regressor (RFR)	Estimators, max_depth, min_samples_leaf, max_features
Extra Trees Regressor (ETR)	
Xtreme Gradient Boosting (XGBOOST)	Estimators, max_depth, min_samples_leaf, learning_rate, subsample, column_sample_tree

The performance metrics of different models used in this study for the $PM_{2.5}$, PM_{10}, and NO_2 pollutants illustrating the values for R^2, MAE, and RSME on training and test data for each pollutant are shown in Fig. 7 for the monitoring station Marienplatz, Stuttgart. The error metric results for the monitoring station Am Neckartor, Stuttgart are presented in Figure A1 in Appendix. Training and test data for each pollutant were applied to the following ML models independently: RIDGE, SVR, RFR, ETR and XGBOOST. Firstly, the R^2 metric for $PM_{2.5}$ pollutant indicated that it was always high when compared with the test data value for all models. Especially the R^2 value for the models RFR and ETR indicated significantly higher values for training data than the test data. Generally, these bagging methods tend to overfit, which was observed in this case. Similarly, this trend was observed for PM_{10} and NO_2 pollutants. However, the RIDGE and SVR models recorded fewer R^2 values for the training data. This behavior was common for all the pollutants measured for the monitoring station Marienplatz, Stuttgart location. For the PM_{10} pollutant, the RIDGE model resulted in higher R^2 value for test data than the training data. Overall the models except for ETR and RFR were found to be underfitting with very low R^2.

The right-side graph of Fig. 7 demonstrates substantial error values for both training and test data for the NO_2 pollutant in comparison with $PM_{2.5}$ and PM_{10} pollutants. This observation was recorded for all the models used in this study at the Marienplatz, Stuttgart location. Also, the difference in error performance metrics for training and test data for all the pollutants was insignificant. This indicates that all models after tuning hyperparameters performed similarly on train and test data. The performance metrics for NO_2 pollutant were found to be considerably more substantial than PM_{10} and $PM_{2.5}$.

In Figure A1, the performance metrics of Am Neckartor are shown. Considering the R^2 score, it was similar to that of PM_{10} and $PM_{2.5}$ at Marienplatz. However, for NO_2, the models performed poorly regarding the test R^2 score. It can be seen that the R^2 score for NO_2 on the test data is negative, a clear scenario of underfitting which indicated that the models did not learn properly and new features were required. On the other hand, the MAE and RMSE for both train and test data for PM_{10} at Neckartor were twice as $PM_{2.5}$ at Am Neckartor. Common observations among both the performance metrics in Fig. 7 and A1 include overfitting of the Bagging-based methods especially for $PM_{2.5}$ and PM_{10}. The performance metrics obtained for NO_2 showed unsatisfactory results when compared to the other two pollutants. The RIDGE model performed poorly compared to the remaining four models in terms of low R^2 score and high MAE and RMSE for both train and test data.

In a related study (Zamani Joharestani et al., 2019), focusing solely on Delhi and employing meteorological parameters for $PM_{2.5}$ estimation, the best performing model, a combination of extra trees and AdaBoost achieved a MAE of 14.3 $\mu g/m^3$. In contrast, the MAE values for Scenario 1 were ranging between 6.2 and 6.4. In terms of NO_2 estimation, the results in Scenario 1 demonstrated better R^2 scores at both Marienplatz and Am Neckartor compared to the models presented in the study done by Zhiyuan et al. (Li et al., 2020).

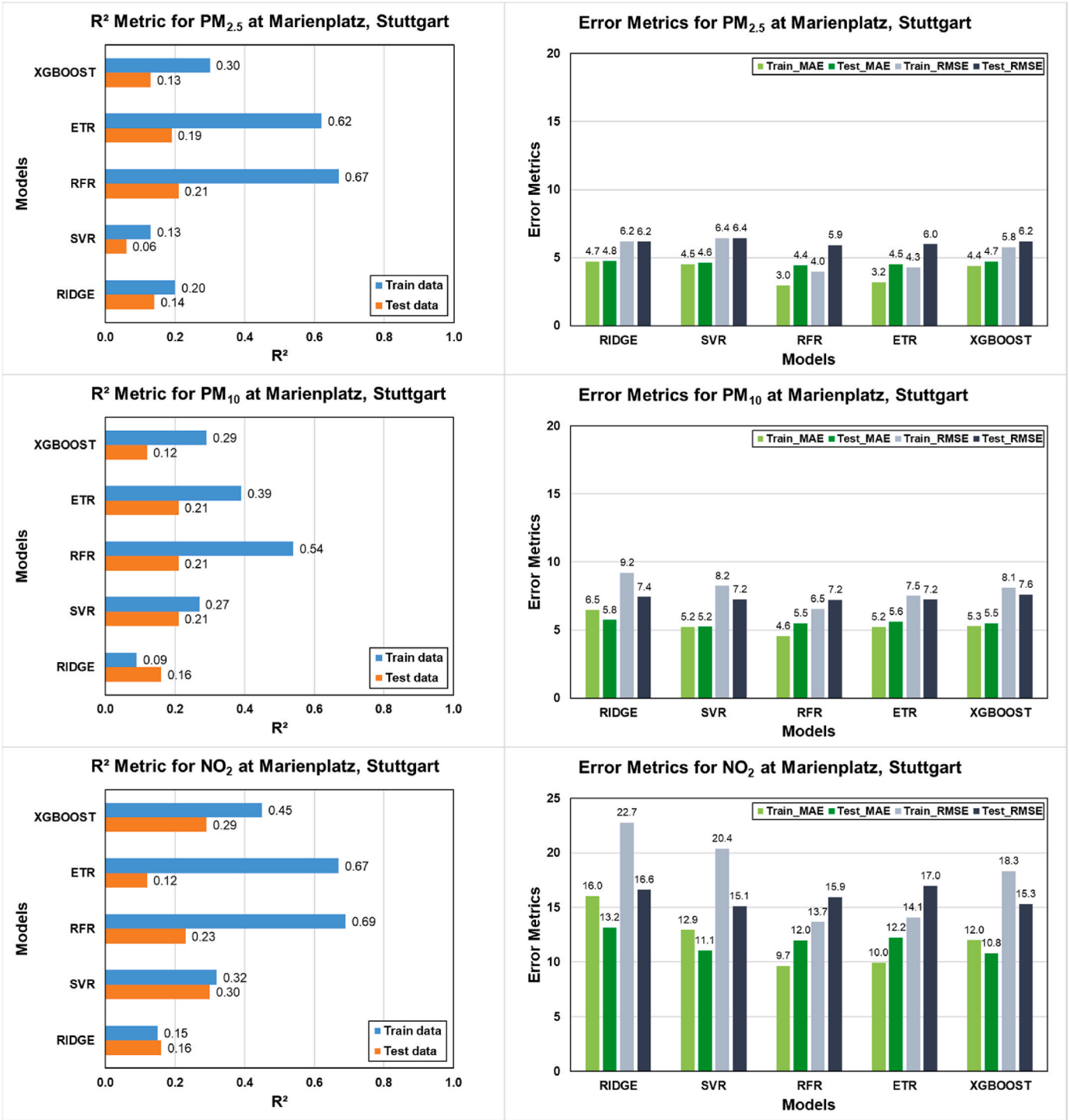


Fig. 7. Scenario 1 performance metrics for pollutants $PM_{2.5}$, PM_{10} and NO_2 at monitoring station Marienplatz, Stuttgart.

3.2. Scenario 2

In Scenario 2, in addition to the temporal and meteorological features, traffic counts, i.e., vehicles passing from the roads adjacent to the monitoring station were also considered. Traffic is one of the contributors to air pollution (Long and Carlsten, 2022). Hence, the models were retrained as mentioned in Scenario 1 with the addition of traffic as a new parameter to the existing meteorological and temporal features.

In Figure A2, presented in Appendix, the performance metrics at Marienplatz after adding the traffic feature are shown. The R^2 score, MAE and RMSE for train and test data for $PM_{2.5}$ and PM_{10} show no improvement compared to Scenario 1. Here too, the test MAE and RMSE were comparable to train MAE and RMSE values. All the ensemble methods still tend to overfit while the other two are underfitting. For NO_2, with respect to R^2 score, an improvement was observed across all the algorithms on both train and test data. However, only a small drop concerning MAE and RMSE values was noticed when compared to Fig. 7.

