Aerosol and Air Quality Research

ORIGINAL RESEARCH
https://doi.org/10.4209/aaqr.230317

Ensemble Machine Learning, Deep Learning, and Time Series Forecasting: Improving Prediction Accuracy for Hourly Concentrations of Ambient Air Pollutants

Valentino Petrić[1], Hussain Hussain[2], Kristina Časni[2], Milana Vuckovic[3], Andreas Schopper[4], Željka Ujević Andrijić[5], Simonas Kecorius[6], Leizel Madueno[7], Roman Kern[2], Mario Lovrić[8,9,10*]

[1] Ascalia d.o.o., Trate 16, HR-40000, Čakovec, Croatia
[2] Know-Center, Sandgasse 36/4, AT-8010 Graz, Austria
[3] European Center for Medium Range Weather Forecasts, Shinfield Park, Reading, UK-RG2 9AX, United Kingdom
[4] Amt der Steiermärkischen Landesregierung, Referat Luftreinhaltung, Landhausgasse 7, 8010 Graz, Austria
[5] University of Zagreb Faculty of Chemical Engineering and Technology, Department of Measurements and Process Control, Savska c. 16, HR-10000 Zagreb, Croatia
[6] Institute of Epidemiology, Helmholtz Zentrum München, Ingolstädter Landstr. 1, DE-85764, Neuherberg, Germany
[7] Leibniz Institute for Tropospheric Research, Leipzig, 04318, Germany
[8] Centre for bioanthropology, Institute for Anthropological Research, Gajeva 32, HR-10000 Zagreb, Croatia
[9] Lisbon Council, 155 rue de la loi, 1040 Brussels, Belgium
[10] Copenhagen Prospective Studies on Asthma in Childhood, Herlev and Gentofte Hospital, University of Copenhagen, Copenhagen, Denmark

ABSTRACT

This study aims to improve the generalisation capabilities of machine learning models for modelling hourly air pollutant concentrations in scenarios where access to high-quality data is limited. A diverse set of techniques was implemented to tackle this challenge, encompassing the utilisation of the prophet, random forest, and three different deep learning architectures: long short-term memory networks, convolutional neural networks, and multilayer perceptrons. A hybrid model of random forest and prophet was also tested. The role of the hybrid model was to combine the forecasting strengths of the Prophet model with the predictive power of the Random Forest model to better capture complex temporal patterns in the data. After testing, the hybrid model demonstrated improved generalization capabilities, achieving statistically significant improvements in R^2 for hourly concentrations of NO (improving by 26%), NO_2 (enhancing by 18%), PM_{10} (with changes ranging from an 8% decline to a 35% improvement), and O_3 (showcasing R^2 coefficients ranging from 0.83 to 0.87) at five sites in Graz, Austria. The utilisation of surface atmospheric ERA5-Land datasets within the models as model features showed high feature post hoc importance in the best (hybrid) models per pollutant and site. Furthermore, error analysis was performed to understand better the conditions under which these models might fail. The results showed that despite the expectations for models to fail with an increasing timeframe (the test set) from March 2019 to March 2020, the models were sufficiently stable for long-term prediction and thus can be used to forecast and predict air pollution.

Keywords: Prophet, Ozone, Air pollution, LSTM, CNN, Random forests

OPEN ACCESS

Received: December 17, 2023
Revised: July 2, 2024
Accepted: September 24, 2024

* **Corresponding Author:**
mario.lovric@inantro.hr

Publisher:
Taiwan Association for Aerosol Research
ISSN: 1680-8584 print
ISSN: 2071-1409 online

1 INTRODUCTION

Particulate matter (PM), gaseous pollutants, nitrogen oxides (NO_x), and ozone (O_3) have impacts on human health. Once inhaled, PM of submicron diameter can penetrate from the lung alveoli into the bloodstream (Anderson *et al.*, 2012; Kim *et al.*, 2015). Therefore, the smaller the particle size, the greater the adverse health effect (Jakovljević *et al.*, 2020; Li *et al.*, 2019; Yang *et al.*, 2020). In addition, NO_x compounds can enter the human body through the airways. People with pre-existing lung diseases are especially vulnerable to NO_x and it can cause chronic respiratory diseases, lung cancer, and cardiovascular diseases, such as heart failure, myocardial ischaemia, infarction, stroke, and arrhythmia (Almetwally *et al.*, 2020). Lung cancer and respiratory diseases are also related to O_3, which is absorbed by the upper respiratory tract and conducted into the intrathoracic airways. Women and children are especially vulnerable to the effects of O_3 because they inhale higher concentrations owing to the smaller size of their airways (Nuvolone *et al.*, 2018).

The application of machine learning (ML) in studies of air pollution, specifically in predicting future values or events, is gaining momentum. This was even more emphasized during the COVID-19 lockdown air quality investigations (Grange *et al.*, 2021; Lovrić *et al.*, 2022, 2021). There are several reasons why ML models are appropriate for air pollution studies: 1) ML models can capture the complex and often nonlinear relationships among various factors influencing air pollution, including meteorological conditions, emissions sources, geographical features, and chemical reactions in the atmosphere; 2) ML models can yield more accurate predictions than traditional statistical methods, especially with large and high-dimensional datasets; 3) ML models can provide real-time or near-real-time predictions once they have been trained; 4) ML models can be automated and scaled, enabling broad geographical coverage and continuous updates as new data become available; and 5) ML models can provide insights into the importance of different features (variables). Therefore, ML models can be useful for identifying the key sources of pollution and targeting interventions. Several factors contribute to the increased application of ML, including increased data availability, advancements in remote sensing, and IoT (internet of things) technology. IoT plays a crucial role here, enabling real-time monitoring of air pollution by connecting sensors and devices to exchange data over the internet. Heightened health concerns further drive the adoption of ML for more effective environmental monitoring and management. Understanding the dynamics of long-term air pollutant concentrations with high spatial coverage is important for risk assessment and environmental policies (Camatini *et al.*, 2017; Segersson *et al.*, 2017; Tobías *et al.*, 2018). ML is being increasingly utilised to analyse air pollutant concentrations and forecast their levels, owing to its numerous benefits over statistical analysis. Such algorithms can process vast amounts of complex and heterogeneous data, including atmospheric and meteorological variables, to identify intricate patterns and relationships that may influence air pollution (Grange and Carslaw, 2019; Šimić *et al.*, 2020). By leveraging ML, researchers can develop highly accurate predictive models that consider various factors, such as emission sources, weather conditions, and geographical features, resulting in more precise and reliable air quality forecasts (Li *et al.*, 2023). In addition, ML techniques can adapt to and learn from new data, allowing models to continuously improve their accuracy over time. ML studies for air pollution prediction can be categorised depending on the pollutants or the ML task (Li *et al.*, 2023). In terms of the pollutants, the primary focus is on PM and gaseous pollutants (Šimić *et al.*, 2020). In a forecasting setting, future predictions of pollutants are estimated using historic measurements and often environmental properties; that is, the task can be seen as an extrapolation of a time series. In the prediction setup, the goal is to predict a pollutant based on measurements from other sources or locations without considering the future development of the target pollutant. This task is equivalent to estimating pollutants for which measurements do not exist or are impossible. Furthermore, this task provides insights into the main influencing factors of specific pollutants.

Methods for forecasting and prediction include traditional ML-based approaches, such as random forest (RF) (Breiman, 2001; Huang *et al.*, 2020) and statistical approaches, such as autoregressive and deep learning (DL) methods (Jiang *et al.*, 2021a; Jiang *et al.*, 2021b). Most approaches mentioned in the literature conducted forecasts on an hourly basis, with a 12-h forecast horizon, such as

Zhou *et al.* (2021), who conducted forecasts to cover quarterly time spans in the Yangtze River Delta, making use of a seasonal nonlinear grey Bernoulli model. Their study highlighted the low rate of change in air pollutants in certain areas. Most existing prediction studies opt for either daily or hourly averages of the target pollutant, with some noteworthy exceptions. For example, Araki *et al.* (2021) predicted various air quality indicators monthly for Japan-based features such as meteorological, land use, and road-related features, achieving a coefficient of determination (R^2) between 0.74 and 0.79. Tree-based approaches are popular, including RF (Bozdağ *et al.*, 2020; Chen *et al.*, 2021; Huang *et al.*, 2021; Zamani Joharestani *et al.*, 2019; Lovrić *et al.*, 2022, 2021; Sun *et al.*, 2021) and boosted trees (Gagliardi and Andenna, 2020; Liu, 2021), as well as the combination of tree-based approaches with other learning algorithms (Qi *et al.*, 2020; Zhang *et al.*, 2020). A wide variety of deep-learning-based approaches have also been explored, including multilayer perceptron (Photphanloet and Lipikorn, 2020; Sayahi *et al.*, 2020), convolutional neural networks (CNNs; Park *et al.*, 2020), recurrent approaches (Huang *et al.*, 2021; Wang *et al.*, 2021), and attention-based approaches (Chen *et al.*, 2021). For evaluation, many authors prefer variants of cross-validation, such as those based on close measurement stations (Wang *et al.*, 2021). Meteorological indicators such as temperature, humidity, wind speed and direction, land-use features, and traffic data are commonly included in the models. The top of the atmosphere and aerosol optical depth have also been explored (Park *et al.*, 2020; Sun *et al.*, 2021; Yan *et al.*, 2021; Zhang *et al.*, 2020) and it has been noted that thorough preprocessing of data is needed.

