Atmos. Chem. Phys., 24, 807–851, 2024
https://doi.org/10.5194/acp-24-807-2024

Atmospheric Chemistry and Physics
Open Access

Improving 3-day deterministic air pollution forecasts using machine learning algorithms

Zhiguo Zhang[1], Christer Johansson[2,3], Magnuz Engardt[3], Massimo Stafoggia[4], and Xiaoliang Ma[1]

[1]Dept. of Civil and Architectural Engineering, KTH Royal Institute of Technology, Stockholm, Sweden
[2]Department of Environmental Science, Stockholm University, Stockholm, Sweden
[3]Environment and Health Administration, SLB-analys, Stockholm, Sweden
[4]Department of Epidemiology, Lazio Region Health Service, Rome, Italy

Correspondence: Christer Johansson (christer.johansson@aces.su.se) and Xiaoliang Ma (liang@kth.se)

Received: 31 January 2023 – Discussion started: 8 February 2023
Revised: 24 September 2023 – Accepted: 28 October 2023 – Published: 19 January 2024

Abstract. As air pollution is regarded as the single largest environmental health risk in Europe it is important that communication to the public is up to date and accurate and provides means to avoid exposure to high air pollution levels. Long- and short-term exposure to outdoor air pollution is associated with increased risks of mortality and morbidity. Up-to-date information on present and coming days' air quality helps people avoid exposure during episodes with high levels of air pollution. Air quality forecasts can be based on deterministic dispersion modelling, but to be accurate this requires detailed information on future emissions, meteorological conditions and process-oriented dispersion modelling. In this paper, we apply different machine learning (ML) algorithms – random forest (RF), extreme gradient boosting (XGB), and long short-term memory (LSTM) – to improve 1, 2, and 3 d deterministic forecasts of PM_{10}, NO_x, and O_3 at different sites in Greater Stockholm, Sweden.

It is shown that the deterministic forecasts can be significantly improved using the ML models but that the degree of improvement of the deterministic forecasts depends more on pollutant and site than on what ML algorithm is applied. Also, four feature importance methods, namely the mean decrease in impurity (MDI) method, permutation method, gradient-based method, and Shapley additive explanations (SHAP) method, are utilized to identify significant features that are common and robust across all models and methods for a pollutant. Deterministic forecasts of PM_{10} are improved by the ML models through the input of lagged measurements and Julian day partly reflecting seasonal variations not properly parameterized in the deterministic forecasts. A systematic discrepancy by the deterministic forecasts in the diurnal cycle of NO_x is removed by the ML models considering lagged measurements and calendar data like hour and weekday, reflecting the influence of local traffic emissions. For O_3 at the urban background site, the local photochemistry is not properly accounted for by the relatively coarse Copernicus Atmosphere Monitoring Service ensemble model (CAMS) used here for forecasting O_3 but is compensated for using the ML models by taking lagged measurements into account.

Through multiple repetitions of the training process, the resulting ML models achieved improvements for all sites and pollutants. For NO_x at street canyon sites, mean squared error (MSE) decreased by up to 60 %, and seven metrics, such as R^2 and mean absolute percentage error (MAPE), exhibited consistent results. The prediction of PM_{10} is improved significantly at the urban background site, whereas the ML models at street sites have difficulty capturing more information. The prediction accuracy of O_3 also modestly increased, with differences between metrics.

Further work is needed to reduce deviations between model results and measurements for short periods with relatively high concentrations (peaks) at the street canyon sites. Such peaks can be due to a combination of non-typical emissions and unfavourable meteorological conditions, which are rather difficult to forecast. Furthermore, we show that general models trained using data from selected street sites can improve the deterministic forecasts of NO_x at the station not involved in model training. For PM_{10} this was only possible using more complex LSTM models. An important aspect to consider when choosing ML algorithms is the computational requirements for training the models in the deployment of the system. Tree-based models (RF and XGB) require fewer computational resources and yield comparable performance in comparison to LSTM. Therefore, tree-based models are now implemented operationally in the forecasts of air pollution and health risks in Stockholm. Nevertheless, there is big potential to develop generic models using advanced ML to take into account not only local temporal variation but also spatial variation at different stations.

1 Introduction

According to the World Health Organization (WHO), air pollution is one of the leading causes of mortality worldwide and is regarded as the single largest environmental health risk (Fuller et al., 2022). Acute effects of air pollution are due to short-term (e.g. daily) exposures that can lead to reduced lung function, respiratory infections, and aggravated asthma (Lee et al., 2021). According to the European air quality directive, information on air quality should be made available to the public. Public information regarding the expected health risks associated with current or the next few days' concentrations of pollutants can be very important for sensitive persons when planning their outdoor activities.

There are different approaches to obtaining information on the spatio-temporal variation of air pollutant concentrations – from complex process-oriented models to different types of statistical models. Gaussian plume models are widely used in urban areas for estimating impacts on atmospheric concentrations from different emission sources and for health risk assessments (Munir et al., 2020; Johansson et al., 2009; Orru et al., 2015; Johansson et al., 2017b). Eulerian chemical transport models that describe emission, transport, mixing, and chemical transformation of trace gases and aerosols, e.g. CHIMERE, EMEP, and MATCH, are part of the Copernicus Atmosphere Monitoring Service (CAMS, https://atmosphere.copernicus.eu/, last access: 20 December 2023) to predict air pollution over Europe (Horálek et al., 2019). The uncertainties in the output of the deterministic models include uncertainties in the input, such as emissions, model algorithms, and parameterizations.

In urban areas, detailed knowledge of the dedicated emission source is often crucial. For example, road traffic, as a main emission source, can be modelled by various levels of emission models (Ma et al., 2012; Keller et al., 2017). To assess the concentration of contaminants, it is often required to combine the models of emission and dispersion processes (e.g. Ma et al., 2014). An alternative approach may derive spatio-temporal distribution of air pollutants without modelling the emission process. For example, using a land use regression model is a popular method to explain spatial contrasts in air pollution concentrations (e.g. Hoek et al., 2008).

Data-driven models using machine learning (ML) have become increasingly popular in predicting outdoor air quality (Rybarczyk and Zalakeviciute, 2018; Iskandaryan et al., 2020). Previous studies predict both hourly and daily average concentrations of particulate matter (PM), as well as gaseous air pollutants, using meteorological and traffic data (e.g. Qadeer et al., 2020; Di et al., 2019; Thongthammachart et al., 2021; Kamińska, 2019; Chuluunsaikhan et al., 2021; Doreswamy et al., 2020; Castelli et al., 2020; Stafoggia et al., 2019, 2020). In addition, a combination of ML, land-use regression (LUR), dispersion modelling, and ground-based and satellite measurements have been used to obtain temporally and spatially distributed concentrations (Shtein et al., 2020; Stafoggia et al., 2019; Brokamp et al., 2017; Di et al., 2019). Recently, Kleinert et al. (2022) conducted a study to forecast O_3 concentrations in a longer-term horizon; meanwhile, a deterministic model was also combined with ML in the study of Hong et al. (2022) to forecast the $PM_{2.5}$ concentration.

This paper aims to demonstrate how ML can improve the 1, 2, and 3 d deterministic forecasts of several critical urban air pollutants: particulate matter (PM_{10}, particles with an aerodynamic diameter less than 10 µm), nitrogen oxides (NO_x), and ozone (O_3). The study covers both urban background and street canyon sites in Stockholm, Sweden. Three ML algorithms were adopted, two based on decision trees (random forest, RF, and extreme gradient boosting, XGB) and one deep neural network model (long short-term memory, LSTM). These models were compared to investigate if there are systematic differences in their prediction performance depending on different pollutants and measurement sites, which can be used to improve current applications in Stockholm. Meanwhile, four methods for feature importance ranking were applied to analyse the effects of different features on the model prediction results.

2 Background

2.1 The Stockholm air quality forecast system

Stockholm city has used an air quality forecast system since 2021. Three different dispersion models are used to forecast concentrations considering emissions and dispersion at the European, urban and street-level scales described by Fig. 1. The CAMS ensemble model, part of the Copernicus Programme, was used to obtain forecasts of long-range transported air pollution from outside of the Greater Stockholm area. Previous assessments have found the ensemble model to be more accurate than any individual model part of CAMS (Meteo-France, 2017; Marécal et al., 2015). CAMS regional ensemble forecasts are published once a day and each forecast covers 96 h (4 d).

The contributions to concentrations due to local emissions in the metropolitan area were performed on a 100 m resolution using a Gaussian dispersion model part of the Airviro system (https://www.airviro.com/airviro/). In this modelling domain (Greater Stockholm, 35 × 35 km) individual buildings and street canyons are not resolved but treated using a roughness parameter (Gidhagen et al., 2005). The Gaussian model is fed with meteorological forecasts from the Swedish Meteorological and Hydrological Institute (SMHI). A diagnostic wind model is used to account for influences of variations in topography and land use on the dispersion parameters input to the Gaussian model. For details regarding uncertainties and validation of local modelling, see Johansson et al. (2017a).

Finally, the Operational Street Pollution Model (OSPM), developed by Berkowicz (2000) and driven by forecasted meteorology from SMHI, is applied to the street canyon sites. It has been applied earlier at Hornsgatan in Stockholm in a number of modelling studies (e.g. Krecl et al., 2021; Ottosen et al., 2015). NO_x and PM_{10} are modelled on all scales, whereas O_3 is only forecasted by the CAMS ensemble model.

For the urban-scale model domain, a detailed emission database is used as input for the local dispersion modelling. The database and its applications and comparisons between modelling and measurements are described in SLB (2022). The total emissions from road traffic are based on emission factors for different vehicle types, including passenger cars, buses, and light- and heavy-duty trucks. Exhaust emission factors of NO_x and particles are based on HBEFA version 3.3 (Keller et al., 2017) depending on the Euro class of the vehicle. The emission factors per vehicle category were weighted according to the national Swedish Transport Administration vehicle registry, but the vehicle composition taken from national vehicle registry has been shown to be similar to the local fleet using real-world number plate recognition measurements at Hornsgatan (Burman and Johansson, 2010; Burman et al., 2019). Non-exhaust emissions of PM due to wearing of brakes, tyres, and roads are calculated using the NORTRIP model (Denby et al., 2013a) forced by the forecasted meteorology from SMHI. Information on shares of studded winter tyres is obtained from manual counting every week during the winter at different locations in the city centre and along highways outside of the city. Road traffic emissions are calculated for all roads with more than 3000 vehicles per day. Other emission sources included in the local emissions database include shipping and private and municipal heating (including burning of waste). More information about the Stockholm air quality forecast system is provided in Engardt et al. (2021).

2.2 Meteorological forecasts

As an integral part of the Stockholm air quality forecast system, meteorological forecasts for a point in central Stockholm are downloaded every morning from the websites of SMHI (https://www.smhi.se/data/oppna-data, last access: 20 December 2023) and MET Norway (https://docs.api.met.no/doc/, last access: 20 December 2023). The meteorological forecasts extend over 10 d and are a combination of output from a number of regional and global numerical weather prediction models. The combination is based on statistical adjustments and manual edits. Initial models of weather-dependent PM emissions and urban and street canyon air quality modelling are driven by meteorology. The forecasted meteorological data are also used as predictors for the models in this study.

3 Methods

3.1 Data and pre-processing

The data used in this study were collected from four monitoring stations in central Stockholm, including one urban background site (Torkel Knutssonsgatan, hereafter called UB or urban) and three street canyon sites (Hornsgatan, HO; Folkungagatan, FO; and Sveavägen, SV). They are all located in central Stockholm (see Fig. 2). Detailed descriptions of measurement methods and sites are provided in Appendix A.

Data from the UB site cover approx. 1000 d (10 April 2019 through 31 December 2021). As the OSPM model became operational at a later date, the street canyon data extend over 500 d (5 August 2020 through 31 December 2021). Pollutant concentration measurements from monitoring stations, pollutant forecasts, and meteorological forecasts from the Stockholm air quality forecast system were aggregated into the following four datasets.

All the data above were collected at 1 h intervals, with details illustrated in Table 1. It should be noted that there are several studies that show the impact of the COVID-19 pandemic on pollutant emissions as a result of some restrictive regulations (Sokhi et al., 2021; Torkmahalleh et al., 2021). The COVID-19 pandemic in Sweden commenced in January 2020 and continued until February 2022, meaning that the majority of the data were collected during this pandemic period.

The pollutant measurements and forecasts from the deterministic model exhibit a missing rate of less than 5 %, with a few inaccurate samples, including outliers and negative values. Appendix B shows the missing status of O_3 in the UB dataset. To accurately represent the extreme values in the real world, outliers were deliberately included in the data because

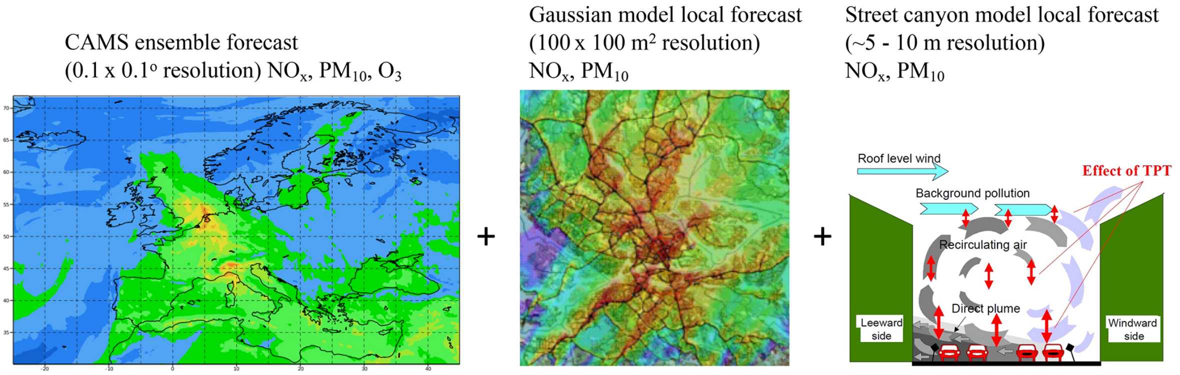


Figure 1. Illustration of the deterministic modelling from European scale at a resolution of 0.1° × 0.1° (ca. 11 km × 6 km) via urban scale (100 m resolution over an area of 35 × 35 km) down to the street canyon sites. The CAMS ensemble forecast map example is taken from https://atmosphere.copernicus.eu/. The map with the Gaussian model local forecast example is output from the Airviro system (https://www.airviro.com/airviro/, last access: 1 February 2023) used in Stockholm. The illustration of a street canyon site is taken from https://www.wikiwand.com/en/Operational_Street_Pollution_Model (last access: 1 February 2023).

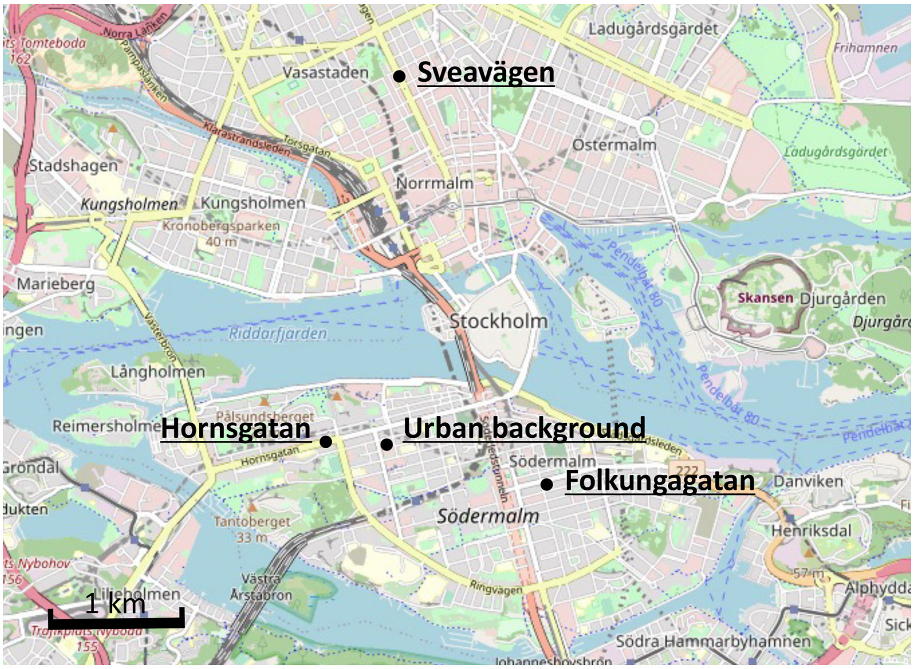


Figure 2. Map of central Stockholm showing the locations of the urban background site and the street canyon traffic sites. Base map credits are as follows: © OpenStreetMap contributors, licensed under the Open Data Commons Open Database License (ODbL) v1.0.

their occurrence is hard to justify. However, negative pollutant samples were eliminated, and missing data were manually interpolated using historical average interpolation (Willmott and Matsuura, 1995).

Frequently employed approaches of interpolating time series data comprise constant interpolation, nearest-neighbour interpolation, and linear interpolation. To keep the temporal relationship, the historical average interpolation is applied based on the periodicity pattern in the data. The periodicity of each feature, denoted by p, is determined by the analysis of the autocorrelation function (ACF) and partial autocorrelation function (PACF) of the data. Subsequently, the missing value $\tilde{p}(t)$ at time t is substituted by the average of the available data from the two preceding periods and their adjacent values:

$$\tilde{p}(t) = \frac{1}{n}\sum\Big(\tilde{p}(t-p),\, \tilde{p}(t-p\pm 1),\, \tilde{p}(t-2p),\, \tilde{p}(t-2p\pm 1)\Big), \tag{1}$$

where n is the number of samples used in Eq. (2). An example result of interpolation is shown in Appendix B.

Table 1. Description of the dataset.

Name	Time Range	Pollutants	Amount	Features
Urban background, UB	10 April 2019–31 December 2021	NO_x, PM_{10}, O_3	23 927	Pollutant measurements
Folkungagatan, FO	5 August 2020–31 December 2021	NO_x, PM_{10}	12 335	Pollutant forecasts
Hornsgatan, HO	5 August 2020–31 December 2021	NO_x, PM_{10}	12 335	Meteorological forecasts
Sveavägen, SV	5 August 2020–31 December 2021	NO_x, PM_{10}	12 335	

3.2 Prediction scheme

This study is to forecast hourly concentrations for the coming 1, 2, and 3 d of data based on historical pollutant measurements and other available information as inputs, which is a time series prediction for multiple time steps, for example, 72 time steps for 3 d prediction. Instead of more complex network structure, multiple single-output ML models are chosen for forecasting different air pollutants for $k = 1$, 2, and 3 d intervals, as shown in Eq. (2).

$$\hat{\rho}_{i,j}(d,t) = \mathrm{ML\_model}\Big(\tilde{\rho}_{i,j}(d-k,t), \overline{\rho}^{\mathrm{S}}_{i,j}(d-k,t), \check{\rho}_{i,j}(d,t), W(d,t), C(d,t)\Big), \tag{2}$$

where $\hat{\rho}_{i,j}(d,t)$ is the forecast of the pollutant j for day d and time t at the location i, and $\tilde{\rho}_{i,j}(d,t)$ is the corresponding real measurement; $\overline{\rho}^{\mathrm{S}}_{i,j}(d,t)$ uses a set S to represent several statistical measures, including maximum, minimum, 25 % quantile, and 75 % quantile of the measured concentration data during the past 24 h until t, and the measurement dataset can be represented by a set, i.e. $\{\tilde{\rho}_{i,j}(d,t), \tilde{\rho}_{i,j}(d,t-1), \tilde{\rho}_{i,j}(d,t-2)\ldots\}$ $\check{\rho}_{i,j}(d,t)$ is the predicted concentration using deterministic model. $W(d,t)$ represents the weather condition predicted for day d and time t.

Figure 3 demonstrates the prediction horizon and lagged information horizon for the case of 1 d prediction. To build consistent statistical ML models with a fixed rolling horizon, a new measurement point at the current time (d,t) will lead to an additional prediction for 1 d ahead, i.e. the predicted value at $(d+1,t)$. In this case, the measurement statistics $\overline{\rho}^{\mathrm{S}}_{i,j}(d,t)$ will be based on 1 d preceding measurement data of (d,t), resulting in a lagged rolling horizon described by Fig. 3.

3.3 Machine learning models

As already mentioned before, two tree-based ML models, RF and XGB, and one deep-learning model, LSTM, are applied to implement the prediction scheme. In addition, an ensemble learning approach based on a general additive model (GAM), aggregating the selected three learning models, is also applied to further optimize the results.

3.3.1 Framework

Figure 4 summarizes the framework of ML models and associated computational experiments for air pollution prediction. The input includes the deterministic forecasts of PM_{10}, NO_x, and O_3 to evaluate how much the deterministic forecasts can be improved by the ML algorithms. In the computational experiments, data-driven forecasting models are trained for one urban background site and three street canyon sites separately. Different ML models are trained and tested separately for predicting various air pollution concentrations in future periods, i.e. 1 d (0–24 h), 2 d (25–48 h), and 3 d (48–72 h).

To make a fair comparison with all models, a vanilla LSTM model in this case is set up to take the same type of input as the other two models. In addition to the measured air pollution time series data itself, the forecasted meteorological conditions for the prediction day d (or $d+1$ or $d+2$) and calendar information such as weekday and hour are also applied as input features. Moreover, the air pollutant concentrations predicted by the deterministic models are also used as inputs to the ML models.

Table 2 presents a detailed explanation of the essential input features that are applied in the computational experiments. During feature engineering, new features are constructed through statistical analysis to expand the feature space and facilitate context extraction. At the same time, temporal attributes are decomposed and encoded to the dataset to reflect the temporal dependence of each sample.

3.3.2 Model setups

All ML models are implemented in *Python* using existing libraries including *scikit-learn* (Bisong et al., 2019) and*pytorch* (Paszke et al., 2019) for conventional ML models and deep-learning models, respectively. The detailed implementation can be seen in the open-source code provided in Zhang and Ma (2023).

The following configurations are applied as the initial models.

- The initial parameters of the two tree-based models (XGB and RF) are the default parameters of *scikit learn*, and the tuned parameters are presented in Appendix C.

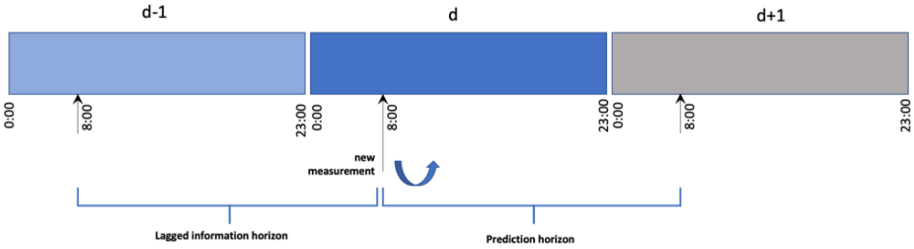


Figure 3. Illustration of the machine learning modelling scheme for 1 d prediction based on available datasets.

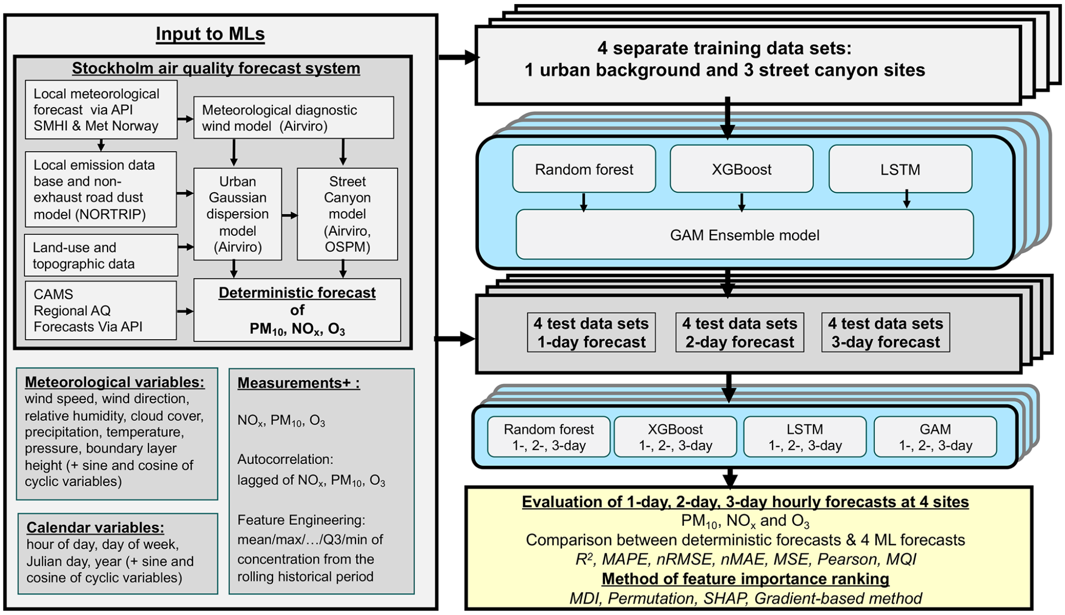

Input to MLs
Stockholm air quality forecast system
Local meteorological forecast via API SMHI & Met Norway
Meteorological diagnostic wind model (Airviro)
Local emission data base and non-exhaust road dust model (NORTRIP)
Urban Gaussian dispersion model (Airviro)
Street Canyon model (Airviro, OSPM)
Land-use and topographic data
CAMS Regional AQ Forecasts Via API
Deterministic forecast of PM_{10}, NO_x, O_3
Meteorological variables: wind speed, wind direction, relative humidity, cloud cover, precipitation, temperature, pressure, boundary layer height (+ sine and cosine of cyclic variables)
Calendar variables: hour of day, day of week, Julian day, year (+ sine and cosine of cyclic variables)
Measurements+ :
NO_x, PM_{10}, O_3
Autocorrelation: lagged of NO_x, PM_{10}, O_3
Feature Engineering: mean/max/…/Q3/min of concentration from the rolling historical period
4 separate training data sets: 1 urban background and 3 street canyon sites
Random forest
XGBoost
LSTM
GAM Ensemble model
4 test data sets 1-day forecast
4 test data sets 2-day forecast
4 test data sets 3-day forecast
Random forest 1-, 2-, 3-day
XGBoost 1-, 2-, 3-day
LSTM 1-, 2-, 3-day
GAM 1-, 2-, 3-day
Evaluation of 1-day, 2-day, 3-day hourly forecasts at 4 sites
PM_{10}, NO_x and O_3
Comparison between deterministic forecasts & 4 ML forecasts
R^2, MAPE, nRMSE, nMAE, MSE, Pearson, MQI
Method of feature importance ranking
MDI, Permutation, SHAP, Gradient-based method

Figure 4. Illustration summarizing input data for modelling 1, 2, and 3 d forecasts of PM_{10}, NO_x, and O_3 using the four models.

