


Managing air quality: Predicting exceedances of legal limits for PM10 and O₃ concentration using machine learning methods

Maryna Krylova^{1,2,3} | Yarema Okhrin⁴ 

¹Research Center Finance and Information Management, Technical University of Munich, Munich, Germany

²Research Center Finance and Information Management, University of Augsburg, Augsburg, Germany

³Research Center Finance and Information Management, University of Bayreuth, Bayreuth, Germany

⁴Department of Statistics, Faculty of Business and Economics, University of Augsburg, Augsburg, Germany

Correspondence

Yarema Okhrin, Department of Statistics, Faculty of Business and Economics, University of Augsburg, Augsburg, Germany
Email: yarema.okhrin@uni-a.de

Abstract

Air pollution imposes great costs on productivity, safety and health of individuals and dictates necessity of a proactive air pollution management. This, in turn, requires powerful tools for air quality modeling. In this article we develop a two-stage procedure for predicting exceedances of the EU legal limits for PM10 and O₃ concentrations using hourly data. Within the first stage we deploy machine learning methods to produce accurate 24-h-ahead forecasts of hourly pollutant concentrations at seven specific locations in the cities of Augsburg and Munich, Germany. The best performance was shown by the Stochastic Gradient Boosting Model—an ensemble tree-based method, especially convenient because of its computational efficiency and robustness to overfitting. Its predictive ability was largely superior to that reported by similar studies. In the second stage, the hourly forecasts were used to predict the exceedances of the EU daily limits for PM10 and O₃ concentrations. For both pollutants we could achieve the average probability of exceedances detection above 80%, while keeping the probability of false alarms at a reasonably low level. Such satisfactory results show that our approach can be successfully applied to anticipate the shocks, which would allow authorities to manage them in the most effective manner.

KEYWORDS

air pollution, air quality forecasting, ozone, particulate matter, stochastic gradient tree boosting

1 | INTRODUCTION

The effects of air pollution on human health have been the focus of many epidemiological studies since the early 1990s. An abundance of evidence has been gathered in support of theory that the long-term exposure to atmospheric pollutants, even in relatively small concentrations, substantially increases the risk of many illnesses including respiratory and cardiovascular diseases as well as different types of cancer (Ayres, 2006; Künzli et al., 2000; Loomis et al., 2013). It has also been proven that a statistically significant link exists between the daily mortality rates and the pollutant concentrations observed during the same or the previous day (see, for instance, Gryparis et al., 2004; Pope & Dockery, 2006). In the recent time the importance of the machine learning tools for modeling and monitoring environmental data was recognized by the scientific community (see Represa et al. (2020)).

Among the variety of known pollutants, the particulate matter (PM), tropospheric ozone (O₃), nitrogen oxides (NO_x), sulfur dioxide (SO₂), carbon monoxide (CO) and lead (Pb) classify as the most dangerous to human health and, therefore,

are seen as the main components determining the air quality in a specific area. The regions with higher concentrations of these pollutants report significantly higher rates of morbidity and mortality, especially among the population with existing preconditions, infants and children, as well as older age groups (Dockery et al., 1993; Pope & Dockery, 2006).

But the economic and societal costs of air pollution go far beyond the negative health effects. More recent studies, in particular, have shown that an exposure to air pollution may also induce behavioral changes and impair cognitive performance. For example, Ebenstein et al. (2016), Zivin and Neidell (2012) and Chang et al. (2016) found that individuals exposed to higher levels of pollution performed worse in their studies or at workplace. These effects can likely be attributed to positive associations that were shown to exist between the higher long-term levels of pollution and the increased reaction time, shorter attention span and lower concentration, decreased perceptual function, and worsened short-term memory of individuals in the affected areas (Chen & Schwartz, 2009). Murphy et al. (2013) and Li et al. (2017), on the other hand, link pollution exposure to the increased levels of serotonin as well as stress hormones—cortisol, cortisone, and epinephrine. This may result in the increased impatience, aggressive or unethical behavior (Lu et al., 2018) as well as higher willingness to take risks (Murphy et al., 2013). All these factors combined help to explain how air pollution may also be responsible for the elevated levels of criminal activity in society (Lu et al., 2018) and the increased number of traffic accidents (Sager, 2019)—areas that were long overlooked by the traditional research on air quality.

Most countries attempt to tackle the pollution problem by defining a set of air quality standards, which impose restrictions on the allowed yearly, daily and hourly mean pollutant concentrations in certain area. In Europe, limits on allowed pollutant concentrations are set by the European Parliament and the Council (EC) and are legally binding for all member states of the European Union (EU). However, given the highly volatile nature of air pollutants and the variety of pollution sources, complying with the legal requirements 100% of the time is not an easy task.

To be able to do so, the authorities need reliable forecasting models that would allow them to anticipate unfavorable situations some time in advance and take preventive action. Such actions may be of various nature. In some cases, local authorities may possess the instruments to shut down plants or regulate the daily traffic volumes, for example, by prohibiting certain types of vehicles to circulate on the road for one day. In cases where residential heating is primarily responsible for peaks in pollution levels, restrictions on heating fuels could be issued for the coming night. Finally, even in the absence of any such instruments, a warning conferred to the public may suffice to encourage people to stay indoors, thus avoiding an exposure to unhealthy air.

Many air pollution models have already been developed and tested in the past. The majority of authors have been focusing on forecasting some kind of daily measure (see, e.g., Chaloulakou et al., 2003; Corani, 2005; McKendry, 2002; Perez & Reyes, 2006) with very few trying to predict the hourly concentrations. This is not surprising, since daily averages are generally less difficult to forecast due to a much smoother behavior. Hourly pollutant concentrations, on the other hand, exhibit substantial variation throughout the day, with peak concentrations being significantly higher than the daily average. These peak concentrations, however, are what we are most interested in predicting since they constitute the real danger to human productivity and health. The hourly forecasts provide the decision-maker with much more detailed information and allow for better planning of the preventive actions. Hence, we consider them to be of a much bigger value.

It is also clear that for hourly forecasts to be of any use, they must be made available at least some time in advance. Yet, among the papers with a focus on hourly instead of daily concentrations, the vast majority have been dealing with the so-called “nowcasting”—that is, predicting concentrations for the coming hour only (see, i.e., Aldrin & Haff, 2005; Arhami et al., 2013; Goulier et al., 2020; Ortiz-García et al., 2010; Schlink et al., 2003). Indeed, we could identify only a few studies that extend their forecasting horizon beyond the next couple of hours (in particular, Ballester et al., 2002; Cai et al., 2009; Fernando et al., 2012; Grivas & Chaloulakou, 2006; Hrust et al., 2009; Paschalidou et al., 2011; Peng et al., 2017).

The type of the models used for forecasting evolved dramatically over the recent years. Earlier the focus lied mainly on classical statistical and econometric tools, such as regression and time series models. These models are flexible enough to capture seasonality and the impact of exogenous variables, but still suffer from the strict functional relationships. In the majority of the cases the impact of regressors is restricted to a linear one. The recent popularity of machine learning tools gave rise to an increasing number of studies deploying such methods to environmental data. As pointed out in an excellent review by Represa et al. (2020) “the data mining paradigm can assist in the study of air quality by providing a structured work methodology that simplifies data analysis.” One stride of the paper attempts to predict the concentration directly. Zhou et al. (2020a), Zhou et al. (2020b), and Valput et al. (2020) rely on artificial neural networks to predict PM_{2.5} concentration whereas Masmoudi et al. (2020) use ensembles of regressor chains-guided feature ranking for predicting several pollutants simultaneously. The other stride of papers attempts to predict the exceedances directly. However, the number of results on limit exceedances is still very limited as noted by Gómez-Losada (2018). The few existing papers attempt to predict the event of exceedance using various machine learning classifiers. Some, such as Yang

et al. (2020) rely on more classical discriminant methods whereas others, for example, Gong and Ordieres-Meré (2016), Gómez-Losada (2018), Yang et al. (2020) use ensemble models such as random forests.

The contributions of this paper are methodological in nature and can be summarized in the light of the above discussion as follows. Generally, we develop a model capable of generating accurate predictions of hourly PM₁₀ (particles with an aerodynamic diameter of 10 μm or smaller) and O₃ concentrations 24 h in advance. Additionally, we wanted to be able to predict if the EU limits for both pollutant concentrations on the next day were going to be exceeded. A two-stage model has been developed to satisfy both of our goals. The model takes hourly observations and provides the probability of the limit exceedance at the output. In the first stage, the hourly pollutant concentrations for the next 24 h are predicted using the past air quality measurements as well as the same-period meteorological and temporal data. For this purpose we consider support vector regression, random forest, extremely randomized trees and stochastic gradient tree boosting. The last three approaches are ensemble-type methods and they appear to be the leading machine learning techniques for forecasting continuous targets. Thus we go beyond the classical modeling using linear regression or seasonal time series models. In the second stage, these hourly forecasts are used to predict the incidents of legal limit exceedances on the coming day using classification tools. The model was tested on a long data on several air quality stations in the cities of Augsburg and Munich, but it may also be applied to other locations with similar climate and pollution sources. Note that the predictive accuracy of the two-stage approach clearly dominates single-step models with temporal aggregation of data.

The rest of this work is organized as follows. Section 2 describes the study area and data used for analysis, the performed data preparation steps and the forecasting model. Section 3 presents the results and compares them with those obtained by similar studies for other regions. Section 5 concludes by summarizing the findings, pointing out limitations and providing outlook for the real-life implementations of the model.

2 | METHODOLOGY

2.1 | Data and study area

The air quality data were obtained from the website of the Bavarian State Office for Environment (Bayerisches Landesamt für Umwelt). It contains information on the mean hourly concentrations of PM₁₀ and ozone for the period between January 1 2008 and December 31 2018. Although, the analysis PM_{2.5} is potentially of greater interest, the data for this pollutant is not available for the whole time span at the needed frequency. The measurements were taken at several stations located in Augsburg and Munich. Both cities lie in the province of Bavaria, south-eastern Germany. We provide the descriptive statistics of these data in the Supplementary Material. We notice that stations that are more heavily affected by traffic emissions (*M/LandshuterAllee*, *M/Stachus*, *A/Karlstraße* and *A/Königsplatz*) experience higher PM₁₀ concentrations than the rest. Traffic also causes higher volatility of PM₁₀ concentrations. For ozone, however, the opposite seems to be the case. The lowest mean O₃ concentration is observed at *M/Stachus* and can probably be explained by a high nitrogen monoxide (NO) concentration at this station. During the night hours NO reacts with ozone, which results in ozone depletion as O₃ gets decomposed into nitrogen dioxide (NO₂) and oxygen (O₂) (Sharma et al., 2016). Consequently, the ozone concentrations are usually lower in busy urban centres and higher in suburban and rural areas, where not as much NO is emitted by traffic. Note, that the minimum concentration is exactly zero for the majority of the stations, implying technical minimum detection limits.

In addition to the air quality data, hourly measurements of 13 different meteorological variables were used as input to predictive model, including air (T_a) and soil temperature (T_s), relative humidity (RH), precipitation (Rf and Rf_bin), cloud cover (CC), visibility (v), sunshine duration (SD), atmospheric pressure (P), and horizontal wind speed (WS) (see also Table 1a for the description of each variable). The data were provided by the German Meteorological Service (Deutscher Wetterdienst, DWD). In case of Augsburg, the data are collected at the local airport which is the only meteorological station in Augsburg and is situated approx. Ten kilometer away from the city center and 9 km away from the nearest air quality monitoring station (*A/Karlstraße*). The farthest-located station is *A/LfU*, which lies about 14 km away. In Munich DWD operates several stations located in the different areas of the city. For our analysis we used the data collected at the *München-Stadt* station as it is the nearest one to all air quality stations. The distance from this station to the four air quality control stations in Munich varies from 1.5 km (*M/Lothstraße*) to ca. Ten kilometre (*M/Johanneskirchen*). Obviously, the exceedances do depend on the intensity of the traffic and the models we use would clearly profit from the traffic-related variables. This data is unfortunately not available at the needed frequency and at the given locations.