In Figure A3 in Appendix section, the performance metrics at Am Neckartor are shown after providing the traffic feature. For $PM_{2.5}$ and PM_{10}, there was no change in performance metrics with respect to Scenario 1. All the performance metrics remained unchanged even after the addition of traffic. However, for NO_2, there was an improvement

observed. It is assumed to be that since the monitoring station is adjacent to the traffic source (federal highway) and the traffic emissions contribute more to NO_2 than particulate matter (LUBW and Landesanstalt für Umwelt Baden-Württemberg, 1999), the NO_2 concentrations can be predicted better than particulate matter. The R^2 score of the training data set increased and a reduction in MAE and RMSE was seen when compared to NO_2 in Figure A1. The R^2 nearly doubled from 0.35 to 0.6, while there was a reduction in train MAE and RMSE by approximately 15%–20%. On the test data set, when the same models were evaluated still the R^2 was negative but a slight decrease was noticed. A similar scenario was observed even with test RMSE and MAE.

The key observations in Scenario 2 were that by adding traffic no significant improvement was noticed in the performance metrics of $PM_{2.5}$ and PM_{10} at both Marienplatz and Neckartor. However, there was an improvement in NO_2 performance metrics at both locations. This positive effect was noticeable more at Am Neckartor than at Marienplatz as Am Neckartor is a traffic hot spot and hence more sensitive to traffic.

In reference to the study performed by Rybarczyk and Zalakeviciute (2017), the authors primarily focused on developing a regression model with weighted coefficients for estimating $PM_{2.5}$ for which an RMSE value of 7.3 was obtained by incorporating meteorological parameters and traffic. Interestingly, it is similar to the results of Scenario 2, where all the models yielded RMSE values ranging from 5.9 to 6.5.

The difference between the two studies is the treatment of traffic, as the current research study used vehicle count data measured through sensors, whereas the study mentioned above (Rybarczyk and Zalakeviciute, 2017) categorized traffic as low, medium, and high.

3.3. Scenario 3

In this Scenario 3, apart from the features mentioned in the previous scenario, a similar pollutant concentration was added as an extra feature from the background station at Bad Cannstatt. Hence, to estimate the $PM_{2.5}$ concentration at Marienplatz and Am Neckartor apart from temporal, meteorological and traffic features, the $PM_{2.5}$ concentration at Bad-Cannstatt was also provided. Similarly, to estimate PM_{10} and NO_2 concentrations at both Marienplatz and Am Neckartor PM_{10} and NO_2 from Bad Cannstatt were provided respectively. The key idea was that air pollutant concentration at nearby stations is correlated with the concentration at the site under consideration due to the dispersion and advection of air pollutants in the area. The distance between the monitoring stations Am Neckartor and Bad Cannstatt is around 4 km, while between Marienplatz and Bad Cannstatt is around 7 km.

Figure A4 in Appendix shows that the performance metrics at Marienplatz improved when compared to the previous two scenarios. For all the pollutants, one common observation was that the effect of underfitting reduced compared to Scenario 2. This can be seen especially by comparing the R^2 scores of all the models for each pollutant. To start with the pollutant $PM_{2.5}$, improved R^2 scores were observed for the train data compared to $PM_{2.5}$ in Scenario 2. The ensemble methods outperformed the RIDGE and SVR, also the MAE and RMSE for the test data were reduced to half. When the same models were applied to the test data, a similar performance was noticed in all performance metrics. The ML models were able to capture the trends and learn better, hence generalizing well on the new unseen test data. The main reason could be attributed to the fact that $PM_{2.5}$ at Marienplatz which was the pollutant to be estimated had a strong co-relation with $PM_{2.5}$ at Bad Cannstatt (0.76). Similarly, for PM_{10} and NO_2, a similar pattern was noticed. In the case of PM_{10}, the MAE, and RMSE were reduced by nearly 40%. However, for NO_2 a reduction of 25% was seen.

Figure A5 in Appendix depicts the performance metrics of Am Neckartor. For $PM_{2.5}$ at Am Neckartor, a decent generalization was observed with respect to training and test data across all models. This can be seen by observing the respective performance metrics of R^2, MAE and RMSE for the train and test data. When compared to Marienplatz, better performance was achieved due to the higher correlation of respective parameters. For PM_{10}, compared to Scenario 2, the MAE and RMSE were reduced by 4 $\mu g/m^3$, also an increase in the R^2 score was noticed. The ensemble models especially XGBOOST showed the best performance across the train and test in all performance metrics. The R^2 scores were positive for NO_2 at Am Neckartor, still, a substantial difference between train and test results existed. The train and test MAE and RMSE were reduced by 20% when compared to performance metrics of NO_2 for Scenario 2. Finally, from Scenario 3 it was observed that providing the background pollutants had a positive effect on the prediction model results, which was highly prominent in the case of $PM_{2.5}$ followed by PM_{10} and then NO_2. Also, this indicated the capabilities of ML models to have a reasonable pollutant concentration estimation with even having pollutant from one monitoring station included as a feature. Thus, the effect of pollutant concentration as an input feature showed a significant impact on predicting the pollutant concentrations.

3.4. Scenario 4

In Scenario 4, to the features mentioned in Scenario 2 (temporal, meteorological and traffic), pollutant concentrations from the monitoring stations as mentioned in Table 1 were provided as input features to check if adding them as input features to the model can further improve the model evaluations. No pollutants from Marienplatz and Am Neckartor were provided as input features as the pollutants from those monitoring stations were to be modeled. To estimate the NO_2 concentration at Am Neckartor and Marienplatz, not only the NO_2 from the remaining stations provided but also PM_{10}, NO, and O_3 concentrations were considered as input features. In this manner, the effect of cross-sensitivity between the pollutants can be established.

In Fig. 8, the performance metrics for all three pollutants for Scenario 4 are displayed. Compared to scenario 3, there is a slight improvement in performance concerning $PM_{2.5}$ and PM_{10} performance metrics. For $PM_{2.5}$ the MAE was below 2 $\mu g/m^3$ and RMSE was below 2.5 $\mu g/m^3$ for all models. Even in the case of PM_{10}, MAE was around 2.5 $\mu g/m^3$ for all models on both train and test data. All the models seem to be performing well for $PM_{2.5}$. For PM_{10} and NO_2, RFR, ETR, and XGBOOST performed better because of their ability to capture non-linearity. Another notable comparison to scenario 3 performance metrics was that NO_2 prediction improved by nearly 20%, keeping the RMSE value below 10 $\mu g/m^3$ for all the models except RIDGE. This further signifies the ability of ensemble methods to capture the non-linear relationship.

The performance metrics for all three pollutants for Am Neckartor are given in Figure A6 in Appendix. Similar to Marienplatz, the $PM_{2.5}$ and PM_{10} performance metrics improved slightly. However, significant improvement was seen in the performance of NO_2 metrics. Compared to NO_2 in Scenario 3, the performance of the models improved considerably. The R^2 metric for train data which was previously around in the range of 0.6–0.8 further increased up to 0.9, whereas for the test data, an improvement was noticed reaching a value between 0.7 and 0.8. Even the phenomenon of underfitting was eliminated, achieving a good amount of generalization. Also, the error metrics MAE and RMSE were halved for the train and test data. Thus, pollutant concentrations from other monitoring stations played an outstanding role in NO_2 concentration estimation.