- The LSTM model architecture consists of two layers of LSTM with 100 neurons and a fully connected layer before the output. The activation function was *tanh*.

- The LSTM model was trained by *Adam* optimizer. The initial learning rate is 0.01 and is dynamically changed using the *ReduceLROnPlateau* algorithm, with a parameter patience of 10, which means that the algorithm will monitor the performance (e.g. validation loss) for 10 consecutive epochs. If there is no improvement, the learning rate will be reduced according to the specified reduction strategy. The initial batch size is set as 72.

The data are split along the time axis with a ratio of 16 : 4 : 5 to achieve non-overlapping sets among training, validation, and test data. Due to the autocorrelation of the air pollutant data, the assumption of independent and identically distributed classical cross-validation is not satisfied. Therefore, to preserve the time-dependent property, the function *TimeSeriesSplit* in *scikit-learn* was chosen as the cross-validation method. In the kth split, the data of the first k folds are set as the training data, whereas the data of the $(k+1)$th fold is the test set. Empirically, the value of k is set to be 5.

Given the inherent uncertainty of the ML models, they are trained by setting different random seeds. Therefore, the final results are presented in terms of statistical means and their confidence intervals, which provide a consistent way to evaluate the robustness of the prediction models. The number of repeated training processes in our experiment is set to 10 for each model.

3.4 Hyperparameter optimization

The grid and greedy search approaches are combined in the hyperparameter tuning process to balance the model optimality and computational cost (Liashchynskyi and Liashchynskyi, 2019). The grid search allows for a systematic investigation of different combinations of hyperparameters, whereas the greedy approach searches local optima for a certain variable iteratively.

Table 3 depicts the strategies of parameter optimization when training the ML models. For each model, a tuning

Table 2. Measured and forecasted air pollutant concentrations used as input data (features) in the ML modelling of pollutant concentrations at the urban background site (UB) and at the street canyon sites (SC). For periodic input data, using sine and cosine values can remove discontinuities and create consistent distance measures, thereby improving model accuracy. The bolded values are explained in column 3 (Description) of Table 2.

Category	Short names	Description
Deterministic features	**NO_x_nday_local** **PM_{10}_nday_local** **$n = 1, 2, 3$**	Deterministic 1, 2, and 3 d forecast of contributions from local emissions based on urban-scale Gaussian modelling
	NO_x_nday_regional **PM_{10}_nday_regional** **O_3_nd_regional** **$n = 1, 2, 3$**	Deterministic 1, 2, and 3 d forecast of contributions based from non-local emissions based on CAMS ensemble model (regional background)
Autocorrelation features	**NO_x_lagXX** **PM_{10}_lagXX** **O_3_lagXX** **XX = 24, 48, 72**	XX h lagged air pollutant concentrations based on autocorrelation and prediction time span.
Statistical features	**NO_x_Sta_dXX** **PM_{10}_Sta_dXX** **O_3_Sta_dXX** **Sta = avg., median, min, max, Q1, Q3** **XX = 24, 48, 72**	Average, median, minimum, maximum, and quantiles 1 and 3 of lagged air pollutant concentrations in rolling XX h periods.
Time features	**Time; Time**_sin; **Time**_cos **Time = year, julianday, month, weekday, day, hour**	Julian day of the year (1, 2, 3, …365), sine, and cosine of $2*\pi/365*$ wind direction. Day of the week (1, 2, 3, …7), sine, and cosine of $2*\pi*$d/7. Hour of the day (0, 1, 2, …23), sine, and cosine of $2*\pi*$h/24. year, month, and day
Meteorological features	wind_direction wind_direction_cos wind_direction_sin	Wind direction [0, 360) at 10 m in central Stockholm, sine and cosine of (2*pi/360)*wind direction
	pressure; temperature; precipitation; cloudiness	Pressure (10 m), temperature (10 m)
	wind_speed	Wind speed (10 m)
	relative_humidity	Relative humidity
	boundary_layer_height	Boundary layer height for central Stockholm

strategy is represented by a combination of a grid search (the searching dimensions are described in { }) and a greedy search (the search sequence is presented by →). The parameter search space and optimal parameter combinations are presented in Appendix C.

For XGB and RF, the most influential parameters are the number of evaluators (*n_estimators*), the number of input features (*max_features*), and the learning rate. Therefore, a grid search is first applied to identify an optimal combination of those parameters. Appendix C shows the results of grid search for *n_estimators* and *learning_rate*. The search spaces for *n_estimators* and *learning_rate* are set to 9 and 12, respectively, resulting in a total of 108 grid points. The optimal model performance is achieved in (60, 0.03). Subsequently, the greedy search strategy is applied sequentially to find the suboptimal combination of the parameters. The model performance is evaluated according to the mean squared error (MSE) on the validation set. For the LSTM model, only the greedy search strategy is applied to optimize the parameters sequentially due to the large search space and computational cost for training the LSTM model.

3.5 Feature importance ranking

ML models used in our study are black-box models, and feature importance analysis plays a key role in understanding the model behaviour and improvement. Feature analysis is carried out by calculating an importance score for each individual feature to quantitatively evaluate how much a feature may contribute to the forecasts.

Table 3. Hyperparameter tuning method and process.

Models	Hyperparameter tuning strategy*
XGBoost	{n_estimators, learning_rate} → max_depth → subsample → colsample_bytree → min_child_weight
RandomForest	{n_estimators, max_features} → max_depth → min_samples_split → min_samples_leaf
LSTM	batch_size → n_steps_in → hidden_size → learning rate

* {} represents the dimension of grid search, and → represents greedy search sequence.

For tree-based models, three methods, namely the mean decrease in impurity (MDI) method, permutation method, and Shapley additive explanations (SHAP) method, are used for feature ranking. For LSTM models, the gradient-based method, permutation method, and SHAP method were frequently employed. Below is a simple explanation of the feature ranking methods for the ML models.

1. *Mean decrease in impurity.* Mean impurity decrease (MDI) is a popular feature importance analysis for tree-based models, such as RF. The implementation of the method is integrated into *scikit-learn*. It calculates the average reduction in impurities using the inclusion of a particular feature as the importance score of this feature. However, the computation of impurity-based importance is based on the training data, so it does not accurately reflect the performance of the features for the test set (Bisong et al., 2019).

2. *Permutation.* The permutation method is defined as the decrease in model performance when a single feature value is randomly shuffled (Breiman, 2001). For the data used in this study, it can be applied to tree-based models but also to neural networks like LSTM. The computation of feature scores allows for the consideration of the impacts of various features on the model prediction capacity. The method has the benefit of circumventing the concerns about the tendency of MDI to favour high-cardinality features.

3. *Gradient-based method.* The gradient-based method explains the local relationship between inputs and outputs by harnessing the gradients of the model prediction with respect to input features as an importance score (Baehrens et al., 2010). It should be noted that the gradients of neural networks depend on both input and output data, and the feature importance for the LSTM model was computed as the average of feature gradient obtained from all samples in test data.

4. *SHAP.* Shapley additive explanations (SHAP) is a general explanatory framework, in which SHAP values represent the average marginal contribution of each feature towards the difference between the model's prediction and a reference prediction. The greatest strength of SHAP is its ability to reflect the influence of each feature on each sample, which is interpreted as a positive or negative influence. The SHAP is an interpretation scheme for almost all ML models. This study uses the *Python* library *shap* to evaluate tree-based models and LSTM, respectively (Lundberg and Lee, 2017; Shrikumar et al., 2017).

3.6 Statistical performance indicators

Several performance metrics have been selected for comparing the prediction results of different ML models including R squared (R^2), mean square error (MSE), and normalized error measures, i.e. mean average error (MAE), mean absolute percentage error (MAPE), root-mean-squared error (RMSE), and Pearson correlation (Pearson). These measures have also been recommended for air quality model benchmarking in the context of the Air Quality Directive 2008/50/EC (AQD) by Janssen and Thunis (2022).

In addition, to properly assess model quality, it is necessary to consider measurement uncertainty. In the Forum for Air Quality Modeling, the modelling quality indicator (MQI) is used to assess if a model fulfils certain objectives (Janssen and Thunis, 2022). It is defined as the ratio between the model bias at a fixed time (i), quantified by the RMSE, and a quantity proportional to the measurement uncertainty as follows:

$$\mathrm{MQI}(i)=\frac{\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_i-\hat{y}_i\right)^2}}{\beta\sqrt{\frac{1}{n}\sum_{i=1}^{n}U(y_i)^2}}=\frac{\mathrm{RMSE}}{\beta\mathrm{RMS}_U},$$

where $U(y_i)$ is the expanded 95th percentile measurement uncertainty and β is a coefficient of proportionality (Janssen and Thunis, 2022). The value of β determines the stringency of the MQI and is set equal to 2, thus allowing deviation between modelled and measured concentrations as twice the measurement uncertainty. The uncertainty of the measurements (RMS_U) was calculated for the mean of the measurement concentrations as follows:

$$U(y_i)=U_r(\mathrm{RV})\sqrt{\left(1-\propto^2\right)y_i^2+\propto^2\mathrm{RV}^2},$$

where $U_r(\mathrm{RV})$ and $\propto$ are parameters that depend on pollutant and RV is a reference value, here taken to be 200, 50, and 120 $\mu\mathrm{g\,m^{-3}}$, corresponding $U_r(\mathrm{RV})$ was 0.24, 0.28, and 0.18 and $\propto$ was 0.20, 0.25, 0.79 for $\mathrm{NO_2}$, $\mathrm{PM_{10}}$, and $\mathrm{O_3}$, respectively (Janssen and Thunis, 2022). In our case we have

Table 4. Performance indicators.

Indicators	Formula	Indicators	Formula
R^2	$R^2(y,\hat{y}) = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2}$	Mean square error	$\mathrm{MSE}(y,\hat{y}) = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2$
Mean absolute percentage error	$\mathrm{MAPE}(y\hat{y}) = \frac{1}{n}\sum_{i=1}^{n}\frac{\lvert y_i - \hat{y}_i\rvert}{\lvert y_i\rvert}$	Root-mean-square error	$\mathrm{RMSE}(y\hat{y}) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2}$
Pearson correlation	$\mathrm{Pearson}(y,\hat{y}) = \frac{\sum_{i=1}^{n}(y_i - \overline{y_i})\left(\hat{y}_i - \overline{\hat{y}_i}\right)}{\sqrt{\sum_{i=1}^{n}(y_i - \overline{y_i})^2}\sqrt{\sum_{i=1}^{n}\left(\hat{y}_i - \overline{\hat{y}_i}\right)^2}}$		

Note that $\hat{y}_i$ is the predicted value of the ith sample, y_i is the corresponding true value, and $\overline{y}$ is the mean value of all n samples. MAE and RMSE were normalized by dividing by the mean of the measured concentrations, hereafter called nMAE and nRMSE.

calculated NO_x, not NO_2, but we used the same settings of the parameters for NO_x as recommended for NO_2.

4 Computational results

The focus of this paper is to compare the deterministic forecasts of NO_x, PM_{10} and O_3 with the forecasts based on the different machine learners which also include the deterministic forecasts as input variables (features). As described above we have made deterministic and ML forecasts for hourly mean concentrations for the coming 72 h, based on 1, 2, and 3 d meteorological forecasts for one urban background site (NO_x, PM_{10}, and O_3) and three street canyon sites (NO_x and PM_{10}). We also compare results separately for the urban background site and the street canyon sites.

4.1 Urban background

4.1.1 Comparison between deterministic forecasts and ML models – urban background

As illustrated in Table 5 and Fig. 5, all statistical performance measures of the deterministic forecasts are improved by the ML models for the pollutants: NO_x, PM_{10}, and O_3. The statistical mean and 95 % confidence intervals are estimated from 10 repeated computational experiments using 10 different random seeds.

Table 5 summarizes the prediction performance of both deterministic and ML models in terms of five selected metrics. For NO_x, the R^2 value increases, from a range between 0.12 and 0.22 for the deterministic forecasts to a range between 0.33 and 0.42 achieved by ML models. The other four metrics, including MAPE, nRMSE, nMAE, and MSE, decrease for all forecasting days. The LSTM model achieves superior performance for almost all the metrics, and XGBoost performs closely in this case. For PM_{10}, R^2 increases, from the range of 0.08–0.21 in the deterministic forecasts to higher values between 0.28 and 0.55 using ML models. Again, there are big reductions in the other four performance measures, among which MSE is decreased by 45 % compared to deterministic forecasts. XGB and RF models are the winners with comparable performance. For O_3 there is about a 40 % drop in MSE for tree-based models, with slight improvements on other metrics for all forecasting days. LSTM also performs equally well and achieves remarkable performance for the 3 d prediction. While the errors of deterministic CAMS modelling for O_3 are quite small when compared to the prediction of NO_x and PM_{10}, MLs demonstrate their capacity to further refine the pollutant prediction.

The width of the confidence interval indicates the reliability of the model prediction results. The two tree-based models (XGB and RF) produce a very small variance, less than 1 %, whereas the LSTM model exhibits a higher variance (but less than 5 %). The higher variance of LSTM model may be due to the random initialization of the weights, which affects the subsequent gradient descent trajectory and model results.

Figure 5 presents statistical mean of 1, 2, and 3 d forecasts by ML and deterministic models. Overall, all the performance metrics, including MQI and Pearson correlation, are consistently improved by ML models for three pollutants, NO_x, PM_{10}, and O_3. The difference in performance metrics achieved by different ML models is less than 30 %.

All MQI results are below 100 %, indicating that the deviation between model results and measurements is smaller than the estimated uncertainties of the measurements. XGBoost seems more efficient in reducing MQI, from 66 % to 52 % for PM_{10}. The LSTM model shows a reduction of around 10 % on MQI for both NO_x and O_3. The Pearson correlation reveals similar behaviour to the R^2 but represents a more pronounced enhancement on improvement.

Figure 6a shows an example time series plot of the forecasts by the GAM and deterministic models during September 2021. Similar plots are also demonstrated for other models in Appendix D. According to the figures, the ML models show better performance in capturing the trends and variation of measured pollutant concentrations, compared to the deterministic forecasts, although they still have obvious deviations from the real measurement. None of the models performs

Table 5. Comparison of 1, 2, 3 d deterministic and ML forecasts for NO_x, PM_{10}, and O_3 for the urban background site. R^2 is R squared, MAPE is mean absolute percentage error, nRMSE is normalized root-mean-square error, nMAE is normalized mean absolute error, and MSE is mean square error. The average performances with their 95 % confidence interval were computed on the test set from 10 experimental repetitions conducted with different random seeds, and the best performances are in bold.

NO_x	R^2			MAPE			nRMSE			nMAE			MSE		
	1 d	2 d	3 d	1 d	2 d	3 d	1 d	2 d	3 d	1 d	2 d	3 d	1 d	2 d	3 d
Det	0.13	0.22	0.12	0.72	0.73	0.88	1.27	1.20	1.28	0.61	0.60	0.69	229.77	205.21	233.25
XGB	0.30 ± 0.01	0.30 ± 0.01	0.30 ± 0.00	**0.37 ± 0.00**	**0.39 ± 0.00**	**0.39 ± 0.00**	1.14 ± 0.00	1.14 ± 0.00	1.14 ± 0.00	**0.40 ± 0.00**	**0.41 ± 0.00**	**0.41 ± 0.00**	184.19 ± 1.58	185.93 ± 1.46	185.91 ± 1.18
RF	0.27 ± 0.00	0.27 ± 0.00	0.27 ± 0.00	0.48 ± 0.00	0.50 ± 0.00	0.50 ± 0.00	1.17 ± 0.00	1.17 ± 0.00	1.17 ± 0.00	0.44 ± 0.00	0.45 ± 0.00	0.45 ± 0.00	192.87 ± 0.53	194.5 ± 0.71	194.66 ± 0.89
LSTM	**0.33 ± 0.05**	**0.41 ± 0.03**	**0.42 ± 0.02**	0.44 ± 0.06	0.41 ± 0.03	0.41 ± 0.03	**1.12 ± 0.04**	**1.04 ± 0.03**	**1.04 ± 0.02**	0.44 ± 0.02	0.43 ± 0.02	0.42 ± 0.02	178.32 ± 12.89	**155.12 ± 8.43**	**153.57 ± 6.19**
GAM	**0.33 ± 0.01**	0.30 ± 0.01	0.34 ± 0.01	0.43 ± 0.01	0.45 ± 0.01	0.45 ± 0.00	**1.12 ± 0.01**	1.14 ± 0.01	1.11 ± 0.01	0.43 ± 0.00	0.44 ± 0.00	0.44 ± 0.00	**176.48 ± 2.52**	184.91 ± 2.60	176.21 ± 2.82

PM_{10}	R^2			MAPE			nRMSE			nMAE			MSE		
	1 d	2 d	3 d	1 d	2 d	3 d	1 d	2 d	3 d	1 d	2 d	3 d	1 d	2 d	3 d
Det	0.21	0.13	0.08	0.54	0.56	0.63	0.67	0.70	0.72	0.46	0.48	0.51	41.67	45.78	48.65
XGB	**0.55 ± 0.00**	**0.49 ± 0.01**	**0.41 ± 0.01**	0.47 ± 0.01	0.53 ± 0.01	0.57 ± 0.01	**0.50 ± 0.00**	**0.53 ± 0.00**	**0.58 ± 0.00**	**0.35 ± 0.00**	**0.38 ± 0.00**	**0.40 ± 0.00**	**23.7 ± 0.26**	**26.75 ± 0.31**	**31.25 ± 0.28**
RF	**0.55 ± 0.00**	0.42 ± 0.00	0.37 ± 0.00	0.44 ± 0.00	0.48 ± 0.00	0.52 ± 0.00	0.51 ± 0.00	0.57 ± 0.00	0.60 ± 0.00	**0.35 ± 0.00**	0.39 ± 0.00	**0.40 ± 0.00**	24.07 ± 0.07	30.63 ± 0.16	33.59 ± 0.12
LSTM	0.37 ± 0.04	0.39 ± 0.04	0.28 ± 0.04	0.57 ± 0.10	0.52 ± 0.04	0.55 ± 0.05	0.60 ± 0.02	0.59 ± 0.02	0.64 ± 0.02	0.43 ± 0.02	0.41 ± 0.01	0.44 ± 0.01	33.41 ± 1.99	32.13 ± 2.34	38.38 ± 1.88
GAM	0.53 ± 0.01	0.36 ± 0.01	0.33 ± 0.01	**0.43 ± 0.01**	**0.47 ± 0.00**	**0.50 ± 0.01**	0.52 ± 0.00	0.60 ± 0.00	0.62 ± 0.00	0.36 ± 0.00	0.41 ± 0.00	0.42 ± 0.00	24.97 ± 0.37	33.72 ± 0.43	35.41 ± 0.34

O_3	R^2			MAPE			nRMSE			nMAE			MSE		
	1 d	2 d	3 d	1 d	2 d	3 d	1 d	2 d	3 d	1 d	2 d	3 d	1 d	2 d	3 d
Det	0.38	0.32	0.19	0.48	0.54	0.59	0.31	0.32	0.35	0.24	0.25	0.28	210.07	231.27	276.85
XGB	**0.65 ± 0.00**	**0.58 ± 0.00**	0.54 ± 0.00	0.40 ± 0.00	0.44 ± 0.00	0.46 ± 0.00	**0.23 ± 0.00**	**0.25 ± 0.00**	0.27 ± 0.00	**0.18 ± 0.00**	**0.20 ± 0.00**	0.21 ± 0.00	**121.14 ± 0.56**	**143.44 ± 1.55**	157.89 ± 1.35
RF	0.62 ± 0.00	0.52 ± 0.00	0.50 ± 0.00	0.42 ± 0.00	0.48 ± 0.00	0.47 ± 0.00	0.24 ± 0.00	0.27 ± 0.00	0.28 ± 0.00	0.19 ± 0.00	0.21 ± 0.00	0.21 ± 0.00	129.52 ± 0.41	164.16 ± 0.38	169.71 ± 0.34
LSTM	0.62 ± 0.01	0.57 ± 0.02	**0.59 ± 0.01**	**0.39 ± 0.01**	**0.42 ± 0.01**	**0.40 ± 0.01**	0.24 ± 0.00	0.26 ± 0.01	**0.25 ± 0.00**	0.19 ± 0.00	**0.20 ± 0.00**	**0.20 ± 0.00**	131.05 ± 3.84	146.17 ± 6.2	**139.51 ± 3.09**
GAM	0.56 ± 0.00	0.43 ± 0.00	0.42 ± 0.00	0.42 ± 0.00	0.50 ± 0.00	0.49 ± 0.00	0.26 ± 0.00	0.30 ± 0.00	0.30 ± 0.00	0.20 ± 0.00	0.23 ± 0.00	0.23 ± 0.00	151.39 ± 0.69	195.4 ± 0.85	196.78 ± 1.19

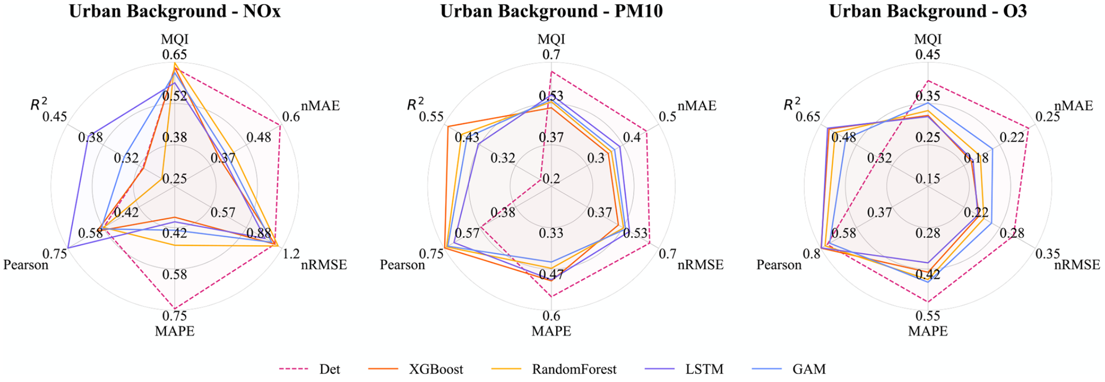


Figure 5. Statistical performances for ML models and the deterministic hourly forecasts for the urban site. Mean of 1, 2, and 3 d forecasts. Note that the ranges are different for different metrics.

well in capturing the peaks of PM_{10}, e.g. on 30 September. Figure 6b demonstrates an example time series plot of the difference between the forecasted concentrations of three pollutants, NO_x, PM_{10}, and O_3, predicted by both deterministic and ML models and the real observation. The graphs illustrate that during some hours all models systematically show large absolute deviations from the observed mean concentrations. Sometimes the hours with large deviations for NO_x coincide with deviations for PM_{10}, indicating some specific meteorological situation or common source that caused this deviation.

Systematic deviations between the observed mean diurnal variations and the deterministic forecast are shown in Appendix D. The deterministic forecasts are significantly improved using the ML models, especially for NO_x and O_3. For O_3 the deterministic forecast systematically overestimates the concentrations, which is mainly due to the fact that the chemical destruction of O_3 in the city centre is not properly accounted for by the regional CAMS model. For NO_x, the concentrations calculated by the deterministic model are systematically shifted 1 h compared to the observed concentration, and this is likely associated with errors in parameterization of traffic emissions, which is the most important source of NO_x in Stockholm. For PM_{10}, concentrations modelled by the deterministic model are too low during the night compared to observations, but this is corrected using RF and XGB but not using GAM.

For the general public, it is important to receive information on future pollution episodes with high concentrations. The plots in Appendix E show that statistical performances for all models are worse when concentrations are higher than when the mean value is analysed. R^2 is somewhat higher for O_3, while NO_x and PM_{10} decreased significantly, with the LSTM model having a relatively higher value among all models for NO_x, while the XGBoost shows higher values for PM_{10}. The nRMSE showed a similar trend to R^2.

4.1.2 Importance of features – urban background

Figure 7 presents the top 10 features obtained by the four feature ranking methods, i.e. MDI, gradient based, permutation, and SHAP. More detailed plots of feature importance ranking are shown in Appendix F, including the results of all models (RF, XGB and LSTM), for all three pollutants (PM_{10}, NO_x, O_3), and for all three forecasting periods (1, 2, and 3 d). It should be noted that the local deterministic models, both Gaussian and OSPM models, use the same meteorological data to forecast hourly pollutant concentrations. So, when the meteorological variables are important features for the ML models, it indicates that the deterministic models do not capture all hidden processes related to those factors. Regarding feature importance ranking for the urban background model, we have the following findings.

1. For the NO_x model, the factors, including temperature, wind speed, calendar data, lagged 24 h mean concentrations, and local deterministic forecasts, are among the top 10 most important variables, but the deterministic forecast is not the most important feature for any model. Among the calendar features, hour is the most important factor, indicating the importance of regular, diurnal variations of traffic emissions. Since both XGB and RF are decision-tree-based algorithms, the top 10 features selected by the three feature ranking methods are basically the same; however, for LSTM, different features are extracted. Among all models, only the permutation model raises the importance of the deterministic forecasts of O_3 and PM_{10}, which reflect the fact that O_3 production is dependent on the status of NO_x (Hagenbjörk et al., 2017) and compensate for the results of other methods of feature importance.