TABLE 1 Additional variables used as input to the model

Name	Description	Measurement units
(a) Meteorological variables		
CC	Total cloud cover	1/8
Rf_bin	Indicator variable for precipitation	binary (0 or 1)
Rf	Hourly precipitation height	mm
T _a	Air temperature 2 m above the ground	°C
ΔT _a	Amplitude of daily air temperature	°C
T _s	Soil temperature in 5 cm depth	°C
ΔT _s	Amplitude of daily soil temperature	°C
RH	Relative humidity 2 m above the ground	%
SD	Hourly sunshine duration	min
P	Mean sea level pressure	hPA
V	Average visibility	m
WS	Average wind speed	m/s
Sin_WD	$\sin(2\pi \cdot \text{WD}/360)$	-1 to 1
CoS_WD	$\cos(2\pi \cdot \text{WD}/360)$	-1 to 1
Variable name	Description	Measurement units
(b) Temporal and persistence variables		
t	Observation number	1 to 95,904
HoD	Hour of the day	0 to 23
DoW	Day of the week	1 to 7
MoY	Month of the year	1 to 12
PM10 (t-1), ..., PM10 (t-24)	Past PM10 concentrations (site-specific)	μg/m ³
O ₃ (t-1), ..., O ₃ (t-24)	Past O ₃ concentrations (site-specific)	μg/m ³

Trigonometric transformations were applied to the original wind direction variable (WD) measured in degrees, resulting in two additional variables: Sin_WD and CoS_WD. ΔT_a and ΔT_s variables, representing the magnitude of daily temperature fluctuations, were furthermore added to the model.

To capture the trend and various seasonalities present in the data, several temporal variables were introduced to the model (see Table 1b). First of all, a variable t was created, which stands for observation number and is supposed to be helpful in capturing the overall historical time trend (decreasing for PM10, increasing for ozone). To account for seasonal effects, we added the categorical month-of-the-year (MoY) variable, which marks the month in which the measurement was taken. The DoW variable was created to capture the special effects associated with the day of the week, whereas the hour-of-the-day variable (HoD) was introduced to account for daily cycles in pollutant concentrations.

Finally, since the pollutant data show high degree of autocorrelation, predictive models are expected to benefit from the inclusion of persistence information (Chaloulakou et al., 2003; Grivas & Chaloulakou, 2006; Peng et al., 2017). Therefore, for each pollutant and each station a set of variables representing the past measurements of the pollutant concentrations (with time lag varying between 1 and 24 h) was further added to the model.

2.2 | Stage 1: Forecasting hourly pollutant concentrations 24 h in advance

In the first stage, the predictions are made iteratively for each next hour—that is, each predicted value is also used as input in predictions for all of the following hours. For example, when predicting the value for $t + 3$, values predicted by

the model for $t + 1$ and $t + 2$ are used as an input together with the actually observed past values: $PM_{10}(t)$ to $PM_{10}(t - 21)$. Note that we always use the last 24 observations (or prediction) to forecast the concentration at the end of the next hour. Such strategy has two main benefits. First, it results in much smoother estimates of hourly concentrations and, second, it can substantially increase the accuracy of predictions for the later hours, for which no actual information on the preceding observations is available.

In order to select the best model for our purpose, we, first, focused on the predictions of the next hour only and tested a series of different machine learning techniques with respect to their ability to predict the next-hour pollutant concentrations at each station. Since the relationship between the independent and dependent variables would be different depending on location and the chemical or physical properties of the respective pollutant, a separate model is required for each station-pollutant combination, which in our case yielded 11 models in total. The following machine learning methods were considered in particular, whereas the last three methods are ensemble-type approaches that rely on trees as weak learners.

- **Support vector machine/regression (SVM/SVR)**: a generalization of the maximal margin classifier which tries to separate the data points belonging to different classes by constructing a set of hyperplanes in a high-dimensional space (James et al., 2013). More precisely, the SVM classifier attempts to maximize the distance from the separating hyperplane and the nearest points in every class. Obviously, the classes in practice are not perfectly separable and one allows for soft margins that penalize points assigned to wrong classes. The next step of generalization is to consider hypersurfaces instead of hyperplanes, that is, we separate the classes by nonlinear functions such as, for example, Radial Basis Function (RBF). In the current study we use the SVM not for classification but for prediction of concentration. In order to do so use a variation of this method—the so-called support vector regression with ϵ -insensitive loss function and l_2 regularization—as implemented by Chang and Lin (2011) based on the general description of the method provided in Vapnik (1998). The RBF was chosen as a kernel due to its ability to capture highly nonlinear patterns. The mathematical formulation of it is given by: $K(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\gamma \|\mathbf{x}_i - \mathbf{x}_j\|^2)$, where $\mathbf{x}_i, \mathbf{x}_j \in R^m$ are the feature vectors of any two training samples i and j , with m standing for the total number of features. The γ parameter determines how much influence a single training sample has on the solution and is typically determined via grid search.
- **Random forest (RF)** (Breiman, 2001; James et al., 2013): an ensemble method that utilizes decision trees as base learners. Each tree consists of a set of splitting rules used to partition the space of all predictor variable values into disjoint regions $R_j, j = 1, 2, \dots, J$, represented by the terminal nodes of the tree. The basic idea is to predict the target variables by a single constant γ_j for the whole region R_j . The regions are typically chosen by minimizing the total MSE with an optimal regularizing penalty. The splitting is done using a top-down, greedy approach, known as recursive binary splitting, which aims at maximal possible variance reduction inside individual regions. A constant γ_j is then assigned to each terminal node, calculated as the mean of the actual target values for all training observations in R_j , and the predictions are made by the rule: if $\mathbf{x}_i \in R_j \Rightarrow \hat{y}_i = \gamma_j$.
A single tree is, however, not flexible enough. In RF multiple trees are grown and, in the end, the ensemble predictions are obtained through averaging of the predictions produced by the individual base learners. Important characteristic of this method is that each individual tree is built from a subsample drawn with replacement from the original training set (a procedure known as bagging and aimed at decreasing the model variance at the expense of a slight increase in its bias). We set the size of this bootstrap sample to only half of the training set size. Additionally, we limited the number of features that are considered at each node of a tree when looking for the best split to \sqrt{m} , where m is the total number of input features. This trick introduces additional randomness into each tree, making them less correlated and, therefore, further reducing the danger of overfitting (James et al., 2013).
- **Extremely randomized trees (ERT)** (Geurts et al., 2006): a meta-learner, similar to random forest, that fits a number of randomized trees (also called extra-trees) to different subsamples of data. The difference to random forest lies in the fact that at each tree node not only the candidate features are drawn at random from the total number of features, but also the thresholds for data splitting are determined completely randomly for each selected feature. The best of these randomly-generated thresholds is finally used as the splitting rule for the node.
- **Stochastic gradient tree boosting (SGTB)** (Friedman, 2001; Friedman, 2002; Hastie et al., 2009): an implementation of a gradient boosting algorithm with trees used as weak learners. This approach solves the key disadvantage of the RF, that the features and the observations are chosen randomly for each tree and independently of the previous ones. The information from a tree with a good predictive power is not used to improve the next trees. SGTB is an extension of boosting and overcomes this problem. The trees are grown sequentially and not in parallel, like in RF, and the predictions are calculated as the sum of the predictions of individual base learners: $\hat{y} = F_B(\mathbf{x}) = \sum_{b=1}^B h_b(\mathbf{x})$, where B is

the total number of trees, and $h_b(\mathbf{x})$ – the value predicted by a tree b . Since this method showed the best performance in the application we provide some technical details on the implementation. A generalized SGTB algorithm consists of the following steps:

1. Initialize the model. $F_0(\mathbf{x})$ is typically given by a single terminal node tree that always predicts some constant (since we use the least absolute deviation as the loss function, our initial model is given by the median value of \mathbf{y}).
2. Calculate the value of the loss function:

$$L_0 = \sum_{i=1}^n L(y_i, F_0(\mathbf{x}_i)) \quad (1)$$

3. For $b = 1, 2, \dots, B$, repeat:

- (a) Fit a tree h_b to the training data so as to minimize the total sum of losses, given the model from the previous step:

$$\mathbf{h}_b = \operatorname{argmin}_h L_b = \operatorname{argmin}_h \sum_{i=1}^n L(y_i, F_{b-1}(\mathbf{x}_i) + h(\mathbf{x}_i)). \quad (2)$$

A first-order Taylor approximation is applied to approximate the value of L :

$$L(y_i, F_{b-1}(\mathbf{x}_i) + h_b(\mathbf{x}_i)) \approx L(y_i, F_{b-1}(\mathbf{x}_i)) + h_b(\mathbf{x}_i)g_i, \quad (3)$$

where g_i is the derivative of the loss with respect to its second parameter, evaluated at $F_{b-1}(\mathbf{x})$ for the sample i :

$$g_i = \left[\frac{\delta L(y_i, F(\mathbf{x}_i))}{\delta F(\mathbf{x}_i)} \right]_{F=F_{b-1}}. \quad (4)$$

For each sample in the training set, the negative gradient, $-g_i$, marks the direction of the most rapid possible decrease in the value of the loss function. Therefore, the most obvious way to reduce the total loss is by fitting \mathbf{h}_b to predict a value that is proportional to the negative gradients of the samples. This strategy is called the steepest descent.

- (b) Update the model:

$$F_b(\mathbf{x}) = F_{b-1}(\mathbf{x}) + h_b(\mathbf{x}). \quad (5)$$

4. Output the boosted model, $F_B(\mathbf{x})$.

The fashion in which the model is built forces each subsequent tree to focus more on samples that are not explained well by the current model. Larger number of estimators, therefore, results in better accuracy of predictions—at least for the training set. To avoid the danger of eventual overfitting on the test set, we apply regularization in form of shrinkage and subsampling. Shrinkage scales the contribution of each individual tree to the model by a factor $0 < \nu < 1$, also called the learning rate, thus slowing down the model training process: $F_b(\mathbf{x}) = F_{b-1}(\mathbf{x}) + \nu \cdot h_b(\mathbf{x})$. Subsampling, on the other hand, combines boosting with the bootstrapping practices typical for forest-like methods. We make sure that each tree is trained on a subsample of the size $n/2$, drawn from the original training set without replacement, and set the maximum number of features to be considered at each split to \sqrt{m} , same as in RF and ERT models.

For comparison purposes we also considered more classical predictive models such as the classical linear regression with its extensions LASSO and ridge. Although, LASSO and ridge are not explicitly likelihood methods, they are regularized extensions of the linear regression. The performance of the linear regression and of ridge/lasso was very weak, so that we decided to exclude them from further analysis. The environmental data exhibits also a strong seasonal pattern and temporal dependence that is typically modeled by an appropriate (seasonal) time series model, for example, SARFIMA. The models applied here are not time series models and do not mimic the memory in the data directly. We include, however, the lagged values of the variables as regressors and, therefore, model the temporal dependence indirectly using nonlinear machine learning tools.

TABLE 2 Cross-validation scores for the 1-h-ahead forecasts of PM10 and O₃ concentrations produced by the four models, averaged across all stations

Method	Best parameters selected by cross-validation ^a	R ²
(a) PM10		
SVM	Kernel = RBF $\epsilon = 0.1, C = 0.1$	0.777
RF ^b	Number of estimators = 750 Maximum tree depth = 32	0.818
ERT ^b	Number of estimators = 1000 Number of estimators = 1500	0.791
SGTB ^b	Learning rate = 0.025 Loss function = $ y - \hat{y} $ (least absolute deviation)	0.829
Method	Best parameters selected by cross-validation ^a	R ²
(b) Ozone		
SVM	Kernel = RBF $\epsilon = 0.1, C = 0.1$	0.941
RF ^b	Number of estimators = 500 Maximum tree depth = 32	0.939
ERT ^b	Number of estimators = 750 Number of estimators = 1500	0.928
SGTB ^b	Learning rate = 0.05 Loss function = $ y - \hat{y} $ (least absolute deviation)	0.945

^a Whenever the difference in R² between 2 or more models was < 0.001, the simplest model was chosen.

