Adv. Stat. Clim. Meteorol. Oceanogr., 9, 121–135, 2023
https://doi.org/10.5194/ascmo-9-121-2023

ASCMO
Open Access

Forecasting 24 h averaged $PM_{2.5}$ concentration in the Aburrá Valley using tree-based machine learning models, global forecasts, and satellite information

Jhayron S. Pérez-Carrasquilla[1,2], Paola A. Montoya[2,3], Juan Manuel Sánchez[2,3], K. Santiago Hernández[2,3], and Mauricio Ramírez[2]

[1]Department of Atmospheric and Oceanic Science, University of Maryland, College Park, USA
[2]Área Metropolitana del Valle de Aburrá, Proyecto Sistema de Alerta Temprana de Medellín y el Valle de Aburrá (SIATA), Medellín, Colombia
[3]Escuela Ambiental, Facultad de Ingeniería, Universidad de Antioquia, Medellín, Colombia

Correspondence: Jhayron S. Pérez-Carrasquilla (jhayron@umd.edu) and Paola A. Montoya (paomon5@gmail.com)

Received: 2 October 2022 – Revised: 23 October 2023 – Accepted: 2 November 2023 – Published: 22 December 2023

Abstract. We develop a framework to forecast 24 h averaged particulate matter ($PM_{2.5}$) concentrations 4 d in advance in ground-based stations over the metropolitan area of the Aburrá Valley, Colombia. The input variables are gathered from a highly diverse set of sources, including in situ real-time $PM_{2.5}$ observations, meteorological forecasts from the Global Forecasting System (GFS), aerosol optical depth (AOD) forecasts from the European Copernicus Atmosphere Monitoring Service (CAMS), and the Moderate Resolution Imaging Spectroradiometer (MODIS) active fire products. We compare the performance of two tree-based machine learning (ML) methods, random forests (RFs) and gradient boosting (GB), with linear regression as a baseline for error metrics. One of the disadvantages of tree-based models is their inability to make skillful predictions out of the domain in which the models were trained. To address that problem, we implement piecewise linear regression learners within the models. Additionally, to enhance the performance of the models, we use a customized loss function that considers the probability distribution of the target values. Tree-based models highly outperform the linear regression, with GB showing the best results in most of the 19 stations used in this study. We also test two approaches for the multi-step output problem, a direct multi-output (MO) scheme and a recursive (RC) scheme, with the GB–MO approach showing the best results. According to the performance analysis, the predictability is less for values away from the mean and decreases between 06:00 LT (local time) and the early afternoon, when the expansion of the boundary layer occurs.

To contribute to understanding the sources of predictability and uncertainty of air quality in the city, we perform a feature importance analysis revealing that the relevance of the different independent variables is a function of the lead time. Particularly, apart from the past concentrations, the variables that most affect the predictability are the forecasted aerosol optical depth (AOD), the integrated fire radiative power over a forecasted back trajectory (BT-IFRP), and the predicted planetary boundary layer height (PBLH). In the testing period, the models showed the ability to forecast poor-air-quality events in the valley with more than 1 d of anticipation. This study serves as a framework for developing and evaluating the ML-based air quality forecasting models over the Andean region.

1 Introduction

Forecasting air quality in urban areas is becoming more relevant every day worldwide. During the last few years, several studies have shown broad evidence of the detrimental effects on human health of high concentrations of aerosols near the surface (Mabahwi et al., 2014), with small-sized particulate matter having particular relevance due to its abundance in densely populated cities (Zhang et al., 2013; Xing et al., 2016) such as the Aburrá Valley metropolitan area, in the Colombian Andean region. Both mortality (Samet et al., 2000; Lepeule et al., 2012) and morbidity increase with high $PM_{2.5}$ concentrations. Atmospheric pollution has been associated with asthma, respiratory inflammation, affectations on lung functions, and even promotion of cancer (Lewis et al., 2005; Orru et al., 2011); outdoors pollution causes between 1.61 and 4.81 million premature deaths worldwide yearly, predominantly in Asia (Lelieveld et al., 2015).

Poor-air-quality episodes that may last several days are frequent in the Aburrá Valley, which has led to the creation of government programs such as the Integral Plan for Air Quality Management in the Aburrá Valley (PIGECA[1]) and the Operational Protocol to deal with Air Pollution Episodes in the Aburrá Valley Metropolitan Area (POECA[2]). With POECA, decision-makers use real-time information to apply emission restrictions in the region for the next few days (traffic, industry, etc.). Therefore, an accurate forecast of pollutant concentrations, especially $PM_{2.5}$, is paramount. The main factors favoring detrimental $PM_{2.5}$ concentration episodes in the urban area of interest are the high population density in the Aburrá Valley (3379 inhabitants per squared kilometer), the influence of external pollution sources such as wildfires, and the topographic characteristics leading to a stagnant near-surface atmosphere (see Fig. 1a) (Herrera-Mejía and Hoyos, 2019; Roldán-Henao et al., 2020). During those episodes, the 24 h averaged $PM_{2.5}$ concentration in several monitoring stations within the city exceeded 35.5 $\mu g\, m^{-3}$, which is unhealthy for sensitive groups according to the air quality index (AQI), or even 55.5 $\mu g\, m^{-3}$, corresponding to an unhealthy AQI.

Several elements must be considered to accurately forecast the air quality conditions a few days in advance. As for the aerosol sources, the two most relevant ones in terms of $PM_{2.5}$ are the urban emissions, mainly resulting from car engines and industrial activity (Herrera-Mejía and Hoyos, 2019), and external biomass burning aerosols from wildfires that occur in northern Colombia, the Orinoco River basin, and, with less impact in terms of local concentrations, in the Amazon River basin (Rincón-Riveros et al., 2020; Hernandez et al., 2019; Mendez-Espinosa et al., 2019; Ballesteros-González et al., 2020; Rodriguez-Gomez et al., 2022). Apart from the emissions, meteorological factors play a crucial role in reducing aerosols concentrations in the valley in two ways: the lifting and subsequent horizontal transport away from the valley of the pollutants, which occurs in response to a thermally unstable lower atmosphere, and wet deposition. Several studies have addressed the relationships between meteorological and air quality conditions in the area of interest during recent years (Isaza Uribe, 2020; Herrera-Mejía and Hoyos, 2019; Hoyos et al., 2020; Roldán-Henao et al., 2020).

Atmospheric stability plays a crucial role (Herrera-Mejía and Hoyos, 2019). Shallow boundary layers that inhibit pollutants' vertical dispersion respond to a relatively low inflow of solar radiation at the surface (Herrera-Mejía and Hoyos, 2019), primarily a product of regional cloudiness and aerosols' radiative forcing (Dubovik et al., 2002; Bond et al., 2013). Also, the dual net effect of rainfall events over the $PM_{2.5}$ concentration must be considered. Roldán-Henao et al. (2020) argue that commonly, afternoon rainfall events have a net detrimental effect on air quality; the contribution to near-surface cooling and the consequent stabilization of the lowest part of the troposphere end up being more relevant than the favorable aerosol reduction due to scavenging. On the other hand, due to their long duration and extensive coverage over the region, nocturnal rainfall events significantly contribute to reducing $PM_{2.5}$ concentrations, thanks to efficient wet deposition (Guo et al., 2014).