2. Regarding PM_{10}, the regional deterministic forecast is the most important feature of all models. Among the meteorological factors, both wind direction and pressure show their importance for prediction. The seasonal variation is reflected in the importance of the Ju-

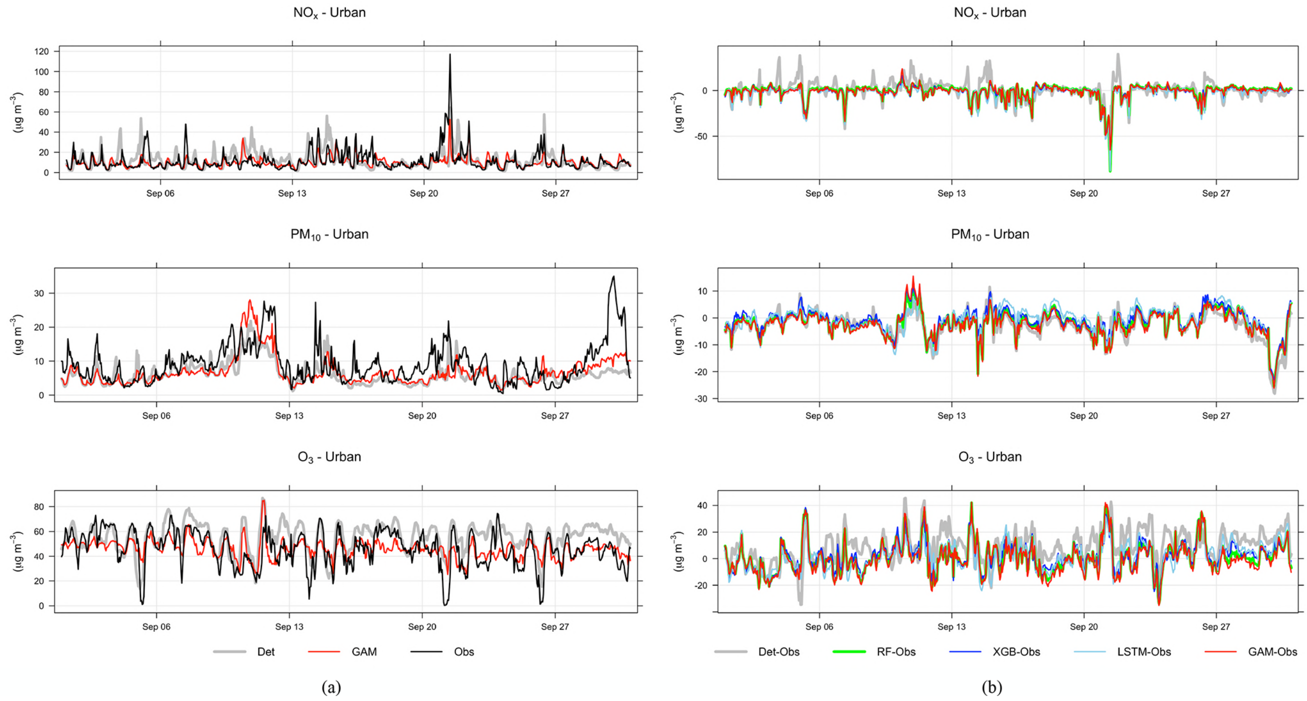


Figure 6. **(a)** Temporal variations of hourly mean concentrations of NO_x, PM_{10}, and O_3 at the urban background site during September 2021 based on mean of 1, 2, and 3 d forecasts for observations, deterministic forecasts, and GAM. **(b)** Absolute deviations of forecasted NO_x, PM_{10}, and O_3 concentrations from observed (Obs) concentrations based on mean of 1, 2, and 3 d forecasts for September 2021. All data are hourly mean concentrations.

lian day. For LSTM, precipitation shows a high importance, indicating the dependence of suspension of dust on surface wetness not being captured by the deterministic forecasts. For redundant features such as hour_sin and hour_cos, the permutation method may calculate lower importance values for both features due to multicollinearity despite being important in reality. In this case, MDI and SHAP can capture those features.

3. For O_3, all models result in similar feature importance rankings. The deterministic forecasts are the dominant features for the models of various forecasting horizons. In addition, the lagged maximum concentration, *O3_max_d24*, demonstrates its higher importance for tree-based models. The high importance of relative humidity (RH) reflects the potential fact that O_3 concentrations may be higher during dry, clear-sky conditions, not completely captured by the deterministic forecasts.

4.2 Street canyon sites

4.2.1 Comparison between deterministic forecasts and ML models – street canyon sites

For all street sites, the forecasts of NO_x are improved by the ML models, which are illustrated in detail for different pollutants in Fig. 8 and Table 6. The improvements in terms of MQI, R squared (R^2), Pearson correlation, MAPE, nRMSE, nMAE, and MSE show similar patterns for the ML models but differ between street sites.

Figure 8 summarizes the improvements, in terms of different statistical performance metrics, for NO_x prediction at all street canyon sites and for different ML models. The error, represented by MAPE, nRMSE, nMAE, and MSE, is reduced by 30 % to 60 %, and the R-squared coefficients are increased by 30 % to 50 %. Similar to urban background, the variation in Pearson correlation is similar to that of R^2, but Pearson correlation tends to be much larger than R^2 for the same model. Also, relative uncertainties decrease using the ML models compared to the deterministic forecast.

It should be noted that the R^2 of some deterministic forecasts is negative in Table 6, which implies that the deterministic forecasts are sometimes worse than simply using the mean of pollutant concentration as the predictor. For Folkungagatan, the GAM model shows a good integration of results from the tree-based model and LSTM, resulting in further improvement of the prediction performance. MSE of the XGBoost model drops by more than 40 % in Sveavägen. Forecasts for Hornsgatan show higher R^2 and lower relative errors compared to the other streets. In addition, LSTM models exhibit greater variability compared to the tree model due to its training process being more susceptible to random influences.

Table 6. Comparison of 1, 2, and 3 d deterministic and ML forecasts for NO_x for the street canyon sites. All data are based on hourly mean values. The average performances with their 95 % confidence interval were computed on the test set from 10 experimental repetitions conducted with different random seeds, and the best performances are given in bold.

	Folkungagatan FO														
	R^2			MAPE			nRMSE			nMAE			MSE		
	1 d	2 d	3 d	1 d	2 d	3 d	1 d	2 d	3 d	1 d	2 d	3 d	1 d	2 d	3 d
Det	−0.08	0.02	0.09	0.84	0.96	0.96	0.97	0.92	0.89	0.62	0.61	0.60	1337.64	1209.23	1125.66
XGB	0.35 ± 0.00	0.34 ± 0.01	0.36 ± 0.01	0.43 ± 0.00	0.44 ± 0.00	0.44 ± 0.01	0.75 ± 0.00	0.75 ± 0.00	0.74 ± 0.00	0.41 ± 0.00	0.42 ± 0.00	0.41 ± 0.00	799.15 ± 6.08	813.34 ± 8.94	791.93 ± 8.65
RF	0.41 ± 0.00	0.40 ± 0.00	0.40 ± 0.00	0.59 ± 0.00	0.63 ± 0.00	0.64 ± 0.00	0.72 ± 0.00	0.72 ± 0.00	0.72 ± 0.00	0.42 ± 0.00	0.43 ± 0.00	0.43 ± 0.00	733.81 ± 2.8	745.93 ± 4.92	741.81 ± 3.91
LSTM	0.46 ± 0.02	0.45 ± 0.03	0.49 ± 0.03	0.45 ± 0.02	0.44 ± 0.02	0.48 ± 0.05	0.68 ± 0.01	0.69 ± 0.02	0.67 ± 0.02	0.40 ± 0.01	0.40 ± 0.01	0.40 ± 0.01	663.36 ± 19.42	680.08 ± 34.83	636.66 ± 39.38
GAM	**0.51 ± 0.01**	**0.49 ± 0.02**	**0.53 ± 0.02**	**0.38 ± 0.01**	**0.40 ± 0.01**	**0.40 ± 0.01**	**0.65 ± 0.01**	**0.66 ± 0.01**	**0.64 ± 0.01**	**0.37 ± 0.00**	**0.38 ± 0.01**	**0.37 ± 0.01**	**604.22 ± 15.79**	**633.22 ± 24.17**	**585.36 ± 22.99**
	Sveavägen SV														
	R^2			MAPE			nRMSE			nMAE			MSE		
	1 d	2 d	3 d	1 d	2 d	3 d	1 d	2 d	3 d	1 d	2 d	3 d	1 d	2 d	3 d
Det	−0.04	0.02	0.03	1.11	1.18	1.00	0.92	0.89	0.88	0.66	0.65	0.62	1620.28	1525.52	1507.64
XGB	**0.49 ± 0.01**	**0.50 ± 0.01**	**0.48 ± 0.00**	**0.50 ± 0.01**	**0.51 ± 0.01**	**0.47 ± 0.01**	**0.64 ± 0.00**	**0.64 ± 0.01**	**0.64 ± 0.00**	**0.37 ± 0.00**	**0.37 ± 0.00**	**0.36 ± 0.00**	**787.94 ± 9.25**	**786.33 ± 17.12**	**804.19 ± 7.64**
RF	0.46 ± 0.00	0.45 ± 0.00	0.43 ± 0.00	0.54 ± 0.00	0.55 ± 0.00	0.54 ± 0.00	0.66 ± 0.00	0.66 ± 0.00	0.68 ± 0.00	0.38 ± 0.00	0.38 ± 0.00	0.38 ± 0.00	847.22 ± 4.06	858.18 ± 3.95	892.8 ± 4.55
LSTM	0.47 ± 0.03	0.42 ± 0.06	0.35 ± 0.08	0.63 ± 0.09	0.62 ± 0.09	0.56 ± 0.07	0.65 ± 0.02	0.68 ± 0.03	0.72 ± 0.04	0.40 ± 0.01	0.42 ± 0.02	0.44 ± 0.03	833.4 ± 53.22	897.53 ± 93.33	1011.5 ± 119.21
GAM	0.46 ± 0.02	0.44 ± 0.02	0.42 ± 0.01	0.54 ± 0.01	0.54 ± 0.01	0.53 ± 0.01	0.66 ± 0.01	0.67 ± 0.01	0.68 ± 0.01	0.40 ± 0.01	0.40 ± 0.01	0.40 ± 0.01	836.73 ± 29.81	873.7 ± 28.96	908.67 ± 23.08
	Hornsgatan HO														
	R^2			MAPE			nRMSE			nMAE			MSE		
	1 d	2 d	3 d	1 d	2 d	3 d	1 d	2 d	3 d	1 d	2 d	3 d	1 d	2 d	3 d
Det	0.29	0.32	0.31	0.59	0.63	0.69	0.77	0.76	0.77	0.47	0.48	0.48	2431.10	2358.07	2387.06
XGB	**0.63 ± 0.00**	**0.63 ± 0.01**	0.65 ± 0.01	**0.42 ± 0.01**	**0.44 ± 0.01**	0.44 ± 0.01	**0.56 ± 0.00**	**0.56 ± 0.01**	0.55 ± 0.01	**0.33 ± 0.00**	**0.33 ± 0.00**	**0.33 ± 0.00**	**1285.99 ± 12.61**	1273.06 ± 33.59	1210.59 ± 30.08
RF	**0.63 ± 0.00**	**0.63 ± 0.00**	**0.66 ± 0.00**	0.45 ± 0.00	0.46 ± 0.00	0.46 ± 0.00	**0.56 ± 0.00**	**0.56 ± 0.00**	**0.54 ± 0.00**	0.34 ± 0.00	0.34 ± 0.00	**0.33 ± 0.00**	1288.93 ± 6.34	**1267.42 ± 5.13**	**1176.98 ± 5.07**
LSTM	0.55 ± 0.04	0.57 ± 0.03	0.61 ± 0.02	0.45 ± 0.09	**0.46 ± 0.07**	**0.38 ± 0.03**	0.62 ± 0.03	0.60 ± 0.02	0.58 ± 0.01	0.36 ± 0.02	0.36 ± 0.01	**0.34 ± 0.01**	1565.54 ± 131.48	1483.74 ± 97.31	1351.91 ± 61.79
GAM	0.60 ± 0.00	0.61 ± 0.01	0.64 ± 0.00	0.47 ± 0.00	0.48 ± 0.01	0.49 ± 0.01	0.58 ± 0.00	0.57 ± 0.00	0.55 ± 0.00	0.35 ± 0.00	0.35 ± 0.00	0.35 ± 0.00	1376.94 ± 13.02	1339.63 ± 20.34	1242.28 ± 11.52

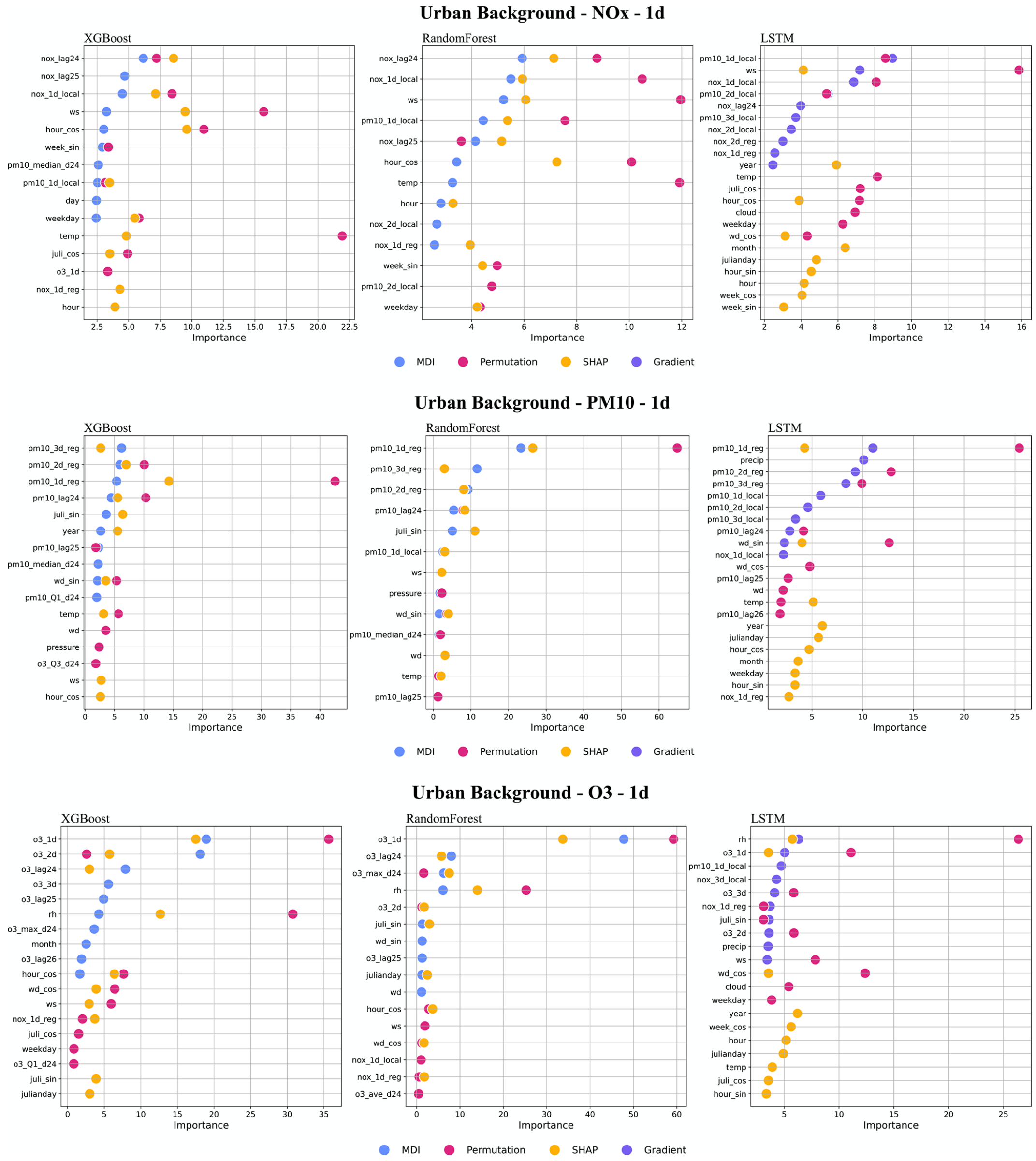


Figure 7. Top 10 important features (%) of all 1 d forecasting models, XGB, RF, and LSTM, for the urban site. All data are hourly mean concentrations.

Comparison between the statistical performance measures of ML models and deterministic forecasts for PM_{10} gives somewhat diverse results, depending on statistical measure, street site, and ML model. MSE decreases slightly in most cases and the normalized RMSE and MAE are lower for most (but not all) ML models and streets, while MAPE often increases using the ML models (Table 7 and Fig. 9).

R^2 and Pearson of LSTM prediction are 10 % to 40 % higher for Folkungagatan and Hornsgatan. However, the prediction results for Sveavägen show little improvement, and tree-based model and GAM give even worse MAPE than the deterministic forecasts. For relative uncertainties represented by MQI, there is no systematic improvement using ML models compared to the deterministic model.

Comparisons between the hourly temporal variations in observations and forecasts of NO_x with the GAM model in October 2022 are shown in Fig. 10. Further details for all models are presented in Appendix G. One can see that the deterministic forecast tends to overestimate concentrations of NO_x during daytime especially for Sveavägen, and this is corrected when the ML model is being applied. Corresponding plots for PM_{10} are shown in Appendix F. In this case, the GAM overestimates concentrations on Hornsgatan during the beginning of October but performs well otherwise.

Table 7. Comparison of 1, 2, and 3 d deterministic and ML forecasts for PM_{10} for the street canyon sites. The average performances with their 95 % confidence interval were computed on the test set from 10 experimental repetitions conducted with different random seeds, and the best performances are bold.

	Folkungagatan FO														
	R^2			MAPE			nRMSE			nMAE			MSE		
	1 d	2 d	3 d	1 d	2 d	3 d	1 d	2 d	3 d	1 d	2 d	3 d	1 d	2 d	3 d
Det	0.12	0.19	**0.19**	**1.17**	1.22	**1.25**	0.81	0.78	**0.78**	0.50	0.51	**0.50**	83.54	77.07	**76.75**
XGB	**0.28 ± 0.01**	0.15 ± 0.01	0.08 ± 0.01	1.23 ± 0.02	1.37 ± 0.02	1.43 ± 0.03	**0.74 ± 0.00**	0.80 ± 0.01	0.83 ± 0.01	**0.47 ± 0.00**	0.54 ± 0.01	0.56 ± 0.01	**69.17 ± 0.94**	81.23 ± 1.02	88.28 ± 1.21
RF	0.18 ± 0.01	0.11 ± 0.01	0.04 ± 0.01	1.59 ± 0.01	1.76 ± 0.02	1.79 ± 0.03	0.79 ± 0.01	0.82 ± 0.01	0.85 ± 0.01	0.52 ± 0.00	0.55 ± 0.01	0.57 ± 0.01	78.83 ± 1.07	84.98 ± 1.43	91.70 ± 1.35
LSTM	0.25 ± 0.22	**0.26 ± 0.08**	**0.16 ± 0.08**	1.35 ± 0.51	**1.16 ± 0.31**	1.31 ± 0.26	**0.74 ± 0.09**	**0.74 ± 0.04**	**0.79 ± 0.04**	0.51 ± 0.1	**0.50 ± 0.03**	**0.52 ± 0.03**	71.55 ± 21.24	**70.96 ± 7.22**	**80.46 ± 8.01**
GAM	0.06 ± 0.04	0.01 ± 0.05	−0.06 ± 0.04	1.30 ± 0.07	1.40 ± 0.07	1.50 ± 0.1	0.84 ± 0.02	0.86 ± 0.02	0.89 ± 0.02	0.51 ± 0.01	0.53 ± 0.01	0.56 ± 0.01	90.24 ± 4.01	94.42 ± 4.35	101.27 ± 4.28

	Sveavägen SV														
	R^2			MAPE			nRMSE			nMAE			MSE		
	1 d	2 d	3 d	1 d	2 d	3 d	1 d	2 d	3 d	1 d	2 d	3 d	1 d	2 d	3 d
Det	0.01	0.08	**0.16**	**1.04**	**1.23**	1.14	0.78	0.75	**0.72**	0.53	0.54	0.51	120.08	111.57	**102.73**
XGB	**0.25 ± 0.01**	0.12 ± 0.02	0.15 ± 0.01	1.19 ± 0.01	1.33 ± 0.02	1.36 ± 0.01	**0.68 ± 0.00**	0.74 ± 0.01	0.73 ± 0.00	**0.47 ± 0.00**	0.52 ± 0.01	0.51 ± 0.00	**91.71 ± 1.10**	107.34 ± 1.85	103.44 ± 1.13
RF	0.21 ± 0.00	0.14 ± 0.01	0.15 ± 0.01	1.40 ± 0.01	1.54 ± 0.01	1.53 ± 0.01	0.70 ± 0.00	0.73 ± 0.00	0.73 ± 0.00	0.48 ± 0.00	0.53 ± 0.00	0.52 ± 0.00	96.62 ± 0.52	105.28 ± 0.93	103.95 ± 0.69
LSTM	0.22 ± 0.06	**0.21 ± 0.05**	**0.11 ± 0.07**	1.08 ± 0.11	**1.24 ± 0.19**	**1.03 ± 0.16**	0.70 ± 0.03	**0.70 ± 0.02**	**0.74 ± 0.03**	0.48 ± 0.01	**0.49 ± 0.02**	**0.50 ± 0.01**	95.35 ± 7.25	**96.84 ± 6.32**	**108.79 ± 9.04**
GAM	0.15 ± 0.03	−0.08 ± 0.04	0.02 ± 0.03	1.23 ± 0.03	1.34 ± 0.05	1.41 ± 0.02	0.73 ± 0.01	0.82 ± 0.02	0.78 ± 0.01	0.51 ± 0.01	0.57 ± 0.01	0.56 ± 0.01	104.08 ± 3.79	131.86 ± 5.05	119.97 ± 3.9

	Hornsgatan HO														
	R^2			MAPE			nRMSE			nMAE			MSE		
	1 d	2 d	3 d	1 d	2 d	3 d	1 d	2 d	3 d	1 d	2 d	3 d	1 d	2 d	3 d
Det	−0.00	−0.21	−0.36	1.09	1.19	1.39	0.94	1.03	1.09	0.67	0.71	0.81	118.50	143.06	160.01
XGB	0.11 ± 0.05	−0.27 ± 0.12	−0.09 ± 0.05	1.05 ± 0.05	1.27 ± 0.06	**1.21 ± 0.02**	0.89 ± 0.03	1.06 ± 0.05	0.98 ± 0.02	0.61 ± 0.02	0.71 ± 0.03	0.67 ± 0.01	105.38 ± 6.19	151.11 ± 13.92	129.35 ± 5.95
RF	**0.18 ± 0.01**	**0.07 ± 0.0**	−0.02 ± 0.01	1.06 ± 0.01	1.28 ± 0.01	1.33 ± 0.01	**0.85 ± 0.00**	**0.91 ± 0.00**	0.95 ± 0.00	0.60 ± 0.00	0.67 ± 0.00	0.70 ± 0.00	**97.74 ± 0.84**	**109.93 ± 0.49**	121.18 ± 1.15
LSTM	0.16 ± 0.06	0.06 ± 0.09	**0.05 ± 0.05**	**0.86 ± 0.16**	**0.99 ± 0.18**	**0.89 ± 0.12**	0.86 ± 0.03	**0.91 ± 0.04**	**0.92 ± 0.03**	**0.56 ± 0.03**	**0.62 ± 0.03**	**0.59 ± 0.02**	99.32 ± 7.14	**111.77 ± 10.36**	**112.97 ± 6.34**
GAM	0.15 ± 0.01	0.07 ± 0.01	−0.02 ± 0.02	1.04 ± 0.01	1.24 ± 0.02	1.32 ± 0.01	0.87 ± 0.01	0.91 ± 0.00	0.95 ± 0.01	0.60 ± 0.00	0.66 ± 0.00	0.69 ± 0.01	100.36 ± 1.65	**109.90 ± 0.89**	120.50 ± 1.92

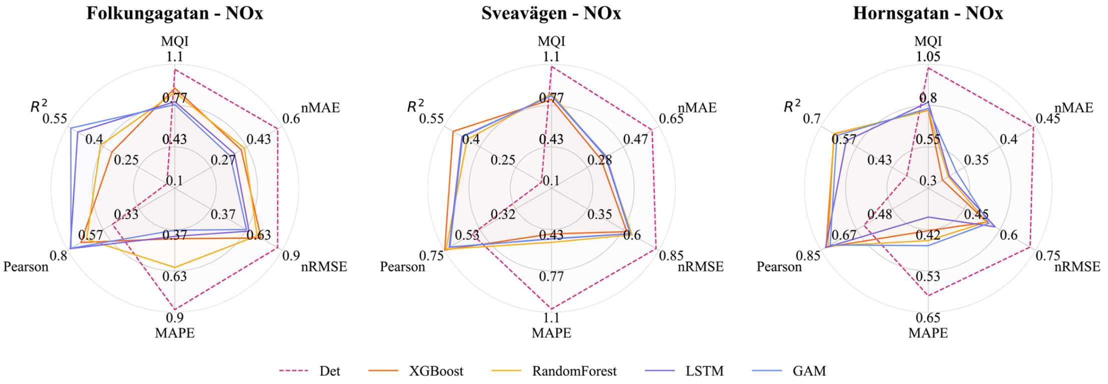


Figure 8. Statistical performances for ML models and the deterministic hourly forecasts of NO_x for the street site. Mean of 1, 2, and 3 d forecasts. Note that the ranges are different for different metrics.