These algorithms often suffer from low data coverage and deterioration over longer periods owing to possible data drift. In recent years, there has been a novel trend of combining ML models with first-principle or physics-based models, known as hybrid or physics-inspired models. This trend has achieved improved model accuracy in many settings (Hoffer *et al.*, 2022; Lovrić *et al.*, 2020). Predicting deterioration over time is a critical consideration in time-series analysis and ML (Vela *et al.*, 2022). Time-series data often exhibit patterns that change over time, making it challenging to maintain accurate predictions over long forecasting periods. If a time series model does not account for these changes, its accuracy may decrease as the forecasting period progresses. Therefore, it is crucial to monitor and assess model predictions. By tracking forecast errors over time, analysts can identify when a model's predictions start to degrade, allowing them to take corrective actions before the model's accuracy deteriorates significantly. Therefore, prediction deterioration versus time is an essential consideration in time-series modelling to ensure accurate and reliable predictions over extended forecasting horizons.

The overall aim of this study was to improve strategies for creating ML models to predict air pollutant concentrations (e.g., PM_{10}, NO, NO_2, and O_3) while utilising meteorological, satellite, and temporal features as predictors. Predictors are values used as inputs in models to help forecast or predict air pollutant concentrations. The objectives of this study were to 1) evaluate data coverage and the choice of both target and predictive variables in ML models to improve model accuracy, 2) compare multiple ML/DL algorithms and forecasters, 3) test hybrid time-series ML models, 4) investigate the limitations of the model using error analysis and model deterioration, and 5) estimate the feature contribution to model performance. The city of Graz in southern Austria was examined as a case study.

2 MATERIALS AND METHODS

2.1 Data Sources

Data collection and site specifications for particulate and gas phase pollutants have been described previously (Lovrić *et al.*, 2021; Moser *et al.*, 2019). This study used hourly data from January 2014 to March 2020, which is an increased temporal resolution compared to (Lovrić *et al.*, 2021). The dataset was based on environmental, pollution, and weather data from publicly available sources provided by the regional government of Styria, Austria. The long-term measurements were taken from five measurement sites (Fig. 1): Süd (eng. South), Nord (eng. North), West (eng. West), Don Bosco, and Ost (eng. East). The latter two are situated on arterial roads with high traffic volumes, especially during the morning and evening rush hours.

A detailed description of the site and potential pollution sources were obtained from a past

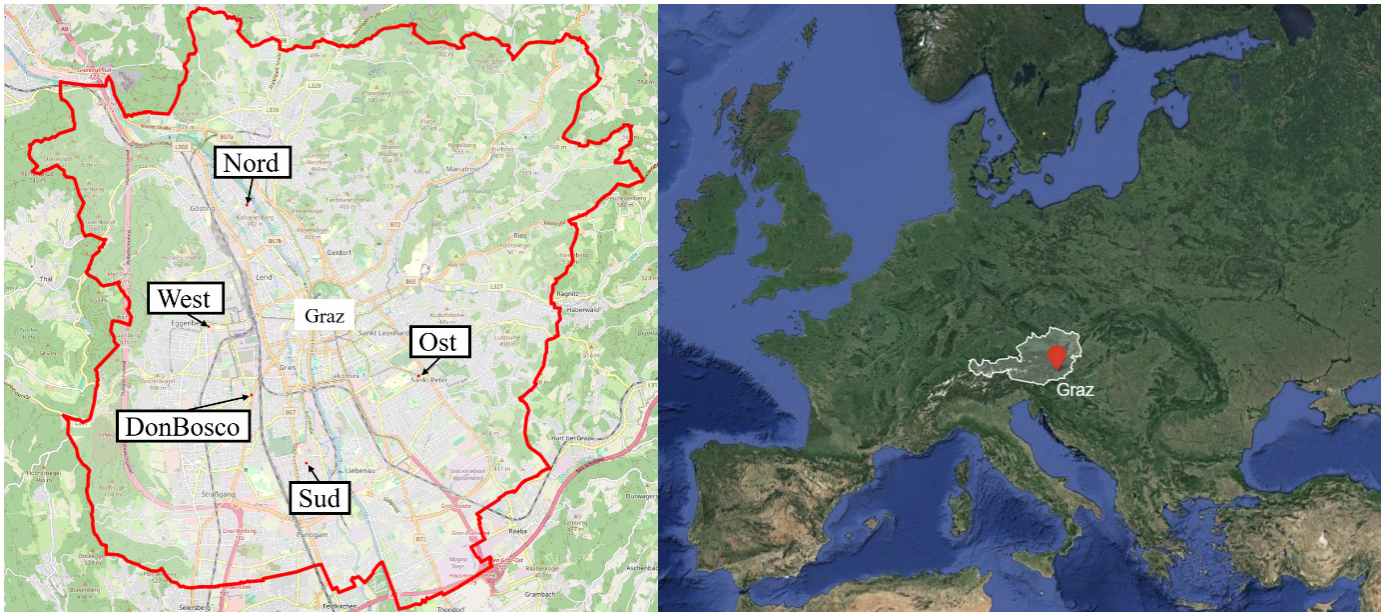

Fig. 1. Locations of the measurement sites in Graz, Austria.

study (Lovrić *et al.*, 2021). The concentrations of NO_2, NO, O_3, and PM_{10} were prepared in a flat table together with temporal information (day of week, days counted from 1st of January 1970 (Julian date), month, year, holiday etc.) and meteorological variables (maximum daily temperature (T), minimal daily T, difference of max and min T, average T, maximum daily pressure (p), minimum daily p, difference of max and min p, average p, maximum daily relative humidity (RH), minimum daily RH, average RH, difference of max and min RH, wind speed, precipitation). To the best of our knowledge, the scope of this study extends beyond prior studies by evaluating an additional set of variables, namely ERA5 reanalysis alongside previously utilised features such as the listed ones. ERA5-Land is a global meteorological reanalysis dataset covering the period from 1950 to the present on a 0.1° × 0.1° horizontal grid (Hersbach *et al.*, 2023; Muñoz Sabater, 2019). Reanalysis combines model data with global observations from across the world into a globally complete and consistent dataset using the laws of physics. ERA5-Land reanalysis does not use observations directly but uses them as inputs to control the simulated land fields and ERA5 atmospheric variables, such as air temperature and air humidity. Hourly reanalysis data were retrieved through API requests from the Copernicus Climate Data Store (Hersbach *et al.*, 2023; Muñoz Sabater, 2019) and aggregated into daily mean, maximum, or minimum values. All analyses were performed using Python (v3.9.16). The libraries and installation environment are provided in the Supplementary Material. The retrieved ERA5 data covered the 1st January 2014 to the 30th April 2022 between 15°E and 16°E and 46.5°N and 47.5°N. The retrieved variables are listed in Table S1 in the Supplementary Material.

2.2 Data Processing

The first data operation involved the removal of outliers. Values with extremely high or low values (such as those affected by fireworks around New Year's Eve) were winsorised based on three-day windows (72 h), where all values above and below four standard deviations in the window were excluded. Missing values were filled in using an iterative imputer. This method models each feature with missing values as a function of the other features, taking turns in a round-robin manner. After filling in the missing values, lag features were created using meteorological data from all five stations (Šimić *et al.*, 2020). Lagging was performed by considering the 12 most recent measurements (12 h window) and the median of the preceding measurements was calculated to represent the value of that hour. To obtain an accurate representation of the wind direction for each measurement, the sinus and cosine were calculated to derive the x and y values in the coordinate system. This was done to avoid situations in which 0° and 360° were interpreted

as different wind directions. The wind speed at 10 m was calculated using the u and v components. Finally, the features were cleaned based on the correlation (above 90%), with all features assigned to all models. The processed dataset consists of 71929 time points from 00:00 on the 1st of January 2014 to the 15th of March 2020, with a total of 123 features, and is provided in a table format within a persistent data repository (Lovrić *et al.*, 2023).