In reference to the study performed by Rybarczyk and Zalakeviciute (2017), the authors observed an improvement in $PM_{2.5}$ concentration prediction after including trace gases as a feature. This outcome aligns with the results of Scenario 4. It is worth mentioning that this study (Rybarczyk and Zalakeviciute, 2017) considered data for two months, however the current research prediction set comprised a complete year. Additionally, the study done by Kumar et al. (2020b) estimated pollutant concentrations using conventional methods such as Inverse Distance Weighting (IDW) and kriging on a monthly basis at various sites in Delhi. The IDW and kriging methods exhibited an average percentage error of around 45% for NO_2. In contrast, the results in Scenario 4 showed an absolute percentage error of approximately 25% for NO_2 at

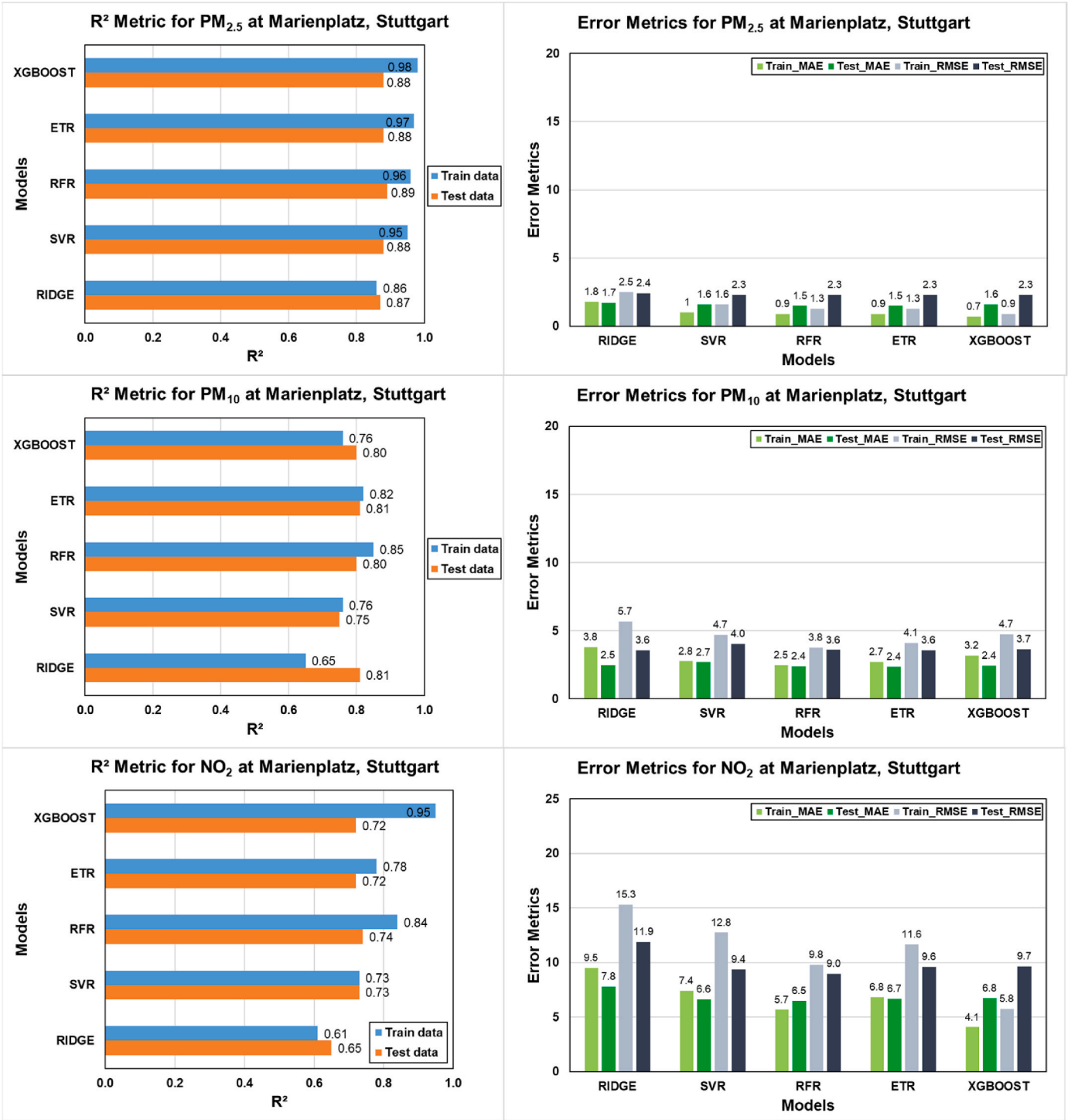


Fig. 8. Scenario 4 performance metrics for pollutants $PM_{2.5}$, PM_{10} and NO_2 at monitoring station Marienplatz, Stuttgart.

Marienplatz and Am Neckartor. Consequently, when compared to conventional methods, the developed approach resulted in an average reduction in percentage error of around 20%.

Since Scenario 4 proved to be the best one compared to other scenarios, a residual error plot between the actual and predicted $PM_{2.5}$, PM_{10} and NO_2 concentrations was made for this scenario that is shown in Fig. 9 at the location Am Neckartor. The results for the location Marienplatz can be seen in Figure A7 in Appendix section. Residuals are the difference between actual and predicted outcomes. The advantage of the residual plot is that the overall range of MAE for individual time steps can be observed. These residual plots are based on the test data. For every time step (hour) three estimations namely minimum, average and maximum were obtained from the models. For the following results, only the average values of all model outputs were considered as the predicted outcome. One reason to use averaged outcomes was that the overall MAE decreased slightly. After averaging the results MAE for $PM_{2.5}$ at Marienplatz and Neckartor was 1.4 and 1.1 $\mu g/m^3$ respectively. For NO_2, at Marienplatz and Neckartor, the average MAE was 6.3 and 5.3 $\mu g/m^3$ respectively. Thus, by averaging a small reduction of MAE was obtained, when compared with individual models test MAE in Scenario 4.

When the predictions were near to actual concentrations, the

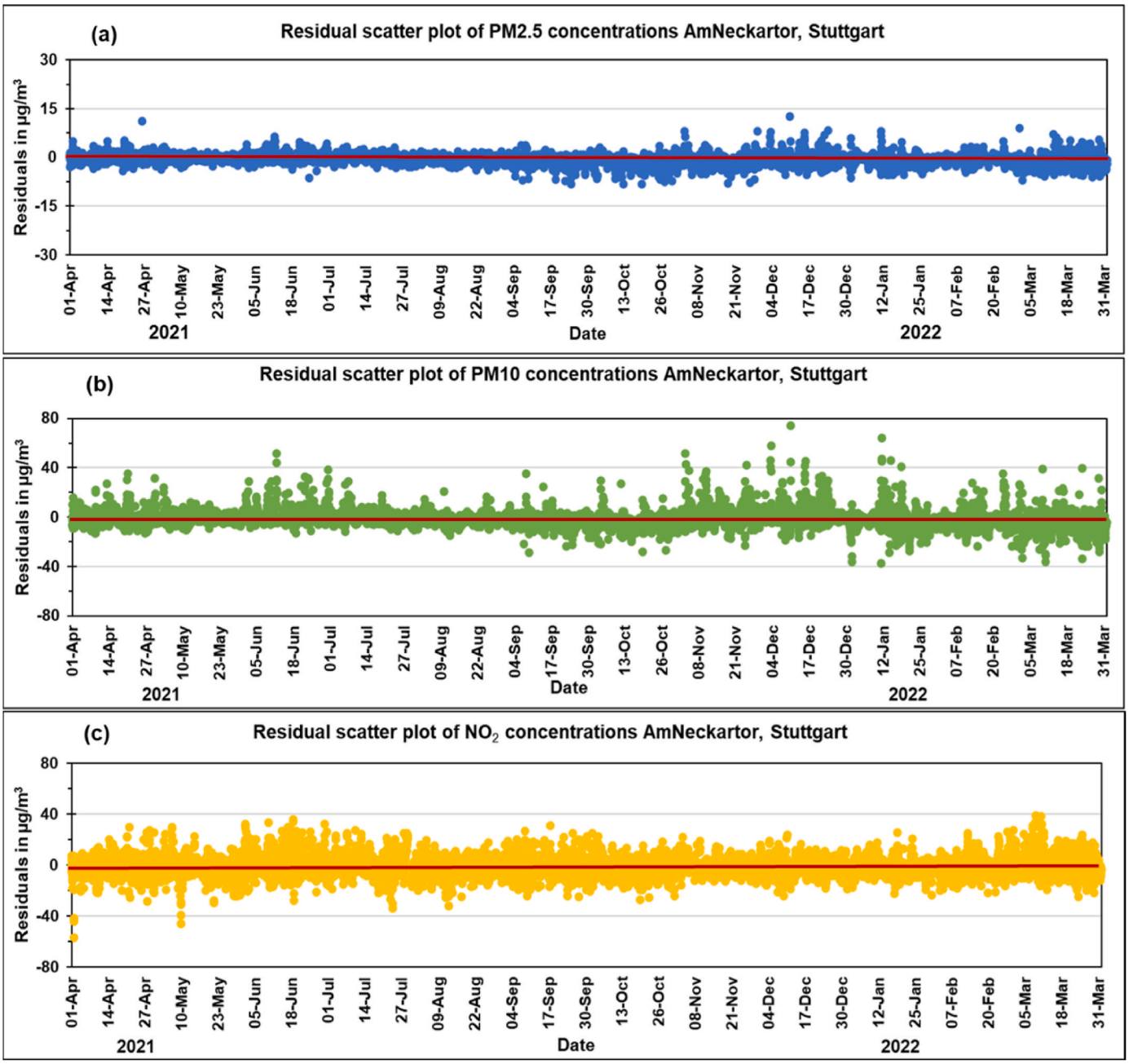


Fig. 9. Residual error plots of pollutants (a) $PM_{2.5}$, (b) PM_{10} and (c) NO_2 at Am Neckartor, Stuttgart.

residual error was close to zero (indicated via the red line). For all three pollutants a common observation was that the residuals were centered around zero, not inclining towards either side heavily, which indicated that models can be used for future evaluation. However, a few observations at the Marienplatz location were seen having residual errors of ± 15 µg/m^3 which could be potentially linked to some specific events that are particular to that location.