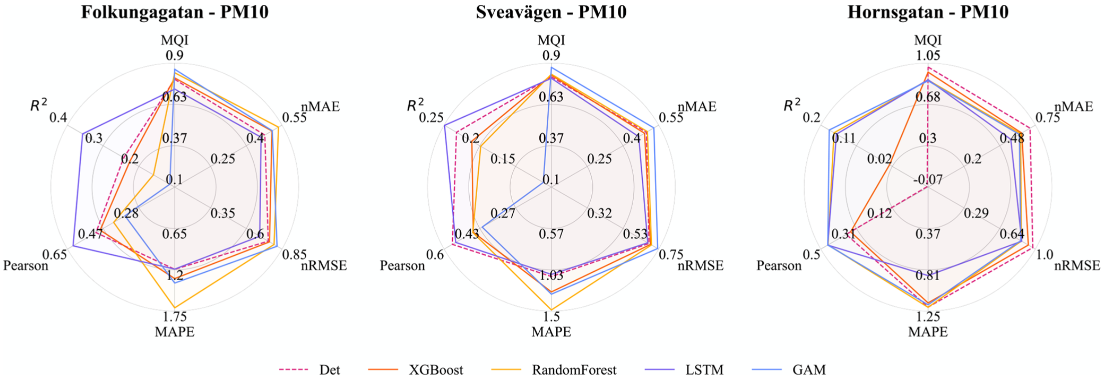


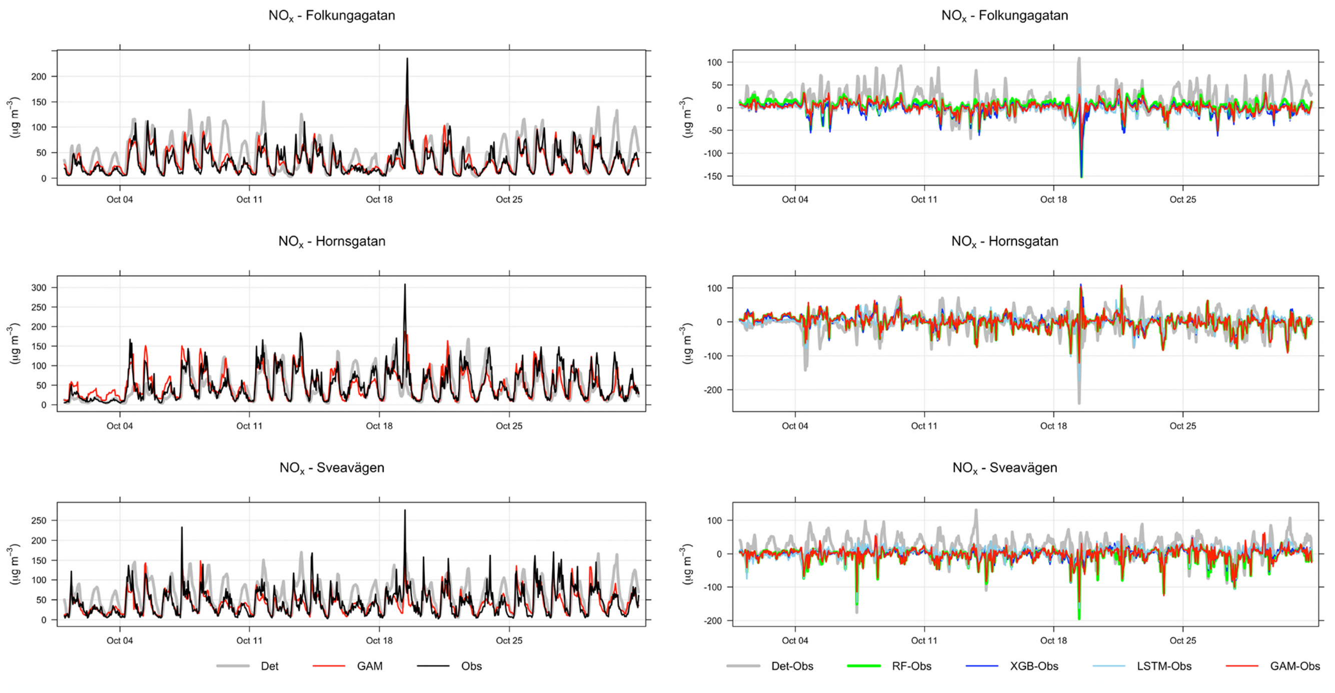
Figure 9. Statistical performances for ML models versus the deterministic hourly forecasts for PM_{10} at the street canyon sites. Mean of 1, 2, and 3 d forecasts. Note that the ranges are different for different metrics.

The improvement of the temporal variations of NO_x and PM_{10} is illustrated by comparing the mean diurnal variations in observations with deterministic model and other models in Appendix G. For all street sites, the deterministic forecasts of both NO_x and PM_{10} concentrations show systematic deviations from observations, which are corrected by applying the ML models, especially for NO_x. The tendency that the GAM model is not as good at capturing variations in PM_{10} at the urban site is also seen here for the street canyon sites.

As pointed out before, it is important to assess statistical performance measures for periods with high concentrations. Similar to what is shown for the urban site, the statistical performance indexes for all models are much worse for the hourly average concentrations that are higher than the mean values, and the pattern is also similar for the almost street sites, as shown in Appendix H. However, the performance of ML models for NO_x maintains the improvement in Hornsgatan, as detailed in Fig. 11, suggesting that the model effectively captures the significant variations in high concentration levels.

4.2.2 Importance of features – street canyon sites

For the street canyon sites, the feature importance rankings are different for PM_{10} and NO_x and also depend on ML models and street sites. Feature importance ranking in Hornsgatan is shown in Fig. 12, and detailed rankings are presented in Appendix I. There are, however, some typical features that tend to be more important. For PM_{10}, Julian day, lagged measurements, and deterministic forecasts are, in most cases, among the top 5 most important features for RF and XGB models, whereas precipitation is an important feature for LSTM models. For NO_x, deterministic forecasts, hour, and weekday are among the most important features, while the features of lagged measurements seem less useful for the ML models. The importance ranking of calendar features of NO_x models indicates the importance of diurnal and weekday variations in traffic emissions not properly captured by the deterministic forecast. The importance of Julian day reflects the seasonal variation of non-exhaust-emission PM_{10}, and the importance of precipitation reflects the impacts of street wetness on the suspension of road dust. Even though there are variations, it is difficult to summarize any



(a)

(b)

Figure 10. **(a)** Temporal variations in hourly mean NO_x concentrations at the street canyon site during October 2021 based on mean of 1, 2, and 3 d forecasts for observations (black), deterministic forecasts (grey) and GAM (red). **(b)** Absolute deviations of forecasted NO_x concentrations from observed (Obs) concentrations at the street canyon site based on a mean of 1, 2, and 3 d forecasts for October 2021.

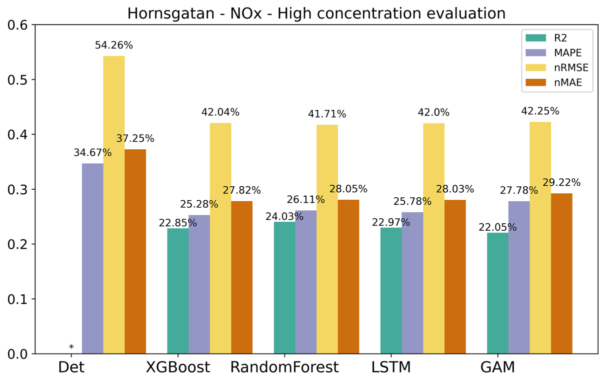


Figure 11. Statistical performance measures for forecasted NO_x hourly mean concentrations higher than the mean values at Hornsgatan, where * represents a negative R^2 value. Mean of 1, 2, and 3 d forecasts.

systematic difference in the features between ML models for the different street sites.

4.3 Generalization of street canyon modelling

Until now, the model performance has been evaluated using training and testing data from three single sites. In Stockholm, as well as in other cities, most of the streets do not have any monitoring stations. This is of course due to resource constraints but also associated with the fact that the EU Air Quality Directives regulate the number of monitoring sites required in a city depending on the level of air pollution and number of inhabitants. The monitoring stations should provide information for both areas where the highest concentrations of air pollutants occur and other areas that are representative of the exposure of the general population. Fewer resources are required if this information can be achieved by accurate enough modelling.

We therefore analyse the generalization capacities of the models, with the expectation that we can achieve certain prediction performances of one site without having any measurement data. Computational experiments were carried out through cross-validation, which combines training and testing data coming from different measurement sites. For the street canyon sites, four combinations of training datasets were applied to evaluate the generalization abilities of different ML models.

Figure 13 shows the mean of 1, 2, and 3 d forecasted NO_x and PM_{10} concentrations on the test set for the three street canyon sites based on training the models on the other streets. It shows that the forecast is improved compared to the deterministic forecast for Hornsgatan and Sveavägen but not so much for Folkungagatan. For Sveavägen the R^2 is 0.14 using the deterministic forecast, whereas the ML models give R^2 between 0.62 and 0.63, and here all errors decrease substantially using the ML models. However, for Folkungagatan

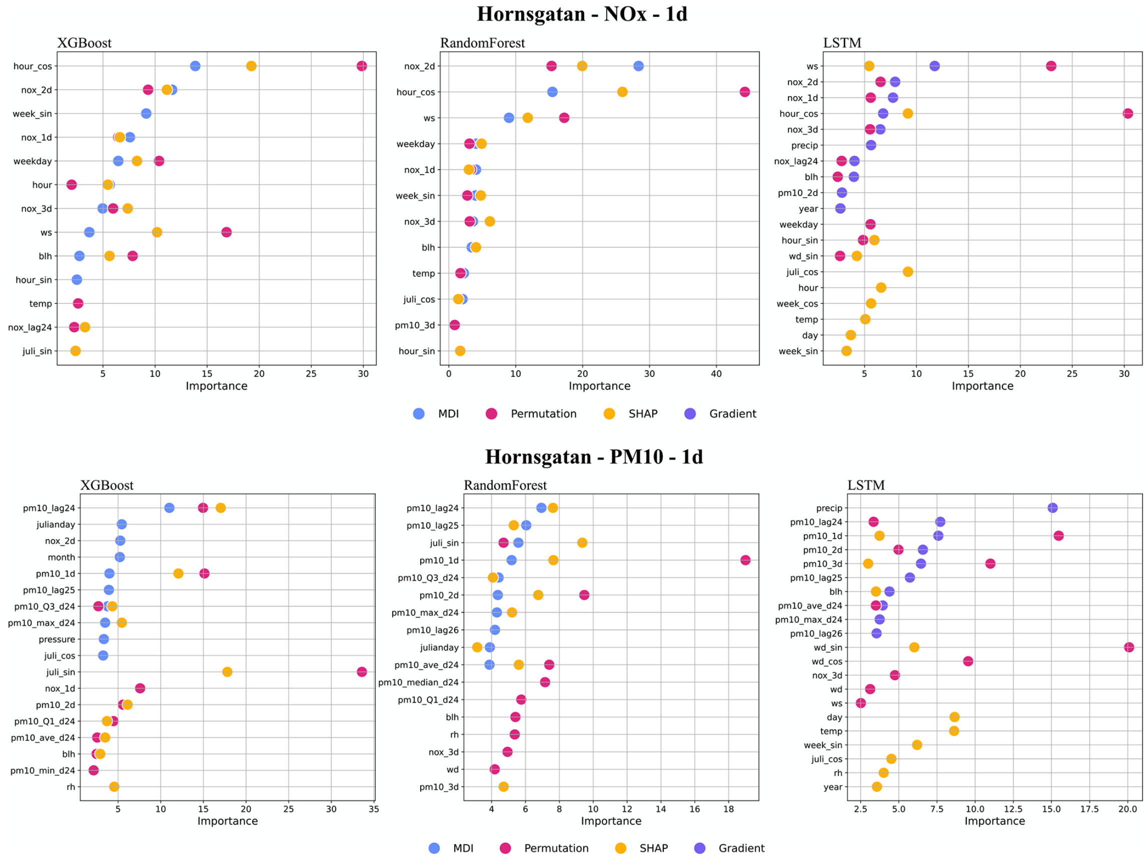


Figure 12. Top 10 most important features (%) for 1 d forecasts using XGB, RF, and LSTM at Hornsgatan. All data are hourly mean concentrations.

the ML models show different results. R^2 is similar or even decreases for tree-based models, whereas errors mostly decrease depending on the ML applied.

The performance of PM_{10} is shown in the right column of Fig. 13. It can be seen that it is not possible to find any major improvement in the deterministic forecast for the streets using RF and XGB. However, with LSTM R^2 increases slightly and errors decrease for Hornsgatan and Sveavägen compared to the deterministic forecasts.

5 Discussion

The performance of the ML models is quite similar for the different sites and forecast days. However, there are large differences in improvements for different pollutants. In general, our results indicate that ML models are more effective at improving NO_x than PM_{10}. For PM_{10} the ML models show slight improvement in R^2 but not much improvement in relative error. This difference in improvement is likely associated with the different processes controlling the concentrations, such as different sources: NO_x concentrations are mainly due to vehicle exhaust emissions, which show regular variations from 1 d to the next depending on day of the week and time of day, while PM_{10} is mainly due to road dust emissions controlled by a combination of variations in vehicle volumes and meteorological conditions that affect suspension of coarse particles from street surfaces (e.g. Denby et al., 2013a; Johansson et al., 2007; Krecl et al., 2021). Road dust accumulates on the road surfaces during wet road surface conditions and is suspended by vehicle-induced turbulence during dry conditions (Denby et al., 2013a).

The improvement of the forecasts of NO_x with ML is partly driven by the calendar, hour, day of the week. and to some degree also Julian day, but different features appear to be important for RF compared to XGB. For PM_{10}, the seasonal variation described by Julian day is the most important feature at the street canyon sites for both RF and XGB. This indicates that the deterministic forecasts are not capable of describing the impacts of meteorology and road dust emissions on PM_{10}, even though parameterizations of these processes are included in the deterministic modelling system. The total mass generated by road wear is a key factor for PM_{10} emissions, and these emissions are strongly controlled by surface moisture conditions. This is taken into account by the NORTRIP model. As pointed out by Denby et al. (2013b), there are periods where surface wetness is not

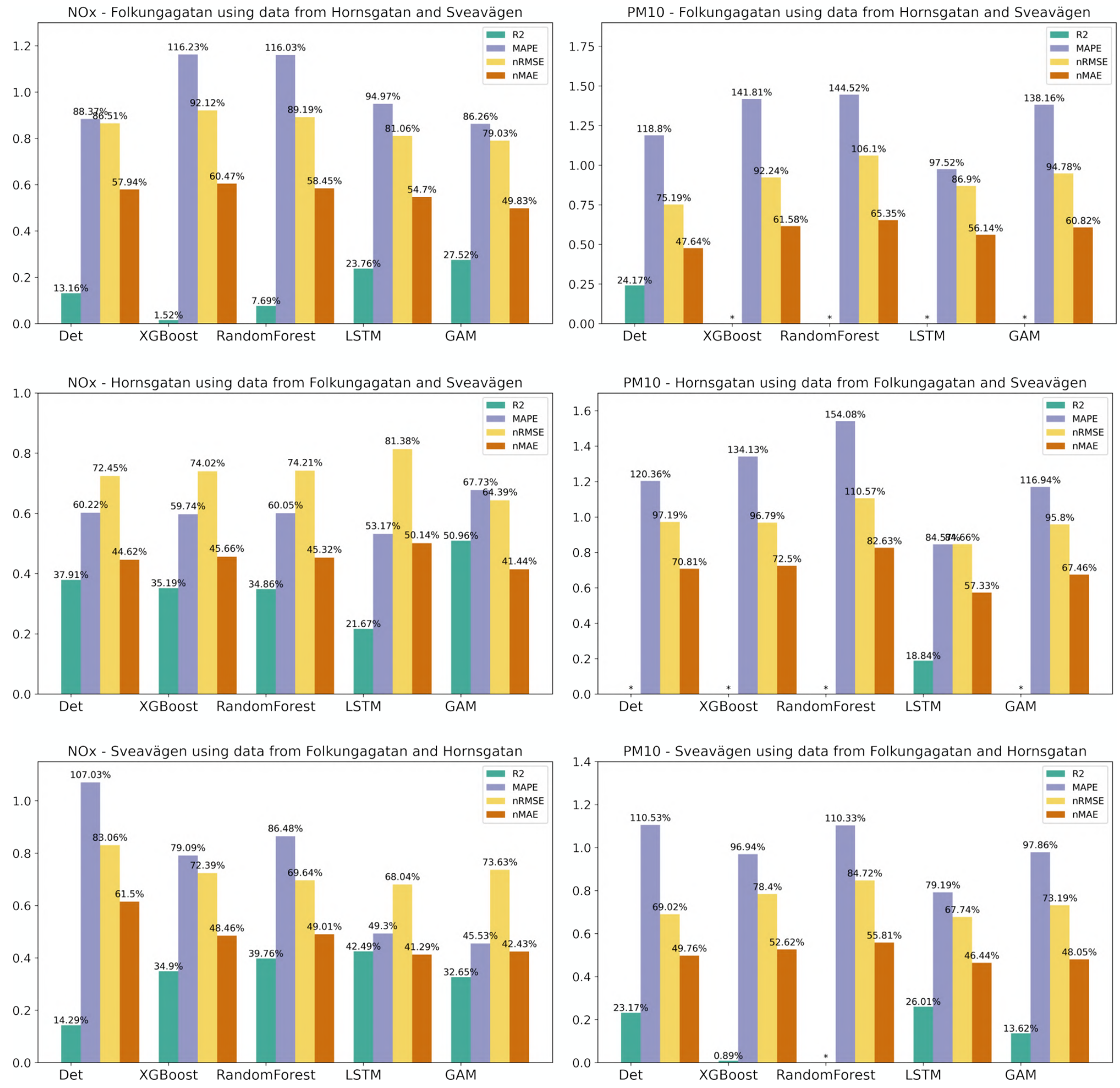


Figure 13. Statistical performances of NO_x and PM_{10} forecasts for the streets in the test set when the ML models are trained using only data from the other streets. Mean of 1, 2, and 3 d forecasts.

well modelled, and it is not known if this is the result of input data, e.g. precipitation, or of the model formulation itself.

It is clear that the deterministic forecast of O_3 underestimates concentrations at the urban site due to the fact that the local emissions of NO_x influencing the photochemistry are not properly considered by the CAMS model, but this is corrected using the ML models. Despite the deterministic forecast being the most important feature for both RF and XGB, lagged measured mean and maximum O_3 concentrations improve the deterministic forecasts.

Despite the fact that the configurations and traffic situations are quite similar for the street canyon sites, the improvements in the deterministic forecasts over ML models differ. For NO_x, the forecasts at Hornsgatan are more accurate (lower errors and higher R^2) than for the other two sites, while for PM_{10} there is no obvious difference between the sites.

The overall model quality according to the recommendations by the Forum for Air Quality Modeling in the context of the air quality directives is improved using the ML models, resulting in uncertainties that are significantly smaller than the measurement uncertainties for all pollutants. However, the forecasts of the highest concentrations, including episodes with high concentrations, are not systematically improved for all pollutants and all performance measures using the ML models.

We have shown that the statistical performances of the deterministic forecasts for concentrations of NO_x at the street canyon sites can be improved using the ML models. However, for PM_{10}, LSTM showed systematic improvements at all sites. Thus, this again accentuates the importance of not testing the models for only one pollutant. Further work is needed to improve deterministic forecasts of PM_{10} based on the training of ML models at a few monitoring stations. As discussed above, the situation in Stockholm is different from cities in central and southern Europe since the road dust contribution is very large. It might be that results for PM_{10} are different in other cities, but we have not found any publication on this matter.

5.1 Comparison of different ML models

Several studies have compared the performances of different machine learners in predicting air quality (Zaini et al., 2021). Assessing forecasts of PM_{10} and $PM_{2.5}$ concentrations, Czernecki et al. (2021) found that XGB performed the best, followed by RF and an artificial neural network model, while stepwise regression performed the worst in four Polish agglomerations. Likewise, Joharestani et al. (2019) found XGB to perform best of three ML models (XGB, RF, and a deep-learning algorithm), in predicting $PM_{2.5}$ in Tehran (Iran). On the contrary, LSTM was shown to outperform XGBoost for forecasting hourly $PM_{2.5}$ concentrations (Qadeer et al., 2020), similar to what was shown by Chuluunsaikhan et al. (2021). Cai et al. (2009) obtained more accurate predictions of CO concentrations using artificial neural network modelling compared to using multiple linear regression and the deterministic California line source dispersion model. On the other hand, Shaban et al. (2016) concluded that a tree-based algorithm (M5P) outperformed artificial neural network modelling when comparing forecasts of different pollutants in Qatar. There are many reasons for the different results presented in the literature, including model formulation and setup, different types of input data, and different atmospheric conditions and source contributions governing the concentrations. In addition, different performance metrics have been used. This makes it hard to draw general conclusions regarding which model to use. However, we find that other factors may be more important to consider than the type of model, such as sources of pollutants and influence of photochemistry, characteristics of the site resulting in different features being of varying importance depending on pollutant type of location. In this context, output of feature importance methods can provide useful information to improve models.

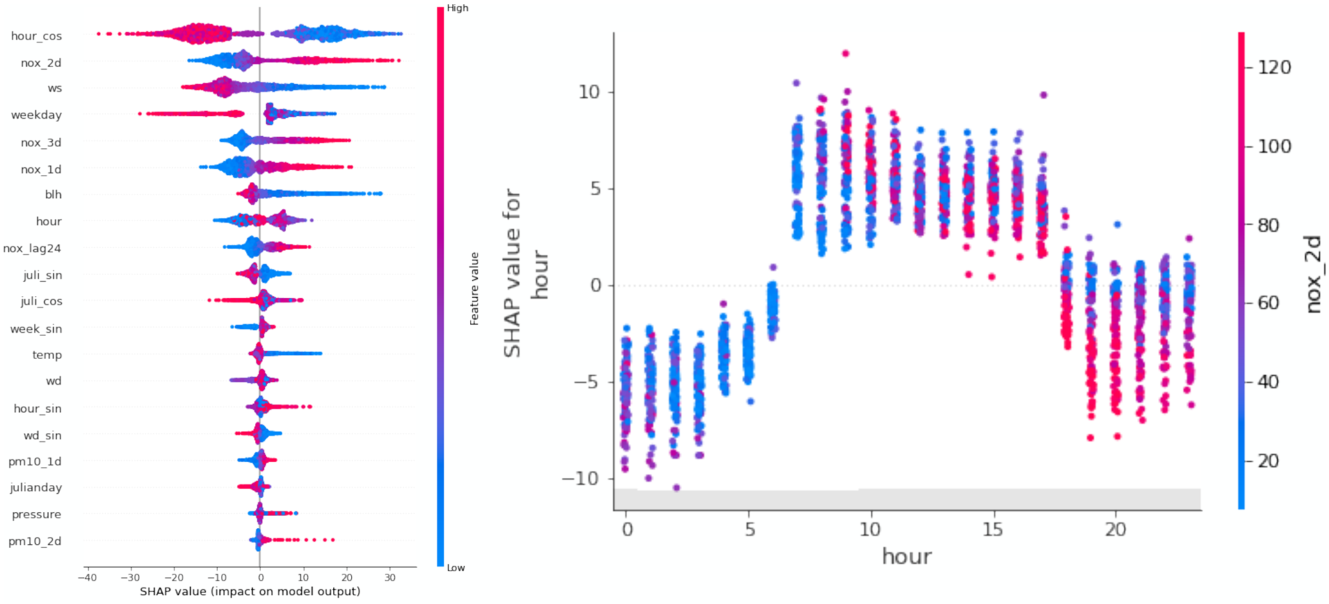
Another more practical aspect to consider when comparing the ML models is the complexity and computer resources required for training the models. In air quality literature, deep-learning models such as standard LSTM and other recurrent neural networks (RNNs) have been explored for their prediction capacities. However, most of the studies have adopted complex neural network structures, such as models of multiple outputs that mainly give convenience for data processing and automated feature handling. Nevertheless, training even a simple LSTM model is computationally much more expensive than the two conventional ML models, i.e. the decision-tree-based models (RF and XGB) in our case. In fact, we have to resort to the high-performance machine, the Swedish Berzelius high-performance computer, to reduce the computational time. For the current practice in our real air quality prediction system, we implemented the two tree-based models instead of LSTM. We are also exploring well-designed deep-learning models, which may replace the conventional models being adopted in the air quality system in the near future, especially due to the ability to deploy a generic model and handle all the modelling processes automatically.

5.2 Temporal dependency of feature importance

The exploration of feature importance is one contribution of the paper for analysing different ML models. In comparison to MDI and permutation methods, SHAP provides a more comprehensive approach to analysing feature importance. The model can compute the importance value of each feature for all data samples but also estimate the feature importance value for each individual sample. This gives us a useful tool to analyse the temporal dependency of feature importance.

Figure 14 illustrates the feature importance analysis using the SHAP method for an XGBoost model of 1 d NO_x prediction. Figure 14a illustrates the feature importance ranking derived from test dataset, employing red dots to denote samples with higher numerical feature values and blue dots to represent lower numerical values. In addition, the dots on the left side of the x axis, i.e. SHAP value < 0, reflect a negative impact for predictions, while dots on the right side suggest a positive impact. Figure 14a revealed a distinct relationship between the feature *hour_cos* and the NO_x predictions. Higher values of *hour_cos*, representing night-time, exhibit a negative impact on the forecasts. Conversely, lower values of *hour_cos* show a positive correlation with the forecasts. Additionally, the wider distribution of this feature indicates its significant influence on the prediction process, suggesting that the model may capture the diurnal pattern of traffic emissions. In Fig. 14b, a more pronounced diurnal pattern emerges. Here, SHAP values of feature *hour* are positive from 07:00 to 17:00 CET, contrasting with the negative values observed at night. Meanwhile, the high concentration of NO_x forecasts *nox_2d* from 2 d deterministic model (red dots) show an evident increase during the heavy traffic period, spanning from 08:00 to 01:00 CET of the next day. This observation reinforces the substantial effect of traffic emissions on NO_x levels.

Figure 15 displays a heatmap of SHAP values, illustrating the temporal variation of feature importance when they are used by a model to forecast. The deterministic forecast *nox_3d* plays an important role in prediction, i.e. executes positive influence (red block) for NO_x predictions with higher numeric value and vice versa. Meanwhile, the weekend, e.g. 2, 9, 16, and 23 October, exhibits negative impacts



(a)

(b)

Figure 14. (a) Feature importance ranking based on the SHAP method of XGBoost model for the 1 d NO_x prediction at the HO site. **(b)** The relationship between feature *hour* and feature *nox_2d* from the results of SHAP method in panel **(a)**. All examples belong to the test set.

(blue block), while the weekday factor provides positive support for model forecasts. The impact of the 24 h lagged values of NO_x, *nox_lag24*, is also evident. For example, the SHAP value at the peak on 19 October has a negative impact, whereas the SHAP value of the next day shows a positive impact, which explains the delay between the predicted peak and real observation.

6 Conclusions

This paper has applied different ML models to improve 1, 2, and 3 d deterministic forecasts of NO_x, PM_{10}, and O_3 concentrations for multiple locations in Stockholm, Sweden. It is shown that the degree of improvement over deterministic forecasts depends more on pollutant and monitoring site than on what ML algorithm is applied. Also, four feature importance methods, namely MDI, permutation, gradient-based, and SHAP, are utilized to identify significant features that are common and robust across models. Notably, deterministic forecasts of NO_x are significantly improved across all sites using all models. R^2 is increased by up to 80 %, and prediction errors are reduced by up to 60 %. For PM_{10}, variable results are achieved, reflecting the more complicated processes controlling the road wear emissions that constitute a large fraction of PM_{10}. For O_3 at the urban background site, the deviation between deterministically modelled absolute level is corrected by the ML models, and nRMSE and nMAE are reduced by on average around 20 %.