Even though it is impossible to have a perfectly accurate model due to the non-linearity of the relationships with the processes and variables responsible for its fluctuations, the mixed natural and anthropogenic origin of its variability, and the chaotic nature of weather phenomena (Lorenz, 1969), a forecasting model for the 24 h $PM_{2.5}$ concentration in the Aburrá Valley is pivotal in supporting decision-makers, especially during periods with high loads of aerosols. The use of ML methods in geosciences is becoming more common every day; their ability to be as accurate as numerical models without the need to know the analytical description of the relations among variables has resulted in several studies applying artificial intelligence (AI), not only for weather applications but also for pollutants' concentration prediction and diagnosis (Wang et al., 2017; Mao et al., 2017; Xu et al., 2020; Lin et al., 2021; Guo et al., 2021; Yang et al., 2021; Lv et al., 2021; Chellali et al., 2016; Perez and Gramsch, 2016; Yang et al., 2020; Perišić et al., 2017; Tian and Chen, 2010).

According to Lee et al. (2020), tree-based algorithms tend to present the best accuracy when trying to predict $PM_{2.5}$ concentrations; they propose a gradient boosting (GB) approach for predicting $PM_{2.5}$ concentration in Taiwan and London. Mao et al. (2017) use aerosol optical depth (AOD), back trajectories, and a multi-layer perceptron (MLP) model for forecasting concentrations 3 d in advance in eastern China, while Wang et al. (2017) perform a hybrid forecasting

[1] https://www.metropol.gov.co/ambiental/calidad-del-aire/Paginas/Gestion-integral/PIGECA.aspx (last access: 14 December 2023)

[2] https://www.metropol.gov.co/ambiental/calidad-del-aire/Paginas/Gestion-integral/POECA.aspx (last access: 14 December 2023)

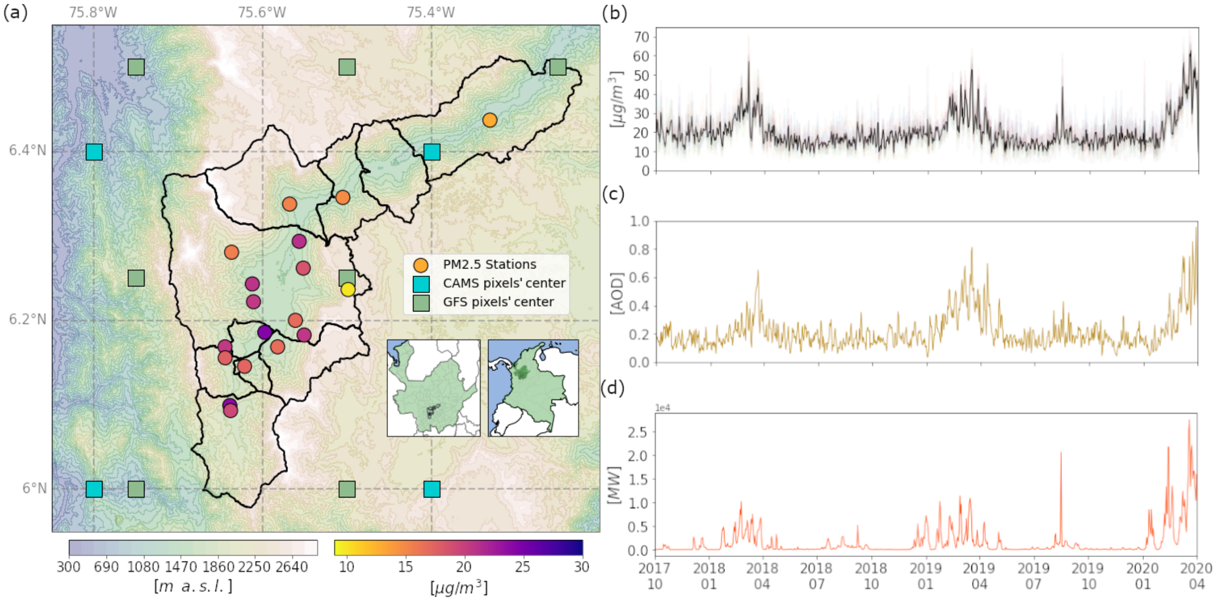

Figure 1. **(a)** Geographical location and digital elevation model of the Aburrá Valley. Filled circles indicate the location of the $PM_{2.5}$ stations used in this study, and their color represents the average concentration during the training period. Green and blue squares indicate the location of the GFS and CAMS pixels' center, respectively. **(b)** Daily-averaged time series of the mean $PM_{2.5}$ concentration from all the stations during the training period; the black line represents the average concentration in the city. **(c)** Same as **(b)** but for the total aerosol optical depth (AOD) most recent forecast from CAMS. **(d)** Same as **(b)** but for the integrated fire radiative power (IFRP) within a radius of 50 km around the 800 hPa 4 d back-trajectory from GFS.

model including support vector regression (SVR) for a 10 d hourly forecast in the same region. Lv et al. (2021) compare the results from a random forest (RF) (Breiman, 2001) and an SVR method with a multiple linear regression (MLR), and they both present a significant relative improvement. Most recently, Ke et al. (2022) developed an automated air quality forecasting system that includes results from MLR, MLP, RF, GB, and SVR; they also include real-time Global Forecasting System (GFS) meteorological forecasts and use stacked generalization (SG) to get a final output from the ensemble.

Considering the advances mentioned above, this study proposes an approach for tackling the challenge of combining and exploiting diverse sources of available and valuable information to forecast $PM_{2.5}$ in a densely populated region. We use data from the monitoring network of the Early Warning System of the Aburrá Valley (SIATA), CAMS (Benedetti et al., 2009; Inness et al., 2019) and GFS (National Centers for Environmental Prediction, National Weather Service, NOAA, U.S. Department of Commerce, 2015) atmospheric forecasting models, and the MODIS active fire products (Justice et al., 2002; Giglio et al., 2003, 2016) to develop a forecast for the 24 h averaged $PM_{2.5}$ concentration in the following 4 d. One of the novelties of this study is the development of a new index used as input for the models, which is based on the satellite-derived fire radiative power within an area determined by low-tropospheric back-trajectories (see Sect. 2.4). This index ends up being crucial for the prediction of detrimental air quality events within the city. We include recent software developments (piecewise linear regression trees and a customized loss function; see Sect. 3) to improve one of the most significant downsides of tree-based models, the low performance of tree-based models over data out of the training domain, and compare the results with an MLR method. Finally, we use eXplainable AI (XAI) methods to learn more about the predictability of the $PM_{2.5}$ concentration within a narrow valley. Section 2 describes the collected data and input variables. Section 3 details the models and the fitting and testing process. In Sect. 4, we show the performance of the models, some examples, and a variable importance analysis. Section 5 presents the conclusions and discussion.

2 Data

One of the possibilities that tree-based algorithms allow is to integrate multiple sources of information and a high number of input variables straightforwardly. We use the following information sources to account for the main forcings of the $PM_{2.5}$ concentration variability (anthropogenic activity, external sources, and meteorology).

2.1 $PM_{2.5}$ monitoring network

We employ in situ hourly observations of $PM_{2.5}$ concentration from the SIATA official monitoring network; 6 BAM1020 and 13 BAM1022 devices by Met One Instruments, Inc. are distributed throughout the Aburrá Valley. The devices measure airborne particulate concentration levels using the principle of beta ray attenuation with methods endorsed by the US Environmental Protection Agency (EPA) and hold international accreditation. Of the 19 stations, 2 are located near roadways: SUR-TRAF and CEN-TRAF.

Most of the stations were installed in October 2017, with continuous measurements of hourly PM$_{2.5}$ concentrations since then. Therefore, we count with 3 years of continuous measurements to train the models. The target of the models is to predict 24 h averaged PM$_{2.5}$ concentrations, which is aligned with the standard that regulates this pollutant in Colombia and thus supports decision-making. A 24 h rolling window is applied to the concentrations, allowing an hourly assessment of the results.