The $PM_{2.5}$ residual errors with respect to pollutants at Am Neckartor showed a better fit compared to $PM_{2.5}$ at Marienplatz. However, for PM_{10} (green line) in December to January, some extreme outliers were noticed, which were not captured by the models. For NO_2, the residual errors observed were within the range of ± 15 µg/m^3.

3.5. Summary

In this section, a comparison of every pollutant across all four scenarios is presented. The predicted outcomes are the minimum, average and maximum pollutant concentrations. Fig. 10 presents the predicted and actual concentrations of $PM_{2.5}$ across all four scenarios on the test data at Marienplatz. Hourly pollutant concentrations were estimated using the ML model, which were averaged for 24 h for these graphs. Since the test data is spread from April 2021 to March 2022, for better understanding, the complete year is divided into four quarters from Q1 to Q4. For scenario 1 and scenario 2, a notable deviation was observed for predicted and actual concentration values. Also, the predicted outcomes were unable to cover sudden increases in concentration values, which is visible in Q2 and Q4. For better understanding, a traceback of train data was performed, where the range of $PM_{2.5}$ during June was found to be between 7 and 12 µg/m^3. So, estimating $PM_{2.5}$ with meteorological and traffic parameters led to a mediocre performance. However, in scenario 3, when pollutant concentration was provided, the trends were captured across all four quarters. However, in

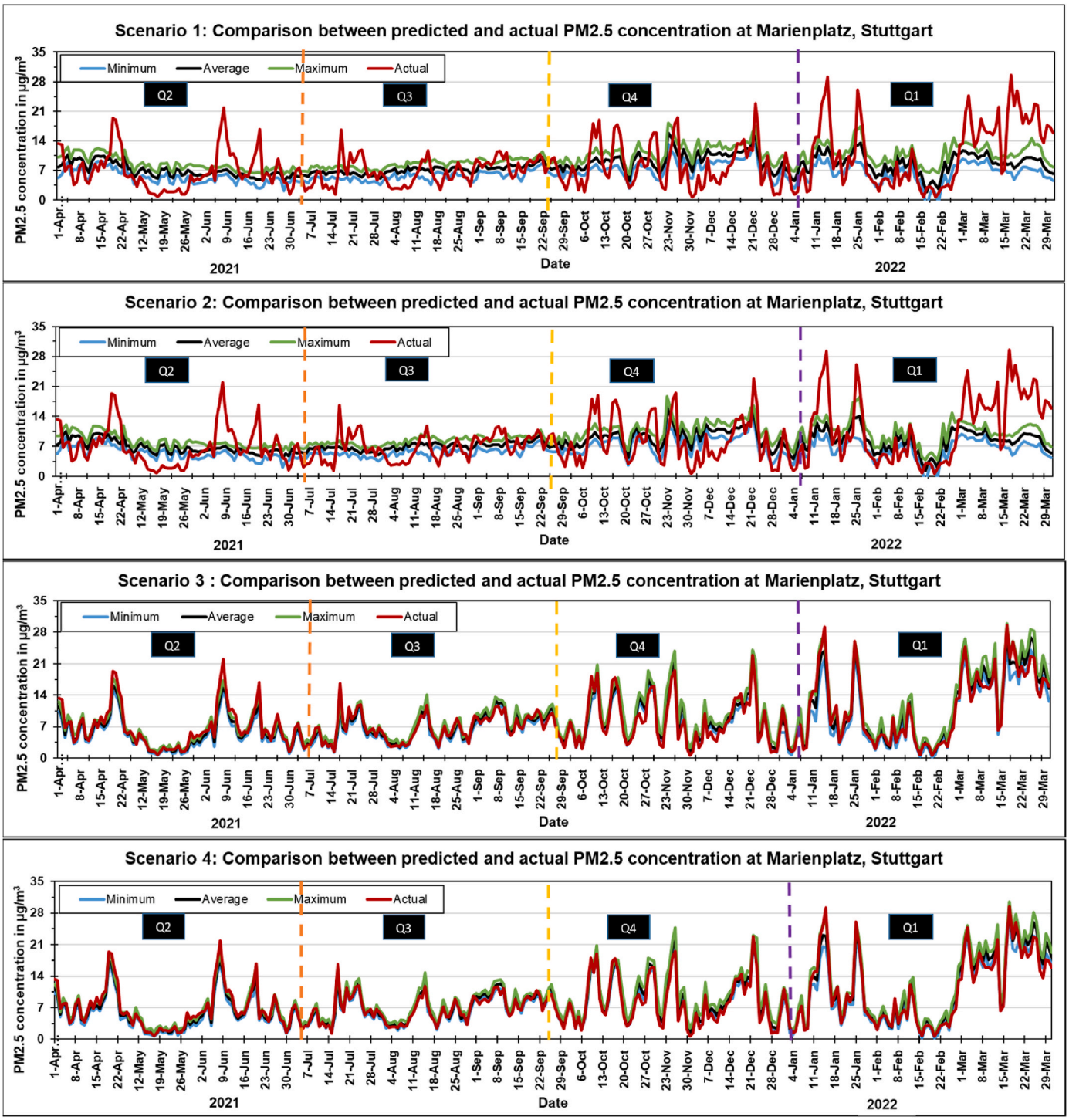


Fig. 10. Comparison of $PM_{2.5}$ across different scenarios at Marienplatz, Stuttgart.

Q4, there was a slight overestimation of pollutant concentration. In scenario 4, even after providing all the pollutants, no significant improvement was seen. Nevertheless, the performance metrics show a small improvement in the R^2 score and an overall decrease in test MAE and RMSE values. A similar phenomenon was noticed concerning $PM_{2.5}$ at Am Neckartor across all four scenarios shown in Figure A8 in Apeendix.

In Fig. 11 and A9 in Appendix PM_{10} is compared across all four scenarios at both the locations of Marienplatz and Am Neckartor respectively. For Scenario 1 and 2, the models were unable to capture the sudden changes similar to the previous results. However, for PM_{10} concentration at Am Neckartor in scenarios 1 and 2, the sudden changes were followed better compared to the PM_{10} concentration at Marienplatz. Still, a deviation with respect to the actual values was detected. By adding the spatial pollutant concentrations, the performance was improved, which can be observed in scenarios 3 and 4.

The NO_2 comparison across all scenarios for locations of Marienplatz and Am Neckartor are presented in Fig. 12 and A10 in Appendix respectively. The impact of providing traffic was found to be minimal at Marienplatz. However, the effect of providing the pollutants from other monitoring stations was widely noticed even in the case of NO_2 at both locations. For the NO_2 results at Am Neckartor, a significant