2.3 Model Training and Algorithms

For air pollution concentration modelling, several approaches were used: 1) RF regression (Breiman, 2001), based on our previous studies (Lovrić *et al.*, 2021; Šimić *et al.*, 2020); 2) Prophet (PRH) (Taylor and Letham, 2018); 3) three DL architectures, namely, multilayer perceptron (MLP) (Rosenblatt, 1958), long short-term memory network (LSTM) network, and a one-dimensional CNN (LeCun *et al.*, 2015); and 4) a hybrid RF model (HYB) inspired by previous works (Hoffer *et al.*, 2022; Lovrić *et al.*, 2020) which uses PRH to generate additional features based on forecasting alongside all other available features. The algorithms are briefly described in the Supplementary Material.

The inspiration for training hybrid models stems from the group's previous works, where different types of models, commonly physical and ML models, were combined and yielded higher accuracies (Hoffer *et al.*, 2022; Lovrić *et al.*, 2021). One way to form a hybrid model is to generate features in one model and utilise them in a subsequent model. In this study, we fed PRH-generated features, such as forecasts and seasonalities, into RF models since RF can't infer those directly from the data. Hence, we trained the PRH models and stored the generated features, which were then utilised alongside meteorological and temporal features to generate predictions for the hybrid models.

The predicted variables (target or outcomes) in this study were the pollutant concentrations at hourly frequencies (PM_{10}, NO, NO_2, and O_3) at all measured locations, while the independent (input) features were the temporal and meteorological variables. A schematic of the model is shown in Fig. 2. The model assumed that the concentrations of PM and gaseous pollutants can be modelled based on temporal and meteorological variables as independent variables. The input of the three DL models at a certain hour consisted of weather and environmental variables, in addition to the ERA5 variables of the past 96 h. With this input, the model predicted the concentration of the target pollutant for the next hour. The data were split into training (from 1st of January 2014 to 14th March 2019) and test (from 15th of March 2019 to 15th of March 2020) sets to determine the extent to which the models degrade over a longer period. A cross-validation

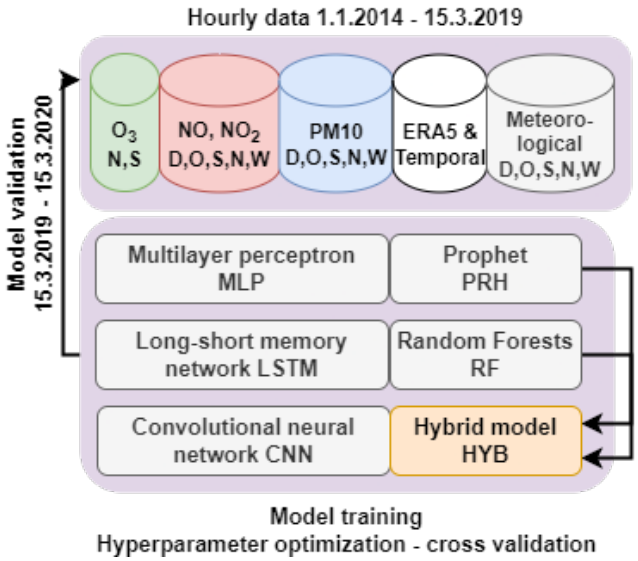

Fig. 2. Overview of the study methodology to detect changes in the relation ships between the (dependent and independent) variables. Deep learning models are colored in grey, while PRH and RF are highlighted since they flow into the HYB model.

was performed using the training set. The details of the validation are provided in the Supplementary Material. Performance was measured by calculating the R^2, normalized mean absolute error (nMAE), and normalized root-mean-square error (nRMSE). R^2 was used as the main metric for standardised intercomparison as it is independent of scale. nMAE and nRMSE were computed by normalizing each metric based on the interquartile range (IQR), which is the difference between the 75th percentile and the 25th percentile.

3 RESULTS

3.1 Model Performance

The model results are depicted in Fig. 3, which shows the performance of all trained models for the test set. The calculation for each metric was performed for each location (Süd, Nord, West, Don Bosco, and Ost) and pollutant (PM_{10}, NO_2, NO, and O_3). The nMAE and nRMSE results revealed notable variations in model performance across different pollutant types and sites (Table S2). For instance, in predicting NO levels at Ost, the HYB and MLP models achieved the lowest nMAE values, indicating superior accuracy. By contrast, LSTM, CNN, and PRH exhibited higher nMAE values, suggesting room for improvement in their predictions. Notably, when assessing NO_2 levels at West, MLP performed exceptionally well, with the lowest nMAE, whereas LSTM struggled, showing relatively higher errors. These variations in performance help reveal which models are suited to predicting specific pollutants in certain regions. The nRMSE results also highlighted distinctions in model performance (Table S3). For instance, MLP consistently demonstrated lower nRMSE values than the other models for predicting PM_{10} concentrations at various sites. Conversely, LSTM and PRH exhibited higher nRMSE values in certain cases, suggesting potential limitations in their predictive capabilities for specific pollutants and sites. The R^2 results provided insights into the overall goodness-of-fit of the models. MLP consistently achieved high R^2 across multiple pollutant types and regions, indicating its capacity to explain a significant portion of the variance in pollutant concentrations (Fig. 3). In contrast, LSTM and PRH exhibited lower R^2 in some cases, suggesting that they did not capture the underlying patterns in the data. The models also showed distinct patterns for specific pollutants. HYB demonstrated remarkable accuracy in predicting NO_2 concentrations, whereas CNN and MLP were superior in forecasting PM_{10} levels. Furthermore, regarding NO prediction, MLP was the most accurate, followed by HYB. For O_3 concentrations, all models except for PRH exhibited high accuracy.

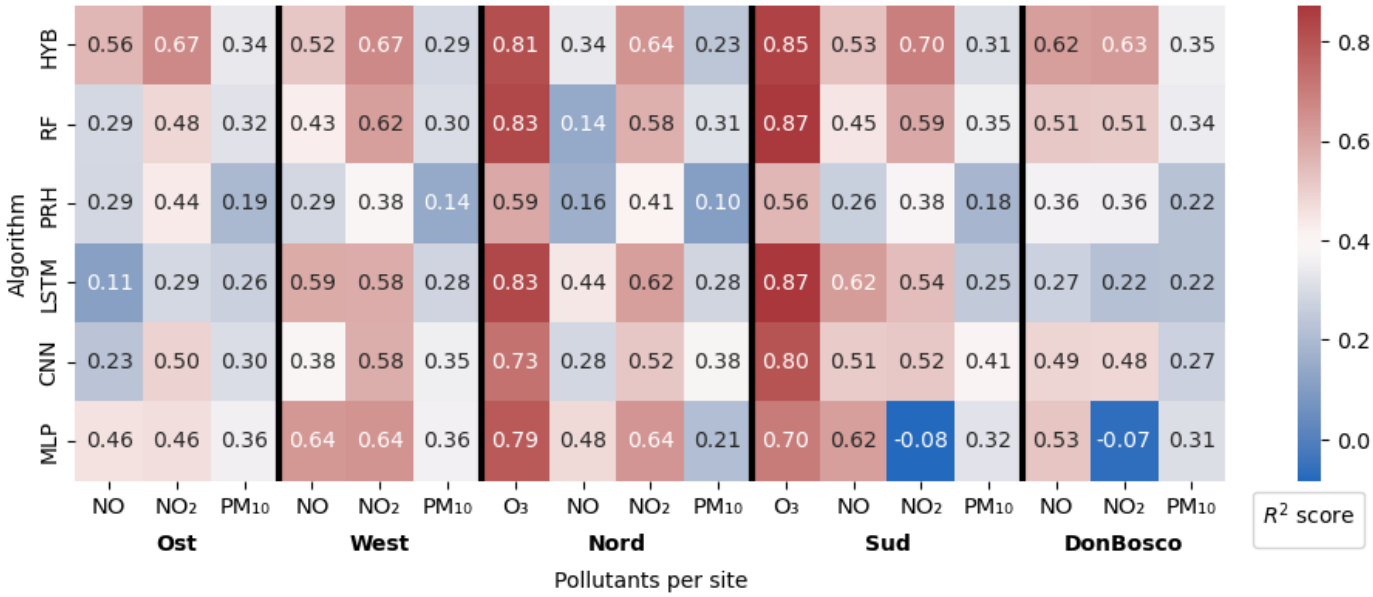

Fig. 3. Model prediction accuracy per pollutant. This figure shows the performance for the different models. The performance was measured by calculating the coefficient of determination (R^2 score).