2.2 AOD from CAMS

To predict episodes in which long-range transport of aerosols affects the air quality in the region of interest, we include the CAMS global atmospheric composition forecast. This model operates twice a day, at 00:00 and 12:00 UTC (https://confluence.ecmwf.int, last access: 14 December 2023), and data for the next 5 d are publicly available around 10 h after the initial run. The model counts with a horizontal resolution of approximately 40 km, and it consists of 56 chemical species of reactive traces gases in the troposphere, stratospheric ozone, and seven different aerosol types. For this analysis, we include the forecasted AOD out to 96 h lead time.

2.3 GFS

To account for the fluctuations of PM$_{2.5}$ concentrations due to meteorological changes, we use outputs from GFS from the National Center for Environmental Predictions (NCEP). This weather model produces forecasts four times a day (00:00, 06:00, 12:00, and 18:00 UTC). It predicts 16 d in advance, with a temporal resolution of 3 h; the horizontal resolution is approximately 28 km; and it has 64 layers in the vertical. In the models developed here, we include data for total cloud cover (TCC), precipitation (TP), and planetary boundary layer height (PBLH).

2.4 MODIS active fire data

The occurrence of fires, commonly related to agricultural activities (Rodriguez-Gomez et al., 2022), causes recurrent increments of PM$_{2.5}$ in the city. Figure 2 shows the correlation coefficient between the daily time series of integrated fire radiative power (IFRP) over 1° × 1° pixels obtained from the MODIS active fire database and the average PM$_{2.5}$ concentration at the valley, the stippling indicates areas where the correlation is statistically significant, based on a p value less than 0.025. We found high correlation coefficients over the Magdalena Basin, northeast of the city, and, more remotely, around the Orinoco Basin, in eastern Colombia and northern Venezuela.

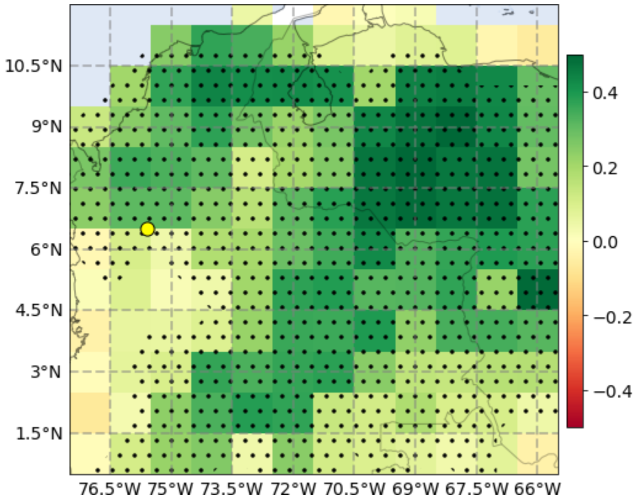

Figure 2. Spearman's correlation coefficient between the daily time series of daily PM$_{2.5}$ concentration in the Aburrá Valley and the daily cumulative IFRP over 1° × 1° pixels from the Aqua satellite (MODIS). The yellow point indicates the location of the Aburrá Valley, and the stippled areas indicate where the correlation is statistically significant, based on a p value less than 0.025.

Given that the area of interest to predict PM$_{2.5}$ depends on the wind field, we develop an approach using back-trajectories (BTs) in conjunction with the MODIS active fire data, acquired in near real time from https://firms.modaps.eosdis.nasa.gov/active_fire/ (last access: 14 December 2023) (documentation available at https://www.earthdata.nasa.gov/learn/find-data/near-real-time/firms/mcd14dl-nrt, last access: 14 December 2023); we compute the IFRP over a buffer of 50 km around a 4 d BT at 800 hPa (near-surface at the valley). The BT approach has been widely used for analysis and forecasting pollution over several places (Cobourn, 2010; Schneider et al., 2021; Wang et al., 2018; Liao et al., 2017; Louie et al., 2005) during recent years. The wind fields for the BT calculation are extracted from the GFS forecasts, and with forecast purposes, we use different wind fields for each future lead time but maintain the hotspots the same; a more detailed description of this process is presented in Appendix A. Figure 1d shows the daily-averaged time series of BT-based IFRP; the correlation coefficient with the time series of PM$_{2.5}$ is 0.59 (p value < 0.025), and good correspondence is found between the seasonal peaks and some high concentration events with the IFRP.

3 Models

One of the challenges when forecasting is to deal with large amounts of data from different sources. Tree-based models, apart from commonly providing accurate forecasts (Kang et al., 2018), allow the inclusion and relatively fast processing of a high number of input variables, which is convenient due to all the factors that affect the air quality within the region of interest. Decision trees (DTs) are the base method used in this study. In DTs, each tree branch node represents a choice between several alternatives, and each leaf node represents a decision (Kang et al., 2018; Quinlan, 1986). The RF and GB methods are sophisticated ways of implementing DTs. On the one hand, *random forest* performs by using bootstrap samples of the training data and random feature selection in tree induction, and prediction is achieved by averaging the predictions of each tree within the ensemble (Cutler et al., 2012). Alternatively, *gradient boosting*, or greedy function approximation, uses fitting functions, loss functions, and gradient descent analysis (Friedman, 2001). GB uses the negative gradient of the loss function as the residual approximation in the fitting tree algorithm and minimizes the loss function by reducing the residual value gradually (Zhang et al., 2021).

The main downside of using tree-based models is their low capacity to make predictions outside the domain in which they were trained (Meyer and Pebesma, 2021). To address that problem, we use the algorithm developed by Tao et al. (2021) and implemented within the LightGBM Python library (Ke et al., 2017). They modify the GB method to use piecewise linear regression trees (PL trees) as base learners. This modification allows the models to capture linear relationships and to extrapolate out of the domain in which the models were trained. In addition to the LightGBM library, the Sklearn Python package (Pedregosa et al., 2011) is used for training the RF models.

One relevant issue faced when aiming to predict pollutant concentrations in cities is that decision-makers and the general public are mainly interested in the forecast of cases in which concentrations of $PM_{2.5}$ are abnormally high or low. Given that the distribution of the target values for training is highly imbalanced, we modify the classic mean squared error (MSE) loss function by weighting the loss values based on the density within the probability distribution of the target concentrations (Steininger et al., 2021). As shown in Fig. 3, the error is multiplied by a penalty value that amplifies similarly to the target values located away from the mean.

Taking into account the data described above, the models include a total of 269 input variables that include the past 48 h $PM_{2.5}$ concentration at the station; CAMS-predicted aerosol optical depth; GFS-predicted total cloud cover, precipitation rate, and boundary layer height; predicted IFRP; day of the week; and hour of the day. The training and validation of the models go from 1 October 2017 to 20 April 2020, and their testing took place from 1 February 2021 to 31 April 2021; the period between 20 April 2020 and 1 February 2021, was excluded from the training, given the anomalous variability and abrupt changes in the emissions due to the COVID-19 pandemic. However, a brief assessment of the performance during COVID is provided in Appendix B. Before training the models, all the inputs for the different ML methods were standardized by removing the median and scaling according to the interquartile range (sklearn.preprocessing.RobustScaler(), https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.RobustScaler.html#sklearn.preprocessing.RobustScaler.transform, last access: 14 December 2023). Considering that the models run operationally and the wide variety of data sources that we use for running the models, it is possible for some input variables to not be available in real time during the operation. In these cases, the missing data points are replaced by zeroes, which would be equivalent to using the median value of the variable during the training period.