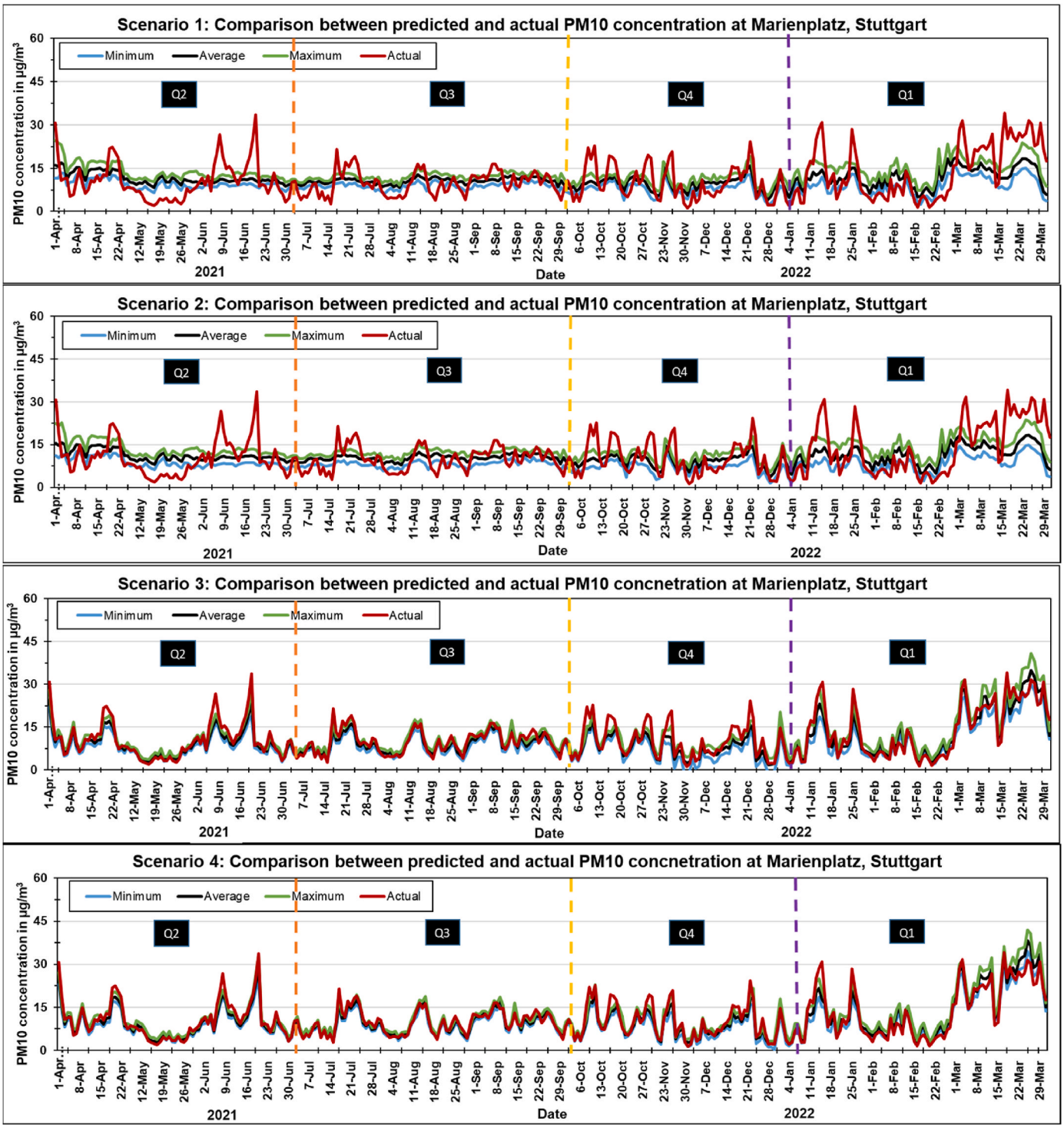


Fig. 11. Comparison of PM_{10} across different scenarios at Marienplatz, Stuttgart.

improvement was observed between the predicted and actual values in every scenario. Finally, the best performance was obtained in scenario 4, where the predicted and actual values showed a strong correlation.

3.6. Feasibility of the proposed concept

In scenario 4, it was observed that adding pollutants from other monitoring stations resulted in enhanced performance of the ML models. A similar concept was applied to a monitoring station Nordwest in Karlsruhe, to check the feasibility of the developed method. The distance between this station and two stations in Stuttgart is around 75 km. Two main reasons for choosing this particular location were the ease of data access and the availability of monitoring stations nearby measuring the required parameters. However, no traffic data was available on this location. Meteorological parameters were available at Karlsruhe Nordwest, which included air temperature, air pressure, precipitation and wind speed. The same pollutants $PM_{2.5}$, PM_{10} and NO_2 were modeled at this location. Table 3 shows the list of monitoring stations in Karlsruhe along with the measured parameters.

To estimate $PM_{2.5}$, PM_{10} and NO_2 at Karlsruhe Nordwest,

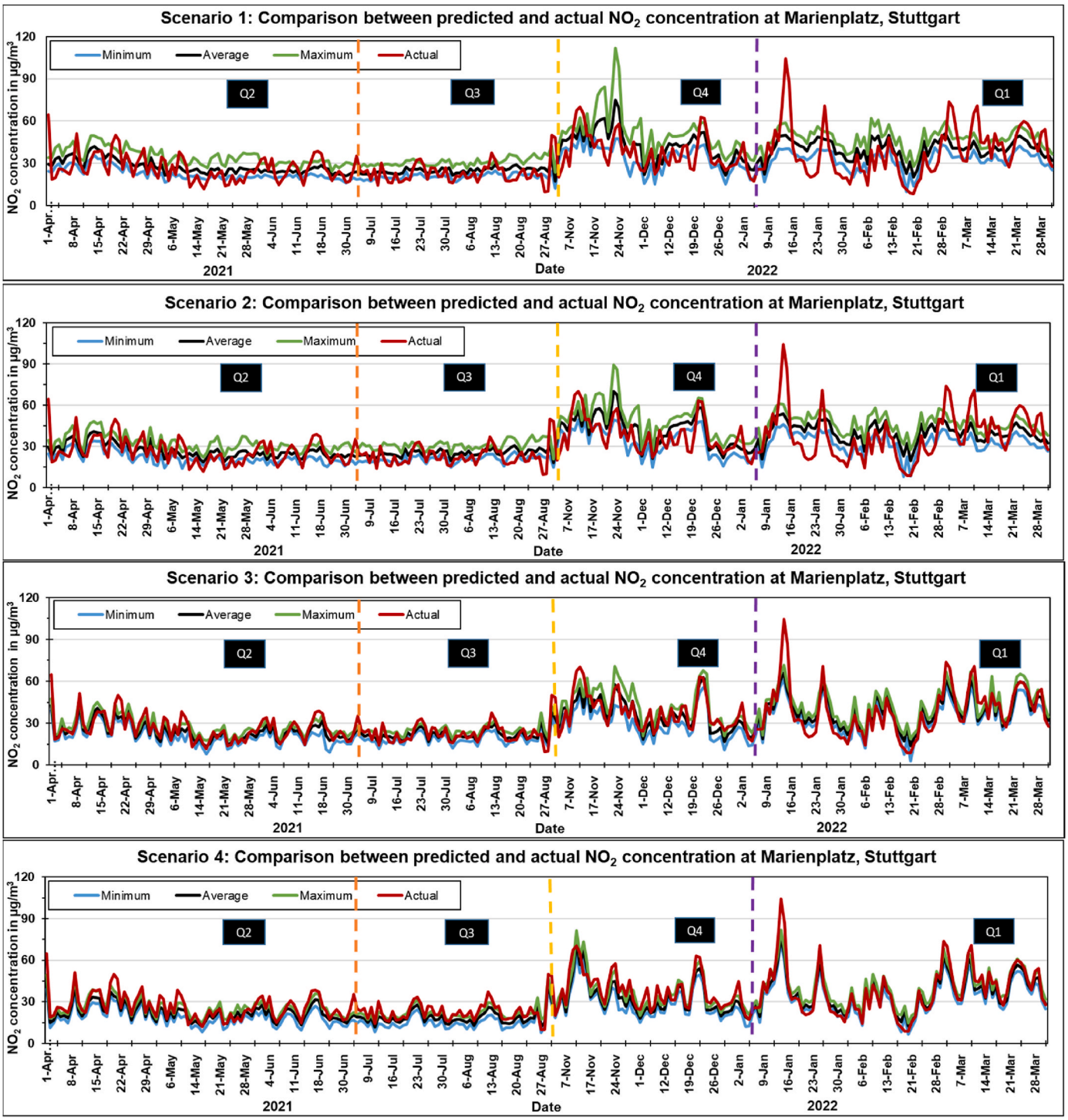


Fig. 12. Comparison of NO_2 across different scenarios at Marienplatz, Stuttgart.

Table 3
LUBW Monitoring stations in Karlsruhe with the list of pollutants being measured.