^b Fraction of the training set used to grow each tree = 0.5. Maximum number of features to be considered at each split = \sqrt{m} , where m stands for the total number of features.

To compare the performance of different methods on training data we apply the 9-fold cross-validation. However, this time the model tuning, or parametrization, was performed in parallel. The best parameter combination for each method was determined as the one resulting in the highest mean cross-validation score—RMSE—from Section 2.4.1 obtained for a given pollutant and method across all years and all stations. The use of cross-validation procedure at this point ensures that our results are not dependent on the specific conditions observed in any given year, thus yielding much more reliable estimates of model performance. The final estimates, corresponding to the best scores achieved by each of the tested methods, are presented in Table 2. The best performance—both for ozone and for PM10—was shown by the SGTB model. Subsequently, this model was retrained using all 9 years of training data and used to produce 24-h-ahead forecasts of hourly PM10 and ozone concentrations—for the training (2008–2016) as well as for the testing (2016–2018) period see also Section 2.2).

For each pollutant and each monitoring station we generate not one, but a total of 24 sets of such forecasts—one at each hour of the day. The first set of forecasts includes those initiated at 12 a.m. for the time period between 1 a.m. of the same day and 12 a.m. of the following day. The second set consists of forecasts made at 1 a.m. each day for hours between 2–1 a.m.^{+1 day}, and so on. We do this, since we expect that forecasts initiated at different hours throughout the day, for example, at nighttime, will follow slightly different patterns. Having 24 forecasts instead of just one will, furthermore, allow us to compute more representative estimates of model performance.

2.3 | Stage 2: Forecasting daily limit exceedances

In stage 2, the hourly forecasts produced in the stage 1 are used to identify episodes of high exposure. At first, hourly values are aggregated into respective daily measures, following the guidelines provided in the Directive 2008/50/EC “on ambient air quality and cleaner air for Europe” (EC, European Council (EC), 2008). For PM10 such measure is

computed as a simple 24-h average of hourly concentrations, for ozone—as a maximum of the 24 8-h moving averages computed throughout the day. Note, that we always calculate the daily averages for the same 24-h periods for which the stage-1 predictions were created. For example, if hourly forecasts were generated at 8 p.m. for the period between 9 p.m. of the same day and 8 p.m. of the following day, then we would also use the same time windows to calculate the daily means. Next, obtained values for each day are compared to limit values set by the Directive, which are $50 \mu\text{g}/\text{m}^3$ for PM10 and $120 \mu\text{g}/\text{m}^3$ for ozone, to identify the days with norm exceedances (the direct approach for exceedances detection).

Unfortunately, most statistical models tend to systematically underpredict the events of extremely high pollutant concentrations (McKendry, 2002). One way to improve the sensitivity of exceedances detection is to lower the threshold value. For example, setting the daily limit to $40 \mu\text{g}/\text{m}^3$ for PM10 and to $110 \mu\text{g}/\text{m}^3$ for ozone would allow us to detect a larger fraction of the observed high-concentration episodes. We shift the focus of our predictions and prioritize the true positives over the true negatives. We decide that in the context of our use case—producing early warnings to the population and the responsible decision makers—the cost of false alarm is much smaller than that of an undetected exceedance. We were also not the first ones to suggest this approach of using lower threshold and follow here the approach of Corani (2005).

Another possibility to achieve better results in predicting incidents of standard violations is to build a second model on top of the first one, that would take daily measures calculated from the stage-1 hourly forecasts as an input and return probabilities of limit exceedance as its output. We have tried using logistic regression model for this purpose.

Logistic regression is a linear model for classification where the probability of an observation belonging to a certain class is modeled using a logistic function. We used a scikit-learn python implementation with “liblinear” solver, which employs a coordinate descent algorithm to solve the optimization problem (Fan et al., 2008; Yu et al., 2011). In order to prevent overfitting, a regularization procedure with l_2 penalty as in ridge and the regularization parameter C chosen by the 9-fold cross validation was implemented. More precisely, we minimize the objective function as defined below

$$\sum_{i=1}^n (v_0 y_i \log \hat{y}_i + v_1 (1 - y_i) \log(1 - \hat{y}_i)) + C \|\mathbf{w}\|^2,$$

where $\hat{y}_i = 1/(1 + e^{-w \cdot x_i})$. Additionally, we set the class weights v_0 and v_1 to be inversely proportional to the observed class frequencies in the input data.

Such approach is recommended for the unbalanced data sets such as ours. Since only a few exceedances of daily limit are observed for each pollutant in each year, the model would normally tend to always predict the nonexceedance. By adjusting the class weights, we can avoid this situation and push the model to try harder to predict the other class as well.

We fit the logistic regression model on daily averages calculated from hourly forecasts of pollutant concentrations in the training period as defined in Section 2.2. Subsequently, the model was applied to make predictions of exceedances in train sample. Using the information about the actually observed exceedances, we constructed the model’s precision-recall curve and used it to find an optimal cutoff probability. Two different approaches were developed for this task:

1. Simply choose such cutoff that maximizes the recall while, at the same time, not allowing the model’s precision to fall below a certain threshold (e.g., 50%, 60%, or 70%).
2. Maximize the F-measure (Hripcsak & Rothschild, 2005) which represents a weighted average between the precision and recall and is universally considered a good metric when looking for the best trade-off between the two: $F_\beta = (1 + \beta^2) \cdot (\text{precision} \cdot \text{recall}) / (\beta^2 \cdot \text{precision} + \text{recall})$. Typically, β is chosen such that the recall is considered β times as important as the precision. We try β values of 2 and 3 since we consider the model’s ability to detect critical cases at least twice as important as its capacity to avoid false alarms. Section 2.4.2 contains more details on these concepts.

In the end, the model is used to predict the probabilities of exceedances in the test sample. The cutoff determined on the train set is then applied to the probabilities of the test set to turn them into final predictions. An algorithmic summary of the Stages I and II is given in the Supplementary Material as Algorithms 1 and 2.

2.4 | Model evaluation

2.4.1 | Evaluation of the hourly forecasts from stage 1

To evaluate the performance of the model in the stage 1, seven metrics were calculated. Additionally to R^2 and $RMSE$ we consider

$$MAE = \frac{1}{n} \sum_{i=1}^{n-1} |y_i - \hat{y}_i|, \quad (6)$$

$$MedAE = \text{median}(|y_1 - \hat{y}_1|, \dots, |y_n - \hat{y}_n|), \quad (7)$$

$$MBE = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i) = \bar{y} - \hat{y}, \quad (8)$$

$$\Delta\sigma = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2} - \sqrt{\frac{1}{n-1} \sum_{i=1}^n (\hat{y}_i - \hat{y})^2} \quad (9)$$

$$d_2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (|y_i - \bar{y}| + |\hat{y}_i - \bar{y}|)^2} \quad (10)$$

where y_i is the actual and \hat{y}_i —the predicted value of i th generic observation and $\hat{y} = \frac{1}{n} \sum_{i=1}^n \hat{y}_i$.

Since the first four measures neglect the direction of the deviance, we calculate the statistic—the mean bias error (MBE)—described by formula 8. To gain some insight on the distributional properties of both the actual and the predicted values, we, furthermore, compute the difference in their standard deviations ($\Delta\sigma$). Finally, we also introduce the index of agreement (d_2)—a measure first proposed by Willmott (1981). Whereas R^2 just reflects how good the model can explain the majority of observed values, d_2 pays somewhat more attention to outliers. If the model is good at predicting the middle-range values but fails to predict the peak concentrations, it will be penalized more by d_2 than by R^2 . Considering that in our case the correct prediction of high pollutant concentrations is much more important than that of moderate or low values, the index of agreement is probably the most important indicator of performance that we have. It is described by Equation 10 and can take values between 0 (worst) and 1 (best).

At first, the above-mentioned metrics were calculated separately for each station and for each of the 24 separate sets of predictions (generated, as you remember, one at each hour of the day). Subsequently, the results were averaged across all possible prediction starting points to produce a single set of performance indicators for each station. These averaged statistics, should provide a robust estimate of model's performance, independent of the specific time at which we choose to generate the forecasts.

2.4.2 | Evaluation of the exceedances predictions from stage 2

To evaluate the performance of the classification at stage 2, we compute the confusion matrix and a number of performance measures for each setting.

A confusion matrix consists of four terms: true positives (TP) which marks the number of correctly forecasted exceedances, true negatives (TN); correctly forecasted nonexceedances, false positives (FP); forecasted but not observed exceedances or false alarm (also called “Type I error”), and false negatives (FN); missed exceedances (observed but not forecasted, “Type II error”). A number of additional performance measures can be calculated from these terms. For example, the overall model accuracy (ACC) is defined as a proportion of instances (both positives and negatives) that were correctly classified by the model. The true positive rate (TPR), often referred to as probability of detection, sensitivity, recall, or power of the model, represents the fraction of correctly identified positives. Another important indicator of performance is the so-called positive predictive value (PPV) or precision of the model, calculated as a ratio of the predicted and observed exceedances to all the forecasted ones. All of the above-mentioned measures assume their best value at 1, while the worst possible value is 0. We also reintroduce the F-measure (14), first mentioned in Section 2.3.

$$ACC = \frac{TP + TN}{TP + FP + TN + FN} \quad (11)$$

$$\text{TPR} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (12)$$

$$\text{PPV} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (13)$$

$$F_\beta = \frac{(1 + \beta^2)(\text{PPV} \times \text{TPR})}{(\beta^2 \times \text{PPV} + \text{TPR})}, \quad \text{with } \beta = 2 \quad (14)$$

$$\text{TSS} = \frac{\text{TP} \times \text{TN} - \text{FP} \times \text{FN}}{(\text{TP} + \text{FN})(\text{TN} + \text{FP})} \quad (15)$$

The last measure that we compute is called the true skill statistic (TSS) and is given by Equation 15. Also known as the Hanssen–Kuipers discriminant or Kuipers' performance index, this metric makes additional adjustments for the correct forecasts that were made by chance. Perfect forecasts receive a score of +1, random (classification by chance) and constant (all observations to only one class) forecasts—a score of 0, and forecasts inferior to random forecasts receive a negative score. The higher is the probability of event occurrence, the lower is the contribution to the skill score resulting from its correct classification by the model. As the likelihood of the event decreases, the contribution of a correct prediction to the skill score, on the contrary, increases. The TSS measure, therefore, has some desirable characteristics for evaluating rare event forecasts. Together with the F-measure, it should play a key role in helping us to make a decision regarding the final choice of the model.

3 | RESULTS AND DISCUSSION

3.1 | Data preparation

The proportion of missing observations in the original air quality data ranged between 0.16% and 4.41% for different stations. The combined meteorological data of Augsburg and Munich contained ca. 4.33% of missing values. Since not all models can handle the presence of missing values in the input data, we have applied an iterative procedure for multivariate feature imputation to fill in the blanks. Such an imputer models each variable in turn as a function of all other variables. In that way, the information contained in those variables can be used to predict the missing values of the target variable. After all missing values have been imputed—that is, each variable has been predicted once—the whole cycle is repeated several more times—either until the maximum number of iterations is reached or until the relative changes in predictions after each round become smaller than some predefined threshold value.

In our case, a Bayesian ridge estimator—regularized linear regression technique (Tipping & Smola, 2001)—was used for modeling, with the maximum number of iterations set to 10.