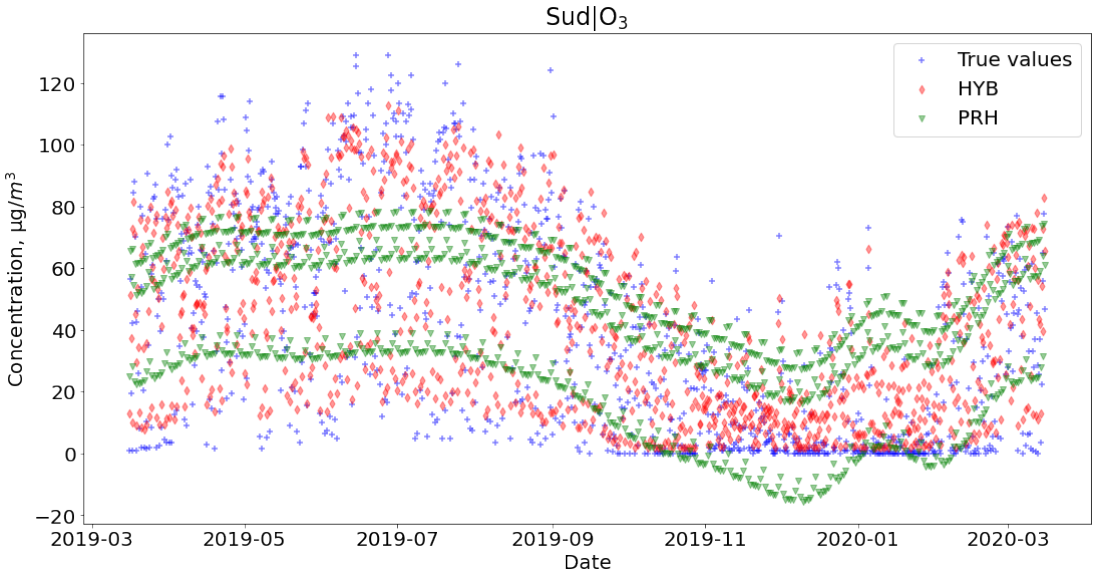

Fig. 4. A comparison of predicted and true values from the test set. This figure shows the true and predicted values of ozone concentration at the Sud measuring site using the HYB model and PRH model, respectively. Each value is the mean of 8 hourly measurements.

3.2 Performance of HYB

The performance of HYB is a noteworthy finding of this study. In previous work (Lovrić *et al.*, 2021), a Random Forest (RF) model was created to predict daily concentrations of air pollutants, using traffic as one of the features. However, in this paper, we used hourly data and hybrid (HYB) models, which showed better performance in predicting air pollutants. The R^2 values for each specific pollutant are better than those in the previous paper. Integrating seasonality features into the RF model induced changes in the model's predictions by incorporating the features generated by PRH. This inclusion resulted in improved predictions for NO_2 and NO across all sites, exhibiting R^2 ranging from 0.521 to 0.618 for NO and from 0.63 to 0.67 for NO_2 (see Fig. S1 in the Supplementary Material). Notably, the most substantial enhancement in NO_2 (18%) and a large increase in NO (26%) was observed at Ost. Conversely, Süd displayed the smallest improvement in these two pollutants, with a 7% increase for NO and a 5% increase for NO_2. In contrast, RF showcased higher R^2 than the HYB model in predicting O_3, ranging from 0.83 to 0.87, and PM_{10} levels, ranging from 0.29 to 0.35. In addition, the incorporation of time-series data led to a decrease in R^2 for PM_{10} prediction, with the most significant decline (8 %) observed at Nord. This decrease can be attributed to the inherent randomness of the models. Similarly, the efficiency of O_3 prediction in HYB decreased by 2% at both Nord and Süd, which can also be attributed to the randomness and influx of features resulting from the integration of the PRH data. Fig. 4 shows an example of HYB model prediction compared against a PRH forecast for the same site and period. PRH predicted negative concentrations, which are physically implausible. This illustrates the poor performance of the PRH model in predicting air pollution concentrations.

3.3 Day- and Night-time Differences in Model Quality

The models were tested against potential limitations to better understand their performance. One of these is the daytime cycle of the predictive quality. This assumes that, owing to missing data on traffic or a lack of available data at the same frequency, there might be differences in daily variations. Fig. 5 shows how the models predicted the daytime and nighttime air pollutant concentrations. These data were created by dividing the time series of the predictions, which included both the true and predicted values for the external test set, into two subsets. The first subset included data from 06:00 until 17:59 (daytime) and the second dataset included data from 18:00 until 05:59 (nighttime). Each dataset had a vector length of 4392. Seasonal changes in daytime and nighttime were not considered since these periods were used to represent commuting

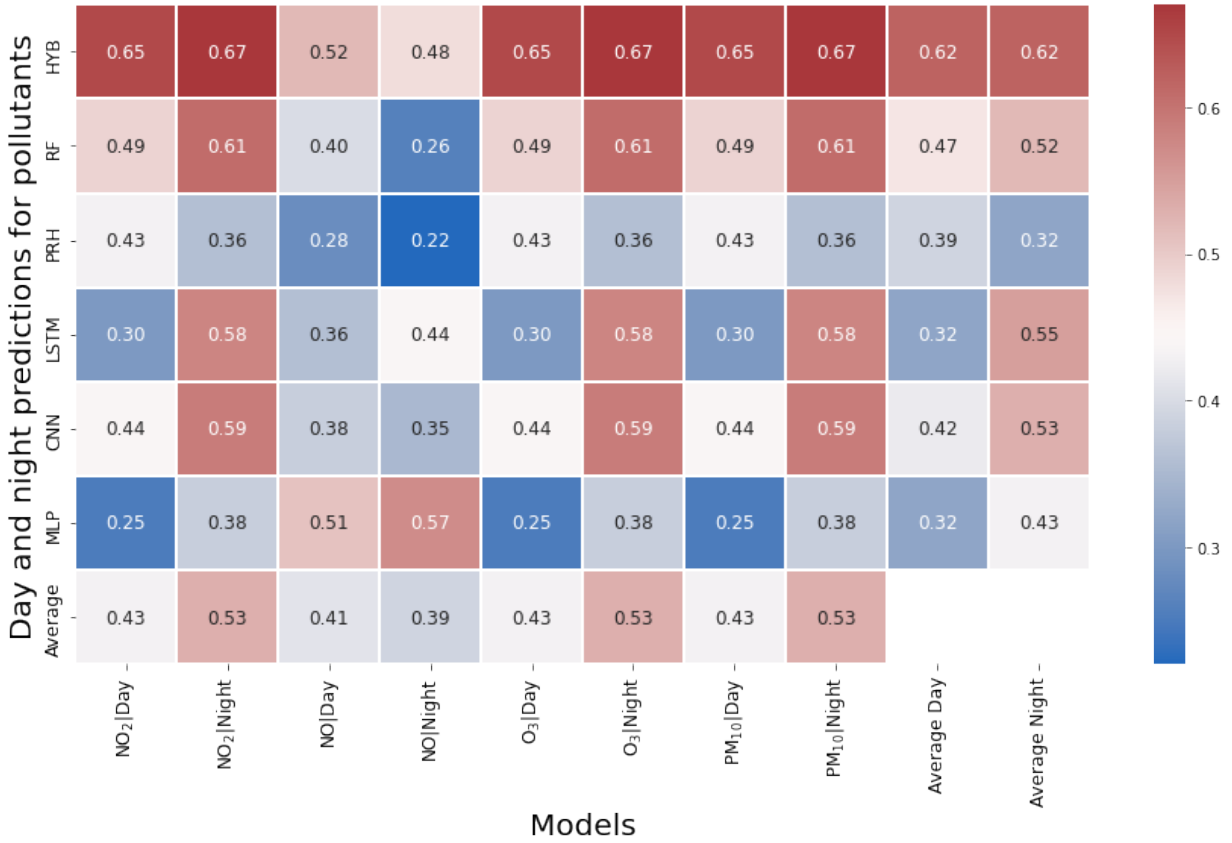

Fig. 5. Heatmap displays the day and night predictions of every pollutant with all six models, using the coefficient of determination as the metric.

and work schedules and not the actual state of the atmosphere and Earth's rotation. After the two data subsets were created, R^2 was calculated for the daytime and nighttime vectors. The best overall model for both vectors was HYB (see Fig. 5), except for NO for nighttime, for which MLP showed better results. Furthermore, a row called" average" was created using the results of each model. Therefore, to calculate the average NO_2 for the daytime, all six mode values for each pollutant and part of the daytime were considered. This average value showed that night-time was usually better predicted than daytime for all pollutants, except for NO. Additionally, "average daytime' and" average night-time' were calculated to analyse how each model predicted daytime and night-time values. For example, to obtain the average HYB daytime, every prediction for every pollutant predicted using HYB during the daytime was considered. All models, except for PRH, had higher accuracies for the night-time concentrations of air pollutants than the daytime concentrations. However, HYB showed similar results for both