Monitoring stations	Pollutants monitored
Karlsruhe Nordwest	$PM_{2.5}$, PM_{10}, NO_2, NO, O_3
Eggenstein	$PM_{2.5}$, PM_{10}, NO_2, NO, O_3
Reinhold Frank Straße	$PM_{2.5}$, NO_2, NO
Pflinztal Karlsruher Straße	NO_2, NO

meteorological parameters and pollutants from the remaining three monitoring stations were given as inputs. To estimate PM_{10} at Karlsruhe Nordwest, all four measured pollutants from Eggenstein, three from Reinhold Frank Straβe and two from Pflinztal Karlsruher Straβe were provided to all the five ML models. In Figure A11 in Appendix section, the performance metrics of the results are shown. The results show that the performance of $PM_{2.5}$ was comparatively better than the other two pollutants because of their homogeneous distribution. For $PM_{2.5}$, all the models obtained an R^2 score of above 0.9 on both train and test data. Performance metrics displayed similar results and a very low MAE in the range of 1–2 $\mu g/m^3$ was obtained on train and test data. For PM_{10}, higher

R^2 values were obtained for SVR and ensembles compared to RIDGE. For all algorithms, the R^2 were in the range of 0.92–0.94 on the train data and 0.83 to 0.84 on the test data. From the PM_{10} performance metrics, the MAE for all the models was around 1.3–1.6 μg/m^3 on the train data and 2.1–2.9 μg/m^3 on the test data. With this performance, the models were neither underfitting nor overfitting and decent generalization was noticed on new unseen data. The performance metrics for NO_2 pollutant showed similar results.

The comparison between the predicted and actual pollutant values is presented in Fig. 13. Here instead of predicted minimum, average and maximum hourly values, for better visualization daily plot is presented. It can be seen from all three subplots that the pollutants were able to learn the trends and also capture the fluctuations during the entire time duration of test data.

In Figure A12 in Appendix, the residual plot of pollutants $PM_{2.5}$, PM_{10} and NO_2 at Nordwest, Karlsruhe is shown. Here, the average values from all 5 models were taken and subtracted from the actual value. Most of the residuals lie in the range of ±7.5 μg/m^3. However, some deviations were noticed during January to March where the models were underperforming, with room for improvement. Also, for PM_{10} and NO_2 the residuals were observed to lie within the range of ±20 μg/m^3. When compared to PM_{10} and NO_2 residual plots of Marienplatz and Am Neckartor, the residuals at Karlsruhe Nordwest were lower.

4. Conclusions

In this research, four different scenarios were explored to investigate the performance of ML models for estimating pollutant concentration. From the results, it can be concluded that the pollutants from other monitoring stations as an input feature, played a significant role in estimation. In each scenario, an improvement in performance was seen with the addition of a new feature.

In scenario 1, a mediocre performance was obtained as the models were unable to capture any fluctuations and were only able to detect simple moving averages. Also, in this scenario overfitting of ensemble models (RF, ETR, XGBoost) and underfitting of RIDGE and SVR were observed, across all the three pollutants at both locations. The MAE for the pollutant $PM_{2.5}$ was similar for both locations, however, the MAE for PM_{10} and NO_2 at Neckartor was twice compared to Marienplatz. To enhance the predicting ability, in scenario 2, traffic data was added as an extra feature to explore its impact. An improvement was observed for pollutant NO_2 at Am Neckartor. However, no improvements were observed for the remaining pollutants at both locations. After adding

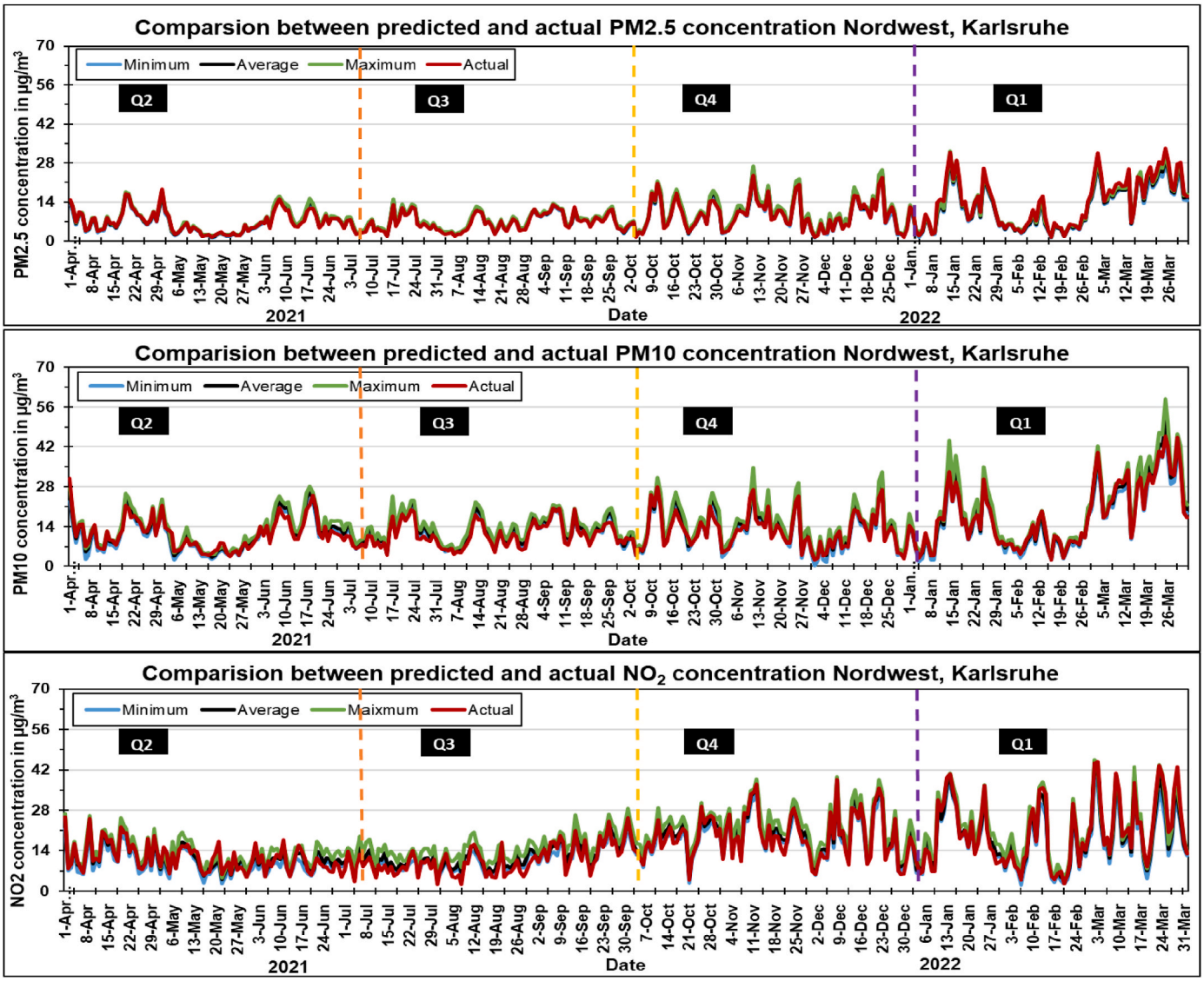


Fig. 13. Comparison of performance metrics for pollutants (a) $PM_{2.5}$, (b) PM_{10} and (c) NO_2 predicted and actual values at Nordwest, Karlsruhe.

pollutants from a background monitoring station (Bad Cannstatt) in scenario 3, significant improvement was observed across all pollutants in both locations. At Am Neckartor the impact of providing the background pollutants was more visible. A similar phenomenon was observed at Marienplatz where the MAE for PM_{10} and NO_2 were reduced. Another notable aspect in this scenario 3, was that the problem of overfitting and underfitting was eliminated. Finally, in scenario 4, when pollutants from other monitoring stations were also added to the existing features in scenario 2, the best possible performance was obtained with the lowest MAE for all the pollutants. The impact was more prominent for NO_2 at both Marienplatz and Am Neckartor. However, in the case of $PM_{2.5}$ and PM_{10} there was only a slight decrease in MAE for both locations was observed. The results from residual plots for scenario 4 showed that the models were able to capture most of the trends and achieve decent generalizing ability. The comparison of this research with existing approaches for pollutant estimation revealed the effectiveness of the developed method in achieving accurate results. Additionally, this technique outperformed conventional methods such as inverse distance weighting and kriging.