After the missing values imputation, the data have been split into two nonoverlapping samples. The first 9 years of data (2008–2016) were classified as the training set and utilized for model selection and tuning, whereas the last 2 years of data (2017 & 2018) were reserved for testing. All data collected on the December 31 and January 1 of each year have been removed from the sample to eliminate the most extreme outliers (thanks to the German tradition of lighting fireworks on New Year's Eve, the PM10 values on these days are sometimes more than 70 times higher than the average observed values). Variation of the splitting data between training and test data had no visible impact on the results. Moreover, two years of test data is sufficient to capture annual trends in the data.

The final step of our data preparation process consisted of removing some of the redundant meteorological variables. Two examples of the redundant variables in our data are the air and the soil temperature. Both variables correlate quite strongly with the target variables (PM10 and ozone), but even more so with each other (Pearson correlation coefficient = 95%). This also applies to the ΔT_a and the ΔT_s variables which are the differences between the maxima and minima of daily air and soil temperatures, respectively.

A simple nonlinear technique—the k-nearest neighbours algorithm (kNN)—was used in order to compare the pairs of variables and choose the most relevant ones for analysis. We set k to 20 and used a 9-fold cross-validation (CV) to estimate the model's performance on training data (Stone, 1974). We opt for this specification, because the training data set we use covers exactly 9 years. We justify the use of CV for time series data by the fact that since we work with hourly data, the changes in the local intraday trends and autocorrelations are almost negligible if we switch from one year to another. Alternatively, one can apply expanding or moving window CV, but due the size of the data set this has little impact on the results, as the preliminary analysis has shown.

At first, the model was trained using the full set of meteorological variables to obtain a baseline estimate of its predictive power. Subsequently, the process was repeated 4 more times, each time excluding one of the temperature variables from the input. The results obtained with each set of variables suggest that for PM10 an exclusion of the air temperature variable leads to the largest decrease in R^2 , as compared to the full model, and the loss of ΔT_a variable has greater impact than the loss of ΔT_s . For ozone, on the other hand, the soil temperature holds the most predictive power, but ΔT_a still seems to be slightly more important than ΔT_s . Therefore, we used the T_a and ΔT_a variables when predicting the PM10 levels, while the T_s and ΔT_a variables were used as input for O_3 predictions.

3.2 | Forecasts of the mean hourly PM10 concentrations

Table 3 shows the averaged statistics, obtained on the 24-h-ahead forecasts of hourly PM10 concentrations. We see that our model was able to achieve an average R^2 score of 0.68 across all stations, which signals good predictive power. The best results were obtained for the *M/Lothstraße* station. Located in a residential area of Munich, this station is characterized by little traffic and low levels of PM10 pollution. It is also the closest station to the site where meteorological data for Munich is collected. The worst results were achieved, on the contrary, for the *M/LandshuterAllee* and *M/Stachus* stations. Both of these stations are located in the immediate vicinity of large city roads and are thus associated with the relatively high overall pollution levels and high amplitude of the observed daily and weekly fluctuations, caused primarily by traffic.

From Table 3 we further notice that the forecasted values show good agreement with the actually observed values (d_2 values ranging between 0.85 and 0.94 across individual stations), however, the standard deviation of the predicted values is lower than that of the actual ones, meaning that the model was not able to capture all of the variability in the original data. The obtained values for RMSE, MAE, and MedAE, furthermore, suggest that the model is very good at predicting the average values, but has difficulties predicting extreme concentrations. That is unfortunate, since extreme concentrations are the ones we are particularly interested in predicting correctly, but, at the same time, understandable since, essentially, all statistical models are aimed at an approximation of the average behavior and thus tend to ignore the outliers. Finally, when looking at MBE values, we notice that most of them do not exceed $1 \mu\text{g}/\text{m}^3$ in absolute terms. Only at two stations—*M/LandshuterAllee* and *M/Stachus*—does our model seem to systematically underpredict the observed concentrations.

Figure 1 illustrates how the accuracy of the produced forecasts is affected by choice of the forecasting period. Recall that hourly predictions are made iteratively for each next hour—that is, each predicted value is also used as input in predictions for all of the following hours. Nevertheless, as the number of actual past observations available to the model decreases, it becomes progressively more difficult to make predictions. As one can see from

TABLE 3 Performance statistics for the 24-h-ahead forecasts of PM10 hourly concentrations produced by the SGTB model

	R^2	RMSE	MAE	MedAE	MBE	$\Delta\sigma$	d_2
M/Johanneskirchen	0.66	7.95	4.94	3.20	0.03	2.60	0.89
M/LandshuterAllee	0.62	10.27	6.66	4.49	1.07	2.05	0.88
M/Lothstraße	0.77	6.19	4.03	2.61	0.52	0.76	0.94
M/Stachus	0.58	11.09	6.44	4.04	2.11	4.03	0.85
A/Karlstraße	0.73	8.47	5.58	3.78	0.08	2.33	0.92
A/Königsplatz	0.70	8.31	5.53	3.86	0.14	3.01	0.90
A/LfU	0.71	7.18	4.88	3.40	0.77	1.97	0.91
Average statistics	0.68	8.49	5.44	3.63	0.67	2.39	0.90

Note: The results presented in this table were computed as averages over the individual results obtained for each of the 24 separate 1-day-ahead forecasts produced at different hours throughout the day. This notion also applies to all of the following tables in this article, unless explicitly stated otherwise.

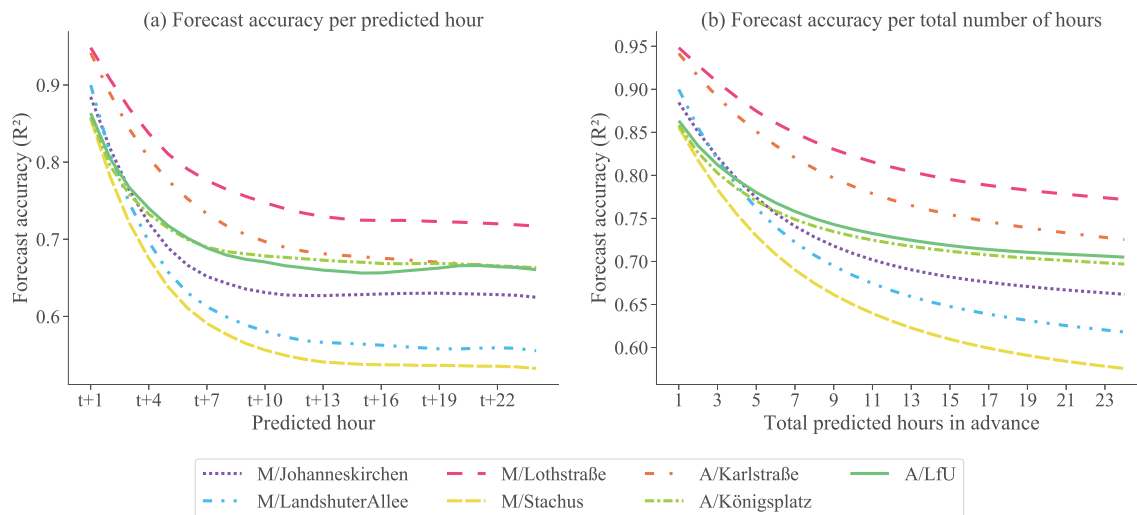


FIGURE 1 Accuracy of the produced PM₁₀ forecasts depending on choice of the forecasting period. The right figure shows the forecast accuracy for averaged over the accuracies up to the corresponding horizon

Figure 1a, the R² scores for the first predicted hour are very high, ranging between 0.86 and 0.95 for individual stations.

After the first hour, the accuracy of hourly predictions drops exponentially for each subsequent hour. Yet, after the first 7 to 10 h the curve becomes rather flat; so, theoretically, even if we did chose a longer forecasting period, the quality of the produced forecasts would not deteriorate that fast. Thanks to the effects of averaging, the accuracy of the aggregate forecasts (see Figure 1b), in any case, decreases more slowly than that of the individual hourly predictions (see Figure 1a).

Until now, we have only been looking at the aggregated results, averaged across all possible initiation points. Yet, the quality of the produced forecasts also depends on the specific time at which they were generated. As one can see from Figure 2, at most stations forecasts produced during the daylight hours tend to be more accurate than those generated in the late evening or night hours. An explanation lies in the fact that the air pollution data is highly autocorrelated, so that the last few actually observed values prior to the start of the forecasting period would always have an unproportionally high impact on the produced forecasts. Since PM₁₀ concentrations at night are typically much lower than during the day, the forecasts generated at night hours might underestimate the overall daily pollution levels. On the other hand, the forecasts produced during the evening rush hours may sometimes overestimate the daily pollution levels and lead to worse results in terms of accuracy.

3.3 | Forecasts of the daily limit exceedances (PM₁₀)

Table 4 presents the results obtained for two different thresholds when using a direct approach to predicting episodes of high exposure from the stage-1 hourly forecasts. 50 µg/m³ is the actual daily limit set by the EU. However, as we see from Table 4a, the probability of the exceedances detection with this threshold amounts to only 71% on average. At the same time, using a threshold of 40 µg/m³ allows us to detect about 89% of all exceedances, which is a very good result. However, such increase in sensitivity happens at a cost of a substantial decrease in precision. At some stations the fraction of predicted positives that correspond to actually observed exceedances is only slightly above 0.5—that is, almost half of the predicted episodes turned out to be a false alarm. That may sound like a lot of noise, yet, in absolute terms such false alarms constitute a very small number of occurrences—between 5 and 9 per year (remember, the test period is 2 years). Furthermore, in the context of our use case—producing early warnings to the population and the responsible decision makers—the cost of false alarm is much smaller than that of an undetected exceedance. Therefore, authorities will likely be willing to accept the higher probability of false alarm for the benefit of better health protection.

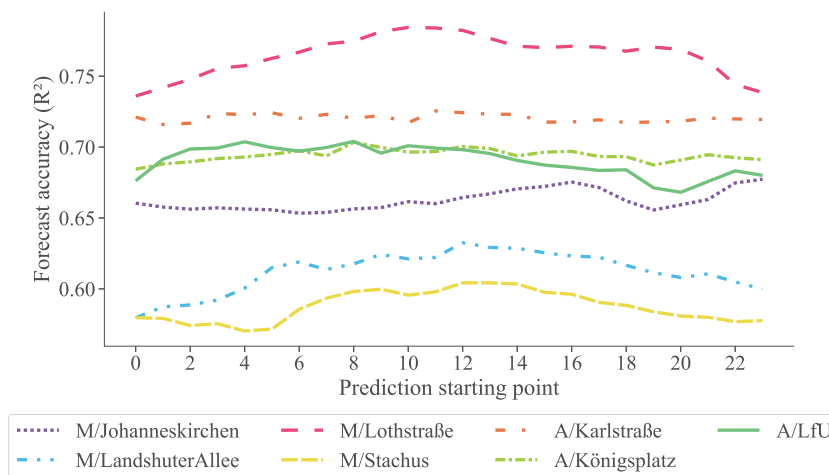


FIGURE 2 Forecast accuracy depending on prediction starting point (PM10)

The results produced by logistic regression are presented in Table 5. As you remember, we have tested two approaches to determine an optimal cutoff probability used to transform the class probabilities returned by a classifier into the binary class predictions. Using the first approach, we select such cutoff that maximizes the probability of exceedances detection (i.e., TPR) in the training set, while, at the same time, not allowing the model's precision (PPV) to fall below a certain threshold. The best results were obtained with a threshold of 0.5 and are reported in Table 5a. The PPV values now range between 0.59 and 0.71, meaning that the model produces fewer false alarms. But its sensitivity to actual exceedances is also slightly decreased, resulting in the average TPR of 0.87 and TSS of 0.85.

Quite similar results were obtained when using the logistic regression with cutoff selected by maximizing F-score with β value set to 2. More interesting are, therefore, the results achieved with $\beta = 3$ that are reported in Table 5b. The average probability of exceedances detection using this method equals 91% and the proportion of correct forecasts amounts to 56%. This method, furthermore, results in high values for F_β and the TSS and, therefore, seems to be the best choice for the task at hand (although, a more cautious decision-maker may, of course, choose even higher β to achieve greater sensitivity at the cost of somewhat lower precision).