We have shown that it is possible to improve deterministic forecasts of NO_x at street canyon sites based on training ML models at other sites. When tested for PM_{10}, only LSTM shows modest improvements compared to the deterministic forecasts.

One contribution of our study is that we compare forecasts based on several pollutants and base our forecasts on a combination of deterministic models, which are based on the underlying physicochemical mechanisms responsible for the emissions and dispersion of the pollutants, and three different ML models with additional variables, such as measurement data, calendar data, and meteorological data. The models are evaluated at different sites and for different pollutants during several months with different meteorological conditions. In addition, by comparing the four feature importance methods, the robust features for associated models are identified, establishing the foundation for model performance analysis and improvement.

There are different aspects that we would like to further improve and extend regarding the models. Investigating the impact of the COVID-19 pandemic on our model's performance is meaningful, especially considering that our dataset predominantly covers this specific time period during the pandemic. Moreover, we will further explore the means to transfer the learning approach to more general models, addressing the challenges posed by the scarcity of monitoring stations in many areas and the representation of spatial correlation of the measurement stations.

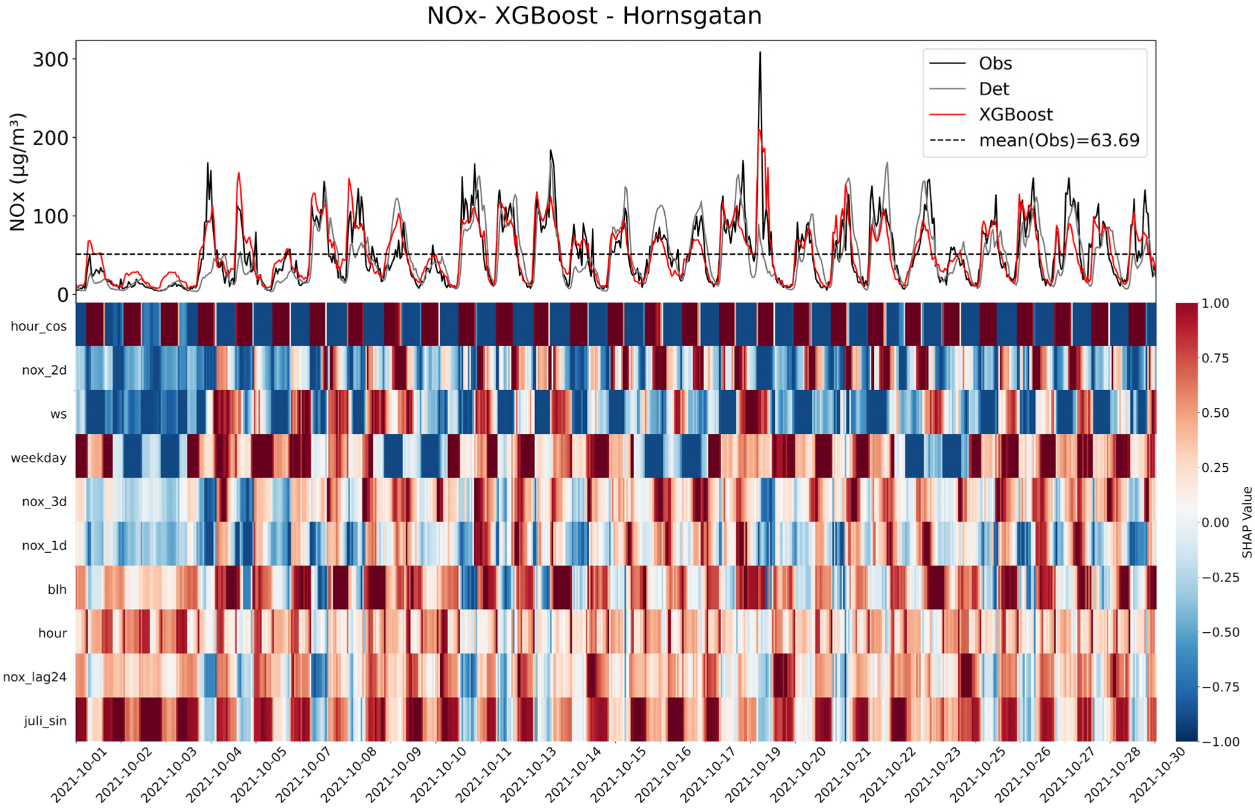


Figure 15. SHAP feature importance analysis of the XGBoost model for the 1 d prediction of NO_x concentrations at the HO street site. All examples belong to the test set. The blue blocks imply a negative impact, while the red blocks are positive.

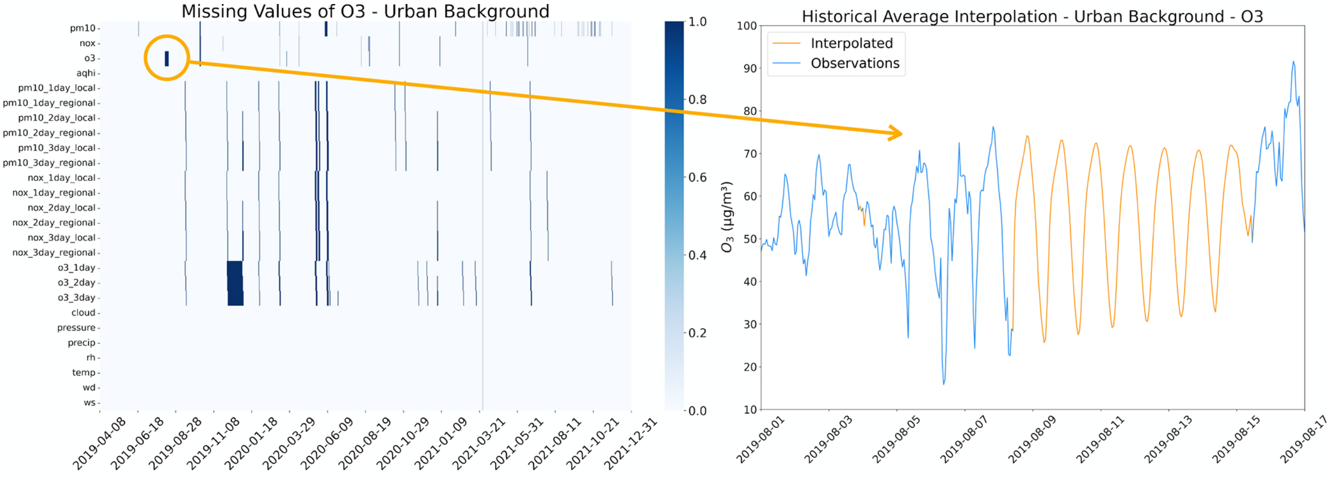
Appendix A: Description of measurement methods and sites

All measurement methods are approved for monitoring according to the EU Air Quality Directives for NO_x, O_3, and PM_{10}. PM_{10} was measured either using an optical particle counter (Hornsgatan: OPC, Grimm EDM 180-MC) or tapered element oscillating microbalance (Sveavägen, Folkungagatan, and urban: TEOM model, 1400AB, Rupprecht & Patashnik, Co.). NO_x was measured using chemiluminescence (AC32M, Environnement S.A.), and O_3 was measured by UV absorption (O342M, Environnement S.A.).

Table A1. Description of monitoring sites.

Site name	Description	Traffic volume	Photo
Hornsgatan	Street canyon site. Measurements of NO_x and PM_{10} on north side of street, 3 m above ground. Street width 24 m and building height 24 m.	23 000 vehicles per day (4 % heavy-duty vehicles). Vehicle composition measured during 4-week campaigns using automatic number plate recognition.	
Sveavägen	Street canyon site. Measurements of NO_x and PM_{10} on west side of street, 3 m above ground. Street width 33 m and building height 24 m.	21 000 vehicles per day (7 % heavy-duty vehicles).	
Folkungagatan	Street canyon site. Measurements NO_x and PM_{10} on west side of street, 3 m above ground. Street width 24 m and building height 24 m.	12 000 vehicles per day (18 % heavy-duty vehicles).	
Torkel Knutssongatan	Urban background. Measurements of NO_x, PM_{10}, ozone, and meteorology on top of a 20 m high building.	ca. 13 000 vehicles on Hornsgatan road 250 m north of site.	

Appendix B: Interpolation



(a)

(b)

Figure B1. **(a)** The missing value of O_3 in the UB dataset, where blue represents missing data and white represents not missing data. **(b)** Interpolation results based on historical averages for O_3 in the UB dataset. The yellow arrows indicate the interpolation results for missing values of O_3 within the yellow circle.

Appendix C: Hyperparameter tuning

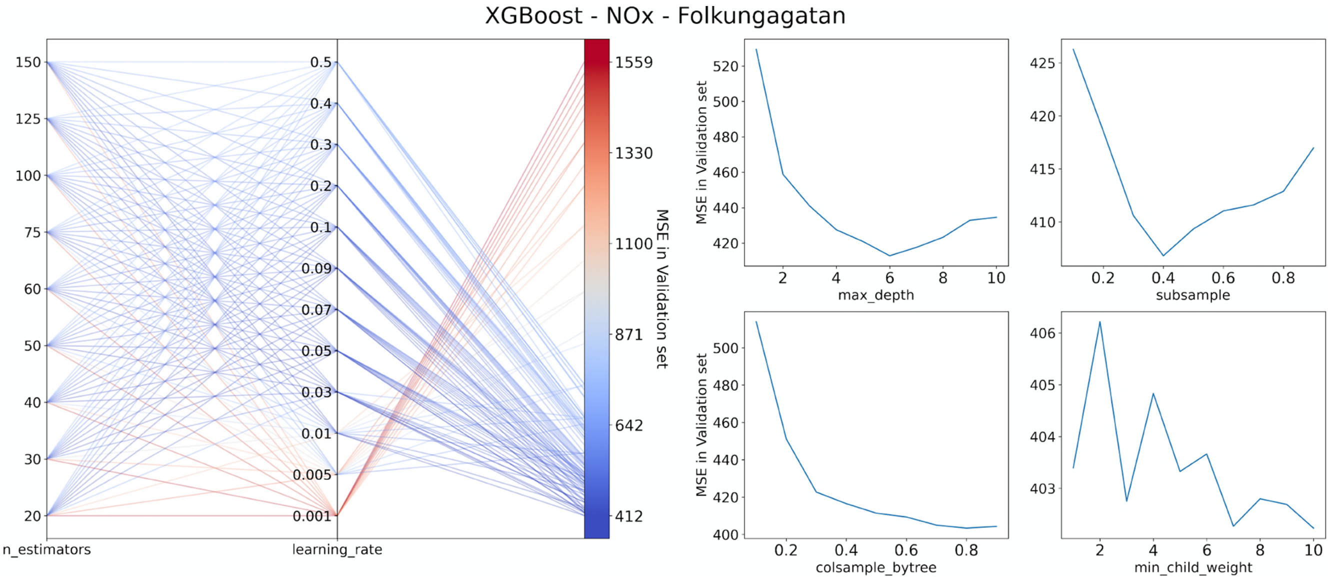


Figure C1. Illustration of the results of hyperparameter tuning for the XGBoost model of NO_x at Folkungagatan.

Table C1. The result of hyperparameter tuning for all models and all sites.

Station	Pollutants	Models	Range of hyperparameters	Best parameters
FO	NO_x	XGBoost	’n_estimators’: [20, 30, 40, 50, 60, 75, 100, 125, 150], ’learning_rate’: [0.005, 0.01, 0.03, 0.05, 0.07, 0.09, 0.1, 0.2, 0.3] ”max_depth”: [1, 2, 3, 4, 5, 6, 7, 8, 9, 10], ”subsample”: [0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9], ”colsample_bytree”: [0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9], ”min_child_weight”: [1, 2, 3, 4, 5, 6, 7, 8, 9, 10].	’n_estimators’: 60, ’max_depth’: 6, ’min_child_weight’: 10, ’colsample_bytree’: 0.8, ’learning_rate’: 0.03, ’subsample’: 0.4.
FO	NO_x	RandomForest	’n_estimators’: [50, 100, 150, 200, 250, 300, 325, 350, 375, 400], ’max_features’: [None, ’sqrt’, ’log2’], ’max_depth’: [None, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10], ’min_samples_split’: [1, 2, 3, 4, 5, 6, 7, 8, 9, 10], ’min_samples_leaf’: [1, 2, 3, 4, 5, 6, 7, 8, 9, 10].	’max_features’: ’sqrt’, ’n_estimators’: 250, ’max_depth’: 7, ’min_samples_split’: 10, ’min_samples_leaf’: 9
FO	NO_x	LSTM	’batch_size’: [24, 48, 72, 96, 120, 144, 168], ’n_steps_in’: [12, 24, 36, 48, 60], ’hidden_size’: [32, 64, 96, 128, 160], ’learning_rate’: $[1e^{-2}, 5e^{-2}, 1e^{-3}, 5e^{-3}, 1e^{-4}]$.	’batch_size’: 168, ’n_steps_in’: 48, ’hidden_size’: 160, ’learning_rate’: 0.001.
FO	PM_{10}	XGBoost	’n_estimators’: [20, 30, 40, 50, 60, 75, 100, 125, 150], ’learning_rate’: [0.005, 0.01, 0.03, 0.05, 0.07, 0.09, 0.1, 0.2, 0.3] ”max_depth”: [1, 2, 3, 4, 5, 6, 7, 8, 9, 10], ”subsample”: [0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9], ”colsample_bytree”: [0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9], ”min_child_weight”: [1, 2, 3, 4, 5, 6, 7, 8, 9, 10].	’learning_rate’: 0.06, ’n_estimators’: 300, ’max_depth’: 2, ’subsample’: 0.5, ’colsample_bytree’: 0.3, ’min_child_weight’: 9.
FO	PM_{10}	RandomForest	’n_estimators’: [50, 100, 150, 200, 300, 400, 425, 450, 475, 500, 550], ’max_features’: [None, ’sqrt’, ’log2’], ’max_depth’: [None, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10], ’min_samples_split’: [1, 2, 3, 4, 5, 6, 7, 8, 9, 10], ’min_samples_leaf’: [1, 2, 3, 4, 5, 6, 7, 8, 9, 10].	’max_features’: None, ’n_estimators’: 475, ’max_depth’: None, ’min_samples_split’: 1, ’min_samples_leaf’: 1.
FO	PM_{10}	LSTM	’batch_size’: [24, 48, 72, 96, 120, 144, 168], ’n_steps_in’: [12, 24, 36, 48, 60], ’hidden_size’: [32, 64, 96, 128, 160], ’learning_rate’: $[1e^{-2}, 5e^{-2}, 1e^{-3}, 5e^{-3}, 1e^{-4}]$.	’batch_size’: 168, ’n_steps_in’: 60, ’hidden_size’: 128, ’learning_rate’: 0.001.
HO	NO_x	XGBoost	’n_estimators’: [20, 30, 40, 50, 60, 75, 100,125, 150], ’learning_rate’: [0.08, 0.085, 0.09, 0.095, 0.1, 0.2, 0.3, 0.4, 0.5], ”max_depth”: [1, 2, 3, 4, 5, 6, 7, 8, 9, 10], ”subsample”: [0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9], ”colsample_bytree”: [0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9], ”min_child_weight”: [1, 2, 3, 4, 5, 6, 7, 8, 9, 10].	’learning_rate’: 0.095, ’n_estimators’: 40, ’max_depth’: 6, ’subsample’: 0.8, ’colsample_bytree’: 0.7, ’min_child_weight’: 6.
HO	NO_x	RandomForest	’n_estimators’: [50, 100, 150, 200, 250, 300, 325, 350, 375, 400], ’max_features’: [None, ’sqrt’, ’log2’], ’max_depth’: [None, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10], ’min_samples_split’: [1, 2, 3, 4, 5, 6, 7, 8, 9, 10], ’min_samples_leaf’: [1, 2, 3, 4, 5, 6, 7, 8, 9, 10].	’max_features’: None, ’n_estimators’: 375, ’max_depth’: None, ’min_samples_split’: 1, ’min_samples_leaf’: 2.
HO	NO_x	LSTM	’batch_size’: [24, 48, 72, 96, 120, 144, 168], ’n_steps_in’: [12, 24, 36, 48, 60], ’hidden_size’: [32, 64, 96, 128, 160], ’learning_rate’: $[1e^{-2}, 5e^{-2}, 1e^{-3}, 5e^{-3}, 1e^{-4}]$.	’batch_size’: 168, ’n_steps_in’: 60, ’hidden_size’: 160, ’learning_rate’:0.005.
HO	PM_{10}	XGBoost	’n_estimators’: [20, 30, 40, 50, 60, 75, 100, 125, 150], ’learning_rate’: [0.08, 0.085, 0.09, 0.095, 0.1, 0.2, 0.3, 0.4, 0.5], ”max_depth”: [1, 2, 3, 4, 5, 6, 7, 8, 9, 10], ”subsample”: [0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9], ”colsample_bytree”: [0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9], ”min_child_weight”: [1, 2, 3, 4, 5, 6, 7, 8, 9, 10].	’learning_rate’: 0.085, ’n_estimators’: 30, ’max_depth’: 4, ’subsample’: 0.6, ’colsample_bytree’: 0.8, ’min_child_weight’: 1.

Table C1. Continued.

Station	Pollutants	Models	Range of hyperparameters	Best parameters
HO	PM_{10}	RandomForest	’n_estimators’: [50, 100, 150, 200, 300, 400, 425, 450, 475, 500, 550], ’max_features’: [None, ’sqrt’, ’log2’], ’max_depth’: [None, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10], ’min_samples_split’: [1, 2, 3, 4, 5, 6, 7, 8, 9, 10], ’min_samples_leaf’: [1, 2, 3, 4, 5, 6, 7, 8, 9, 10].	’max_features’: ’sqrt’, ’n_estimators’: 450, ’max_depth’: None, ’min_samples_split’: 4, ’min_samples_leaf’: 1.
HO	PM_{10}	LSTM	’batch_size’: [24, 48, 72, 96, 120, 144, 168], ’n_steps_in’: [12, 24, 36, 48, 60], ’hidden_size’: [32, 64, 96, 128, 160], ’learning_rate’: [$1e^{-2}, 5e^{-2}, 1e^{-3}, 5e^{-3}, 1e^{-4}$].	’batch_size’: 168, ’n_steps_in’: 60, ’hidden_size’: 32, ’learning_rate’: 0.001.
SV	NO_x	XGBoost	’n_estimators’: [20, 30, 40, 50, 60, 75, 100, 125, 150], ’learning_rate’: [0.001, 0.005, 0.01, 0.03, 0.05, 0.07, 0.09,0.1, 0.2, 0.3, 0.4, 0.5], ”max_depth”: [1, 2, 3, 4, 5, 6, 7, 8, 9, 10], ”subsample”: [0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9], ”colsample_bytree”: [0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9], ”min_child_weight”: [1, 2, 3, 4, 5, 6, 7, 8, 9, 10].	’learning_rate’: 0.09, ’n_estimators’: 60, ’max_depth’: 6, ’subsample’: 0.8, ’colsample_bytree’: 0.6, ’min_child_weight’: 10.
SV	NO_x	RandomForest	’n_estimators’: [50, 100, 150, 200, 250, 300, 325, 350, 375, 400], ’max_features’: [None, ’sqrt’, ’log2’], ’max_depth’: [None, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10], ’min_samples_split’: [1, 2, 3, 4, 5, 6, 7, 8, 9, 10], ’min_samples_leaf’: [1, 2, 3, 4, 5, 6, 7, 8, 9, 10].	’max_features’: ’log2’, ’n_estimators’: 375, ’max_depth’: None, ’min_samples_split’: 8, ’min_samples_leaf’: 5
SV	NO_x	LSTM	’batch_size’: [24, 48, 72, 96, 120, 144, 168], ’n_steps_in’: [12, 24, 36, 48, 60], ’hidden_size’: [32, 64, 96, 128, 160], ’learning_rate’: [$1e^{-2}, 5e^{-2}, 1e^{-3}, 5e^{-3}, 1e^{-4}$].	’batch_size’: 168, ’n_steps_in’: 12, ’hidden_size’: 64, ’learning_rate’: 0.001.
SV	PM_{10}	XGBoost	’n_estimators’: [30, 40, 50, 100, 150, 200, 250, 300, 350, 400, 450, 500], ’learning_rate’: [0.001, 0.005, 0.01, 0.02, 0.03, 0.04, 0.05, 0.06, 0.07, 0.08, 0.09, 0.1, 0.2, 0.3, 0.4], ”max_depth”: [1, 2, 3, 4, 5, 6, 7, 8, 9, 10], ”subsample”: [0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9], ”colsample_bytree”: [0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9], ”min_child_weight”: [1, 2, 3, 4, 5, 6, 7, 8, 9, 10].	’learning_rate’: 0.02, ’n_estimators’: 50, ’max_depth’: 3, ’subsample’: 0.2, ’colsample_bytree’: 0.9, ’min_child_weight’: 1
SV	PM_{10}	RandomForest	’n_estimators’: [50, 100, 150, 200, 300, 400, 425, 450, 475, 500, 550], ’max_features’: [None, ’sqrt’, ’log2’], ’max_depth’: [None, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10], ’min_samples_split’: [1, 2, 3, 4, 5, 6, 7, 8, 9, 10], ’min_samples_leaf’: [1, 2, 3, 4, 5, 6, 7, 8, 9, 10].	’max_features’: ’log2’, ’n_estimators’: 500, ’max_depth’: 8, ’min_samples_split’: 3, ’min_samples_leaf’: 1
SV	PM_{10}	LSTM	’batch_size’: [24, 48, 72, 96, 120, 144, 168], ’n_steps_in’: [12, 24, 36, 48, 60], ’hidden_size’: [32, 64, 96, 128, 160], ’learning_rate’: [$1e^{-2}, 5e^{-2}, 1e^{-3}, 5e^{-3}, 1e^{-4}$].	’batch_size’: 168, ’n_steps_in’: 48, ’hidden_size’: 96, ’learning_rate’:0.01.
UB	NO_x	XGBoost	’n_estimators’: [20, 30, 40, 50, 60, 75, 100, 125, 150], ’learning_rate’: [0.001, 0.005, 0.01, 0.02, 0.03, 0.04, 0.05, 0.07, 0.09, 0.1, 0.2, 0.3, 0.4], ”max_depth”: [1, 2, 3, 4, 5, 6, 7, 8, 9, 10], ”subsample”: [0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9], ”colsample_bytree”: [0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9], ”min_child_weight”: [1, 2, 3, 4, 5, 6, 7, 8, 9, 10].	’learning_rate’: 0.02, ’n_estimators’: 150, ’max_depth’: 6, ’subsample’: 0.8, ’colsample_bytree’: 0.6, ’min_child_weight’: 3.
UB	NO_x	RandomForest	’n_estimators’: [50, 100, 150, 200, 225, 250, 275, 300, 325, 350, 375, 400], ’max_features’: [None, ’sqrt’, ’log2’], ’max_depth’: [None, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10], ’min_samples_split’: [1, 2, 3, 4, 5, 6, 7, 8, 9, 10], ’min_samples_leaf’: [1, 2, 3, 4, 5, 6, 7, 8, 9, 10].	’max_features’: ’sqrt’, ’n_estimators’: 275, ’max_depth’: 10, ’min_samples_split’: 1, ’min_samples_leaf’: 7.
UB	NO_x	LSTM	’batch_size’: [24, 48, 72, 96, 120, 144, 168], ’n_steps_in’: [12, 24, 36, 48, 60], ’hidden_size’: [32, 64, 96, 128, 160], ’learning_rate’: [$1e^{-2}, 5e^{-2}, 1e^{-3}, 5e^{-3}, 1e^{-4}$].	’batch_size’: 168, ’n_steps_in’: 60, ’hidden_size’: 160, ’learning_rate’: 0.001.
UB	PM_{10}	XGBoost	’n_estimators’: [50, 75, 100, 200, 300, 400, 500, 600], ’learning_rate’: [0.01, 0.03, 0.04, 0.05, 0.06, 0.07, 0.09, 0.1, 0.2, 0.3, 0.4], ”max_depth”: [1, 2, 3, 4, 5, 6, 7, 8, 9, 10], ”subsample”: [0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9], ”colsample_bytree”: [0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9], ”min_child_weight”: [1, 2, 3, 4, 5, 6, 7, 8, 9, 10].	’learning_rate’: 0.04, ’n_estimators’: 600, ’max_depth’: 6, ’subsample’: 0.4, ’colsample_bytree’: 0.8, ’min_child_weight’: 1.

Table C1. Continued.

Station	Pollutants	Models	Range of hyperparameters	Best parameters
UB	PM_{10}	RandomForest	’n_estimators’: [50, 100, 150, 200, 250, 300, 325, 350, 375, 400], ’max_features’: [None, ’sqrt’, ’log2’], ’max_depth’: [None, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10], ’min_samples_split’: [1, 2, 3, 4, 5, 6, 7, 8, 9, 10], ’min_samples_leaf’: [1, 2, 3, 4, 5, 6, 7, 8, 9, 10].	’max_features’: None, ’n_estimators’: 250, ’max_depth’: None, ’min_samples_split’: 6, ’min_samples_leaf’: 5.
UB	PM_{10}	LSTM	’batch_size’: [24, 48, 72, 96, 120, 144, 168], ’n_steps_in’: [12, 24, 36, 48, 60], ’hidden_size’: [32, 64, 96, 128, 160], ’learning_rate’: $[1e^{-2}, 5e^{-2}, 1e^{-3}, 5e^{-3}, 1e^{-4}]$.	’batch_size’: 168, ’n_steps_in’: 24, ’hidden_size’: 96, ’learning_rate’: 0.001.
UB	O_3	XGBoost	’n_estimators’: [50, 100, 150, 200, 250, 275, 300, 325, 350, 400], ’learning_rate’: [0.02, 0.03, 0.04, 0.05, 0.06, 0.08, 0.2, 0.3, 0.4], ”max_depth”: [1, 2, 3, 4, 5, 6, 7, 8, 9, 10], ”subsample”: [0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9], ”colsample_bytree”: [0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9], ”min_child_weight”: [1, 2, 3, 4, 5, 6, 7, 8, 9, 10].	’learning_rate’: 0.04, ’n_estimators’: 300, ’max_depth’: 4, ’subsample’: 0.7, ’colsample_bytree’: 0.7, ’min_child_weight’: 10.
UB	O_3	RandomForest	’n_estimators’: [50, 100, 200, 300, 350, 375, 400, 425, 450, 500, 550, 600], ’max_features’: [None, ’sqrt’, ’log2’], ’max_depth’: [None, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10], ’min_samples_split’: [1, 2, 3, 4, 5, 6, 7, 8, 9, 10], ’min_samples_leaf’: [1, 2, 3, 4, 5, 6, 7, 8, 9, 10].	’max_features’: None, ’n_estimators’: 400, ’max_depth’: None, ’min_samples_split’: 1, ’min_samples_leaf’: 7.
UB	O_3	LSTM	’batch_size’: [24, 48, 72, 96, 120, 144, 168], ’n_steps_in’: [12, 24, 36, 48, 60], ’hidden_size’: [32, 64, 96, 128, 160], ’learning_rate’: $[1e^{-2}, 5e^{-2}, 1e^{-3}, 5e^{-3}, 1e^{-4}]$.	’batch_size’: 168, ’n_steps_in’: 24, ’hidden_size’: 128, ’learning_rate’: 0.0001.