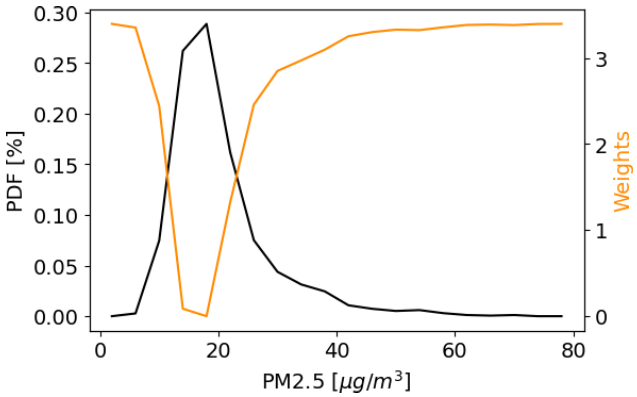

Figure 3. Probability distribution function of the 24 h averaged $PM_{2.5}$ concentrations at the station MED-LAYE (black line) and the weights associated with each bin that are used to modify the loss function for training the models (orange line).

The algorithms used have relatively low computational costs. The models were trained by a computer with an 8-core CPU and 16 GB of RAM. With this architecture, the models for all stations took around 4 h to be trained. However, once trained, these models allow $PM_{2.5}$ predictions 4 d ahead in just 20 min. That includes pre-processing of the input information, generation of predictions, and post-processing.

3.1 Direct multi-output and recursive schemes

In this study, we approach the problem of having multiple outputs in two ways. On the one hand, we use the direct multi-output strategy (MO), in which we train independent models to predict each of the 96 lead time hours. On the other hand, aiming to exploit dependencies among the target values to improve the skill of the models, we try a recursive scheme (RC) (Spyromitros-Xioufis et al., 2016); in this case, each model, except for the first hour of the forecast, also depends on the predicted 24 h averaged concentration before the objective lead time. For example, the predicted values for the lead times 1 and 2 are input variables for the third forecasted hour.

3.2 Hyperparameter selection

We performed a hyperparameter grid search with a 5-fold validation during the first period to find the optimal set (Ke et al., 2022). The final choice was based on the average root mean squared error (RMSE) of the 96 predicted hours. Table C1 shows the possible values tried for every hyperparameter. The selected set of hyperparameters for a model that predicts the mean 24 h $PM_{2.5}$ concentration in the city is highlighted.

4 Results

We use the root mean squared error (RMSE), the correlation coefficient, and the mean absolute percentage error (MAPE) as evaluation metrics for the trained models; the RMSE and MAPE computation includes multiplying by the weight values described in Sect. 3 and shown in Fig. 3. The RMSE penalizes large errors more, which is relevant due to the additional importance given to cases with high concentrations. The correlation coefficient indicates covariance between the predicted and true values, and the MAPE allows a fair comparison among different stations. Figure 4a and b show the RMSE and correlation coefficient averaged over all stations for each lead time for the four developed models, and a fifth one with the same input variables but using linear regression (LR), the emergence of collinearity, due to the high number of inputs (Gregorich et al., 2021), was addressed in the LR by implementing L2 regularization (ridge regression; McDonald, 2009). According to Fig. 4a, b, GB outperforms LR at all lead times, and GB presents lower RMSE and higher correlation coefficients than RF before hour 24. Although the performance of RF is similar to GB after day 1, it performs worse than LR and GB before the first 8 forecasted hours. GB–MO has the best performance for most lead times and stations. The RMSE for GB–MO goes from 1.3 to around 7 $\mu g\, m^{-3}$ during the 96 h lead time, and the correlation goes from 1 to 0.3. Although average correlation coefficients are low (below 0.5) after 2 d of forecast, the performance of the trained models is still better than the original CAMS $PM_{2.5}$ forecast, which presents correlation coefficients lower than 0.3 and RMSEs higher than 9 $\mu g\, m^{-3}$ starting from day 0 for most of the stations. Following these results, GB–MO will be used for the rest of the paper.

According to Figs. 4a and c and 8, most of the system's memory attributable to past $PM_{2.5}$ concentrations is lost by around 24 h. The relatively low MAPE in the first 24 h could be a consequence of the use of $PM_{2.5}$ observations in forecasts, which greatly improve the results near initialization, but their influence falls as the lead time increases (see Fig. 8). After the first 24 h, other input variables like IFRP and AOD from CAMS contribute the most to the predictability. For the first lead times, MAPE varies between 8 % and 12 % among stations. After hour 24, it ranges from 18 % to 40 %. Overall, the best performance is achieved over stations EST-HOSP, MED-ALTA, ITA-CJUS, ITA-CONC, ENV-HOSP, and MED-TESO, located south of the valley. The stations that show the worst performance are SUR-TRAF and CEN-TRAF, two stations in which traffic is highly relevant, and CAL-JOAR, MED-BEME, and COP-CVID. According to Hernández et al. (2022), small-scale urban dynamics can significantly affect the representation of wind and temperature regimes in the center of the valley and to the north, where the valley is also narrower and global models such as GFS and CAMS may have a worse performance. A poor representation of these features may explain the spatial differences in error; however, a deeper analysis of the particularities of each station is necessary to assess the causes of the higher errors. Small-scale dynamics or emissions may also cause a reduction in the models' performance.

Figure 5 shows the scatter plots between the predicted and observed 24 h averaged $PM_{2.5}$ concentration for lead times of 1, 12, 24, and 48 h at the stations MED-SELE, MED-LAYE, and SUR-TRAF. Thanks to the customized loss function and the inclusion of the PL trees as base learners, the models accurately represent concentrations that are significantly separated from the mean. Comparisons were made with a model trained with MSE as a loss function and traditional GB (without the PL trees; not shown), and the conventional model struggled to predict high or low concentrations relative to the mean, even from the first lead times of forecast. One of the limitations of the models is that, after hour 24, they exhibit a positive bias (see Fig. 5 at hours 24 and 48), and, as expected, the representation of events with significantly high or low concentrations gets worse. One of the reasons for the inadequate estimation of these events could be the lack of those values during the training period. Additionally, although the modifications to the loss function and the GB method helped, the inability of tree-based models to extrapolate is a vulnerability for the long-term sustainability of this strategy. However, given the low computational power needed, the models can be re-fitted frequently (e.g., monthly or every 3 months) as more data are available. Other sources of error include the initial and boundary conditions used for CAMS and GFS, which are later used as inputs in our models. The use of other model architectures, or the improvement of the inputs representing the meteorology in the region by implementing high-resolution numerical modeling, could complement these results and lead to a more robust and reliable forecast, especially given the high number of new regulations and governmental decisions that could influence air quality in urban areas.

Despite these performance limitations, the model is still helpful for decision-makers in the city. During the testing period, the night of the 6 March and early morning of 7 March presented an event with generalized poor conditions in terms of air quality over the metropolitan area of the Aburrá Valley.