3.4 Weekday Differences in Model Quality

The objective of this subsection is to address the limitations of the models caused by weekday disparities, including potential fluctuations that may be crucial for model generalisation, presumably linked to the absence of traffic data. To better comprehend the effect of weekdays on pollutant levels, R^2 was calculated by splitting the prediction vectors and their corresponding true values into seven subsets, representing each day of the week. Model quality metrics were computed for each subset, considering individual air pollutants and consolidating the outcomes across multiple locations. Only the best model was included in the analysis of the effect of weekdays on pollutants. The examination commenced by focusing on PM_{10}, which is the most challenging pollutant to predict using the model. R^2 ranged from 0.24 on Sunday to 0.31 on Thursday (see Fig. S2). The model provided more accurate predictions on weekdays than on weekends. For NO, the model exhibited improved performance, with R^2 ranging from 0.42 to 0.52. The weakest predictions were

observed on Saturdays and Sundays, indicating that the model performs better on workdays. However, NO_2 displayed different outcomes compared to the other pollutants, with R^2 ranging from 0.61 to 0.68, the highest values recorded on Saturday and Sunday, and strong predictions on Friday. Lastly, for O_3, the prediction was consistent throughout the week, with R^2 ranging from 0.82 to 0.84. Overall, no significant variations were identified when pollutant predictions across different weekdays were compared.

3.5 Prediction Deterioration over Time

The models were expected to deteriorate owing to commonly observed drifts in the covariance matrix between the predictors and targets, sometimes also referred to as concept drift (Lu *et al.*, 2019). This can occur due to meteorological changes or other events in an urban setting, such as long-term road closure, new regulations causing changes in traffic, and other government policies, such as the COVID-19 lockdowns (Lovrić *et al.*, 2022, 2021; Stipaničev *et al.*, 2022). In concept drift, the models deteriorate, allowing the generalisation ability of models to be determined over a longer period. Model deterioration was analysed by predicting a one-year time series and inspecting the model quality. Fig. 6 shows the true versus predicted values for each pollutant and

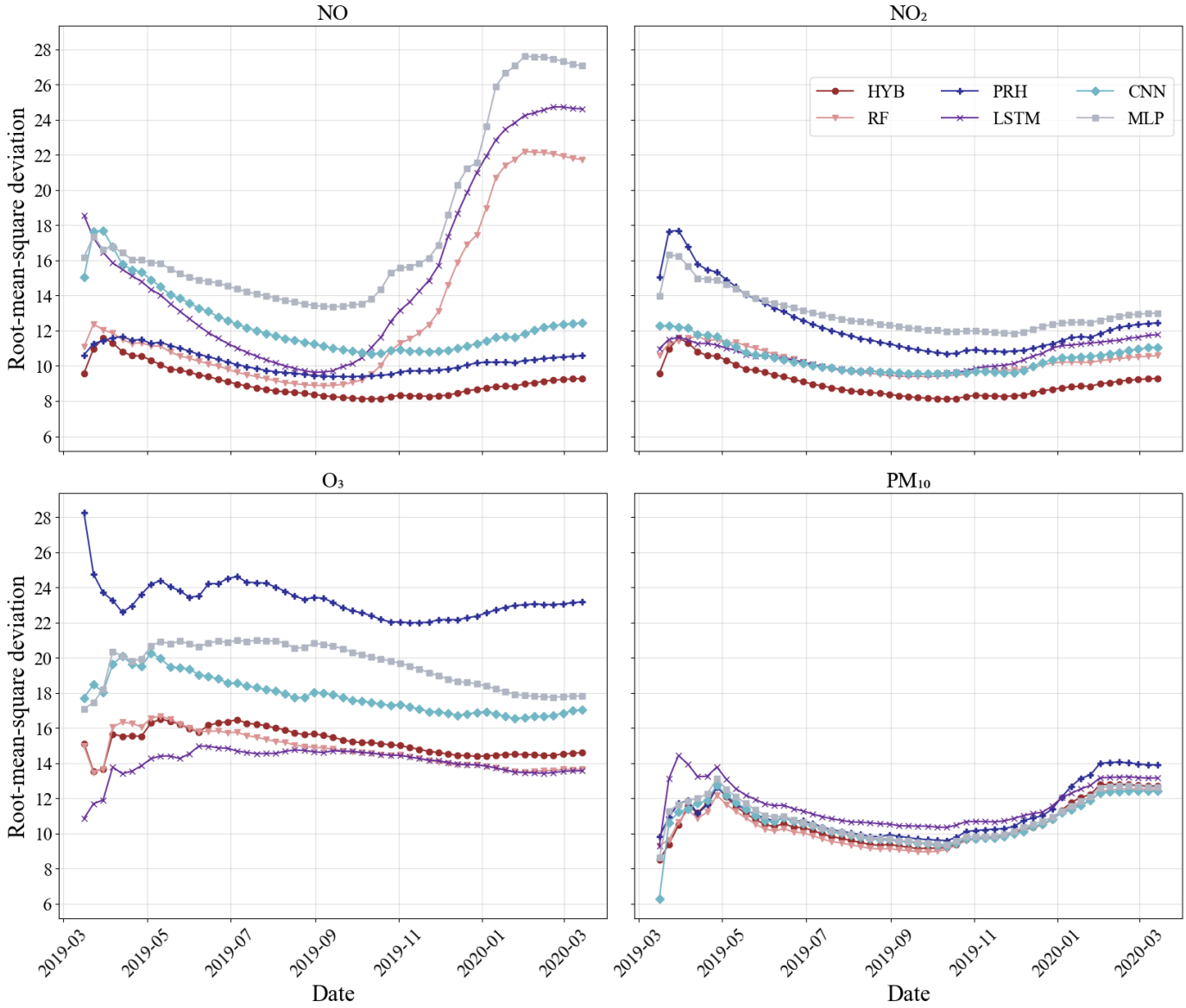

Fig. 6. Root-mean-square deviation of models per pollutant over one year of a test set.

model. Fig. 6 was generated by calculating the root-mean-square error (RMSE) deviation for each pollutant at each of the five sites and then averaging the scores across the five sites to represent the model predictions for the pollutant. The same process was repeated every hour of the day. For visualisation purposes, the data were smoothed the rolling mean of seven days was used. The error in the NO_2 level worsened during the colder months of the year. However, the models were generally stable. A similar pattern was observed for PM_{10}. For NO, there was deterioration in some of the algorithms over time, whereas our emphasised (HYB) model showed good stability. O_3 remained relatively stable over time compared to the other pollutants but had more variance when comparing the models. For all pollutants, considering RMSE deviation, HYB performed well, indicating its ability for long-term predictions.

Additionally, in Fig. 6, some models (MLP, LSTM, and RF) show greater deterioration in performance for NO compared to others from autumn 2019 to spring 2020. Several factors could lead to this. First, the complexity and architecture of the models play a significant role. Different models, such as LSTM and CNN, have varying capabilities in capturing temporal and spatial dependencies compared to traditional machine learning models like RF. Additionally, the choice of hyperparameters during training can significantly affect model performance, with some models being better tuned for the characteristics of NO data. Data handling and feature engineering also contribute to the differences, as each model may preprocess the data and incorporate features differently, affecting their ability to capture underlying patterns in NO concentrations. Furthermore, issues of overfitting and underfitting impact model performance. Lastly, temporal variations and external factors, such as weather conditions and traffic patterns, can cause variations in RMSE over time, with some models being more sensitive to these factors and thus showing varying performance across different periods.

3.6 Feature Analysis

To explain model prediction, we utilised permutation importance (explained in the Supplementary Material), as used in previous studies (Lovrić *et al.*, 2021; Šimić *et al.*, 2020). The trained ML models are nonlinear black boxes; hence, one of the ways to comprehend them is to utilise post-hoc methods model explainability methods such as permutation importance. The final model, HYB, showed the best results. The permutation importance was calculated for each model and pollutant three times, meaning that each feature was shuffled three times and then the model dependence was tested. Because many features were used per model, we reported only the five most important features for each. Out of the 17 models, 14 had a PRH prediction feature as the most important feature: either yhat or yh (a common naming for predicted vectors, here of the PRH model) or its upper/lower values (yh-up, yh-low). Furthermore, 15 out of the 17 models had at least two PRH features among the five most important features. All models had a windspeed feature among the five most important features, either from local meteorology or ERA5 (WS, peek WS). Total precipitation (tp) and 10 m u-component of wind(u10) were in the top five for seven of the models. There were three types of wind features. Although these did not correlate, there was a strong presence of wind driving the concentration variation of the pollutants. Only three models had temperature (T) as an included feature. NO models were strongly weighted by their PRH predictions, showing strong seasonality patterns and less dependence on other factors, probably being driven by traffic, which was not represented in the data. NO_2 models showed a similar pattern, with Süd having a PRH feature in the top five and Don Bosco being strongly weighted by PRH NO prediction. Interestingly, for PM_{10}, the presence of PRH NO_2 predictions was clear. This points to NO_2 partially explaining PM_{10}, probably via being a surrogate for traffic (Gilbert *et al.*, 2003). The two models for O_3, both had a strong dependence on their PRH predictions show. The results are shown in Table 1.