Appendix D: Temporal variations in hourly mean NO_x, PM_{10}, and O_3 concentrations at the urban background

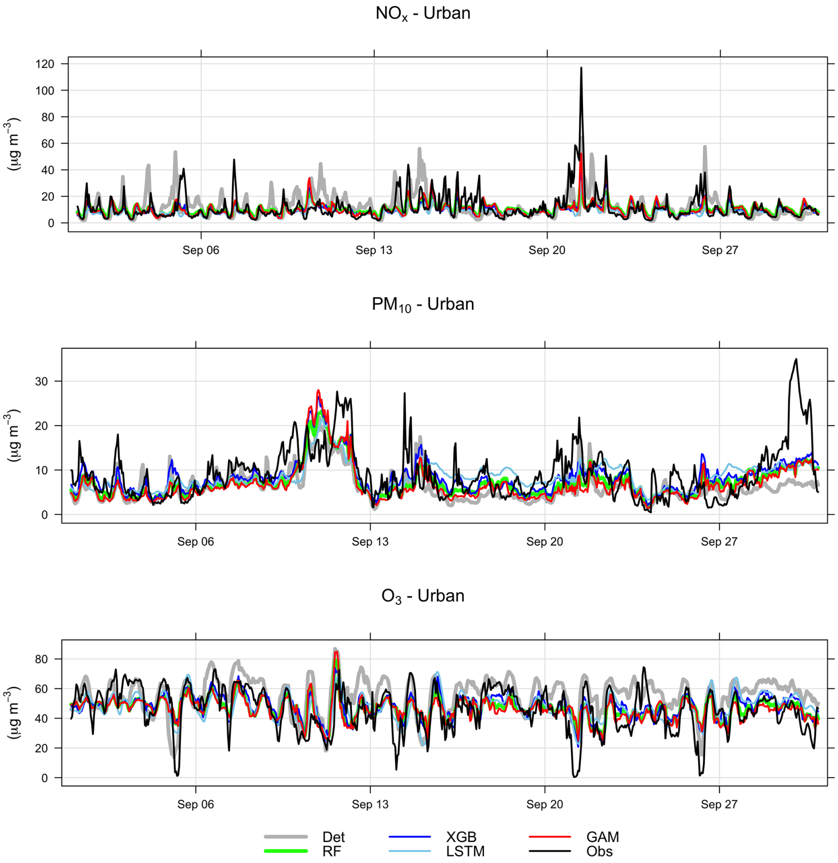


Figure D1. Temporal variations of deterministic and ML-forecasted NO_x, PM_{10}, and O_3 concentrations together with corresponding measured concentrations at the urban background site for September 2021. Mean of 1, 2, and 3 d forecasts.

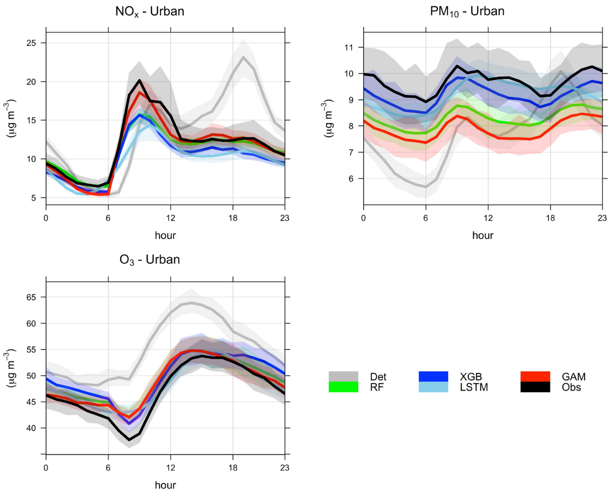


Figure D2. Mean diurnal variations in measured and forecasted concentrations of NO_x, PM_{10}, and O_3 at the urban site. Mean of 1, 2, and 3 d forecasts for June–December 2021.

Appendix E: Statistical performance measures for forecasts higher than the hourly mean concentrations at the urban site

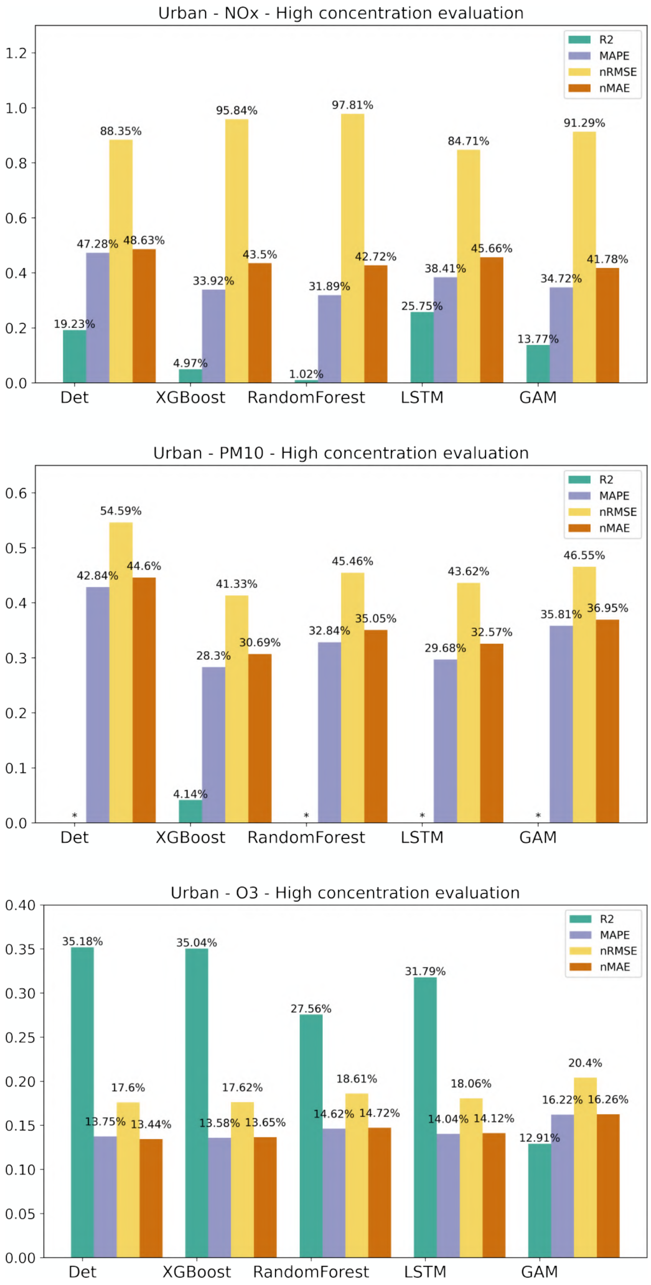


Figure E1. Statistical performance measures for concentrations of NO_x, PM_{10}, and O_3 higher than the hourly mean value at the urban site, where * represents a negative R^2 value. Mean of 1, 2, and 3 d forecasts.

Appendix F: Importance of features – urban

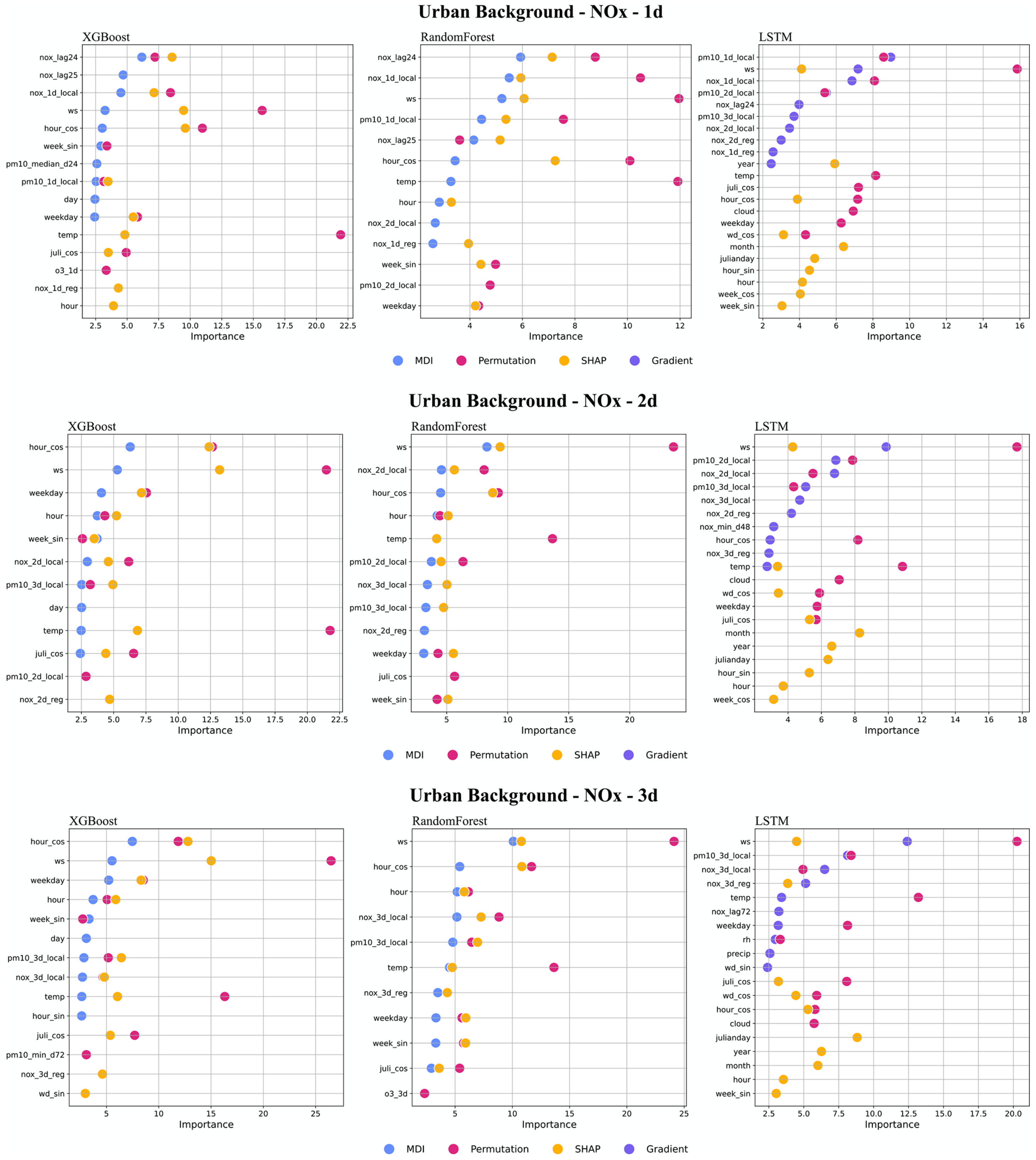


Figure F1. Top 10 most important features (%) for NO_x forecasts using XGB, RF, and LSTM at the urban site.

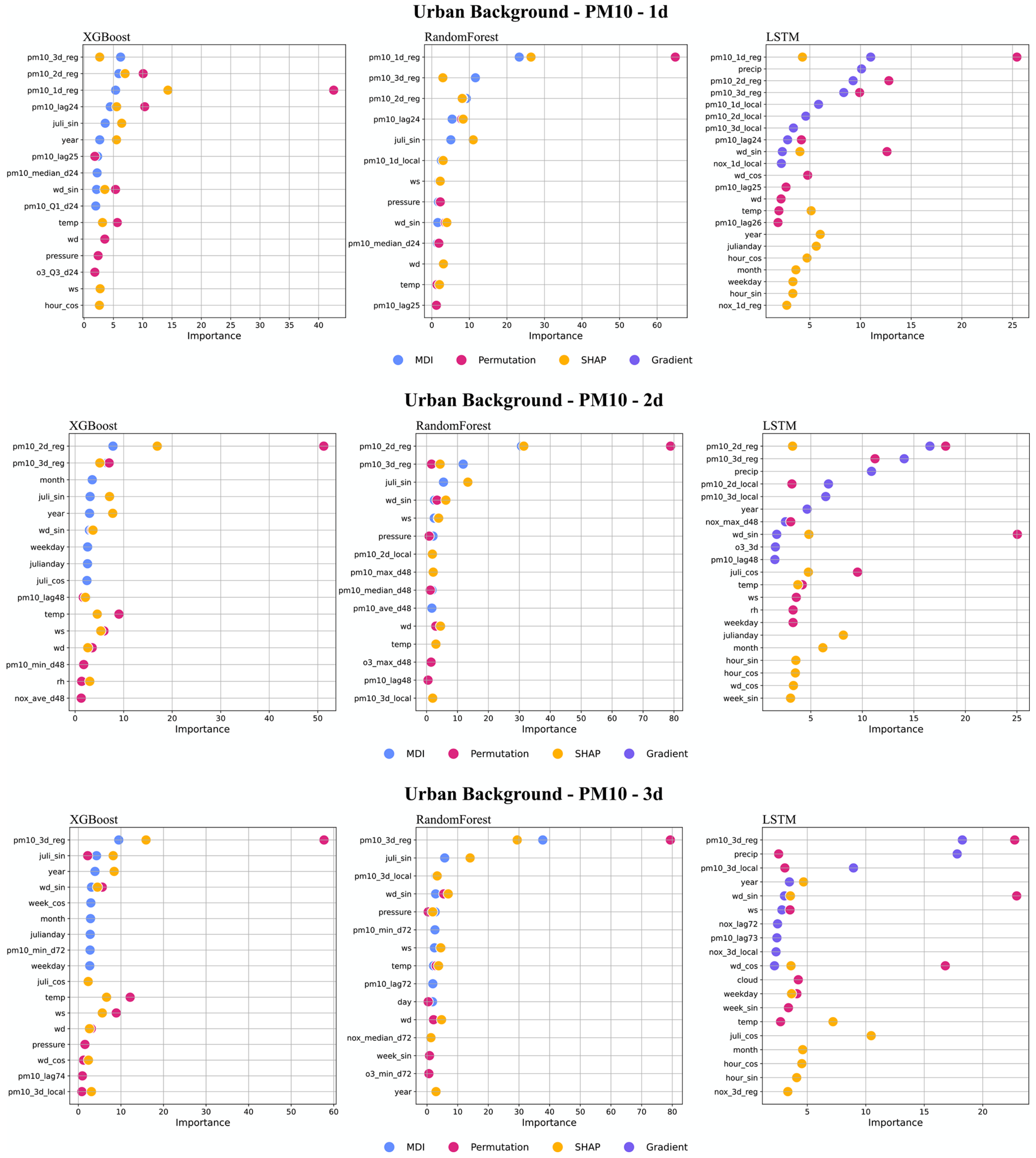


Figure F2. Top 10 most important features (%) for PM_{10} forecasts using XGB, RF, and LSTM at the urban site.

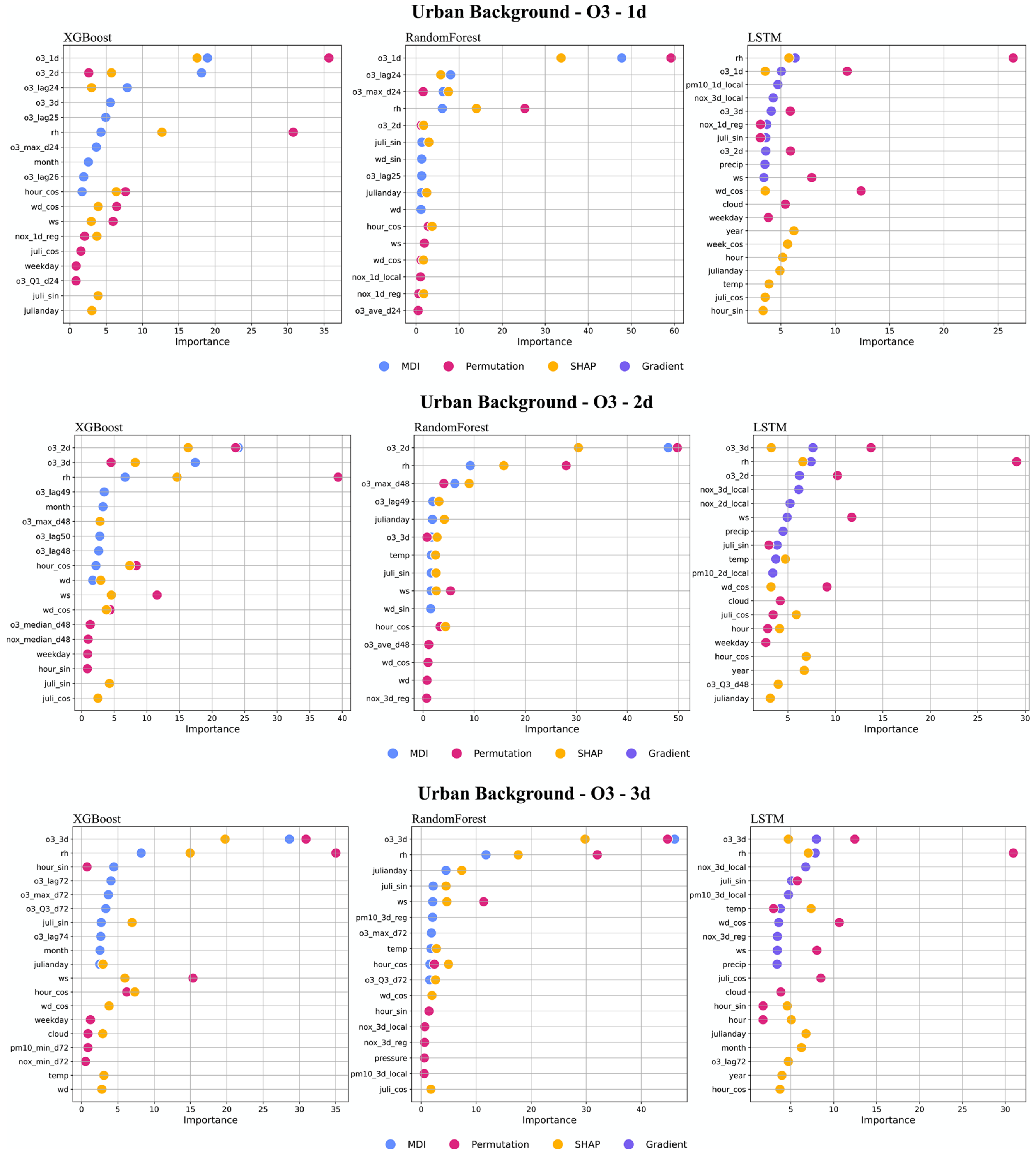


Figure F3. Top 10 most important features (%) for O_3 forecasts using XGB, RF, and LSTM at the urban site.

Appendix G: Temporal variations in hourly mean NO_x, PM_{10}, and O_3 concentrations at the street canyon sites

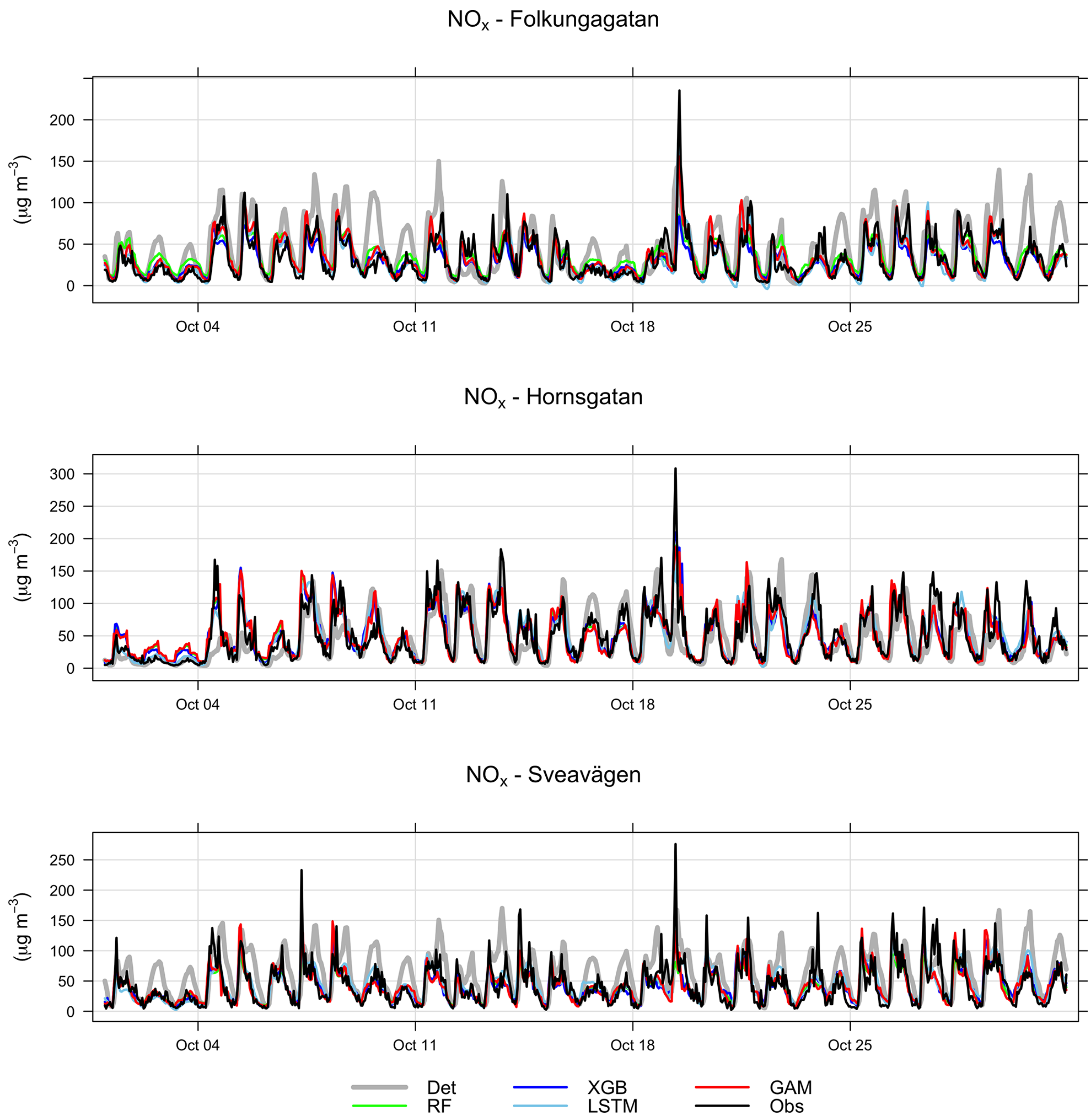


Figure G1. Temporal variations of hourly deterministic and ML-forecasted NO_x concentrations together with corresponding measured concentrations at street canyon sites for October 2021. Mean of 1, 2, and 3 d forecasts.

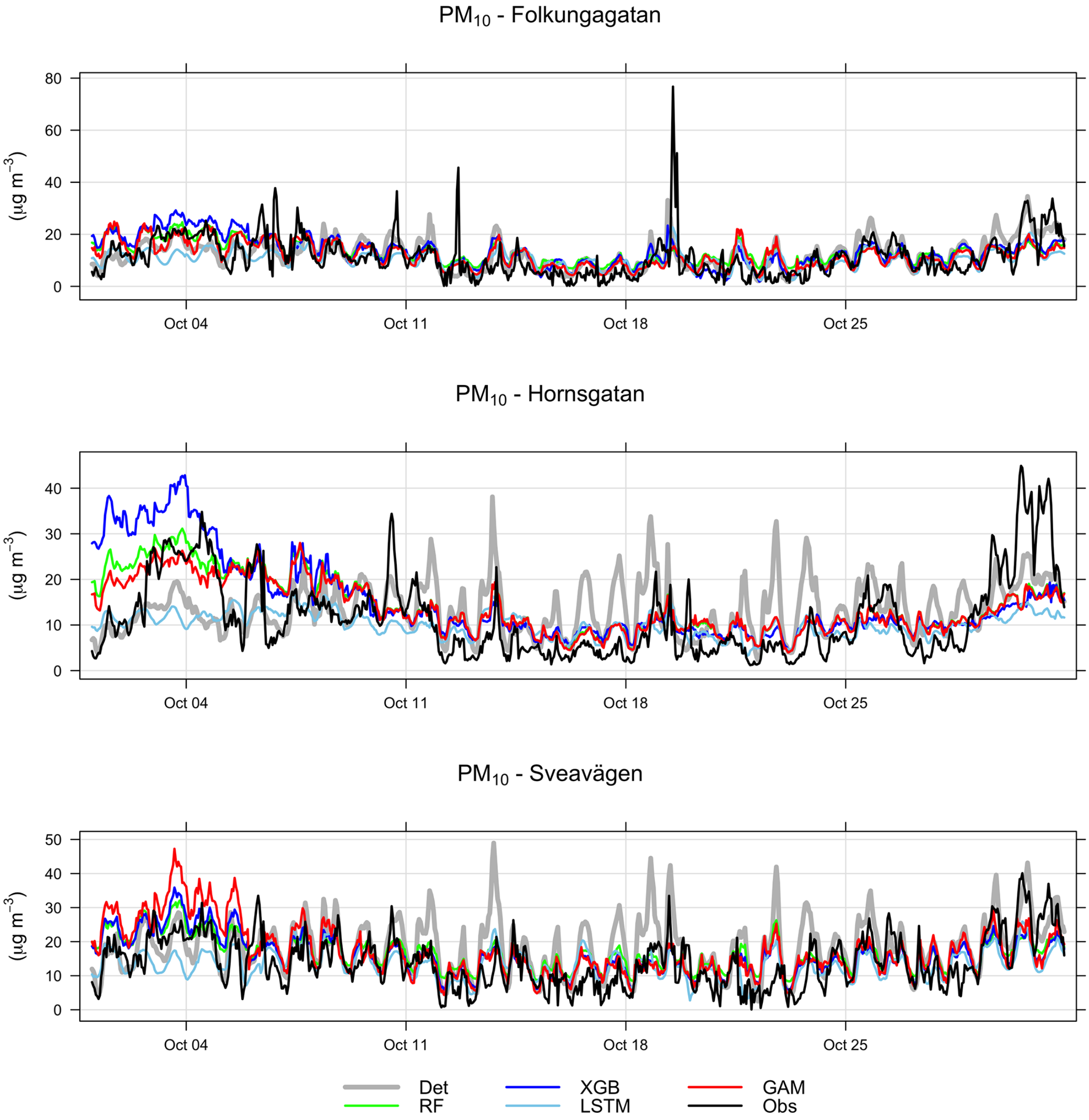


Figure G2. Temporal variations of hourly deterministic and ML-forecasted PM_{10} concentrations together with corresponding measured concentrations at the street canyon sites for October 2021. Mean of 1, 2, and 3 d forecasts.