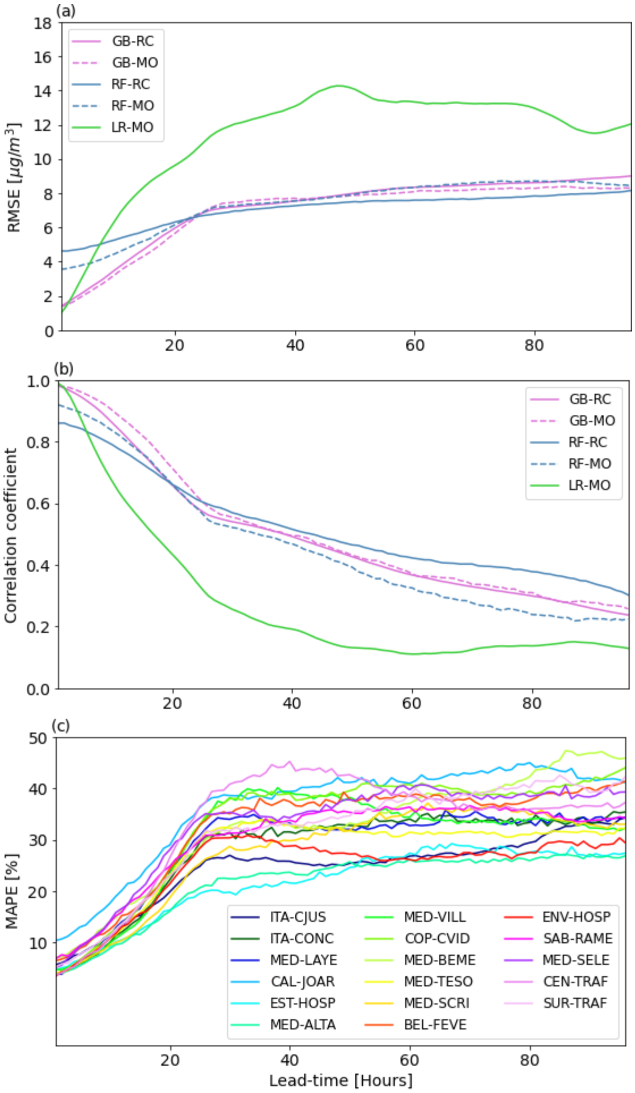

Figure 4. **(a)** RMSE at different lead times for the four developed tree-based models and for a fifth one with the same input variables but with a linear regression. **(b)** Same as **(a)** but for the correlation coefficient. **(c)** Shows the MAPE for the GB–MO model at all the stations of interest.

Two population-representative stations presenting the worse concentrations were CAL-JOAR and ITA-CJUS. Figure 6 shows a sudden increment in the 24 h average concentration at both stations of around 15 $\mu g\,m^{-3}$, leading to unhealthy conditions for sensitive groups over a prolonged time. The sudden increment was observable in the two stations' forecasts more than 24 h in advance (see blue lines in Fig. 6). To check the relevance of some variables for predicting this event, we trained the same model but excluded one of the inputs each time. AODs from CAMS and BT-IFRP were crucial to simulate the high concentrations in advance; in the tests excluding them (see green and orange lines in Fig. 6), the models did not capture the high magnitude of the peak.

Figure 7 shows how the error (MAPE and RMSE) varies during the day for different lead times. While there are no marked differences among different hours of the day for the first 12 h within the forecast range, a gradient in error starts developing in the early morning at a lead time of around 18 h, with a peak in error that starts in the early morning and shifts towards the afternoon as the lead time increases. The gradient becomes the sharpest at around 30 h. This pattern repeats again for the third, fourth, and fifth days of forecasts, but for these days, the peak in error presents at around 10:00. After noon, the error decreases slightly again in the hour-of-the-day dimension. As recent studies about air quality in the Aburrá Valley mentioned (Herrera-Mejía and Hoyos, 2019; Hoyos et al., 2020; Roldán-Henao et al., 2020), one of the most significant sources of uncertainty is the time of the boundary layer's expansion, which is also closely related to the turbulent processes occurring near the surface. Compared to global models (like GFS), regional models such as the Weather Research and Forecasting (WRF) model have proven a better performance in representing localized meteorological conditions (Posada-Marín et al., 2019; Gutowski et al., 2020). This is especially important in complex terrain regions such as the Aburrá Valley (Henao et al., 2020). Thus, the improvement in the representation of dynamics and thermodynamics by WRF simulations could provide a better representation of small-scale processes responsible for the intra-diurnal variability of the $PM_{2.5}$ concentration and improve air quality forecasts in the region.

4.1 Variable importance

Aiming to have a better understanding of the sources of predictability and uncertainty for the forecasts, we use two complementary approaches to analyze the variables' importance: (i) the intrinsic method from GB or the mean decrease in the loss function (MDLF). This method quantifies feature importance based on the improvement in the loss function contributed by each feature across all splits in the tree (Ke et al., 2017). (ii) We also perform a permutation feature importance analysis, which measures how much the forecast error increases when randomly modifying each input variable. For the second approach, we repeated the computation 10 times during the evaluation period for each variable in order to get more robustness. The MDLF approach is computationally efficient, but it suffers from biases, especially when the cardinality of the input variables is high (Lundberg et al., 2018). On the other hand, despite the permutation importance method being generally accepted as reliable (Loecher, 2022), it is significantly more computationally expensive than the MDLF approach. Both methods struggle when dealing with input variables that are correlated among them; however, we performed the same analyses using principal component analysis to reduce the dimensionality of the inputs, and the features' importance remained the same.

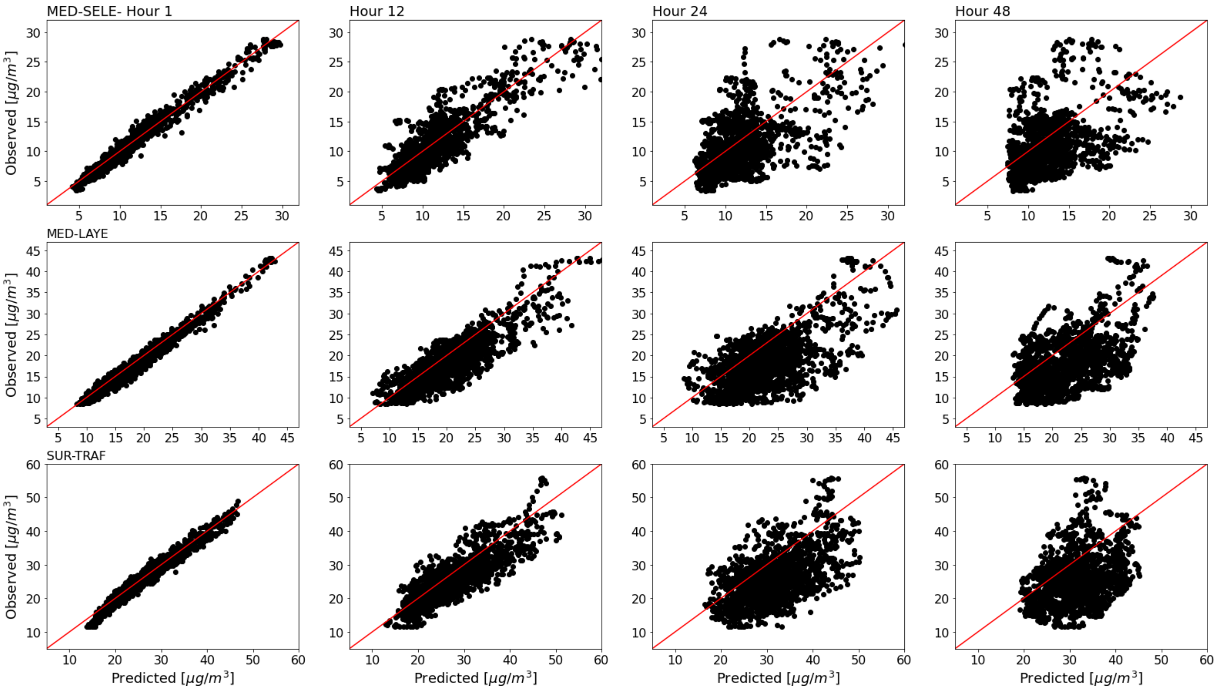

Figure 5. Scatter plots between the observed and predicted 24 h averaged PM$_{2.5}$ concentration at three different stations (MED-SELE, MED-LAYE, SUR-TRAF) and four different lead times (1, 12, 24, and 48 h.)