4 DISCUSSION

Overall, the models had the best generalisation for O_3 followed by NO_2, NO, and PM_{10}, which is in line with previous studies (Lovrić *et al.*, 2021; Šimić *et al.*, 2020). The present study went further than past studies by incorporating more complex methods. The DL methods of LSTM,

Table 1. Feature importance ranked by means of the Permutation Importance method with Feature 1 being the most important followed by others in descending order.

Target	Feature1 set	Feature2 set	Feature3 set	Feature4 set	Feature5 set
Ost|NO	NO (pr,yh)	NO (pr,yh-low)	WS	u10	peek WS
Ost|NO_2	NO_2 (pr,yh-up)	peek WS	NO_2 (pr,yh-low)	RH	str
Ost|PM_{10}	PM_{10} (pr,yh-low)	peek WS	tp	sshf	NO_2 (pr,ys)
West|NO	NO (pr,yh)	NO (pr,yh-low)	tp	u10	WS
West|NO_2	NO_2 (pr,yh)	peek WS	tp	NO_2 (pr,yh-up)	WS
West|PM_{10}	NO_2 (pr,ys)	PM_{10} (pr,yh-low)	tp	T	peek WS
Nord|O_3	O_3 (pr,yh)	RH	peek WS	T	O_3 (pr,yh-up)
Nord|NO	NO (pr,yh)	NO (pr,yh-low)	u10	Rad	peek WS
Nord|NO_2	NO_2 (pr,yh)	peek WS	u10	NO_2 (pr,yh-up)	NO_2 (pr,yh-low)
Nord|PM_{10}	NO_2 (pr,ys)	PM_{10} (pr,yh-low)	slhf	peek WS	tp
Sud|O_3	RH	WS	O_3 (pr,yh)	O_3 (pr,yh-low)	T
Sud|NO	WS	u10	NO (pr,yh)	NO (pr,yh-low)	tp
Sud|NO_2	WS	NO_2 (pr,yh)	peek WS	RH	u10
Sud|PM_{10}	NO_2 (pr,ys)	peek WS	NO_2 (pr,yh)	tp	PM_{10} (pr,yh-low)
DonBosco|NO	NO (pr,yh)	peek WS	u10	NO (pr,yh-low)	WS
DonBosco|NO_2	NO_2 (pr,yh-up)	peek WS	NO_2 (pr,yh-low)	WS	NO (pr,ws)
DonBosco|PM_{10}	PM_{10} (pr,yh-low)	tp	peek WS	smlt	T

CNN, and MLP achieved competitive R^2 scores for PM_{10}, O_3, and NO. It is expected that forecasting methods and DL will be superior because of their capability to inherently capture temporal dependencies and complex patterns in multivariate time-series data, which can be a significant advantage over RF. However, the results showed that the gap between the RF/HYB and DL predictions was marginal. This similarity in performance has been reported in the literature for time-series forecasting (Makridakis *et al.*, 2023), particularly for short-term forecasting. Indeed, the literature shows that well-designed statistical methods can outperform complicated state-of-the-art DL methods (Elsayed *et al.*, 2021; Makridakis *et al.*, 2023). In general, when the gain in the performance of DL methods is marginal, as in our case, more efficient and explainable approaches, such as RF and HYB, are preferred. PRH adeptly captures time series nuances, including trends and seasonality. Whereas RF excels in detecting nonlinear relationships and intricate feature interactions. Together, they offer a more comprehensive modelling approach with improved prediction accuracy. Similar results were observed in past studies (Hoffer *et al.*, 2022; Lovrić *et al.*, 2021). The results showed that the predicted night-time pollution concentration values better agreed with the true values than the daytime values. This may be because pollutant concentrations during the night are less variable than those during the day due to the lack of traffic, industrial operations, and other human activities. Additionally, stable atmospheric conditions, such as lower wind speeds and less atmospheric mixing, contribute to this reduced variability (Dobson *et al.*, 2021; Singh *et al.*, 2020). Daytime pollutant concentrations started to increase with increasing daily human activity, which usually started at 06:00 (Kecorius *et al.*, 2017). Between 06:00 and 18:00, pollutant concentrations suddenly increased (due to peak hours), then slightly decreased at noon because of greater mixing in the planetary boundary layer, increased again because of evening rush hour, and stabilised during the nighttime because of the lack of emission sources. In comparison, nighttime pollution concentrations remained relatively stable, which may have been easier to capture by the models. In our previous study (Lovrić *et al.*, 2021), O_3 was easier to predict than NO_2 and PM_{10}, as were the daily averages. A better model representation of the O_3 concentration may be the result of a distinct O_3 concentration profile during the daytime. Ozone is produced by a chemical reaction between natural and anthropogenic pollutants (gases) involving sunlight (Steinfeld, 1998). Therefore, tropospheric O_3 concentrations highly depend on human daily activity patterns and sunlight, both of which peak between 06:00 and 18:00. This causes a monotonous daily O_3 concentration increase and decrease during the day and night, respectively (Singh *et al.*, 2020). The nitric oxides have features similar to those in equilibrium (Steinfeld, 1998). In our previous work, we showed that PM_{10} is challenging to predict (Lovrić *et al.*, 2021). There

are many potential sources of PM_{10}, some of which are difficult to model, such as sand events (Federal Office: MeteoSwiss, 2020), sun flares, local construction, burning, earthquakes (Lovrić *et al.*, 2022), and other weather patterns. The major unknowns in such analyses are traffic and other emission factors from vehicular fleets, chemistry and physics, emission factors from industry, and atmospheric transport.

Regarding the important features, the results suggested a clear pattern in which 14 out of 17 models identified PRH prediction as the most crucial feature, either in the form of yhat or its upper/lower values. The majority (15 of 17) of models had at least two PRH features among the top five most significant features. This showcases a strong dependence on temporal information and seasonality which is commonly missing for models like RF but can be added either through lag features (Šimić *et al.*, 2020) or PRH-generated features. The five sites exhibited unique pollution sources and environmental conditions that influenced the features chosen for air pollutant concentration modelling. Feature selection is affected by the data availability, which varies across sites. Correlations among features can lead to the exclusion of some features to prevent multicollinearity. Techniques such as boosting aggregation in random trees can be used to ensure debiasing and address collinearity. Additionally, temporal changes, such as seasonal variations, can shift feature importance. Overall, multiple factors, from site-specific conditions to modelling techniques, determine the feature selection; hence, different features may be chosen.

These findings underscore the significant impact of model selection on the accuracy of air pollution predictions, highlighting that different models may yield varying performances depending on the specific pollutants and locations under examination. Further research is warranted to delve into the underlying factors contributing to these disparities and identify optimal strategies for model selection in different air pollution scenarios, as well as mechanisms and sources.

5 CONCLUSION

This study evaluated several factors for modelling the hourly concentrations of commonly monitored air pollutants in urban settings, such as PM_{10}, NO, NO_2, and O_3. These factors were the a) choice of the ML algorithm, b) type of features used, c) model deterioration, and d) sensitivity to temporal events. The algorithms were based on DL (LSTM, CNN, and MLP), time-series forecasting (PRH), and ensembles (RF). We also proposed HYB, which fed RF into the PRH forecast. HYB performed best for all pollutants, regardless of the time of day, except for NO during the night, for which MLP performed best. RF, LSTM, and CNN better predicted NO_2, O_3, and PM_{10} during the night compared to daytime than the other models. PRH performed poorly for almost every pollutant but performed better for the daytime dataset than the nighttime dataset. Overall, the models showed the best generalisation for O_3, followed by NO_2, NO, and PM_{10}. In addition to the more accurate results, HYB exhibited low model deterioration over a one-year prediction period. We further inspected which predictive features had a high weight in the predictions of the best models using permutation importance. The models were mostly driven by seasonality and predictions from the PRH in HYB. Another highly relevant driver was wind speed, which was represented by different features. The results showed that temporal variations are strong predictors of air pollutant concentrations and indicate missing key features, which represent sources such as heating and traffic. In the absence of unobserved confounders, the predictions can be improved by adding temporal features to act as surrogates.