Finally, to estimate $PM_{2.5}$, PM_{10} and NO_2 at Karlsruhe Nordwest, a similar set of features as described in scenario 4 was employed without traffic data as it was unavailable. After evaluating the results, it was concluded that the ML models were able to estimate pollutant concentrations at other locations as well. Hence, the developed technique can be transferred to any location where pollutant prediction is required.

The scope of this research in the field of air quality monitoring can be significant as by applying this method, the monitoring stations can be replaced with ML models, creating a virtual monitoring station. In this manner dependency on the monitoring stations can be reduced and high costs can be avoided. This can also assist the respective authorities to identify the minimum number of monitoring stations to achieve maximum coverage within a city. Furthermore, this study can benefit from incorporating other neural networks such as CNN-LSTM (Convolution Neuron Network - Long Short Term Memory) in capturing temporal dependencies and patterns in data, which could further enhance the accuracy of pollutant estimation. The limitation of this study is that the forecasting of pollutant concentration is not possible as the data from other monitoring stations is required for prediction. Further research can be done in this regard so that the pollutant concentrations can be forecasted for the next days which would enable proactive measures and decision-making based on anticipated air quality conditions.

CRediT authorship contribution statement

A. Samad: Conceptualization, Methodology, Software, Data curation, Investigation, Writing – original draft, Supervision, Writing – review & editing. **S. Garuda:** Conceptualization, Methodology, Software, Data curation, Investigation, Writing – original draft, Formal analysis, Visualization, Writing – review & editing. **U. Vogt:** Conceptualization, Methodology, Supervision, Writing – review & editing. **B. Yang:** Supervision, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.atmosenv.2023.119987.

References

Akoglu, H., 2018. User's guide to correlation coefficients. Turkish journal of emergency medicine, Bd. 18 (Nr. 3, S), 91–93.

Analitis, A., Barratt, B., Green, D., Beddows, A., Samoli, E., Schwartz, J., Katsouyanni, K., 2020. Prediction of pm2.5 concentrations at the locations of monitoring sites measuring pm10 and nox, using generalized additive models and machine learning methods: a case study in london. Atmospheric Environment, Bd. 240 (S), 117757 [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1352231020304891.

Balamurugan, V., Balamurugan, V., Chen, J., 2022. Importance of ozone precursors information in modelling urban surface ozone variability using machine learning algorithm. Scientific Reports, Bd. 12 (S), 5646.

Baumbach, G., 1996. Air Quality Control. Formation and Sources, Dispersion, Characteristics and Impact of Air Pollutants - Measuring Methods, Techniques for Reduction of Emissions and Regulations for Air Quality Control. Springer Berlin Heidelberg (Environmental Engineering), Berlin, Heidelberg.

Botchkarev, A., 2018. Performance Metrics (Error Measures) in Machine Learning Regression, Forecasting and Prognostics: Properties and Typology" arXiv preprint arXiv:1809.03006.

Chelani, A., 2009. Prediction of daily maximum ground ozone concentration using support vector machine. Environmental monitoring and assessment, Bd. 162 (S), 169–76.

Chen, G., Li, S., Knibbs, L.D., Hamm, N., Cao, W., Li, T., Guo, J., Ren, H., Abramson, M.J., Guo, Y., 2018. A machine learning method to estimate pm2.5 concentrations across China with remote sensing, meteorological and land use information. Science of The Total Environment, Bd. 636 (S), 52–60. https://www.sciencedirect.com/science/article/pii/S0048969718314281.

Demertzis, K., Bougoudis, I., Liadis, L., 2015. "Hisycol a hybrid computational intelligence system for combined machine learning: the case of air pollution modeling in athens". Neural Computing and Applications, Bd 27, 5.

Dong, G., Liu, H. (Eds.), 2018. Feature Engineering for Machine Learning and Data Analytics, first ed. CRC Press. https://doi.org/10.1201/9781315181080.

Du, P., Bai, X., Tan, K., Xue, Z., Samat, A., Xia, J., Li, E., Su, H., Liu, W., 2020. Advances of four machine learning methods for spatial data handling: a review. J geovis spat anal 4, 13. https://doi.org/10.1007/s41651-020-00048-5, 2020.

Duyzer, J., van den Hout, D., Zandveld, P., van Ratingen, S., 2015. Representativeness of air quality monitoring networks. Atmos. Environ. 104, 88–101. https://doi.org/10.1016/j.atmosenv.2014.12.067, 2015.

EU, 2015a. Consolidated Text: Directive 2008/50/EC of the European Parliament and of the Council of 21 May 2008 on Ambient Air Quality and Cleaner Air for Europe. http://data.europa.eu/eli/dir/2008/50/2015-09-18.

EU, 2015b. Commission Directive (EU) 2015/1480 of 28 August 2015 Amending Several Annexes to Directives 2004/107/EC and 2008/50/EC of the European Parliament and of the Council Laying Down the Rules Concerning Reference Methods, Data Validation and Location of Sampling Points for the Assessment of Ambient Air Quality. Available online: http://data.europa.eu/eli/dir/2015/1480/oj.

European Environment Agency (EEA), 2022. Air Quality in Europe 2022 Report. Publications Office. https://doi.org/10.2800/488115. ISBN: 978-92-9480-515-7, ISSN: 1977-8449.

Filonchyk, M., Hurynovich, V., Yan, H., Yang, S., 2020. Atmospheric pollution assessment near potential source of natural aerosols in the South Gobi Desert region, China. GIScience Remote Sens. 57 (2), 227–244. https://doi.org/10.1080/15481603.2020.1715591.

Geurts, P., Ernst, D., Wehenkel, L., 2006. Extremely randomized trees. Machine learning, Bd. 63 (Nr. 1, S), 3–42.

Ghaemi, Z., Alimohammadi, A., Farnaghi, M., 2018. Lasvm-based big data learning system for dynamic prediction of air pollution in tehran. Environmental Monitoring and Assessment, Bd. 190, 4.

Goodfellow, I., Bengio, Y., Courville, A., 2016. Deep Learning. MIT press. ISBN: 9780262035613.

Han, Y., Wu, J., Zhai, B., Pan, Y., Huang, G., Wu, L., Zeng, W., 2019. Coupling a bat algorithm with xgboost to estimate reference evapotranspiration in the arid and semiarid regions of China. Advances in Meteorology, Bd. (S), 1–16, 102019.

He, L., Cheng, Y., Li, Y., Li, F., Fan, K., Li, Y., 2021. An improved method for soil moisture monitoring with ensemble learning methods over the Tibetan plateau. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, Bd. PP, S. 1–1.

Hu, K., Sivaraman, V., Bhrugubanda, H., Kang, S., Rahman, A., 2016. Svr based dense air pollution estimation model using static and wireless sensor network. In: 2016 IEEE SENSORS, pp. 1–3. S.

Hubert, M., Vandervieren, E., 2008. An adjusted boxplot for skewed distributions. Computational Statistics Data Analysis, Bd. 52 (Nr. 12, S), 5186–5201 [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167947307004434.

Isam Drewil, G., Jabbar Al-Bahadili, R., 2022. Air pollution prediction using LSTM deep learning and metaheuristics algorithms. Measurement: Sensors 24 (2022), 100546. https://doi.org/10.1016/j.measen.2022.100546. ISSN 2665-9174.

Khodakarami, J., Ghobadi, P., 2016. Urban pollution and solar radiation impacts. Renewable and Sustainable Energy Reviews, Bd. 57. S. 965–976.

Khomenko, S., Cirach, M., Pereira-Barboza, E., Mueller, N., Barrera-Gómez, J., Rojas-Rueda, D., de Hoogh, K., Hoek, G., Nieuwenhuijsen, M., 2021. Premature mortality due to air pollution in European cities: a health impact assessment. The Lancet Planetary Health, Bd. 5, e121–e134. Nr. 3, S. https://www.sciencedirect.com/science/article/pii/S2542519620302722.

Kumar, S., Mishra, S., Singh, S.K., 2020a. A machine learning-based model to estimate pm2.5 concentration levels in Delhi's atmosphere. Heliyon, Bd. 6 (Nr. 11, S), e05618

[Online]. Available: https://www.sciencedirect.com/science/article/pii/S2405844020324610.

Kumar, A., Dhakhwa, S., Dikshit, A.K., 2020b. Comparative evaluation of fitness of interpolation techniques of ArcGIS using leave-one-out scheme for air quality mapping. J geovis spat anal 6, 9. https://doi.org/10.1007/s41651-022-00102-4, 2022.