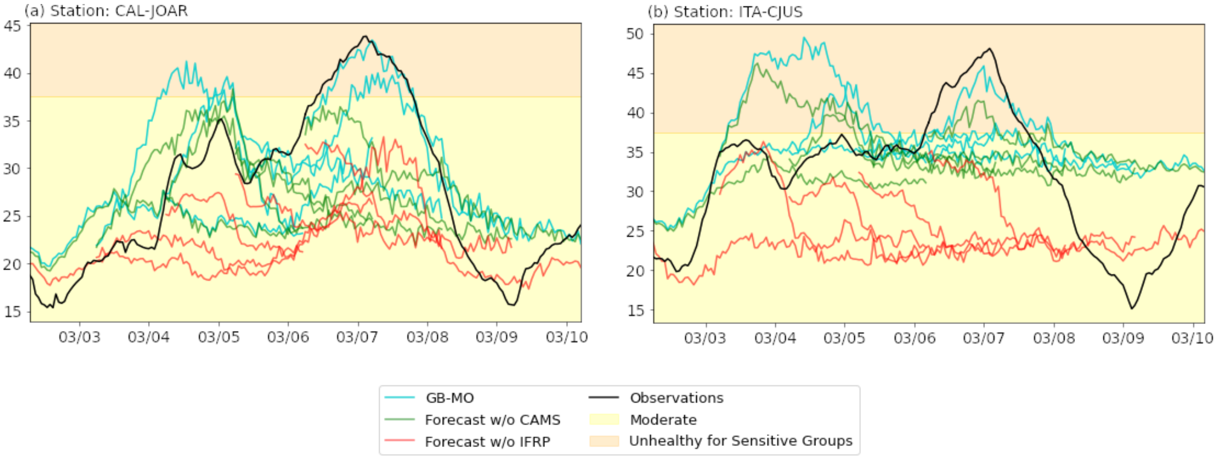

Figure 6. Observed 24 h averaged PM$_{2.5}$ concentration at the stations CAL-JOAR and ITA-CJUS (black line) and forecasts every 24 h before the peak concentration on 21 March 2021 at 06:00 UTC−5. The green and red lines indicate the forecast without considering CAMS and the IFRP index, respectively. The background colors indicate the corresponding air quality index (AQI) levels of concern.

Given the high number of inputs and stations, we summed the values corresponding to eight different macro-variables for each lead time: past PM$_{2.5}$ concentrations, AOD, TCC, precipitation rate (PRATE), height of the PBL (HPBL), IFRP, day of the week, and hour of the day. Figure 8a and d show the importance of each category from the two mentioned methods averaged over all stations. As expected, both methods agree that the relevance of past concentrations is maximum during the first hours of the forecast and decreases quickly until around the 30th hour. Also, the relevance of CAMS AOD increases with time, and it becomes the most important variable at around the 40th hour. Apart from those two variables, both methods coincide in that the BT-based IFRP index is the most useful variable between the 20th and the 40th hour of the forecast.

Relative to the past concentrations, CAMS AOD, and IFRP, the importance of the other variables considered in this study is significantly lower. The relative magnitude of

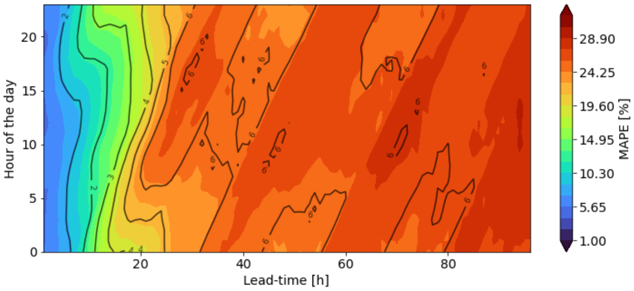

Figure 7. MAPE (filled contours) and RMSE (black contours) for the different lead times at different hours during the day averaged over all stations.

the importance of these other variables depends on the station, and there is a high level of agreement between the two methods. The analyses show that precipitation from GFS is significantly relevant after the 20th hour of the forecast for stations ITA-CONC, CAL-JOAR, EST-HOSP, MED-ALTA (see Fig. 8b), MED-VILL, MED-SCRI, CEN-TRAF, and SUR-TRAF. Similarly, planetary boundary layer height is relevant for stations CEN-TRAF, MED-ARAN, MED-TESO, MED-BEME, MED-VILL, and ITA-CJUS. Cloud cover was shown to be relevant after the 60th hour of the forecast for stations MED-LAYE, COP-CVID, MED-TESO, and MED-SELE (see Fig. 8c). Finally, the day of the week showed high importance in ITA-CJUS, MED-ALTA, MED-BEME, CEN-TRAF, and SUR-TRAF. These results correspond with the environmental characteristics and location of each station. In MED-SELE (Fig. 8c), the variables that affect the near-surface concentration at the base of the valley, such as HPBL or PRATE, are not as important, which is consistent with MED-SELE being significantly higher than the rest of stations. For SUR-TRAF and CEN-TRAF, whose variability is dominated mainly by vehicular emissions, and for other stations near the region's urban and commercial center, the day of the week showed predominance in the lead times after day 2.

5 Conclusions

Given the negative effects of poor-air-quality episodes in highly populated areas, forecasting $PM_{2.5}$ concentrations is paramount for decision-makers in cities, especially in developing countries like Colombia. In this study, we developed a framework for forecasting 24 h averaged $PM_{2.5}$ concentrations in the Aburrá Valley. In order to exploit the variety of useful information sources available, we combine tree-based machine learning algorithms with multiple data sources that are accessible in real time. Along the process, we developed a novel input variable for the models based on the radiative characteristics of fires occurring within an area of interest indicated by back-trajectories. Finally, we included a feature importance analysis that contributes to the understanding of the uncertainties and physical processes influencing short-term predictability over a complex terrain and densely populated region.

Both RF and GB performed significantly better than linear regression. However, the GB–MO strategy presented the best performance at almost all stations. The models were able to predict the episode with the poorest air quality during the testing period with more than a day of anticipation. Additionally, while one of the main limitations of tree-based methods is the low skill when dealing with predictions out of the domain of the training dataset, using a density-weighted loss function and implementing piecewise linear regression trees as base learners for the GB method helped overcome the issue.

According to the feature importance analysis, both the identification of external sources, such as wildfires, and local meteorological changes, especially those related to the boundary layer, proved to be relevant in predicting the occurrence of poor-air-quality events and the intra-diurnal changes in $PM_{2.5}$ concentration, respectively. The BT-based IFRP approach accomplished the objective of including near-real-time observations in the forecast and adding predictability by complementing the CAMS forecast. There is room for improvement regarding the precision of the fires' intensity, the error in the predicted wind fields, and the temporal evolution of fires.

Future work on the forecasts should be mainly focused on improving the representation of the diurnal variability within the models and reducing the biases after the first day within the lead times. A better understanding of the boundary layer processes in the urban area and the inclusion of more detailed meteorological forecasts could lead to performance improvements, mainly in the morning hours. An inadequate representation of the boundary layer could lead to false alarms when underestimating the expansion of the boundary layer and to missing high concentrations when overestimating it. Another factor that should be considered in the future is the formation of secondary particulate matter. Apart from the inclusion of $PM_{2.5}$ precursors that could lead to better accuracy, studies in the region that identify and characterize those compounds are also needed.

ML models show great potential to simplify, integrate, and apply large amounts of information for forecasting air quality conditions, which ultimately ends up as an excellent tool for decision-makers in the city. Furthermore, the knowledge acquired in the SIATA monitoring network during the past few years was vital to selecting the correct inputs and looking at the most conducive variables. In addition, without maintaining the records that span several years at the studied $PM_{2.5}$ stations, models' robustness would not have been possible, which motivates continued data collection at the respective locations.