The model shows a nonuniform performance at different stations. The worst results are obtained for *M/Stachus* station, which comes across naturally since already at stage 1 this station was among the hardest ones to make predictions for. The best scores on average are achieved using the data from *M/Johanneskirchen* and *A/Königsplatz* stations, followed by the *M/Lothstraße*, *A/Karlstraße* and *A/LfU* – which is, again, largely consistent with the results seen at stage 1. When looking at the total number of correctly predicted limit exceedances in relation to the time of the day when predictions were made (as depicted in Figure 3), we notice a substantial variability as well. In part, it is too caused by the differing quality of the stage-1 hourly forecasts since those are used as basis for exceedances predictions. So, for example, at the *M/Lothstraße* station both the hourly forecasts and the exceedances predictions generated during the day tend to be more accurate than the ones produced at night hours. Same is true for *M/LandshuterAllee* and *M/Stachus*, whereas at *A/LfU*, on the contrary, best results in both stages are achieved if predictions are generated at night.

The perceived variability may be, of course, used as guidance when selecting the best initiation point. Yet, practicability is even more important criterion if we want forecasts to be used in the proactive air pollution management. We must ensure that the responsible authorities have enough time to inform the population and, eventually, to implement the countermeasures. Based on these considerations, evening hours or the early morning seem to be the most appropriate time for the generation of forecasts. For example, Table 6 presents the performance statistics of forecasts created at 7 p.m. using logistic regression with a modified cutoff, selected by maximizing the $F_{\beta=3}$.

Finally, we would like to compare our findings to those reported by similar studies at other locations around the world. This is not an easy task since different authors use different target measures of PM10 concentrations and different thresholds, depending on properties of the analysed data and the existing country-specific air quality standards.

TABLE 4 Performance statistics for the forecasts of PM10 daily limit exceedances produced using the direct approach

	TN, FP, FN, TP	ACC	TPR	PPV	F_{β}	TSS
(a) Threshold value = 50 $\mu\text{g}/\text{m}^3$						
M/Johanneskirchen	(707, 2, 3, 11)	0.99	0.76	0.87	0.78	0.76
M/LandshuterAllee	(645, 6, 11, 25)	0.97	0.69	0.81	0.71	0.68
M/Lothstraße	(704, 3, 4, 13)	0.99	0.77	0.81	0.77	0.76
M/Stachus	(664, 6, 12, 21)	0.97	0.63	0.77	0.66	0.62
A/Karlstraße	(683, 4, 12, 19)	0.98	0.61	0.81	0.64	0.60
A/Königsplatz	(698, 4, 6, 17)	0.99	0.75	0.80	0.76	0.74
A/LfU	(705, 2, 4, 10)	0.99	0.74	0.83	0.75	0.74
Averaged statistics	–	0.98	0.71	0.81	0.72	0.70
	TN, FP, FN, TP	ACC	TPR	PPV	F_{β}	TSS
(b) Threshold value = 40 $\mu\text{g}/\text{m}^3$						
M/Johanneskirchen	(698, 11, 1, 13)	0.98	0.94	0.56	0.83	0.93
M/LandshuterAllee	(632, 18, 5, 31)	0.97	0.86	0.63	0.80	0.83
M/Lothstraße	(692, 16, 1, 16)	0.98	0.93	0.51	0.79	0.91
M/Stachus	(654, 16, 8, 25)	0.97	0.76	0.60	0.72	0.73
A/Karlstraße	(673, 14, 4, 27)	0.97	0.87	0.65	0.81	0.85
A/Königsplatz	(686, 16, 1, 22)	0.98	0.96	0.57	0.85	0.94
A/LfU	(698, 9, 1, 12)	0.99	0.92	0.58	0.82	0.91
Averaged statistics	–	0.98	0.89	0.59	0.80	0.87

Note: Because of the rounding effects as well as due to the fact that the days on which more than 25% of the hourly measurements were missing in the actual data are excluded from analysis, the total number of days in the confusion matrix may differ between stations and/or between tables. This notion also applies to all of the following tables in this section and the corresponding section for ozone (3.5), unless explicitly stated otherwise.

Moreover, different performance indicators are used to evaluate the models, which often makes a direct comparison impossible. Nevertheless, we did our best to collect information on each study on PM10 exceedances forecasting that was conducted in the last couple of decades and, for the purpose of better comparability, summarized the findings in terms of the performance measures used in the current study, whenever such translation was possible. You can find an overview of the results in Table 7.

In terms of TSS score, no other study was able to match our results. With respect to TPR and PPV values, however, Chaloulakou et al. (2003) outperform our model by far. The reason for such outstanding performance may lie in the fact that the mean PM10 concentration during the testing period of the respective study in Athens was, in fact, equal to 79 $\mu\text{g}/\text{m}^3$, which is higher than the applicable limit value for PM10. Such circumstance, of course, makes the prediction task much easier since, as we have already pointed out, statistical models are better at predicting concentrations around the mean. This could, at least partially, also explain the good results presented in the study of air quality in Milan, conducted by Corani (2005), where the average daily PM10 concentration corresponds to around 45 $\mu\text{g}/\text{m}^3$ and the threshold is set to 50 $\mu\text{g}/\text{m}^3$. Finally, we would also like to mention Perez and Reyes (2006), who used a linear model and an ANN to produce forecasts of PM10 limit exceedances in Santiago, Chile, and managed to achieve quite satisfactory results in terms of both TPR and PPV values.

3.4 | Forecasts of the mean hourly ozone concentrations

The performance statistics calculated from the 24-h-ahead ozone predictions for each station are summarized in Table 8. We notice that the accuracy of the produced forecasts is much higher for ozone than it was for PM10. The

TABLE 5 Performance statistics for the forecasts of PM10 daily limit exceedances produced by logistic regression

	TN, FP, FN, TP	ACC	TPR	PPV	F_β	TSS
(a) Minimum threshold value for precision = 0.5						
M/Johanneskirchen	(703, 6, 1, 13)	0.99	0.91	0.71	0.86	0.91
M/LandshuterAllee	(628, 23, 5, 32)	0.96	0.87	0.59	0.79	0.83
M/Lothstraße	(698, 10, 2, 16)	0.98	0.90	0.61	0.82	0.88
M/Stachus	(657, 14, 8, 24)	0.97	0.75	0.64	0.72	0.73
A/Karlstraße	(674, 13, 5, 25)	0.97	0.83	0.66	0.79	0.81
A/Königsplatz	(688, 15, 1, 22)	0.98	0.96	0.60	0.85	0.94
A/LfU	(702, 5, 2, 11)	0.99	0.84	0.70	0.80	0.83
Averaged statistics	–	0.98	0.87	0.64	0.81	0.85
	TN, FP, FN, TP	ACC	TPR	PPV	F_β	TSS
(b) $\beta = 3$						
M/Johanneskirchen	(700, 8, 1, 13)	0.99	0.94	0.60	0.85	0.93
M/LandshuterAllee	(616, 35, 3, 33)	0.94	0.91	0.50	0.78	0.85
M/Lothstraße	(694, 13, 1, 16)	0.98	0.92	0.56	0.81	0.91
M/Stachus	(653, 18, 8, 25)	0.96	0.77	0.59	0.72	0.74
A/Karlstraße	(661, 26, 2, 29)	0.96	0.94	0.53	0.81	0.90
A/Königsplatz	(684, 18, 1, 22)	0.97	0.97	0.55	0.84	0.95
A/LfU	(697, 10, 1, 13)	0.99	0.94	0.59	0.83	0.93
Averaged statistics	–	0.97	0.91	0.56	0.81	0.89

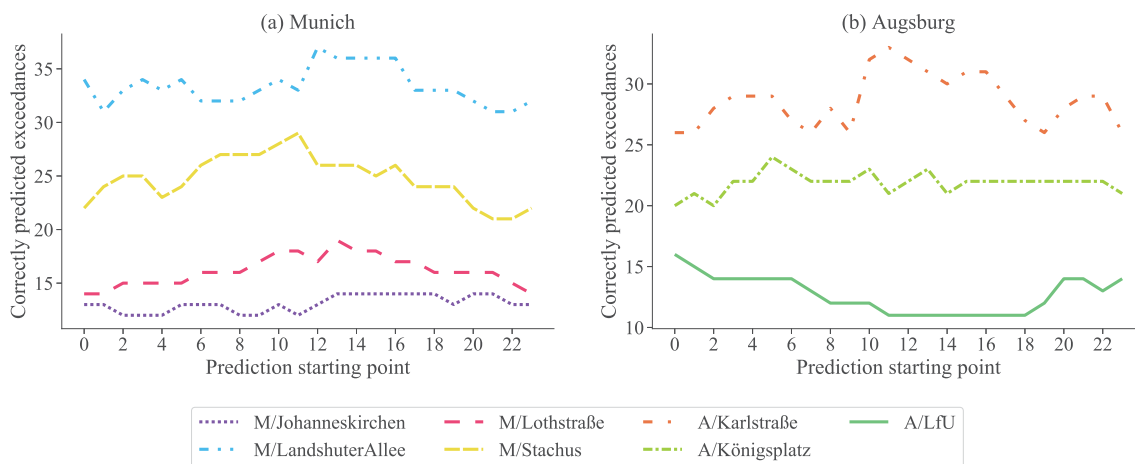


FIGURE 3 Total number of correctly predicted limit exceedances depending on forecast initiation point (PM10, logistic regression)

coefficient of determination reaches values between 0.83 and 0.87, and at all four locations the predicted values show good agreement with the observed ones, resulting in d_2 values around 0.96. Such results are consistent to the ones reported by Hrust et al. (2009) and are notably better than the results demonstrated by Ballester et al. (2002) and Peng et al. (2017).

The relationship between the RMSE, MAE, and MedAE values, once again, suggests that the model can predict average concentrations better than the extremes. Furthermore, the MBE values indicate that the model tends to overestimate the ozone levels at the *A/LfU* station and underestimate them at all three Munich stations. That is because the O_3 concentrations at these stations have been rising at a much higher pace than at *A/LfU* station, resulting in hourly concentrations

TABLE 6 Performance statistics for the next-day forecasts of PM10 daily limit exceedances produced at 7 p.m. using logistic regression with a modified cutoff probability

	TN, FP, FN, TP	ACC	TPR	PPV	F_{β}	TSS
M/Johanneskirchen	(700, 8, 0, 13)	0.99	1.00	0.62	0.89	0.99
M/LandshuterAllee	(618, 32, 4, 33)	0.95	0.89	0.51	0.77	0.84
M/Lothstraße	(690, 18, 1, 16)	0.97	0.94	0.47	0.78	0.92
M/Stachus	(657, 16, 5, 24)	0.97	0.83	0.60	0.77	0.80
A/Karlstraße	(670, 19, 3, 26)	0.97	0.90	0.58	0.81	0.87
A/Königsplatz	(686, 17, 0, 22)	0.98	1.00	0.56	0.87	0.98
A/LfU	(697, 9, 2, 12)	0.98	0.86	0.57	0.78	0.84
Averaged statistics	–	0.97	0.92	0.56	0.81	0.89

TABLE 7 Summary of the results from recent studies on PM10 limit exceedances forecasting in urban areas, using statistical models

Study ^a	Target measure	Threshold	TPR	PPV	TSS
Current study ^b	daily mean	50 $\mu\text{g}/\text{m}^3$	0.91	0.56	0.89
Zickus et al. (2002)	daily mean	50 $\mu\text{g}/\text{m}^3$	0.63	0.73	–
Chaloulakou et al. (2003)	daily mean	75 $\mu\text{g}/\text{m}^3$	0.93	0.87	0.82
Corani (2005)	daily mean	50 $\mu\text{g}/\text{m}^3$	0.82	0.84	0.76
Hooyberghs et al. (2005)	daily mean	100 $\mu\text{g}/\text{m}^3$	0.73	0.46	0.66
Grivas and Chaloulakou (2006)	hourly mean	$\bar{y} + 2\sigma$ ^c	0.58	0.68	0.55
Perez and Reyes (2006)	max 24h MA ^d	150 $\mu\text{g}/\text{m}^3$	0.81	0.70	0.76
Paschalidou et al. (2011)	hourly mean	$\bar{y} + 2\sigma$ ^c	0.70	–	–
Perez (2012)	max 24h MA ^d	195 $\mu\text{g}/\text{m}^3$	0.76	0.59	0.70

^aThe results correspond to statistics obtained on testing set across all studied locations. In case several models were tested in the paper, only statistics corresponding to the best-performing model are reported.