ADDITIONAL INFORMATION

Funding

M.L., H.H., K.Č., and V.P. are partially funded by the EU-Commission Grant Nr. 101057497 - EDIAQI. Know-Center is a COMET Centre within the COMET – Competence Centers for Excellent Technologies Programme and funded by BMK, BMAW as well as the co-financing provinces Styria, Vienna, and Tyrol. COMET is managed by FFG.

Data Availability

The data is deposited at https://zenodo.org/deposit/7959116

Supplementary Material

Supplementary material for this article can be found in the online version at https://doi.org/10.4209/aaqr.230317

REFERENCES

Almetwally, A.A., Bin-Jumah, M., Allam, A.A. (2020). Ambient air pollution and its influence on human health and welfare: An overview. Environ. Sci. Pollut. Res. Int. 27, 24815–24830. https://doi.org/10.1007/s11356-020-09042-2

Anderson, J.O., Thundiyil, J.G., Stolbach, A. (2012). Clearing the air: A review of the effects of particulate matter air pollution on human health. J. Med. Toxicol. 8, 166–175. https://doi.org/10.1007/s13181-011-0203-1

Araki, S., Hasunuma, H., Yamamoto, K., Shima, M., Michikawa, T., Nitta, H., Nakayama, S.F., Yamazaki, S. (2021). Estimating monthly concentrations of ambient key air pollutants in Japan during 2010–2015 for a national-scale birth cohort. Environ. Pollut. 284, 117483. https://doi.org/10.1016/j.envpol.2021.117483

Bozdağ, A., Dokuz, Y., Gökçek, Ö.B. (2020). Spatial prediction of PM_{10} concentration using machine learning algorithms in Ankara, Turkey. Environ. Pollut. 263, 114635. https://doi.org/10.1016/j.envpol.2020.114635

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

Camatini, M., Gualtieri, M., Sancini, G. (2017). Impact of the Airborne Particulate Matter on the Human Health, in: Tomasi, C., Fuzzi, S., Kokhanovsky, A. (Eds.), Atmospheric Aerosols, Wiley, pp. 597–643. https://doi.org/10.1002/9783527336449.ch10

Chen, B., You, S., Ye, Y., Fu, Y., Ye, Z., Deng, J., Wang, K., Hong, Y. (2021). An interpretable self-adaptive deep neural network for estimating daily spatially-continuous $PM_{2.5}$ concentrations across China. Sci. Total Environ. 768, 144724. https://doi.org/10.1016/j.scitotenv.2020.144724

Dobson, R., Siddiqi, K., Ferdous, T., Huque, R., Lesosky, M., Balmes, J., Semple, S. (2021). Diurnal variability of fine-particulate pollution concentrations: data from 14 low- and middle-income countries. Int. J. Tuberc. Lung Dis. 25, 206–214. https://doi.org/10.5588/ijtld.20.0704

Elsayed, S., Thyssens, D., Rashed, A., Jomaa, H.S., Schmidt-Thieme, L. (2021). Do we really need deep learning models for time series forecasting? arXiv:2101.02118 https://doi.org/10.48550/arXiv.2101.02118

Federal Office: MeteoSwiss (2020). Saharan dust events - MeteoSwiss. https://www.meteoswiss.admin.ch/home/climate/the-climate-of-switzerland/specialties-of-the-swiss-climate/saharan-dust-events.html (accessed 31 July 2020).

Gagliardi, R.V., Andenna, C. (2020). A machine learning approach to investigate the surface ozone behavior. Atmosphere 11, 1173. https://doi.org/10.3390/atmos11111173

Gilbert, N.L., Woodhouse, S., Stieb, D.M., Brook, J.R. (2003). Ambient nitrogen dioxide and distance from a major highway. Sci. Total Environ. 312, 43–46. https://doi.org/10.1016/S0048-9697(03)00228-6

Grange, S.K., Carslaw, D.C. (2019). Using meteorological normalisation to detect interventions in air quality time series. Sci. Total Environ. 653, 578–588. https://doi.org/10.1016/j.scitotenv.2018.10.344

Grange, S.K., Lee, J.D., Drysdale, W.S., Lewis, A.C., Hueglin, C., Emmenegger, L., Carslaw, D.C. (2021). COVID-19 lockdowns highlight a risk of increasing ozone pollution in European urban areas. Atmos. Chem. Phys. 21, 4169–4185. https://doi.org/10.5194/acp-21-4169-2021

Hersbach, H., Bell, B., Berrisford, P., Biavati, G., Horányi, A., Muñoz Sabater, J., Nicolas, J., Peubey, C., Radu, R., Rozum, I., Schepers, D., Simmons, A., Soci, C., Dee, D., Thépaut, JN. (2023): ERA5 hourly data on single levels from 1940 to present. Copernicus Climate Change Service (C3S) Climate Data Store (CDS). https://doi.org/10.24381/cds.adbb2d47

Hoffer, J.G., Ofner, A.B., Rohrhofer, F.M., Lovric, M., Kern, R., Lindstaedt, S., Geiger, B.C. (2022). Theory-inspired machine learning—towards a synergy between knowledge and data. Weld. World 66, 1291–1304. https://doi.org/10.1007/S40194-022-01270-Z

Huang, G., Li, X., Zhang, B., Ren, J. (2021). $PM_{2.5}$ concentration forecasting at surface monitoring sites using GRU neural network based on empirical mode decomposition. Sci. Total Environ. 768, 144516. https://doi.org/10.1016/j.scitotenv.2020.144516

Huang, Y., Zhou, J.L., Yu, Y., Mok, W.C., Lee, C.F.C., Yam, Y.S. (2020). Uncertainty in the impact of the COVID-19 pandemic on air quality in Hong Kong, China. Atmosphere 11, 914. https://doi.org/10.3390/atmos11090914

Jakovljević, I., Sever Štrukil, Z., Godec, R., Bešlić, I., Davila, S., Lovrić, M., Pehnec, G. (2020). Pollution sources and carcinogenic risk of PAHs in PM_1 particle fraction in an urban area. Int. J. Environ. Res. Public. Health 17, 9587. https://doi.org/10.3390/ijerph17249587

Jiang, F., Qiao, Y., Jiang, X., Tian, T. (2021a). MultiStep ahead forecasting for hourly PM_{10} and $PM_{2.5}$ based on two-stage decomposition embedded sample entropy and group teacher optimization algorithm. Atmosphere 12, 64. https://doi.org/10.3390/atmos12010064

Jiang, S., Zhao, C., Fan, H. (2021b). Toward understanding the variation of air quality based on a comprehensive analysis in Hebei Province under the influence of COVID-19 lockdown. Atmosphere 12, 267. https://doi.org/10.3390/atmos12020267

Kecorius, S., Madueño, L., Vallar, E., Alas, H., Betito, G., Birmili, W., Cambaliza, M.O., Catipay, G., Gonzaga-Cayetano, M., Galvez, M.C., Lorenzo, G., Müller, T., Simpas, J.B., Tamayo, E.G., Wiedensohler, A. (2017). Aerosol particle mixing state, refractory particle number size distributions and emission factors in a polluted urban environment: Case study of Metro Manila, Philippines. Atmos. Environ. 170, 169–183. https://doi.org/10.1016/j.atmosenv.2017.09.037

Kim, K.H., Kabir, E., Kabir, S. (2015). A review on the human health impact of airborne particulate matter. Environ. Int. 74, 136–143. https://doi.org/10.1016/j.envint.2014.10.005

LeCun, Y., Bengio, Y., Hinton, G. (2015). Deep learning. Nature 521, 436–444. https://doi.org/10.1038/nature14539

Li, N., Chen, G., Liu, F., Mao, S., Liu, Y., Hou, Y., Lu, Y., Liu, S., Wang, C., Xiang, H., Guo, Y., Li, S. (2019). Associations of long-term exposure to ambient PM_1 with hypertension and blood pressure in rural Chinese population: The Henan rural cohort study. Environ. Int. 128, 95–102. https://doi.org/10.1016/j.envint.2019.04.037

Li, Y., Sha, Z., Tang, A., Goulding, K., Liu, X. (2023). The application of machine learning to air pollution research: A bibliometric analysis. Ecotoxicol. Environ. Saf. 257, 114911. https://doi.org/10.1016/j.ecoenv.2023.114911

Liu, J. (2021). Mapping high resolution national daily NO_2 exposure across mainland China using an ensemble algorithm. Environ. Pollut. 279, 116932. https://doi.org/10.1016/j.envpol.2021.116932

Lovrić, M., Meister, R., Steck, T., Fadljević, L., Gerdenitsch, J., Schuster, S., Schiefermüller, L., Lindstaedt, S., Kern, R. (2020). Parasitic resistance as a predictor of faulty anodes in electro galvanizing: a comparison of machine learning, physical and hybrid models. Adv. Model. Simul. Eng. Sci. 7, 46. https://doi.org/10.1186/s40323-020-00184-z