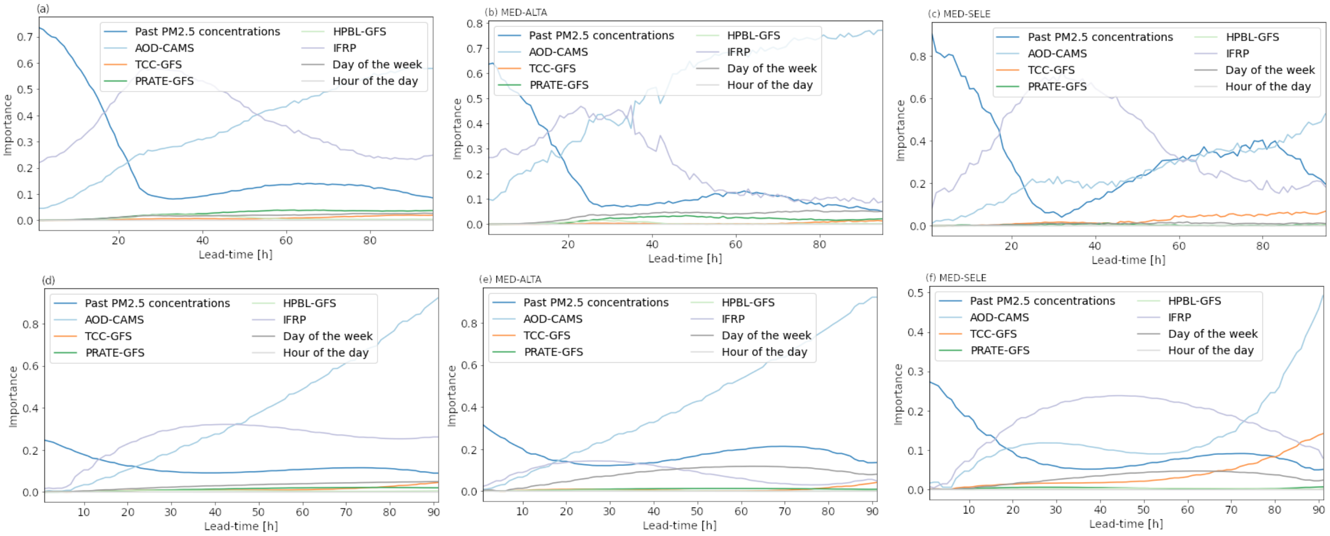

Figure 8. MDLF **(a–c)** and permutation **(d–f)** feature importance for the different groups of variables and lead times included in the forecast. Panels **(a)** and **(d)** show the feature importance when averaged for all stations. Panels **(b)** and **(e)** and **(c)** and **(f)** show the feature importance for stations MED-ALTA and MED-SELE, respectively.

Appendix A: Back-trajectory (BT) methodology for the BT-based integrated radiative power (IFRP)

Lagrangian analyses, and specifically back-trajectories, help analyze the origin of pollutants spatially, mainly regional long-range transport (Cobourn, 2010; Schneider et al., 2021; Wang et al., 2018; Liao et al., 2017; Louie et al., 2005). In the case of the Aburrá Valley, as shown in Sect. 2, fires occurring hundreds of kilometers away from the city have a significant and recurrent impact over the PM$_{2.5}$ concentration in the stations of interest.

The BT tracking is based on sequentially computing the past position of the parcel of air of interest in a given past time. The past coordinates of the parcel are computed as

$$\boldsymbol{X}(t-\Delta t)=\boldsymbol{X}(t)-\boldsymbol{V}(\boldsymbol{X},t-\Delta t)\cdot\Delta t, \tag{A1}$$

where $\boldsymbol{X}$ is the 3D spatial location of the parcel, t is the current time, Δt is the time step, and $\boldsymbol{V}(\boldsymbol{X},t-\Delta t)$ is the wind speed in the three directions.

In our case, the ending point of the parcels – or position at the time of interest (t_f) – is located horizontally near the valley's center and at a pressure slightly lower than the average of the region's surface, that is, longitude, latitude, and pressure of

$$\boldsymbol{X}(t_\mathrm{f})=\begin{bmatrix}75^\circ\,\mathrm{W}\\ 6.25^\circ\,\mathrm{N}\\ 800\,\mathrm{hPa}\end{bmatrix}. \tag{A2}$$

The algorithm computes the parcel's position 4 d back in time at a temporal resolution (Δt) of 1 h, and the forecasted wind fields for the trajectory calculation are extracted from the latest available GFS forecast.

After computing the BTs corresponding to parcels arriving at the valley at different lead times, we create a buffer of 50 km around the obtained trajectory. Finally, we integrate the radiative power of the fires located within the buffer to get the BT-IFRP index used as input for forecasting PM$_{2.5}$ concentrations.

Appendix B: Performance of the models during the COVID period

In order to further elucidate the models' limitations, we selected two additional evaluation periods: (i) a 10 d period during the COVID-19 pandemic, from 20 to 30 April 2020, for a comparison with the same 10 d of the test period in 2021, and (ii) a 30 d period (from 20 April to 20 May 2020) to get more robust statistics. On average, the 24 h PM$_{2.5}$ concentration for the 10 d test period during the pandemic is 3.7 µg m^{-3} lower than the corresponding 10 d within the study's test period (20 to 30 April 2021), which corresponds to a difference of around 20 %. Due to these differences, we used MAPE as the evaluation metric. Figure B1 shows that errors during the pandemic are comparable to those in the 2021 testing period and even lower for lead times of more than 24 h. Additionally, correlation coefficients are also comparable between the two periods (see Appendix Table B1).

Altogether, these results highlight the ability of the forecasts to generalize the predictions under different conditions. Nonetheless, future studies can perform longer training of the models, including data from subsequent years, which could result in an improvement of the forecast's skill.

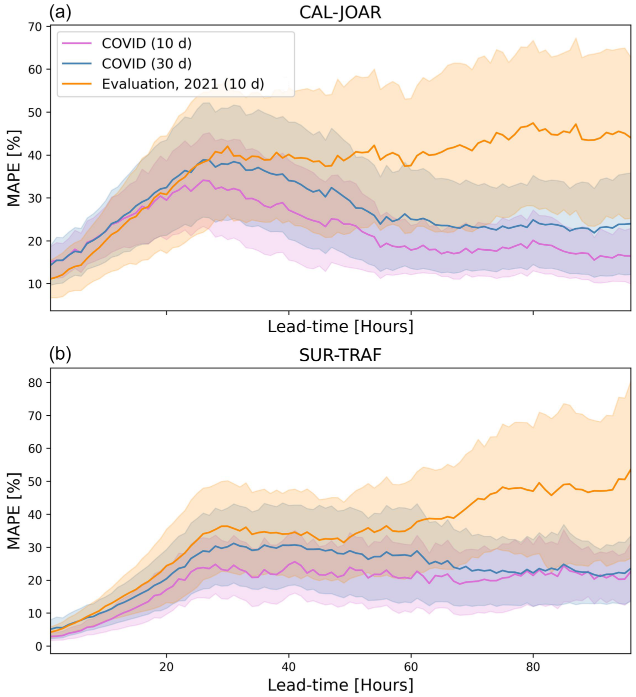

Figure B1. Mean absolute percentage error (MAPE) as a function of lead time for 10 d of the validation period (orange), the corresponding 10 d during the COVID pandemic (magenta), and the 30 d COVID evaluation period (blue). Panel **(a)** shows the results for the CAL-JOAR station, while panel **(b)** is for the SUR-TRAF station. Shading represents 1 standard deviation error.