^bPresented numbers correspond to average statistics reported in the Table 5b.

^c2 standard deviations above the average measured PM10 concentration levels at each site.

^dMaximum value of the 24-h moving average PM10 concentration measured on any given day.

becoming on average 16% to 39% higher during the testing period as compared to the training period. Our model was able to capture some of this increase but not all of it as changes happened too fast.

Similar as it was with the predictions of PM10, we observe an exponential decrease in the quality of the produced forecasts as we extend the forecasting horizon further into the future (see Figure 4). For the first predicted hour the R^2 values range between 0.92 and 0.95. These results were compared to the findings of multiple nowcasting studies and were found to be superior in every single instance (see, for example, Ortiz-García et al., 2010; Arhami et al., 2013; Goulier et al., 2020).

The quality of the produced forecasts (measured in terms of R^2) also varies depending on the prediction starting point, with forecasts produced in the morning and during the day showing worse performance than the ones generated at night (see Figure 5). However, the fluctuations are of a much smaller magnitude than the ones observed for PM10 and can generally be neglected.

3.5 | Forecasts of the daily limit exceedances (O_3)

Table 9a,b present the results obtained for two different thresholds when using the direct approach to predicting the daily limit exceedances. Once again, we find the results produced with lower threshold to be more satisfactory. At the same

TABLE 8 Performance statistics for the produced 24-h-ahead forecasts of O₃ hourly concentrations averaged across all prediction starting points

	R ²	RMSE	MAE	MedAE	MBE	$\Delta\sigma$	d ₂
M/Johanneskirchen	0.87	12.49	9.37	7.09	1.37	0.59	0.96
M/Lothstraße	0.85	12.76	9.59	7.38	1.90	1.35	0.96
M/Stachus	0.83	11.30	8.59	6.71	1.48	3.09	0.95
A/LfU	0.86	12.86	9.71	7.43	-2.19	-0.56	0.96
Average statistics	0.85	12.35	9.31	7.15	0.64	1.12	0.96

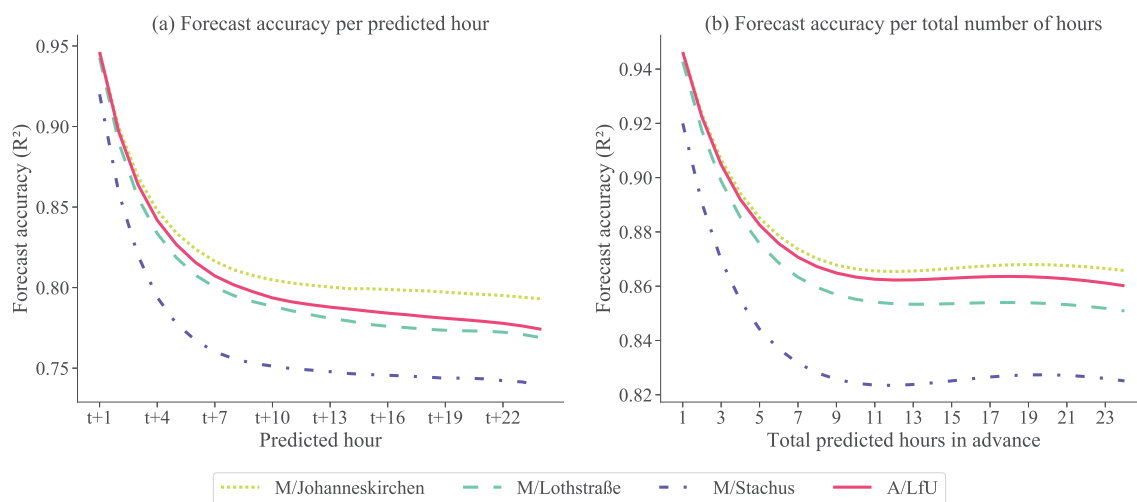


FIGURE 4 Accuracy of the produced O₃ forecasts depending on choice of the forecasting period

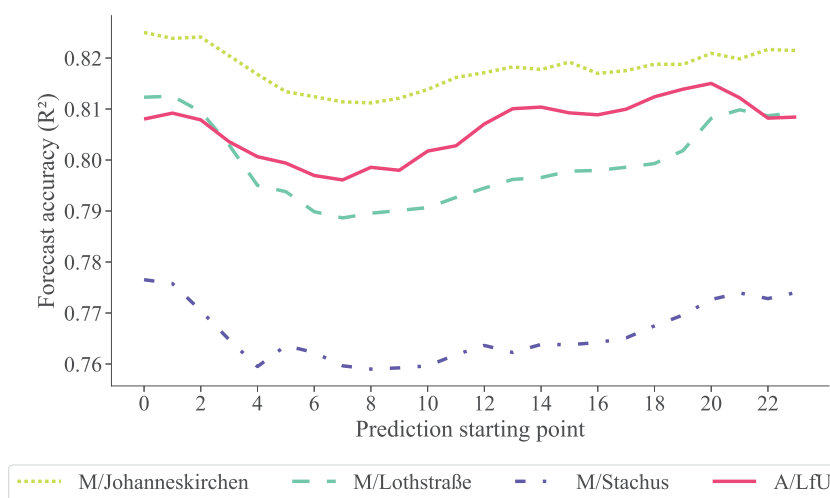


FIGURE 5 Forecast accuracy depending on prediction starting point (O₃)

time, we notice that these results are not as good as the ones obtained for PM₁₀. That is somewhat surprising since the quality of hourly forecasts produced at stage 1 was substantially better for ozone. Yet, such results are in line with the findings of other comparable studies, as we should see later on.

The indirect approach to the prediction of daily limit exceedances for ozone is identical to the one we have used for the prediction of PM₁₀ exceedances. However, to attain a reasonable balance between the model's sensitivity and the PPV value achieved on test sample, we were forced to adopt a higher minimum threshold for precision and select a lower β , when determining an optimal cutoff. A threshold of 0.6, for example, ensures probability of exceedances

TABLE 9 Performance statistics for the forecasts of O₃ daily limit exceedances produced using the direct approach

	TN, FP, FN, TP	ACC	TPR	PPV	F _β	TSS
(a) Threshold value = 120 µg/m ³						
M/Johanneskirchen	(639, 13, 15, 55)	0.96	0.79	0.81	0.79	0.77
M/Lothstraße	(648, 11, 16, 44)	0.96	0.74	0.81	0.75	0.72
M/Stachus	(604, 1, 8, 1)	0.99	0.13	0.65	0.16	0.13
A/LfU	(617, 35, 9, 58)	0.94	0.86	0.62	0.80	0.81
Averaged statistics	–	0.96	0.63	0.72	0.63	0.61
	TN, FP, FN, TP	ACC	TPR	PPV	F _β	TSS
(b) Threshold value = 110 µg/m ³						
M/Johanneskirchen	(595, 57, 3, 67)	0.92	0.96	0.54	0.83	0.87
M/Lothstraße	(626, 32, 6, 54)	0.95	0.91	0.63	0.83	0.86
M/Stachus	(601, 3, 5, 4)	0.99	0.44	0.56	0.45	0.44
A/LfU	(567, 85, 3, 64)	0.88	0.95	0.43	0.77	0.82
Averaged statistics	–	0.93	0.81	0.54	0.72	0.75

TABLE 10 Performance statistics for the forecasts of O₃ daily limit exceedances produced by logistic regression

	TN, FP, FN, TP	ACC	TPR	PPV	F _β	TSS
(a) Minimum threshold value for precision = 0.6						
M/Johanneskirchen	(604, 48, 4, 66)	0.93	0.94	0.58	0.83	0.86
M/Lothstraße	(627, 31, 6, 54)	0.95	0.90	0.64	0.83	0.86
M/Stachus	(603, 1, 8, 1)	0.99	0.15	0.62	0.17	0.14
A/LfU	(574, 78, 4, 64)	0.89	0.94	0.45	0.77	0.83
Averaged statistics	–	0.94	0.73	0.57	0.65	0.67
	TN, FP, FN, TP	ACC	TPR	PPV	F _β	TSS
(b) β = 2						
M/Johanneskirchen	(609, 43, 5, 65)	0.93	0.93	0.61	0.84	0.86
M/Lothstraße	(634, 24, 9, 51)	0.95	0.85	0.69	0.81	0.82
M/Stachus	(603, 2, 7, 2)	0.99	0.23	0.56	0.26	0.23
A/LfU	(573, 79, 4, 64)	0.88	0.94	0.45	0.77	0.82
Averaged statistics	–	0.94	0.74	0.58	0.67	0.68

detection above 90% and TSS scores above 0.8 at 3 out of 4 stations. Similar results are obtained with β set to 2 (see Table 10).

We notice, however, that none of the approaches is capable to predict the exceedances at *M/Stachus* station to any satisfying degree. That is because this station is somewhat different from the rest examined in this study. Much lower levels of ozone had been registered there, and during the whole study period only 16 cases of exceedances were documented in total—half of them found in the training and half in the testing data. Furthermore, even when exceedances were reported, they were not substantial—in 5 out of 8 cases occurred during the testing period and in all cases of the training period the corresponding O₃ concentrations were found to be in range between 120–130 µg/m³. Since already the stage-1 model notably underestimates the hourly ozone concentrations, we cannot expect to detect many true exceedances by means of simple comparison with the threshold. But also the logistic regression fails to recognize incidents of high ozone exposure at this station. The only way to achieve a higher rate of exceedances detection would be to further

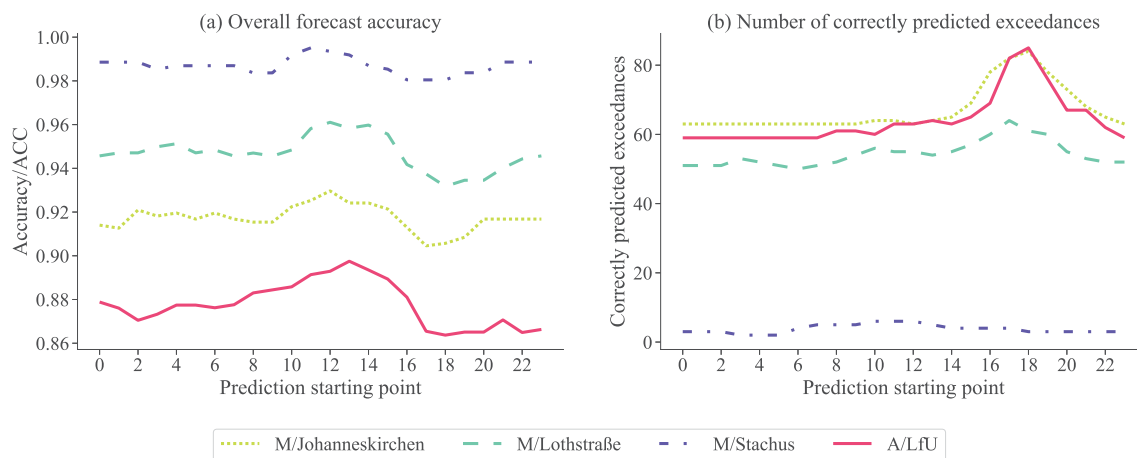


FIGURE 6 The quality of the O₃ exceedances predictions depending on forecast initiation point (direct approach with a threshold of 110 µg/m³)

TABLE 11 Performance statistics for the forecasts of O₃ daily limit exceedances generated at 7 p.m. using the direct approach

	TN, FP, FN, TP	ACC	TPR	PPV	F _β	TSS
M/Johanneskirchen	(577, 63, 3, 78)	0.91	0.96	0.55	0.84	0.86
M/Lothstraße	(611, 37, 10, 60)	0.93	0.86	0.62	0.80	0.80
M/Stachus	(601, 3, 7, 3)	0.98	0.30	0.50	0.33	0.30
A/LfU	(546, 94, 3, 76)	0.87	0.96	0.45	0.78	0.82
Averaged statistics	–	0.92	0.77	0.53	0.69	0.69

Note: A threshold of 110 µg/m³ was used to classify limit exceedances at all stations.

sacrifice model's precision. This could be done by selecting a higher β or by lowering the minimum threshold value for precision.