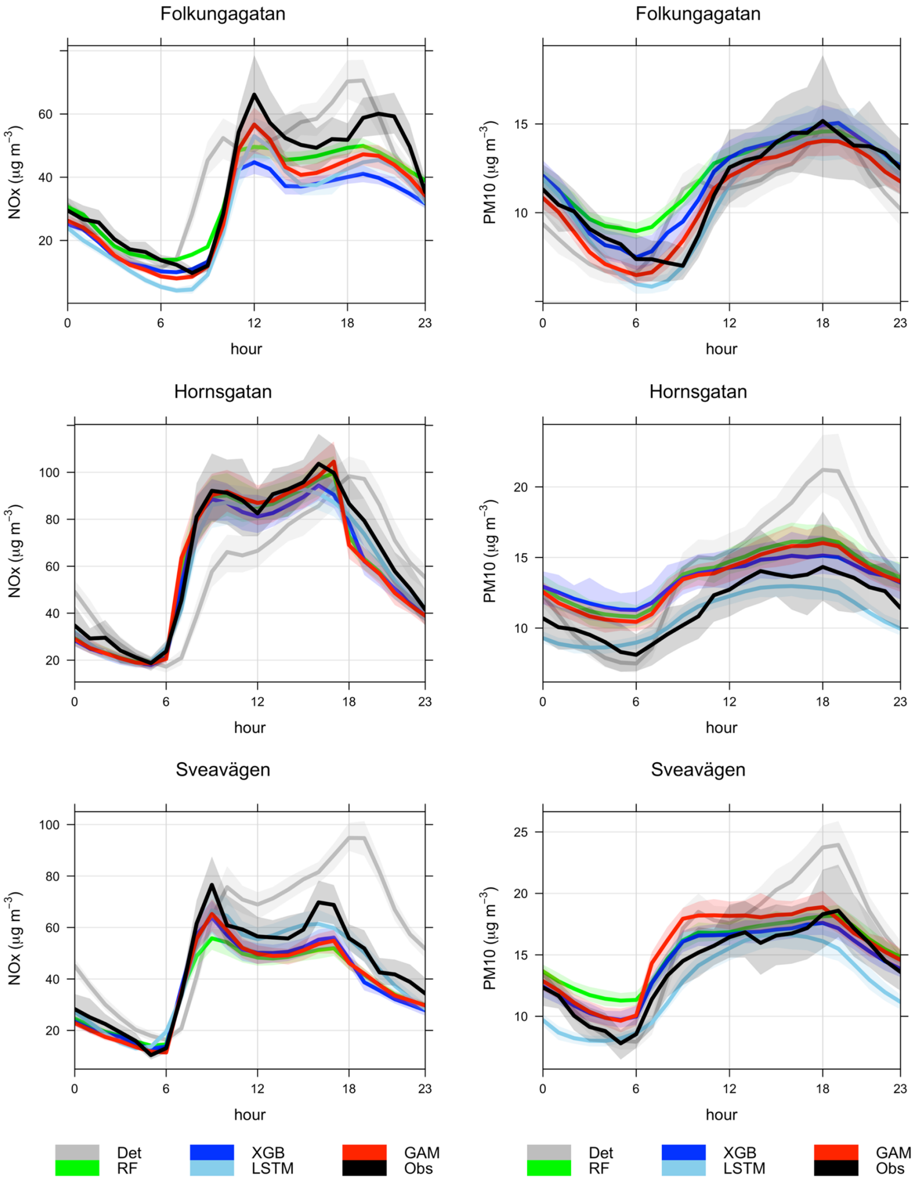


Figure G3. Mean diurnal variations in measured and forecasted concentrations of NO_x and PM_{10} at the street canyon sites. Mean of 1, 2, and 3 d forecasts for September–December 2021. Shaded areas are 95 % confidence intervals.

Appendix H: Statistical performance measures for forecasted hourly mean concentrations higher than the mean values at the street canyon sites

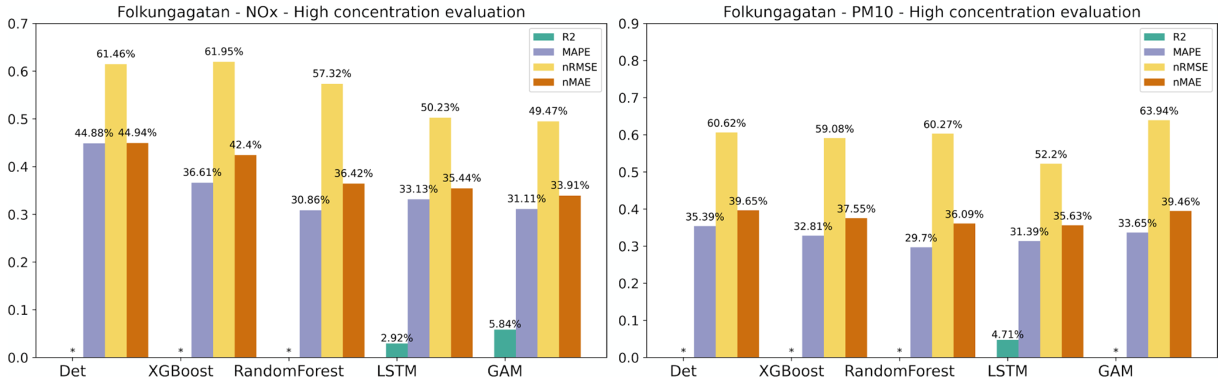


Figure H1. Statistical performance measures for forecasted NO_x and PM_{10} hourly mean concentrations higher than the mean values at Folkungagatan, where * represents a negative R^2 value. Mean of 1, 2, and 3 d forecasts.

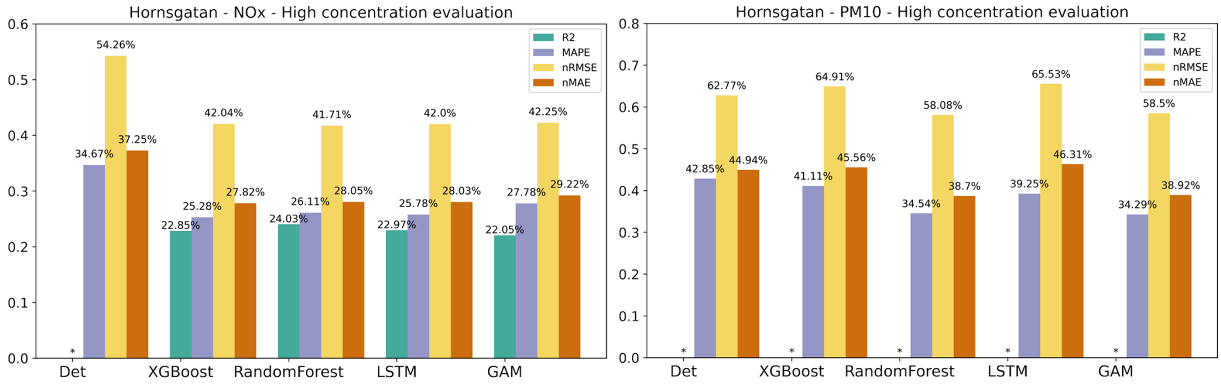


Figure H2. Statistical performance measures for forecasted NO_x and PM_{10} hourly mean concentrations higher than the mean values at Hornsgatan, where * represents a negative R^2 value. Mean of 1, 2, and 3 d forecasts.

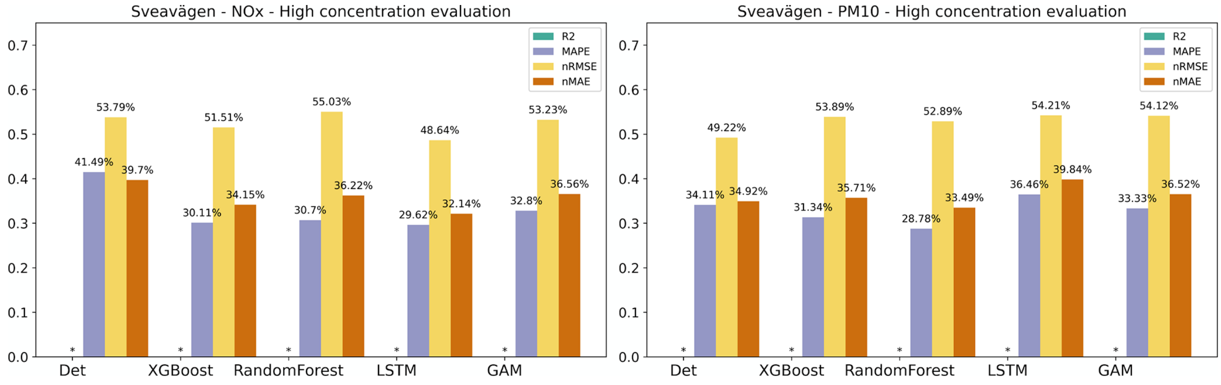


Figure H3. Statistical performance measures for forecasted NO_x and PM_{10} hourly mean concentrations higher than the mean values at Sveavägen, where * represents a negative R^2 value. Mean of 1, 2, and 3 d forecasts.

Appendix I: Importance of features – street canyon sites

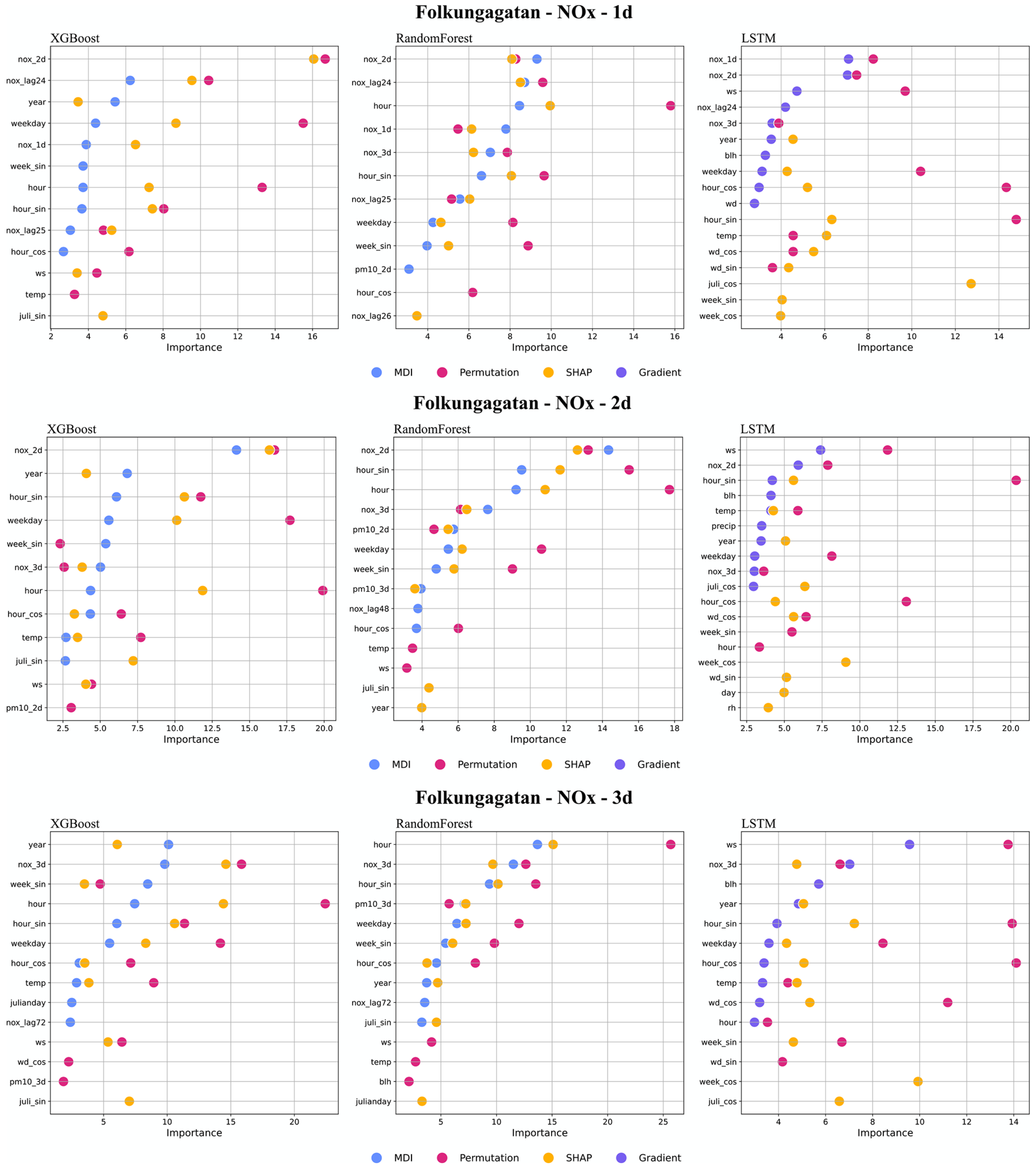


Figure I1. Top 10 most important features (%) for NO_x forecasts using RF, XGB, and LSTM at Folkungagatan.

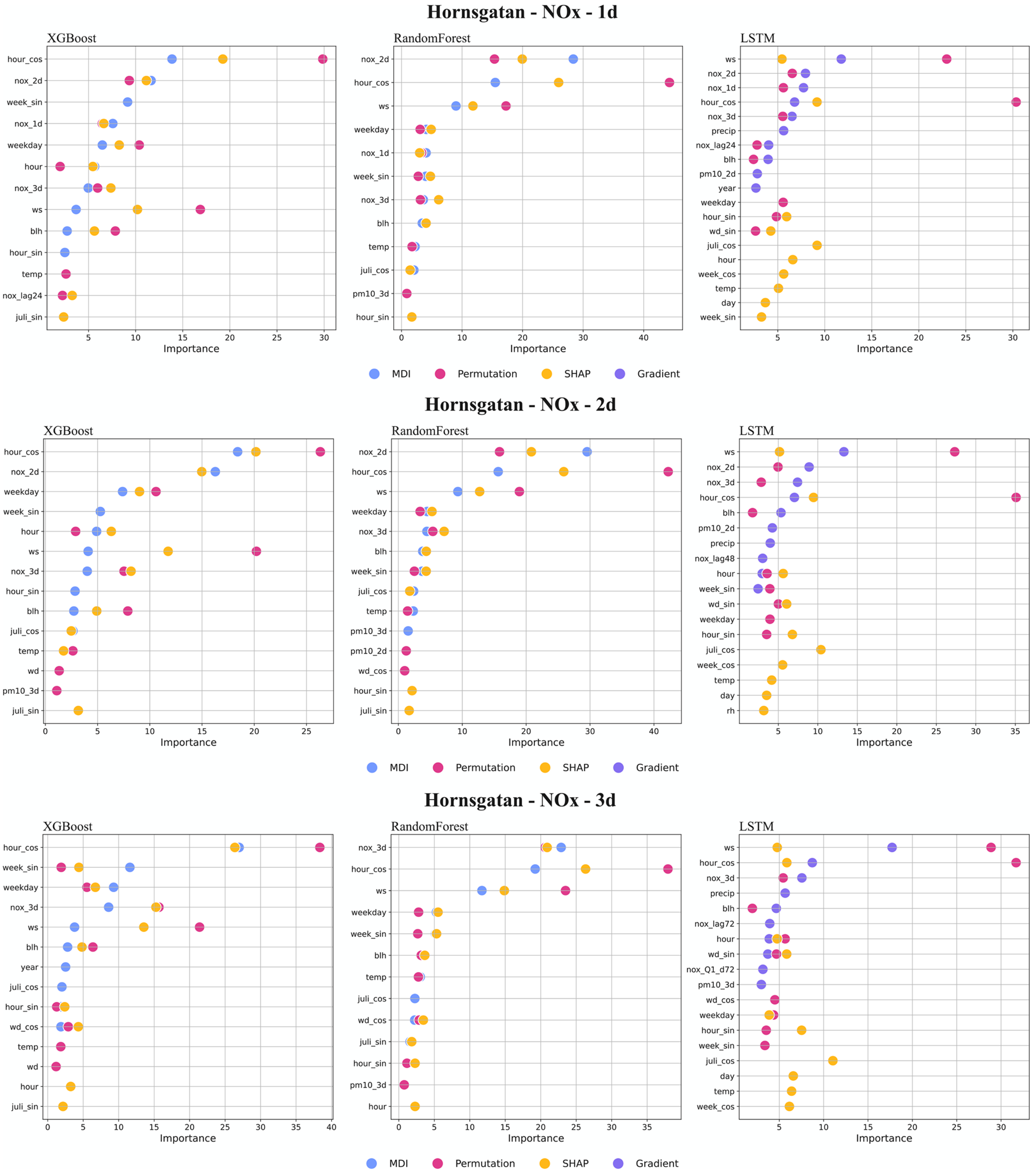


Figure I2. Top 10 most important features (%) for NO_x forecasts using RF, XGB, and LSTM at Hornsgatan.

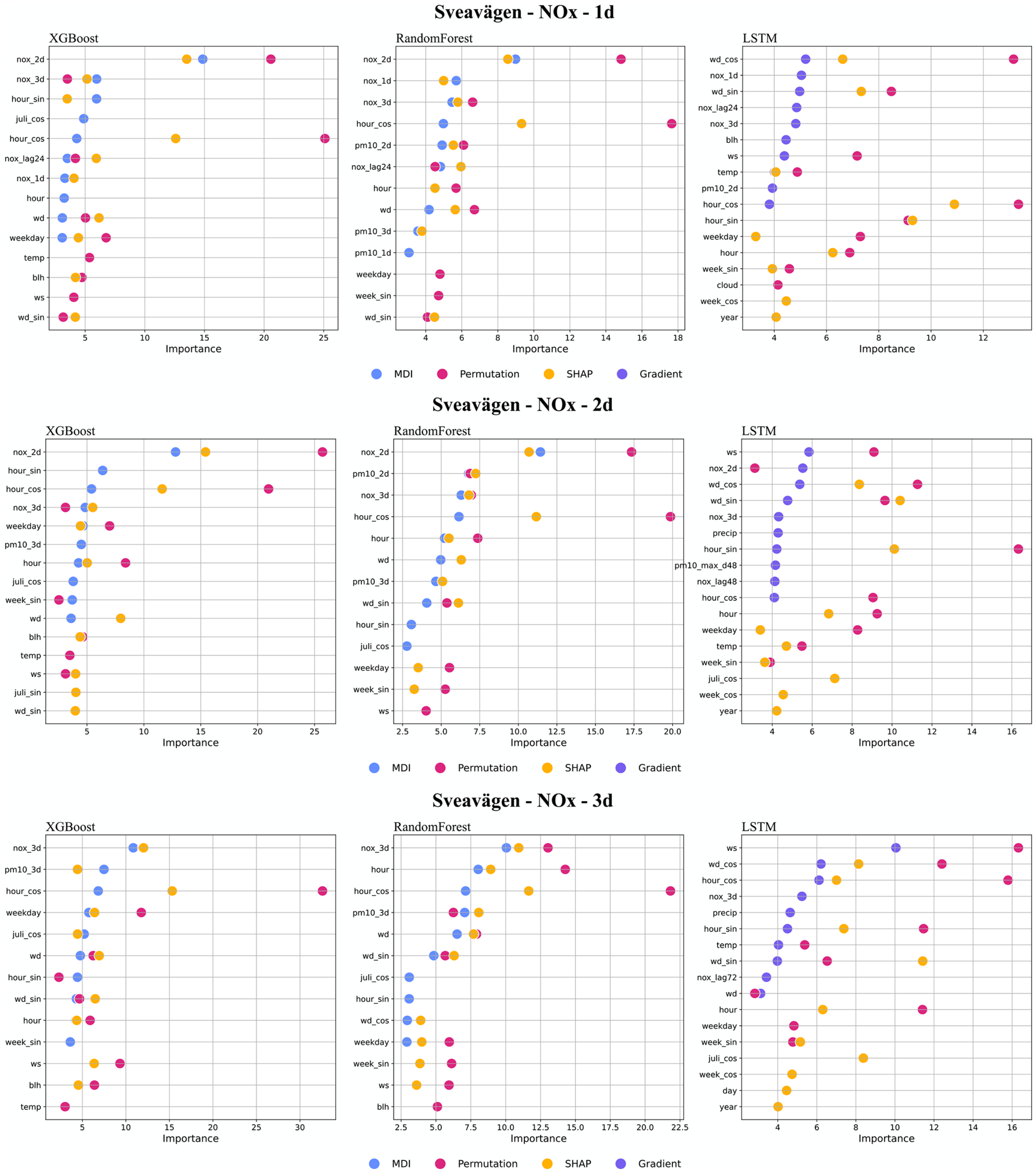


Figure I3. Top 10 most important features (%) for NO_x forecasts using RF, XGB, and LSTM at Sveavägen.

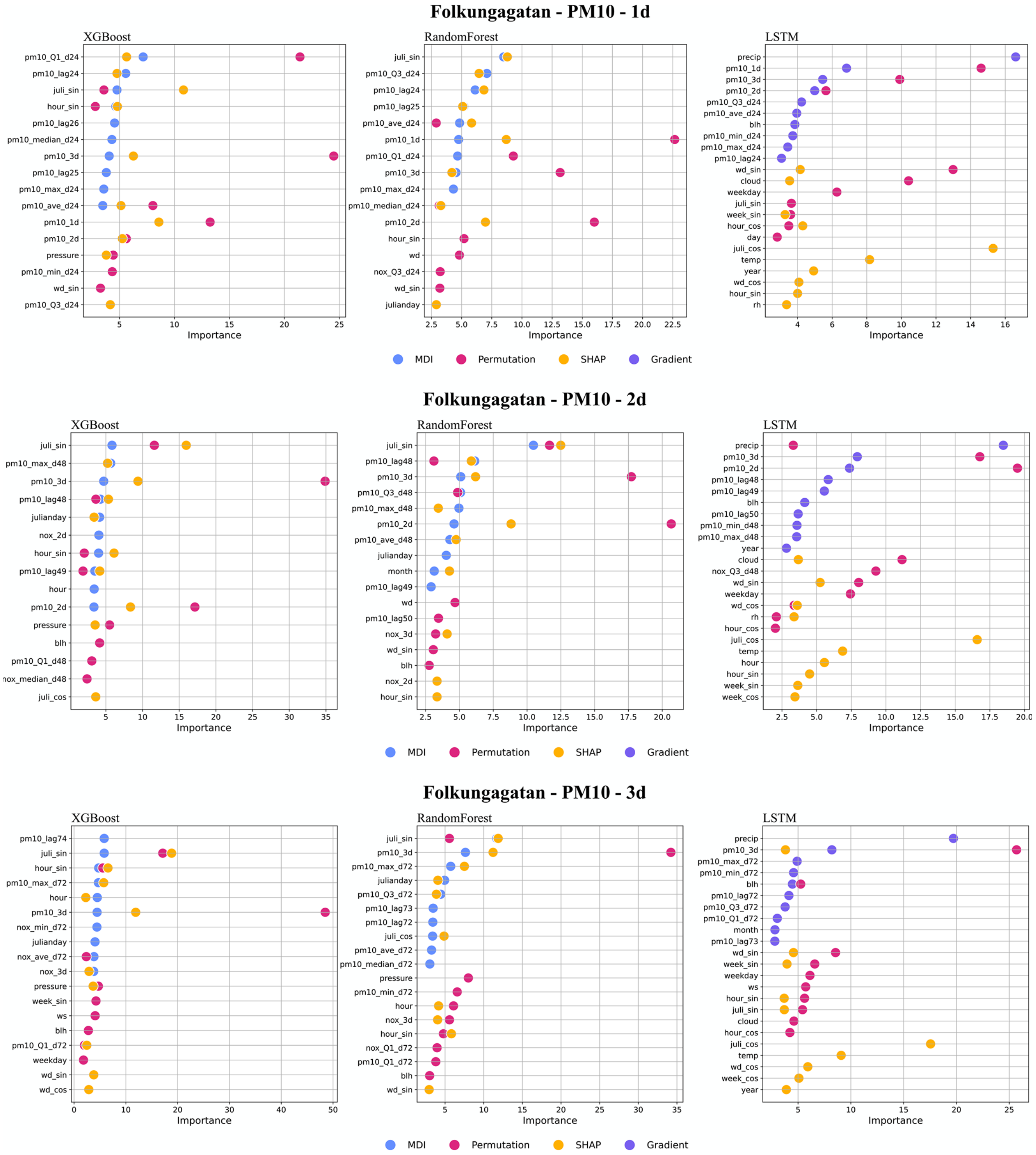


Figure I4. Top 10 most important features (%) for PM_{10} forecasts using RF, XGB, and LSTM at Folkungagatan.