Latini, G., Grifoni, R.C., Passerini, G., 2002. Influence of meteorological parameters on urban and suburban air pollution. WIT Transactions on Ecology and the Environment, Bd. 53.

Li, Z., Yim, S.H.-L., Ho, K.-F., 2020. High temporal resolution prediction of street-level pm2.5 and nox concentrations using machine learning approach. Journal of Cleaner Production, Bd. 268 (S), 121975 [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0959652620320229.

Liaw, A., Wiener, M., 2001. Classification and regression by randomforest. Forest, Bd. 23, 11.

Long, E., Carlsten, C., 2022. Controlled human exposure to diesel exhaust: results illuminate health effects of traffic-related air pollution and inform future directions. Particle and Fibre Toxicology, Bd. 19 (Nr. 1, S), 1–35.

LUBW, Landesanstalt für Umwelt Baden-Württemberg, 1999. Wirkungen von Emissionen des Kfz Verkehrs auf Pflanzen und die Umwelt“ [Online; accessed October 26, 2021]. [Online]. Available: https://pudi.lubw.de/detailseite/-/publication/12203-Wirkungen_von_Emissionen_des_Kfz-Verkehrs_auf_Pflanzen_und_die_Umwelt_-_Literaturstudie.pdf.

Maqsood, I., Khan, M., Abraham, A., 2004. An ensemble of neural networks for weather forecasting“. Neural Computing and Applications, Bd 13 (S), 112–122.

McDonald, G.C., 2009. Ridge regression. Wiley Interdisciplinary Reviews: Computational Statistics, Bd. 1 (Nr. 1, S), 93–100.

Mosley, S., 2014. Environmental History of Air Pollution and Protection, the Basic Environmental History. Springer, pp. 143–169, 2014, S.

Rybarczyk, Y., Zalakeviciute, R., 2017. Regression Models to Predict Air Pollution from Affordable Data Collections“. IntechOpen. https://doi.org/10.5772/intechopen.71848. Kap. 2. [Online]. Available:

Saithanu, K., Mekparyup, J., 2014. Using multiple linear regression to predict pm 10 concentration in chonburi, Thailand. Global Journal of Pure and Applied Mathematics, Bd. (10), 835–839, 122014.

Samad, A., Vogt, U., 2020. Assessing the effect of traffic density and cold airflows on the urban air quality of a city with complex topography using continuous measurements. Modern Environmental Science and Engineering, Bd. 6 (S), 529–541.

Song, Z., Chen, B., Huang, Y., Dong, L., Yang, T., 2021. Estimation of pm2.5 concentration in China using linear hybrid machine learning model. Atmospheric Measurement Techniques, Bd. 14 (Nr. 8, S), 5333–5347 [Online]. Available: https://amt.copernicus.org/articles/14/5333/2021/.

Spangl, W., Schneider, J., Moosmann, L., Nagl, C., 2007a. Representativeness and Classification of Air Quality Monitoring Stations. Umweltbundesamt GmbH, Vienna, Austria. Available online: https://www.umweltbundesamt.at/fileadmin/site/publikationen/REP0121.pdf. (Accessed 10 August 2019).

Spangl, W., Schneider, J., Moosmann, L., Nagl, C., 2007b. Representativeness and Classification of Air Quality Monitoring Stations. Umweltbundesamt.

Stuttgart, Stadtklima 21, 2008. Grundlagen zum Stadtklima und zur Planung Stuttgart 21. Amt für Umweltschutz, Abt. Stadtklimatologie.

Sun, Z., Zhu, D., 2019. Exposure to outdoor air pollution and its human health outcomes: a scoping review. PLOS ONE, Bd. 14 (Nr. 5, S), 1–18. https://doi.org/10.1371/journal.pone.0216550 [Online]. Available:

Talebizadeh, M., Moridnejad, A., 2011. Uncertainty analysis for the forecast of lake level fluctuations using ensembles of ann and anfis models. Expert Systems with Applications, Bd. 38 (Nr. 4, S), 4126–4135 [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0957417410010328.

Tian, X., Cui, K., Sheu, H.-L., Hsieh, Y.-K., Yu, F., 2021. Effects of rain and snow on the air quality index, PM2.5 levels, and dry deposition flux of PCDD/fs. Aerosol and Air Quality Research, Bd. 21 (Nr. 8, S), 210158, 10.4209%2Faaqr.210158.

Vapnik, V., Golowich, S., Smola, A., 1996. Support vector method for function approximation, regression estimation and signal processing. In: Mozer, M., Jordan, M., Petsche, T., Bd, Hg (Eds.), Advances in Neural Information Processing Systems, vol. 9. MIT Press [Online]. Available: https://proceedings.neurips.cc/paper/1996/file/4 f284803bd0966cc24fa8683a34afc6e-Paper.pdf.

Verghese, S., Nema, A.K., 2022. Optimal design of air quality monitoring networks: a systematic review. Stoch. Environ. Res. Risk Assess. S. 1–16.

Vlasenko, A., Matthias, V., Callies, U., 2021. Simulation of chemical transport model estimates by means of a neural network using meteorological data. Atmospheric Environment, Bd. 254 (S), 118236 [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1352231021000546.

Wichard, J., Ogorzalek, M., 2004. Time series prediction with ensemble models. S.. In: 2004 IEEE International Joint Conference on Neural Networks (IEEE Cat. No.04CH37541), Bd., vol. 2, pp. 1625–1630. vol. 2.

Wolpert, D.H., 1992. “Stacked generalization“. Neural networks, Bd 5 (Nr. 2, S), 241–259.

Wong, P.-Y., Hsu, C.-Y., Wu, J.-Y., Teo, T.-A., Huang, J.-W., Guo, H.-R., Su, H.-J., Wu, C.-D., Spengler, J.D., 2021. Incorporating land-use regression into machine learning algorithms in estimating the spatial-temporal variation of carbon monoxide in taiwan. Environmental Modelling Software, Bd. 139 (S), 104996 [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1364815221000396.

Xing, J., Zheng, S., Ding, D., Kelly, J.T., Wang, S., Li, S., Qin, T., Ma, M., Dong, Z., Jang, C., et al., 2020. Deep learning for prediction of the air quality response to emission changes. Environmental science & technology, Bd. 54 (Nr. 14, S), 8589–8600.

Xu, Y., Liu, X., Cao, X., Huang, C., Liu, E., Qian, S., Liu, X., Wu, Y., Dong, F., Qiu, C.-W., Qiu, J., Hua, K., Su, W., Wu, J., Xu, H., Han, Y., Fu, C., Yin, Z., Liu, M., Roepman, R., Dietmann, S., Virta, M., Kengara, F., Zhang, Z., Zhang, L., Zhao, T., Dai, J., Yang, J., Lan, L., Luo, M., Liu, Z., An, T., Zhang, B., He, X., Cong, S., Liu, X., Zhang, W., Lewis, J.P., Tiedje, J.M., Wang, Q., An, Z., Wang, F., Zhang, L., Huang, T., Lu, C., Cai, Z., Wang, F., Zhang, J., 2021. Artificial intelligence: a powerful paradigm for scientific research. The Innovation, Bd. 2, 100179. Nr. 4, S. https://www.sciencedirect.com/science/article/pii/S2666675821001041.

Zamani Joharestani, M., Cao, C., Ni, X., Bashir, B., Talebiesfandarani, S., 2019. PM2.5 prediction based on random forest, xgboost, and deep learning using multisource remote sensing data. Atmosphere, Bd. 10 (Nr. 7) [Online]. Available: https://www.mdpi.com/2073-4433/10/7/373.

Zhan, Y., Luo, Y., Deng, X., Zhang, K., Zhang, M., Grieneisen, M., di, B., 2018. Satellitebased estimates of daily no2 exposure in China using hybrid random forest and spatiotemporal kriging model. Environmental Science Technology, Bd. 52, 3.

Zhan, J., Liu, Y., Ma, W., Zhang, X., Wang, X., Bi, F., Zhang, Y., Wu, Z., Li, H., 2022. Ozone formation sensitivity study using machine learning coupled with the reactivity of volatile organic compound species. Atmospheric Measurement Techniques, Bd. 15 (Nr. 5, S), 1511–1520 [Online]. Available: https://amt.copernicus.org/articles/15/1511/2022/.

Zhang, C., Ma, Y., 2012. Ensemble Machine Learning: Methods and Applications. Springer Publishing Company, Incorporated.