Overall, after comparing the results presented in Tables 9 and 10, we come to a conclusion that, in case of ozone, building an additional model for classification purposes was not justified. Simple comparison with the adjusted threshold of 110 µg/m³ allows us to correctly identify more than 90% of limit exceedances at 3 stations. Admittedly, the PPV values are rather low at all stations except for *M/Lothstraße*. Yet, we should not forget that the only short-term preventive action that could be taken by authorities in view of the expected episodes of high ozone exposure is issuing an early warning to alert the population. The cost of a false alarm under such circumstances (people changing their plans and choosing to stay indoors for several hours, even if unnecessary) is indisputably lower than that of failing to predict an episode of high exposure (and thus endangering the people's health and even risking people's lives). Therefore, once again, we put more weight on model's ability to predict exceedances than on its capacity to avoid false alarms.

Figure 6 illustrates how the quality of the produced exceedances predictions varies depending on time when predictions were made. We notice that the forecasts produced during the day, when ozone concentration is at its highest, tend to have higher accuracy, but the largest number of limit exceedances are detected by forecasts generated somewhere between 4 and 8 p.m.—that is, in the hours immediately following the daily ozone peaks. Affected by the last few observations, the model tends to predict higher O₃ concentrations around that time, thus capturing more actual exceedances. However, this also results in higher number of false positives, leading to a sudden fall in the overall accuracy of the forecasts.

Naturally, if the model was to be put in operational use, the forecasts would most likely not be generated at each hour of the day but only once per day. In that case we would prefer an evening hour as initiation point. Table 11, for example, reflects the performance statistics for the forecasts produced at 7 p.m.

4 | DISCUSSION AND COMPARISON

In this section we compare the results of this paper with those in the literature. Note that the results are not directly comparable, since we different data, different time span and different methods than other authors. This the comparison is very general.

The obtained results are similar to those achieved by Hrust et al. (2009) and are better than the results reported by Fernando et al. (2012). Both teams apply methodology similar to ours but in combination with an ANN method and achieve an R^2 of 0.66–0.72 and 0.39 respectively. Cai et al. (2009), on the other hand, obtain an R^2 of about 0.83 ($\rho = 0.91$) on their more than 10-h-ahead forecasts for three stations in China, which is only comparable to the results we got for *M/Lothstraße* and *A/Karlstraße* (if we would consider the same forecasting horizon, see Figure 1b). The better results can be explained by shorter testing period and by the fact that, Cai et al. (2009) additionally use the traffic volume and several geographical variables as input to their models. Inclusion of traffic data into the developed model would likely substantially improve the quality of forecasts at traffic-impacted stations, such as *M/LandshuterAllee* and *M/Stachus*. Yet, in the absence of access to such data, we had to rely on various temporal variables described in Section 2.1, which, at least in part, may also be seen as proxies for the absent traffic variable.

Kukkonen (2003), Grivas and Chaloulakou (2006), as well as Paschalidou et al. (2011) use somewhat different methodology, by predicting each hour using only such persistence information that was available 24 h before the predicted time (e.g., $PM_{10}(t-24)$, $PM_{10}(t-25)$ and $PM_{10}(t-26)$) in addition to meteorological and temporal variables observed at time for which predictions were made. Paschalidou et al. (2011) achieve an average R^2 of 0.68 across four stations on Cyprus, whereas Kukkonen (2003) and Grivas and Chaloulakou (2006) demonstrate substantially worse results, with mean R^2 amounting to 0.37 and 0.60 respectively and d_2 values ranging between 0.73 and 0.89.

This result is comparable to the findings of Arhami et al. (2013), who report an R^2 of 0.87 on their next-hour predictions of PM_{10} concentrations in Tehran, Iran. The results published in other similar studies are not as convincing, with R^2 taking values between 0.29 and 0.72 (see, e.g., Aldrin & Haff, 2005; Goulier et al., 2020).

With respect to ozone, We could identify nine studies in total, conducted between 2003–2010, that attempt forecasts of high O_3 concentrations in urban areas. Same as it was the case with PM_{10} , we see a large variety of target measures and thresholds adopted by the authors. Four studies in total (Corani, 2005; Hogrefe et al., 2007; Schlink et al., 2003; Tsai et al., 2009) focus on the same forecasting target as us (the maximum 8-h moving average), but only Schlink et al. (2003) also use the same threshold. The study by Tsai et al. (2009) is included in the table twice. That is because these authors, similar to us, had several stations with a very small number of daily exceedances in their data. The developed model produced much worse results at these stations than at other locations. Hence, we also report the average results calculated without the contribution of such stations to avoid possible misrepresentation. Still, all of the above-mentioned papers demonstrate worse results than we do, with Schlink et al. (2003) coming closest in terms of TSS value and Corani (2005) in terms of TPR and PPV values.

Predicting daily maximum O_3 concentration seems to be a more easy task than predicting the maximum 8-h moving average since the results obtained by authors in the lower part of the table are generally better. Remarkably good results are reported by Slini et al. (2002) and Kumar and de Ridder (2010). The best TPR and TSS values were achieved by Dutot et al. (2007), who focus on hourly instead of daily concentrations and predict exceedances of hourly maximum O_3 values in Orléans, France (it is possible, however, that the authors meant daily maximum of hourly concentrations while using the term “hourly maximum”).

5 | CONCLUSIONS

A 2-stage model has been developed to produce short-term forecasts of PM_{10} and O_3 concentrations and tested using the data collected in the cities of Augsburg and Munich, Germany. In the first stage, the mean hourly pollutant concentrations for each of the next 24 h were predicted from meteorological, temporal and persistence data, using Stochastic Gradient Tree Boosting as the prediction algorithm. In the second stage, these predictions were used to forecast exceedances of daily limit concentrations set by the EU. We were able to achieve very good results at both stages. Consistent with findings in the literature, the accuracy of the hourly forecasts was substantially better for ozone, with average R^2 of 0.85 and d_2 of 0.96 computed across all stations. The R^2 values obtained for PM_{10} vary between 0.58 and 0.77 depending on location, with worst results achieved on data from traffic-impacted stations. An inclusion of traffic variables into the input data could significantly improve the quality of the predictions for these stations.

In terms of exceedances detection performed at stage 2, the better overall results were achieved for PM₁₀. Using logistic regression with a modified cutoff probability, we can predict ca. 91% of all limit exceedances occurred at seven stations during the testing period. About 44% of episodes predicted by the model were false positives. Yet, in absolute terms less than 18 false alarms per year were registered at each station. The TSS score of 0.89, unmatched by any of the compared studies, is another powerful indicator of the outstanding model performance.

In case of ozone, satisfactory results were obtained also without the logistic regression model. Instead, hourly forecasts from the stage 1 were used to calculate a set of daily measures, which were then compared to a specified threshold value to identify the exceedances. Since all statistical models tend to underpredict the extreme values, we replaced the original threshold of 120 µg/m³ with a lower one to ensure better exceedances detection. Using a threshold of 110 µg/m³ we could identify more than 90% of actually occurred exceedances and achieve TSS values of above 0.8 at 3 stations out of 4. The *M/Stachus* station turned out to be much more difficult to make predictions for due to the extremely small total number of observed O₃ exceedances. The average precision attained across all stations was only slightly above 50%. Yet, the cost of issuing a false alarm is very small whereas the cost of an undetected exceedance, on the other hand, is arguably high. Therefore, a public authority would likely be willing to trade a higher rate of false alarms for the benefit of better health protection.

Overall, the good performance of the model proves that it can be a useful tool for the short-term air quality management. An important advantage of our model lies in the very low requirements it puts on computational resources. Run on a private computer, the algorithm can process more than a decade of hourly data and generate predictions in a matter of minutes. Hence it can easily be adopted by a governmental agency without the need of any additional investment into the hosting infrastructure. Still, there are some limitations. One of them lies in the fact that we constructed the model using the actual meteorological data. And although numerical weather forecasts of all meteorological variables used as input are available on routinely basis, their uncertainty would contribute to the total uncertainty of the model and may result in slightly worse performance. There are also limitations in terms of time period and location. Because the relationship between the independent and dependent variables at different locations around the world is never exactly the same, a separate model should be developed for each measurement site to ensure the best possible performance. Moreover, since the underlying patterns in data can change with time, the model needs to be periodically recalibrated and retrained using several most recent years of observations. However, the general approach would always remain the same.

The model can be extended in many directions. First, the availability of traffic data is expected to improve the predictive performance of the model at both stages. Second, it would be of interest to deploy machine learning classification tools in place of logistic regression at the stage 2. Third, spatial models that take into account observations not only at a given but also at the nearby stations may provide insights into the interconnectedness of the stations and further improve the model's predictive power. These issues are left for further research.

DATA AVAILABILITY STATEMENT

The data used in this project is publicly available upon request from the German Weather Service (Deutscher Wetterdienst, <https://www.dwd.de>) and from the Bavarian State Office for Environment (Bayerisches Landesamt für Umwelt, <https://www.lfu.bayern.de>).

ORCID

Yarema Okhrin  <https://orcid.org/0000-0003-4704-5233>

REFERENCES

- Aldrin, M., & Haff, I. (2005). Generalised additive modelling of air pollution, traffic volume and meteorology. *Atmospheric Environment*, *39*, 2145–2155.
- Arhami, M., Kamali, N., & Rajabi, M. M. (2013). Predicting hourly air pollutant levels using artificial neural networks coupled with uncertainty analysis by Monte Carlo simulations. *Environmental Science and Pollution Research International*, *20*, 4777–4789.
- Ayres, J. G. (2006). *Cardiovascular disease and air pollution: A report by the committee on the medical effects of air pollutants (COMEAP)*. Health Protection Agency (Great Britain).
- Ballester, E. B., i Valls, G. C., Carrasco-Rodriguez, J. L., Olivas, E. S., & del Valle-Tascon, S. (2002). Effective 1-day ahead prediction of hourly surface ozone concentrations in eastern Spain using linear models and neural networks. *Ecological Modelling*, *156*, 27–41.
- Breiman, L. (2001). Random forests. *Machine Learning*, *45*, 5–32.
- Cai, M., Yin, Y., & Xie, M. (2009). Prediction of hourly air pollutant concentrations near urban arterials using artificial neural network approach. *Transportation Research Part D: Transport and Environment*, *14*, 32–41.
- Chaloulakou, A., Saisana, M., & Spyrellis, N. (2003). Comparative assessment of neural networks and regression models for forecasting summertime ozone in Athens. *Science of the Total Environment*, *313*, 1–13.