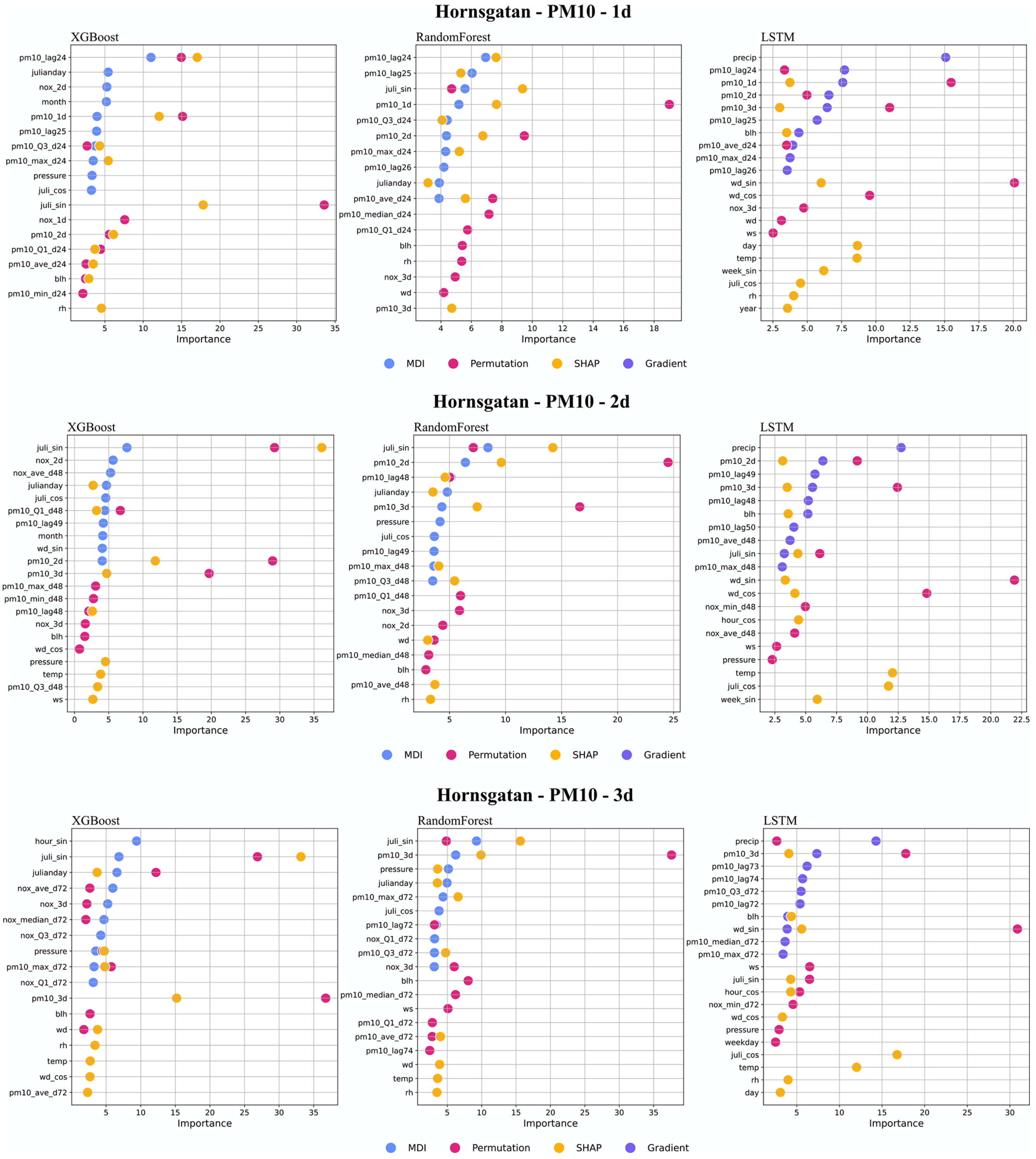


Figure I5. Top 10 most important features (%) for PM_{10} forecasts using RF, XGB, and LSTM at Hornsgatan.

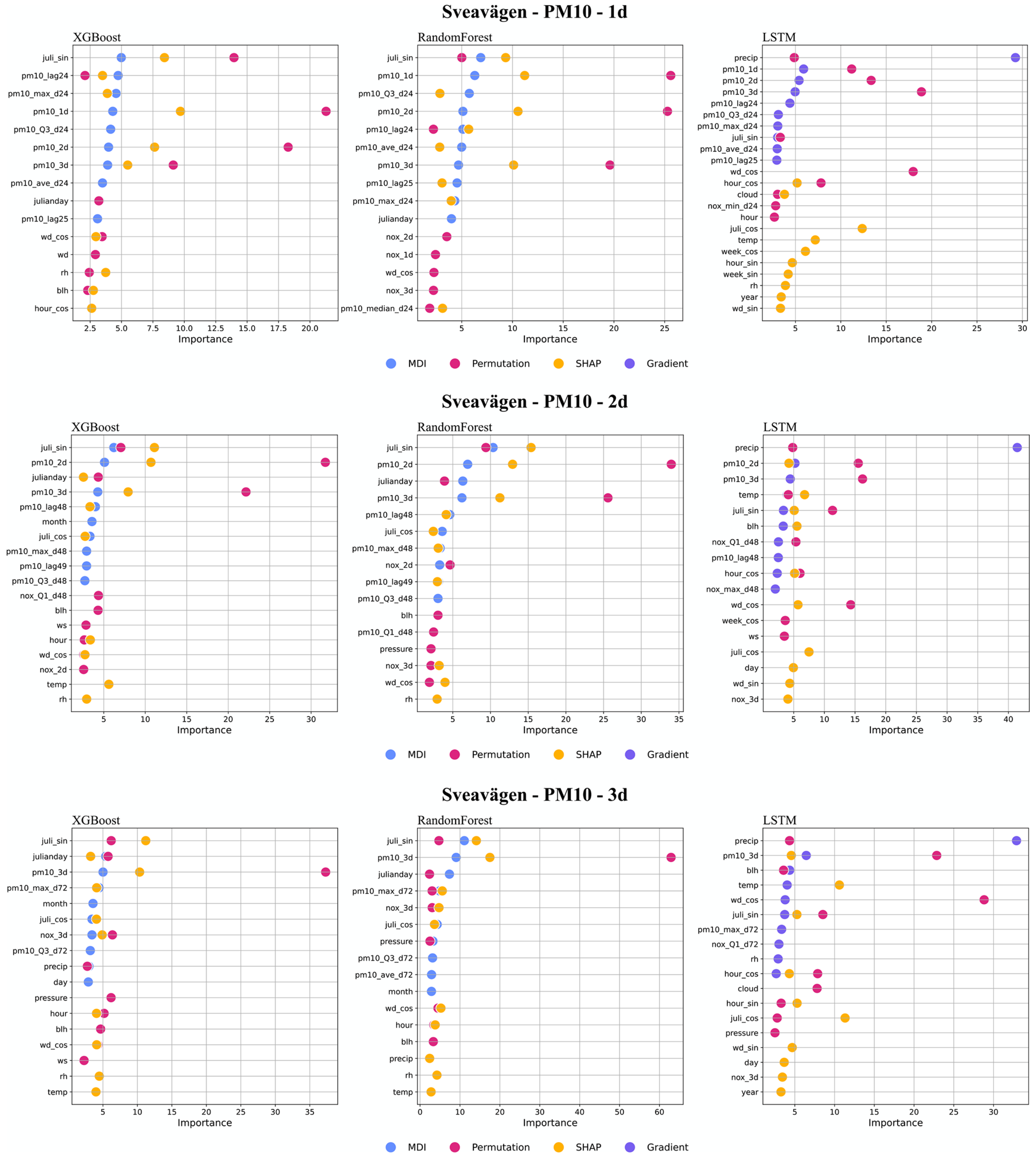


Figure I6. Top 10 most important features (%) for PM_{10} forecasts using RF, XGB, and LSTM at Sveavägen.

Code and data availability. Python code and all datasets in this paper are publicly accessible on Zenodo at https://doi.org/10.5281/zenodo.8433033 (Zhang and Ma, 2023).

Author contributions. CJ and MS provided initial input on the air quality prediction modelling approaches. ME was responsible for deterministic modelling and providing monitoring data and meteorological forecasts. ZZ and XM were responsible for the ML modelling and optimization, feature importance analysis, and statistical calculations. CJ, XM, and ME initiated and planned the project that supported this study. All authors contributed to analysing the data and writing the manuscript.

Competing interests. The contact author has declared that none of the authors has any competing interests.

Acknowledgements. The authors acknowledge the financial support of the ICT – The Next Generation and Digital Future at KTH Royal Institute of Technology.

Financial support. This research has been supported by the ICT – The Next Generation and KTH Royal Institute of Technology (KTH Kungliga Tekniska högskolan; grant no. VF 2021-0082).

The article processing charges for this open-access
publication were covered by Stockholm University.

Review statement. This paper was edited by Daniel Varon and reviewed by John Schreck and one anonymous referee.

References

Baehrens, D., Schroeter, T., Harmeling, S., Kawanabe, M., Hansen, K., and Muller, K.-R.: How to Explain Individual Classification Decisions, J. Mach. Learn. Res., 11, 1803–1831, 2010.

Berkowicz, R.: OSPM – A parameterised street pollution model, Environ. Monit. Assess., 65, 323–331, 2000.

Bisong, E. and Bisong, E.: Introduction to Scikit-learn. Building Machine Learning and Deep Learning Models on Google Cloud Platform: A Comprehensive Guide for Beginners, 215–229, https://doi.org/10.1007/978-1-4842-4470-8, 2019.

Breiman, L.: Random forests, Mach. Learn., 45, 5–32, 2001.

Brokamp, C., Jandarov, R., Rao, M. B., LeMasters, G., and Ryan, P.: Exposure assessment models for elemental components of particulate matter in an urban environment: A comparison of regression and random forest approaches, Atmos. Environ, 151, 1–11, 2017.

Burman, L. and Johansson, C.: Emissions and Concentrations of Nitrogen Oxides and Nitrogen Dioxide on Hornsgatan Street, Evaluation of Traffic Measurements during Autumn 2009, SLB Report 7, https://www.slb.nu/slb/rapporter/pdf8/slb2010_007.pdf (last access: 20 December 2023), 2010 (in Swedish).

Burman, L., Elmgren, M., and Norman, M.: Fordonsmätningar på Hornsgatan år 2017, https://www.slb.nu/slb/rapporter/pdf8/slb2019_002.pdf (last access: 20 December 2023), 2019 (in Swedish).

Cai, M., Yin, Y., and Xie, M.: Prediction of hourly air pollutant concentrations near urban arterials using artificial neural network approach, Transport Res. D-Tr. E. 14, 32–41, https://doi.org/10.1016/j.trd.2008.10.004, 2009.

Castelli, M., Clemente, F. M., Popovič, A., Silva, S., and Vanneschi, L.: A Machine Learning Approach to Predict Air Quality in California, Complexity, 2020, 8049504, 23 pp., https://doi.org/10.1155/2020/8049504, 2020.

Chuluunsaikhan, T., Heak, M., Nasridinov, A., and Choi, S.: Comparative Analysis of Predictive Models for Fine Particulate Matter in Daejeon, South Korea, Atmosphere, 12, 1295, https://doi.org/10.3390/atmos12101295, 2021.

Czernecki, B., Marosz, M., and Jędruszkiewicz, J.: Assessment of Machine Learning Algorithms in Short-term Forecasting of PM_{10} and $PM_{2.5}$ Concentrations in Selected Polish Agglomerations, Aerosol Air Qual. Res., 21, 200586, https://doi.org/10.4209/aaqr.200586, 2021.

Denby, B. R., Sundvor, I., Johansson, C., Pirjola, L., Ketzel, M., Norman, M., Kupiainen, K., Gustafsson, M., Blomqvist, G., and Omstedt, G.: A coupled road dust and surface moisture model to predict non-exhaust road traffic induced particle emissions (NORTRIP). Part 1: road dust loading and suspension modelling, Atmos. Environ., 77, 283–300, 2013a.

Denby, B. R., Sundvor, I., Johansson, C., Pirjola, L., Ketzel, M., Norman, M., Kupiainen, K., Gustafsson, M., Blomqvist, G., and Omstedt, G.: A coupled road dust and surface moisture model to predict non-exhaust road traffic induced particle emissions (NORTRIP). Part 2: surface moisture and salt impact modelling, Atmos. Environ., 81, 485–503, 2013b.

Di, Q., Amini, H., Shi, L., Kloog, I., Silvern, R., Kelly, J., Sabath, M. B., Choirat, C., Koutrakis, P., Lyapustin, A., Wang, Y., Mickley, L. J., and Schwartz, J.: An ensemble-based model of $PM_{2.5}$ concentration across the contiguous United States with high spatiotemporal resolution, Environ. Int., 130, 104909, https://doi.org/10.1016/j.envint.2019.104909, 2019.

Doreswamy, H. K. S., Yogesh, K. M., and Gad, I.: Forecasting Air Pollution Particulate Matter ($PM_{2.5}$) Using Machine Learning Regression Models, Procedia Comput. Sci., 171, 2057–2066, 2020.

Engardt, M., Bergström, S., and Johansson, C.: Luften du andas - nu och de kommande dagarna, Utveckling av ett automatiskt prognossystem för luftföroreningar och pollen, SLB 36:2021, 33 pp., https://www.slbanalys.se/slb/rapporter/pdf8/slb2021_036.pdf (last access: 20 December 2023), 2021 (in Swedish).

Fuller, R., Landrigan, P. J., Balakrishnan, K., Bathan, G., Bose-O'Reilly, S., Brauer, M., Caravanos, J., Chiles, T., Cohen, A., Corra, L., Cropper, M., Ferraro, G., Hanna, J., Hanrahan, D., Hu, H., Hunter, D., Janata, G., Kupka, R., Lanphear, B., Lichtveld, M., Martin, K., Mustapha, A., Sanchez-Triana, E., Sandilya, K., Schaefli, L., Shaw, J., Seddon, J., Suk, W., María Téllez-Rojo, M., and Yan, C.: Pollution and health: a progress update, The Lancet Planetary Health, 6, e535–e547, https://doi.org/10.1016/S2542-5196(22)00090-0, 2022.

Gidhagen, L., Johansson, C., Langner, J., and Foltescu, V. L.: Urban scale modeling of particle number concentration in Stockholm, Atmos. Environ., 39, 1711–1725, 2005.

Hagenbjörk, A., Malmqvist, E., Mattisson, K., Sommar, N. J., and Modig, L.: The spatial variation of O_3, NO, NO_2 and NO_x and the relation between them in two Swedish cities, Environ. Monit. Assess., 189, 1–12, 2017.

Hoek, G., Beelen, R., de Hoogh, K., Vienneau, D., Gulliver, J., Fischer, P., and Briggs, D.: A review of land-use regression models to assess spatial variation of outdoor air pollution, Atmos. Environ., 42, 7561–7568, https://doi.org/10.1016/j.atmosenv.2008.05.057, 2008.

Hong, H., Choi, I., Jeon, H., Kim, Y., Lee, J.-B., Park, C. H., and Kim, H. S.: An Air Pollutants Prediction Method Integrating Numerical Models and Artificial Intelligence Models Targeting the Area around Busan Port in Korea, Atmosphere 13, 1462, https://doi.org/10.3390/atmos13091462, 2022.

Horálek, J., Hamer, P., Schreiberová, M., Colette, A., Schneider, P., and Malherbe, L.: Potential use of CAMS modelling results in air quality mapping under ETC/ATNI, Eionet Report – ETC/ATNI 2019/17, European environment Agency, European Topic Centre on Air Pollution, transport, noise and industrial pollution, ISBN 978-82-93752-21-9, 2019.

Iskandaryan, D., Ramos, F., and Trilles, S.: Air Quality Prediction in Smart Cities Using Machine Learning Technologies based on Sensor Data: A Review, Appl. Sci., 10, 2401, https://doi.org/10.3390/app10072401, 2020.

Janssen, S. and Thunis, P.: FAIRMODE Guidance Document on Modelling Quality Objectives and Benchmarking (version 3.3), EUR 31068 EN, Publications Office of the European Union, Luxembourg, JRC129254, ISBN 978-92-76-52425-0, https://doi.org/10.2760/41988, 2022.

Johansson, C., Norman, M., and Gidhagen, L.: Spatial & temporal variations of PM_{10} and particle number concentrations in urban air, Environ. Monit. Assess., 127, 477–487, 2007.

Johansson, C., Burman, L., and Forsberg, B.: The effects of congestions tax on air quality and health, Atmos. Environ., 43, 4843–4854, 2009.

Johansson, C., Eneroth, K., Lövenheim, B., Silvergren, S., Burman, L., Bergström, S., Norman, M., Engström Nylén, A., Hurkmans, J., Elmgren, M., Brydolf, M., and Täppefur, M.: Luftkvalitetsberäkningar för kontroll av miljökvalitetsnormer, SLB 11:2017 ver2, https://www.slbanalys.se/slb/rapporter/pdf8/slb2017_011.pdf (last access: 20 December 2023), 2017a (in Swedish with English summary).

Johansson, C., Lövenheim, B., Schantz, P., Wahlgren, L., Almström, P., Markstedt, A., Strömgren, M., Forsberg, B., and Nilsson Sommar, J.: Impacts on air pollution and health by changing commuting from car to bicycle, Sci. Total Environ., 584–585, 55–63, 2017b.

Joharestani, M. Z., Cao, C., Ni, X., Bashir, B., and Talebiesfandarani, S.: $PM_{2.5}$ Prediction Based on Random Forest, XGBoost, and Deep Learning Using Multisource Remote Sensing Data, Atmosphere, 10, 373, https://doi.org/10.3390/atmos10070373, 2019.

Kamińska, J. A.: A random forest partition model for predicting NO_2 concentrations from traffic flow and meteorological conditions, Sci. Total Environ., 651, 475–483, 2019.

Keller, M., Hausberger, S., Matzer, C., Wüthrich, P., and Notter, B.: HBEFA 3.3, Update of NOx Emission Factors of Diesel Passenger Cars-Background Documentation, https://www.umweltbundesamt.de/sites/default/files/medien/2546/dokumente/hbefa33_documentation_20170425.pdf (last access: 20 December 2023), 2017.

Kleinert, F., Leufen, L. H., Lupascu, A., Butler, T., and Schultz, M. G.: Representing chemical history in ozone time-series predictions – a model experiment study building on the MLAir (v1.5) deep learning framework, Geosci. Model Dev., 15, 8913–8930, https://doi.org/10.5194/gmd-15-8913-2022, 2022.

Krecl, P., Harrison, R. M., Johansson, C., Targino, A. C., Beddows, D. C., Ellermann, T., Lara, C., and Ketzel, M.: Long-term trends in nitrogen oxides concentrations and on-road vehicle emission factors in Copenhagen, London and Stockholm, Environ. Pollut., 290, 118105, https://doi.org/10.1016/j.envpol.2021.118105, 2021.

Lee, Y.-G., Lee, P.-H., Choi, S.-M., An, M.-H., and Jang, A.-S.: Effects of Air Pollutants on Airway Diseases, Int. J. Env. Res. Pub. He., 18, 9905, https://doi.org/10.3390/ijerph18189905, 2021.

Liashchynskyi, P. and Liashchynskyi, P.: Grid search, random search, genetic algorithm: a big comparison for NAS, arXiv [preprint], https://doi.org/10.48550/arXiv.1912.06059, 12 December 2019.

Lundberg, S. M. and Lee, S.-I.: A unified approach to interpreting model predictions, Advances in neural information processing systems, arXiv [preprint], https://doi.org/10.48550/arXiv.1705.07874, 25 November 2017.

Ma, X., Lei, W., Andréasson, I., and Chen, H.: An evaluation of microscopic emission models for traffic pollution simulation using on-board measurement, Environ. Model. Assess., 17, 375–387, 2012.

Ma, X., Huang, Z., and Koutsopoulos, H.: Integrated traffic and emission simulation: a model calibration approach using aggregate information, Environ. Model. Assess., 19, 271–282, 2014.

Marécal, V., Peuch, V.-H., Andersson, C., Andersson, S., Arteta, J., Beekmann, M., Benedictow, A., Bergström, R., Bessagnet, B., Cansado, A., Chéroux, F., Colette, A., Coman, A., Curier, R. L., Denier van der Gon, H. A. C., Drouin, A., Elbern, H., Emili, E., Engelen, R. J., Eskes, H. J., Foret, G., Friese, E., Gauss, M., Giannaros, C., Guth, J., Joly, M., Jaumouillé, E., Josse, B., Kadygrov, N., Kaiser, J. W., Krajsek, K., Kuenen, J., Kumar, U., Liora, N., Lopez, E., Malherbe, L., Martinez, I., Melas, D., Meleux, F., Menut, L., Moinat, P., Morales, T., Parmentier, J., Piacentini, A., Plu, M., Poupkou, A., Queguiner, S., Robertson, L., Rouïl, L., Schaap, M., Segers, A., Sofiev, M., Tarasson, L., Thomas, M., Timmermans, R., Valdebenito, Á., van Velthoven, P., van Versendaal, R., Vira, J., and Ung, A.: A regional air quality forecasting system over Europe: the MACC-II daily ensemble production, Geosci. Model Dev., 8, 2777–2813, https://doi.org/10.5194/gmd-8-2777-2015, 2015.

Meteo-France: Regional Production, Description of the operational models and of the ENSEMBLE system, Copernicus Atmosphere Monitoring Service, https://atmosphere.copernicus.eu/sites/default/files/2018-02/CAMS50_factsheet_201610_v2.pdf (last access: 20 December 2023), 2017.

Munir, S., Mayfield, M., Coca, D., Mihaylova, L. S., and Osammor, O.: Analysis of Air Pollution in Urban Areas with Airviro Dispersion Model – A Case Study in the City of Sheffield, United Kingdom, Atmosphere, 11, 285, https://doi.org/10.3390/atmos11030285, 2020.

Orru, H., Lövenheim, B., Johansson, C., and Forsberg, B.: Estimated health impacts of changes in air pollution exposure associated with the planned by-pass Förbifart Stockholm, J. Expo. Sci. Env. Epid., 25, 524–531, 2015.

Ottosen, T.-B., Kakosimos, K. E., Johansson, C., Hertel, O., Brandt, J., Skov, H., Berkowicz, R., Ellermann, T., Jensen, S. S., and Ketzel, M.: Analysis of the impact of inhomogeneous emissions in the Operational Street Pollution Model (OSPM), Geosci. Model Dev., 8, 3231–3245, https://doi.org/10.5194/gmd-8-3231-2015, 2015.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., and Desmaison, A.: Pytorch: An imperative style, high-performance deep learning library, arXiv [preprint], https://doi.org/10.48550/arXiv.1912.01703, 3 December 2019.

Qadeer, K., Rehman, W. U., Sheri, A. M., Park, I., Kim, H. K., and Jeon, M.: A Long Short-Term Memory (LSTM) Network for Hourly Estimation of $PM_{2.5}$ Concentration in Two Cities of South Korea, Appl. Sci., 10, 3984, https://doi.org/10.3390/app10113984, 2020.

Rybarczyk, Y. and Zalakeviciute, R.: Machine Learning Approaches for Outdoor Air Quality Modelling: A Systematic Review, Appl. Sci., 8, 2570, https://doi.org/10.3390/app8122570, 2018.

Shaban, K. B., Kadri, A., and Rezk, E.: Urban Air Pollution Monitoring System With Forecasting Models, IEEE Sens. J., 16, 2598–2606, 2016.

Shrikumar, A., Greenside, P., and Kundaje, A.: Learning important features through propagating activation differences, in: Proceedings of the 34th International conference on machine learning, Sydney, Australia, 6–11 August 2017, 3145–3153, 2017.

Shtein, A., Kloog, I., Schwartz, J., Silibello, C., Michelozzi, P., Gariazzo, C., Viegi, G., Forastiere, F., Karnieli, A., Just, A. C., and Stafoggia, M.: Estimating Daily $PM_{2.5}$ and PM_{10} over Italy Using an Ensemble Model, Environ. Sci. Technol., 54, 120–128 https://doi.org/10.1021/acs.est.9b04279, 2020.

SLB: Methods for calculating air pollution concentrations in relation to the limit values. Report, Environment and Health Administration of Stockholm, SLB analys, Stockholm, Sweden, Report no. 50:2021, https://www.slbanalys.se/slb/rapporter/pdf8/slb2021_050.pdf (last access: 30 November 2022), 2022 (in Swedish with English summary).

Sokhi, R. S., Singh, V., Querol, X., Finardi, S., Targino, A. C., de Fatima Andrade, M., Pavlovic, R., Garland, R. M., Massagué, J., Kong, S., and Baklanov, A.: A global observational analysis to understand changes in air quality during exceptionally low anthropogenic emission conditions, Environ. Int., 157, 106818, https://doi.org/10.1016/j.envint.2021.106818, 2021.

Stafoggia, M., Bellander, T., Bucci, S., Davoli, M., de Hoogh, K., de Donato, F., Gariazzo, C., Lyapustin, A., Michelozzi, P., Renzi, M., Scortichini, M., Shtein, A., Viegi, G., Kloog, I., and Schwartz, J.: Estimation of daily PM_{10} and $PM_{2.5}$ concentrations in Italy, 2013–2015, using a spatiotemporal land-use random-forest model, Environ. Int., 124, 170–179, 2019.

Stafoggia, M., Johansson, C., Glantz, P., Renzi, M., Shtein, A., de Hoogh, K., Kloog, I., Davoli, M., Michelozzi, P., and Bellander, T.: A Random Forest Approach to Estimate Daily Particulate Matter, Nitrogen Dioxide, and Ozone at Fine Spatial Resolution in Sweden, Atmosphere, 11, 239, https://doi.org/10.3390/atmos11030239, 2020.

Thongthammachart, T., Araki, S., Shimadera, H., Eto, S., Matsuo, T.. and Kondo, A.: An integrated model combining random forests and WRF/CMAQ model for high accuracy spatiotemporal $PM_{2.5}$ predictions in the Kansai region of Japan, Atmos. Environ., 262, 118620, https://doi.org/10.1016/j.atmosenv.2021.118620, 2021.

Torkmahalleh, M. A., Akhmetvaliyeva, Z., Darvishi Omran, A., Darvish Omran, F., Kazemitabar, M., Naseri, M., Naseri, M., Sharifi, H., Malekipirbazari, M., Kwasi Adotey, E., and Gorjinezhad, S.: Global air quality and COVID-19 pandemic: do we breathe cleaner air?, Aerosol Air Qual. Res., 21, 200567. https://doi.org/10.4209/aaqr.200567, 2021.

Willmott, C. J. and Matsuura, K.: Smart interpolation of annually averaged air temperature in the United States, J. Appl. Meteorol. Clim., 34, 2577–2586, 1995.

Zaini, N., Ean, L.W., Ahmed, A.N., Malek, M.A.: A systematic literature review of deep learning neural network for time series air quality forecasting. Environmental Science and Pollution Research, https://doi.org/10.1007/s11356-021-17442-1, 2021.

Zhang, Z. and Ma, X.: ACP-2023-38 paper submission support: code and data for 3-days prediction of Air Quality using Machine Learning algorithms, Version 4, Zenodo [data set/code], https://doi.org/10.5281/zenodo.8433033, 2023.