Lovrić, M., Pavlović, K., Vuković, M., Grange, S.K., Haberl, M., Kern, R. (2021). Understanding the true effects of the COVID-19 lockdown on air pollution by means of machine learning. Environ. Pollut. 274, 115900. https://doi.org/10.1016/j.envpol.2020.115900

Lovrić, M., Antunović, M., Šunić, I., Vuković, M., Kecorius, S., Kröll, M., Bešlić, I., Godec, R., Pehnec, G., Geiger, B.C., Grange, S.K., Šimić, I. (2022). Machine Learning and Meteorological Normalization for Assessment of Particulate Matter Changes during the COVID-19 Lockdown in Zagreb, Croatia. Int. J. Environ. Res. Public. Health 19, 6937. https://doi.org/10.3390/ijerph19116937

Lovrić, M., Petrić, V., Pavlović, K., Schopper, A., Vuckovic, M. (2023). Hourly air pollution data for Graz, Austria. Zenodo https://doi.org/10.5281/ZENODO.7959116

Lu, J., Liu, A., Dong, F., Gu, F., Gama, J., Zhang, G. (2019). Learning under concept drift: A review. IEEE Trans. Knowl. Data Eng. 31, 2346–2363. https://doi.org/10.1109/TKDE.2018.2876857

Makridakis, S., Spiliotis, E., Assimakopoulos, V., Semenoglou, A.A., Mulder, G., Nikolopoulos, K.

(2023). Statistical, machine learning and deep learning forecasting methods: Comparisons and ways forward. J. Oper. Res. Soc. 74, 840–859. https://doi.org/10.1080/01605682.2022.2118629

Moser, F., Kleb, U., Katz, H. (2019). Statistische Analyse der Luftqualitätin Graz anhand von Feinstaub und Stickstoffdioxid. Joanneum Research, Graz.

Muñoz Sabater, J. (2019). ERA5-Land hourly data from 1950 to present. Copernicus Climate Change Service (C3S) Climate Data Store (CDS). https://doi.org/10.24381/cds.e2161bac

Nuvolone, D., Petri, D., Voller, F. (2018). The effects of ozone on human health. Environ. Sci. Pollut. Res. 25, 8074–8088. https://doi.org/10.1007/s11356-017-9239-3

Park, Y., Kwon, B., Heo, J., Hu, X., Liu, Y., Moon, T. (2020). Estimating $PM_{2.5}$ concentration of the conterminous United States via interpretable convolutional neural networks. Environ. Pollut. 256, 113395. https://doi.org/10.1016/j.envpol.2019.113395

Photphanloet, C., Lipikorn, R. (2020). PM_{10} concentration forecast using modified depth-first search and supervised learning neural network. Sci. Total Environ. 727, 138507. https://doi.org/10.1016/j.scitotenv.2020.138507

Qi, C., Zhou, W., Lu, X., Luo, H., Pham, B.T., Yaseen, Z.M. (2020). Particulate matter concentration from open-cut coal mines: A hybrid machine learning estimation. Environ. Pollut. 263, 114517. https://doi.org/10.1016/j.envpol.2020.114517

Rosenblatt, F. (1958). The perceptron: A probabilistic model for information storage and organization in the brain. Psychol. Rev. 65, 386–408. https://doi.org/10.1037/H0042519

Sayahi, T., Garff, A., Quah, T., Lê, K., Becnel, T., Powell, K.M., Gaillardon, P.E., Butterfield, A.E., Kelly, K.E. (2020). Long-term calibration models to estimate ozone concentrations with a metal oxide sensor. Environ. Pollut. 267, 115363. https://doi.org/10.1016/j.envpol.2020.115363

Segersson, D., Eneroth, K., Gidhagen, L., Johansson, C., Omstedt, G., Nylén, A.E., Forsberg, B. (2017). Health impact of PM_{10}, $PM_{2.5}$ and black carbon exposure due to different source sectors in Stockholm, Gothenburg and Umea, Sweden. Int. J. Environ. Res. Public. Health 14, 11–14. https://doi.org/10.3390/ijerph14070742

Šimić, I., Lovrić, M., Godec, R., Kröll, M., Bešlić, I. (2020). Applying machine learning methods to better understand, model and estimate mass concentrations of traffic-related pollutants at a typical street canyon. Environ. Pollut. 263, 114587. https://doi.org/10.1016/j.envpol.2020.114587

Singh, V., Singh, S., Biswal, A., Kesarkar, A.P., Mor, S., Ravindra, K. (2020). Diurnal and temporal changes in air pollution during COVID-19 strict lockdown over different regions of India. Environ. Pollut. 266, 115368. https://doi.org/10.1016/j.envpol.2020.115368

Steinfeld, J.I. (1998). Atmospheric chemistry and physics: from air pollution to climate change. Environ. Sci. Policy Sustain. Dev. 40, 26–26. https://doi.org/10.1080/00139157.1999.10544295

Stipaničev, D., Repec, S., Vucić, M., Lovrić, M., Klobučar, G. (2022). COVID-19 lockdowns - effect on concentration of pharmaceuticals and illicit drugs in two major Croatian rivers. Toxics 10, 241. https://doi.org/10.3390/TOXICS10050241

Sun, J., Gong, J., Zhou, J. (2021). Estimating hourly $PM_{2.5}$ concentrations in Beijing with satellite aerosol optical depth and a random forest approach. Sci. Total Environ. 762, 144502. https://doi.org/10.1016/j.scitotenv.2020.144502

Taylor, S.J., Letham, B. (2018). Forecasting at scale. Am. Stat. 72, 37–45. https://doi.org/10.1080/00031305.2017.1380080

Tobías, A., Rivas, I., Reche, C., Alastuey, A., Rodríguez, S., Fernández-Camacho, R., Sánchez de la Campa, A.M., de la Rosa, J., Sunyer, J., Querol, X. (2018). Short-term effects of ultrafine particles on daily mortality by primary vehicle exhaust versus secondary origin in three Spanish cities. Environ. Int. 111, 144–151. https://doi.org/10.1016/j.envint.2017.11.015

Vela, D., Sharp, A., Zhang, R., Nguyen, T., Hoang, A., Pianykh, O.S. (2022). Temporal quality degradation in AI models. Sci. Rep. 12, 11654. https://doi.org/10.1038/s41598-022-15245-z

Wang, B., Yuan, Q., Yang, Q., Zhu, L., Li, T., Zhang, L. (2021). Estimate hourly $PM_{2.5}$ concentrations from Himawari-8 TOA reflectance directly using geo-intelligent long short-term memory network. Environ. Pollut. 271, 116327. https://doi.org/10.1016/j.envpol.2020.116327

Yan, X., Zang, Z., Jiang, Y., Shi, W., Guo, Y., Li, D., Zhao, C., Husi, L. (2021). A Spatial-Temporal Interpretable Deep Learning Model for improving interpretability and predictive accuracy of satellite-based $PM_{2.5}$. Environ. Pollut. 273, 116459. https://doi.org/10.1016/j.envpol.2021.116459

Yang, M., Guo, Y.M., Bloom, M.S., Dharmagee, S.C., Morawska, L., Heinrich, J., Jalaludin, B., Markevychd, I., Knibbsf, L.D., Lin, S., Hung Lan, S., Jalava, P., Komppula, M., Roponen, M., Hirvonen, M.R., Guan, Q.H., Liang, Z.M., Yu, H.Y., Hu, L.W., Yang, B.Y., *et al.* (2020). Is PM_1 similar to $PM_{2.5}$? A new insight into the association of PM_1 and $PM_{2.5}$ with children's lung function. Environ. Int. 145, 106092. https://doi.org/10.1016/j.envint.2020.106092

Zamani Joharestani, M., Cao, C., Ni, X., Bashir, B., Talebiesfandarani, S. (2019). $PM_{2.5}$ prediction based on random forest, XGBoost, and deep learning using multisource remote sensing data. Atmosphere 10, 373. https://doi.org/10.3390/atmos10070373

Zhang, T., Geng, G., Liu, Y., Chang, H.H. (2020). Application of bayesian additive regression trees for estimating daily concentrations of $PM_{2.5}$ components. Atmosphere 11, 1233. https://doi.org/10.3390/atmos11111233

Zhou, H., Zhang, F., Du, Z., Liu, R. (2021). Forecasting $PM_{2.5}$ using hybrid graph convolution-based model considering dynamic wind-field to offer the benefit of spatial interpretability. Environ. Pollut. 273, 116473. https://doi.org/10.1016/j.envpol.2021.116473