- Chang, C.-C., & Lin, C.-J. (2011). LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2, 1–27.
- Chang, T., Graff Zivin, J., Gross, T., & Neidell, M. (2016). Particulate pollution and the productivity of pear packers. *American Economic Journal: Economic Policy*, 8, 141–169.
- Chen, J.-C., & Schwartz, J. (2009). Neurobehavioral effects of ambient air pollution on cognitive performance in US adults. *Neurotoxicology*, 30, 231–239.
- Corani, G. (2005). Air quality prediction in Milan: Feed-forward neural networks, pruned neural networks and lazy learning. *Ecological Modelling*, 185, 513–529.
- Dockery, D. W., Pope, C. A., Xu, X., Spengler, J. D., Ware, J. H., Fay, M. E., Ferris, B. G., & Speizer, F. E. (1993). An association between air pollution and mortality in six U.S. cities. *The New England Journal of Medicine*, 329, 1753–1759.
- Dutot, A.-L., Rynkiewicz, J., Steiner, F. E., & Rude, J. (2007). A 24-h forecast of ozone peaks and exceedance levels using neural classifiers and weather predictions. *Environmental Modelling & Software*, 22, 1261–1269.
- Ebenstein, A., Lavy, V., & Roth, S. (2016). The long-run economic consequences of high-stakes examinations: Evidence from transitory variation in pollution. *American Economic Journal: Applied Economics*, 8, 36–65.
- European Council (EC). (2008). Directive 2008/50/EC of the European parliament and of the council of 21 May 2008 on ambient air quality and cleaner air for Europe: 02008L0050 - EN - 18.09.2015 - 001.002. *Official Journal, L 152*, 1–56.
- Fan, R.-E., Chang, K.-W., Hsieh, C.-J., Wang, X.-R., & Lin, C.-J. (2008). LIBLINEAR: A library for large linear classification. *Journal of Machine Learning Research*, 9, 1871–1874.
- Fernando, H. J. S., Mammarella, M. C., Grandoni, G., Fedele, P., Di Marco, R., Dimitrova, R., & Hyde, P. (2012). Forecasting PM10 in metropolitan areas: Efficacy of neural networks. *Environmental Pollution*, 163, 62–67.
- Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29, 1189–1232. www.jstor.org/stable/2699986
- Friedman, J. H. (2002). Stochastic gradient boosting. *Computational Statistics & Data Analysis*, 38, 367–378.
- Geurts, P., Ernst, D., & Wehenkel, L. (2006). Extremely randomized trees. *Machine Learning*, 63, 3–42.
- Gómez-Losada, A. (2018). Forecasting ozone threshold exceedances in urban background areas using supervised classification and easy-access information. *Atmospheric Pollution Research*, 9, 1052–1061.
- Gong, B., & Ordieres-Meré, J. (2016). Prediction of daily maximum ozone threshold exceedances by preprocessing and ensemble artificial intelligence techniques: Case study of Hong Kong. *Environmental Modelling & Software*, 84, 290–303.
- Goulier, L., Paas, B., Ehrnsperger, L., & Klemm, O. (2020). Modelling of urban air pollutant concentrations with artificial neural networks using novel input variables. *International Journal of Environmental Research and Public Health*, 17, 1–22.
- Grivas, G., & Chaloulakou, A. (2006). Artificial neural network models for prediction of PM10 hourly concentrations, in the greater area of Athens, Greece. *Atmospheric Environment*, 40, 1216–1229.
- Gryparis, A., Forsberg, B., Katsouyanni, K., Analitis, A., Touloumi, G., Schwartz, J., Samoli, E., Medina, S., Anderson, H. R., Niciu, E. M., Wichmann, H.-E., Kriz, B., Kosnik, M., Skorkovsky, J., Vonk, J. M., & Dörtdubad, Z. (2004). Acute effects of ozone on mortality from the "Air pollution and health: A European approach" project. *American Journal of Respiratory and Critical Care Medicine*, 170, 1080–1087.
- Hastie, T., Tibshirani, R., & Friedman, J. H. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed.). Springer.
- Hogrefe, C., Hao, W., Civerolo, K., Ku, J.-Y., Sistla, G., Gaza, R. S., Sedefian, L., Schere, K., Gilliland, A., & Mathur, R. (2007). Daily simulation of ozone and fine particulates over new york state: Findings and challenges. *Journal of Applied Meteorology and Climatology*, 46, 961–979.
- Hooyberghs, J., Mensink, C., Dumont, G., Fierens, F., & Brasseur, O. (2005). A neural network forecast for daily average pm concentrations in belgium. *Atmospheric Environment*, 39, 3279–3289.
- Hripcsak, G., & Rothschild, A. S. (2005). Agreement, the F-measure, and reliability in information retrieval. *Journal of the American Medical Informatics Association: JAMIA*, 12, 296–298.
- Hrust, L., Klaić, Z. B., Križan, J., Antonić, O., & Hercog, P. (2009). Neural network forecasting of air pollutants hourly concentrations using optimised temporal averages of meteorological variables and pollutant concentrations. *Atmospheric Environment*, 43, 5588–5596.
- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning: With applications in R* (Vol. 103). Springer.
- Kukkonen, J. (2003). Extensive evaluation of neural network models for the prediction of NO₂ and PM10 concentrations, compared with a deterministic modelling system and measurements in central Helsinki. *Atmospheric Environment*, 37, 4539–4550.
- Kumar, U., & de Ridder, K. (2010). GARCH modelling in association with FFT-ARIMA to forecast ozone episodes. *Atmospheric Environment*, 44, 4252–4265.
- Künzli, N., Kaiser, R., Medina, S., Studnicka, M., Chanel, O., Filliger, P., Herry, M., Horak, F., Puybonnieux-Textier, V., Quénel, P., Schneider, J., Seethaler, R., Vergnaud, J.-C., & Sommer, H. (2000). Public-health impact of outdoor and traffic-related air pollution: A European assessment. *The Lancet*, 356, 795–801.
- Li, H., Cai, J., Chen, R., Zhao, Z., Ying, Z., Wang, L., Chen, J., Hao, K., Kinney, P. L., Chen, H., & Kan, H. (2017). Particulate matter exposure and stress hormone levels: A randomized, double-blind, crossover trial of air purification. *Circulation*, 136, 618–627.
- Loomis, D., Grosse, Y., Lauby-Secretan, B., Ghissassi, F. E., Bouvard, V., Benbrahim-Tallaa, L., Guha, N., Baan, R., Mattock, H., & Straif, K. (2013). The carcinogenicity of outdoor air pollution. *The Lancet Oncology*, 14, 1262–1263.
- Lu, J. G., Lee, J. J., Gino, F., & Galinsky, A. D. (2018). Polluted morality: Air pollution predicts criminal activity and unethical behavior. *Psychological Science*, 29, 340–355.
- Masmoudi, S., Elghazel, H., Taieb, D., Yazar, O., & Kallel, A. (2020). A machine-learning framework for predicting multiple air pollutants' concentrations via multi-target regression and feature selection. *Science of the Total Environment*, 715, 136991.

- McKendry, I. G. (2002). Evaluation of artificial neural networks for fine particulate pollution (PM₁₀ and PM_{2.5}) forecasting. *Journal of the Air & Waste Management Association*, 52(9), 1096–1101.
- Murphy, S. R., Schelegle, E. S., Miller, L. A., Hyde, D. M., & van Winkle, L. S. (2013). Ozone exposure alters serotonin and serotonin receptor expression in the developing lung. *Toxicological Sciences*, 134, 168–179.
- Ortiz-García, E. G., Salcedo-Sanz, S., Pérez-Bellido, Á., Portilla-Figueras, J. A., & Prieto, L. (2010). Prediction of hourly O₃ concentrations using support vector regression algorithms. *Atmospheric Environment*, 44, 4481–4488.
- Paschalidou, A. K., Karakitsios, S., Kleanthous, S., & Kassomenos, P. A. (2011). Forecasting hourly PM₁₀ concentration in Cyprus through artificial neural networks and multiple regression models: Implications to local environmental management. *Environmental Science and Pollution Research International*, 18, 316–327.
- Peng, H., Lima, A. R., Teakles, A., Jin, J., Cannon, A. J., & Hsieh, W. W. (2017). Evaluating hourly air quality forecasting in Canada with nonlinear updatable machine learning methods. *Air Quality, Atmosphere, & Health*, 10, 195–211.
- Perez, P. (2012). Combined model for PM₁₀ forecasting in a large city. *Atmospheric Environment*, 60, 271–276.
- Perez, P., & Reyes, J. (2006). An integrated neural network model for PM₁₀ forecasting. *Atmospheric Environment*, 40, 2845–2851.
- Pope, C. A., & Dockery, D. W. (2006). Health effects of fine particulate air pollution: Lines that connect. *Journal of the Air & Waste Management Association*, 56, 709–742.
- Represa, N. S., Fernández-Sarría, A., Porta, A., & Palomar-Vázquez, J. (2020). Data mining paradigm in the study of air quality. *Environmental Processes*, 7, 1–21.
- Sager, L. (2019). Estimating the effect of air pollution on road safety using atmospheric temperature inversions. *Journal of Environmental Economics and Management*, 98, 102250.
- Schlink, U., Dorling, S., Pelikan, E., Nunnari, G., Cawley, G., Junninen, H., Greig, A., Foxall, R., Eben, K., Chatterton, T., Vondracek, J., Richter, M., Dostal, M., Bertuccio, L., Kolehmainen, M., & Doyle, M. (2003). A rigorous inter-comparison of ground-level ozone predictions. *Atmospheric Environment*, 37, 3237–3253.
- Sharma, S., Sharma, P., Khare, M., & Kwatra, S. (2016). Statistical behavior of ozone in urban environment. *Sustainable Environment Research*, 26, 142–148.
- Slini, T., Karatzas, K., & Moussiopoulos, N. (2002). Statistical analysis of environmental data as the basis of forecasting: An air quality application. *Science of the Total Environment*, 288, 227–237.
- Stone, M. (1974). Cross-validation and multinomial prediction. *Biometrika*, 61, 509–515.
- Tipping, M. E., & Smola, A. E. (2001). Sparse bayesian learning and the relevance vector machine. *Journal of Machine Learning Research*, 1, 211–244.
- Tsai, C.-H., Chang, L.-C., & Chiang, H.-C. (2009). Forecasting of ozone episode days by cost-sensitive neural network methods. *Science of the Total Environment*, 407, 2124–2135.
- Valput, D., Navares, R., & Aznarte, J. (2020). Forecasting hourly NO₂ concentrations by ensembling neural networks and mesoscale models. *Neural Computations and Applications*, 32, 9331–9342.
- Vapnik, V. N. (1998). *Adaptive and learning systems for signal processing, communications, and control*. In H. Simon, ed. *Statistical learning theory*. Wiley.
- Willmott, C. J. (1981). On the validation of models. *Physical Geography*, 2, 184–194.
- Yang, J., Zhao, C., Yang, J., Wang, J., Li, Z., Wan, X., Guo, G., Lei, M., & Chen, T. (2020). Discriminative algorithm approach to forecast Cd threshold exceedance probability for rice grain based on soil characteristics. *Environmental Pollution*, 261, 114211.
- Yu, H.-F., Huang, F.-L., & Lin, C.-J. (2011). Dual coordinate descent methods for logistic regression and maximum entropy models. *Machine Learning*, 85, 41–75.
- Zhou, Y., Chang, F.-J., Chen, H., & Li, H. (2020a). Exploring copula-based Bayesian model averaging with multiple ANNs for PM_{2.5} ensemble forecasts. *Journal of Cleaner Production*, 263, 121528.
- Zhou, Y., Chang, L.-C., & Chang, F.-J. (2020b). Explore a multivariate Bayesian uncertainty processor driven by artificial neural networks for probabilistic PM_{2.5} forecasting. *Science of the Total Environment*, 711, 134792.
- Zickus, M., Greig, A. J., & Niranjani, M. (2002). Comparison of four machine learning methods for predicting PM₁₀ concentrations in Helsinki, Finland. *Water, Air & Soil Pollution: Focus*, 2, 717–729.
- Zivin, J. G., & Neidell, M. (2012). The impact of pollution on worker productivity. *The American Economic Review*, 102, 3652–3673.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of this article.

How to cite this article: Krylova, M., & Okhrin, Y. (2021). Managing air quality: Predicting exceedances of legal limits for PM₁₀ and O₃ concentration using machine learning methods. *Environmetrics*, e2707. <https://doi.org/10.1002/env.2707>