Table B1. Pearson correlation coefficient computed for the different testing periods: the 10 d from 2021 (Eval10d), the 10 d from 2020 (COV10d), and the 30 d from 2020 (COV30d). The metric was calculated as a function of lead time (hours 12 and 24) and for different stations.

Station	Hour 12			Hour 24		
SUR-TRAF	Eval10d: 0.88	COV10d: 0.82	COV30d: 0.82	Eval10d: 0.62	COV10d: 0.52	COV30d: 0.41
CAL-JOAR	Eval10d: 0.86	COV10d: 0.78	COV30d: 0.72	Eval10d: 0.67	COV10d: 0.43	COV30d: 0.35
MED-SELE	Eval10d: 0.69	COV10d: 0.77	COV30d: 0.84	Eval10d: 0.34	COV10d: 0.19	COV30d: 0.52
MED-LAYE	Eval10d: 0.77	COV10d: 0.79	COV30d: 0.88	Eval10d: 0.43	COV10d: 0.39	COV30d: 0.57
ITA-CJUS	Eval10d: 0.79	COV10d: 0.70	COV30d: 0.71	Eval10d: 0.21	COV10d: 0.16	COV30d: 0.26
EST-HOSP	Eval10d: 0.86	COV10d: 0.77	COV30d: 0.82	Eval10d: 0.64	COV10d: 0.50	COV30d: 0.34
MED-VILL	Eval10d: 0.83	COV10d: 0.69	COV30d: 0.86	Eval10d: 0.42	COV10d: 0.11	COV30d: 0.57
MED-ARAN	Eval10d: 0.65	COV10d: 0.72	COV30d: 0.82	Eval10d: 0.02	COV10d: 0.24	COV30d: 0.51

Appendix C: The explored set of hyperparameters

Table C1 shows the hyperparameters explored during the optimization process. The selected set of hyperparameters for each model is highlighted.

Table C1. Grid of hyperparameters used in the hyperparameter optimization process. The best options based on the output average error in the 96 h of the forecast are in bold.

Method	Hyperparameter grid
GB–MO	Bagging fraction: (**1**, 0.8, 0.6, 0.4) Maximum number of leaves in one tree: (**5**, 10, 20, 40, 80) Minimum number of data in one leaf: (5, 10, 20, 40, 80,**160**, 320) Maximum depth of trees: (**0**, 5, 10, 20, 40) Extra trees: (“true”,**“false”**)
GB–RC	Bagging fraction: (**1**, 0.8, 0.6, 0.4) Maximum number of leaves in one tree: (5, **10**, 20, 40, 80) Minimum number of data in one leaf: (**5**, 10, 20, 40, 80, 160, 320) Maximum depth of trees: (0, 5, **10**, 20, 40) Extra trees: (“true”, **“false”**)
RF–MO	Bagging fraction: (1, 0.8, **0.6**, 0.4) Maximum depth of trees: (5, **10**, 20, 40, 80) Maximum number of features in tree: (5, 10, 20, 40, **80**)
RF-RC	Bagging fraction: (1, **0.8**, 0.6, 0.4) Maximum depth of trees: (5, **10**, 20, 40, 80) Maximum number of features in tree: (5, **10**, 20, 40, 80)

Code availability. The developed Python code is available at https://github.com/jhayron-perez/ForecastPM2.5-SIATA (last access: 19 December 2023; DOI: https://doi.org/10.5281/zenodo.10383573, Pérez-Carrasquilla, 2023).

Data availability. Data necessary for the training and evaluation of the models can be accessed at https://doi.org/10.5281/zenodo.7091239 (Pérez-Carrasquilla, 2022). For specific details, please contact the corresponding author or solicitud.datos@siata.gov.co.

Author contributions. JSPC and MR designed the research framework. PAM handled GFS and MODIS data. JMS handled CAMS data. JSPC handled in situ data and performed the training and evaluation of ML models. KSH contributed to implementing the customized loss function for the ML models and to assessing some of the most critical comments from the reviewers, including one comment related to the performance of the models during the COVID period. All authors contributed to writing and plotting the figures for the manuscript.

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

Disclaimer. Publisher’s note: Copernicus Publications remains neutral with regard to jurisdictional claims made in the text, published maps, institutional affiliations, or any other geographical representation in this paper. While Copernicus Publications makes every effort to include appropriate place names, the final responsibility lies with the authors.

Acknowledgements. We would like to thank the Sistema de Alerta Temprana de Medellín y el Valle de Aburrá (SIATA), a project of the Area Metropolitana del Valle de Aburrá (AMVA) managed by Universidad EAFIT, for the financial support. The authors acknowledge Carlos Hoyos and Laura Herrera for their work developing the SIATA monitoring network, their immense contributions to understanding aerosols’ behavior in the region, and the early ideas that finally led to this work. The authors would also like to thank Natalia Roldán, Maria Molina, Manuel Zuluaga, Alexis Hoffman, the two anonymous reviewers, and the handling editor, Chris Forest, for their insightful comments on the research framework and the manuscript. The authors acknowledge the technical support from Camilo Trujillo and the developers of Sklearn (Pedregosa et al., 2011), LightGBM (Ke et al., 2017), numpy (Harris et al., 2020), SciPy (Virtanen et al., 2020), and Matplotlib (Hunter, 2007) for the valuable software tools.

Financial support. This research has been supported by the Sistema de Alerta Temprana de Medellín y el Valle de Aburrá (SIATA), a project of the Area Metropolitana del Valle de Aburrá (AMVA) managed by Universidad EAFIT.

Review statement. This paper was edited by Chris Forest and reviewed by Alexis Hoffman and two anonymous referees.

References

Ballesteros-González, K., Sullivan, A. P., and Morales-Betancourt, R.: Estimating the air quality and health impacts of biomass burning in northern South America using a chemical transport model, Sci. Total Environ., 739, 139755, https://doi.org/10.1016/j.scitotenv.2020.139755, 2020.

Benedetti, A., Morcrette, J.-J., Boucher, O., Dethof, A., Engelen, R., Fisher, M., Flentje, H., Huneeus, N., Jones, L., Kaiser, J., Razinger, M., Schulz, M., Serrar, S., Simmons, A. J., Sofiev, M., Suttie, M., Tompkins, A. M., and Untch, A.: Aerosol analysis and forecast in the European centre for medium-range weather forecasts integrated forecast system: 2. Data assimilation, J. Geophys. Res.-Atmos., 114, D06206, https://doi.org/10.1029/2008JD011235, 2009.

Bond, T. C., Doherty, S. J., Fahey, D. W., Forster, P. M., Berntsen, T., DeAngelo, B. J., Flanner, M. G., Ghan, S., Kärcher, B., Koch, D., Kinne, S., Kondo, Y., Quinn, P. K., Sarofim, M. C., Schultz, M. G., Schulz, M., Venkataraman, C., Zhang, H., Zhang, S., Bellouin, N., Guttikunda, S. K., Hopke, P. K., Jacobson, M. Z., Kaiser,J. W., Klimont, Z., Lohmann, U., Schwarz, J. P., Shindell, D., Storelvmo, T., Warren, S. G., and Zender, C. S.: Bounding the role of black carbon in the climate system: A scientific assessment, J. Geophys. Res.-Atmos., 118, 5380–5552, 2013.

Breiman, L.: Random forests, Machine Learning, 45, 5–32, 2001.

Chellali, M., Abderrahim, H., Hamou, A., Nebatti, A., and Janovec, J.: Artificial neural network models for prediction of daily fine particulate matter concentrations in Algiers, Environ. Sci. Pollut. R., 23, 14008–14017, 2016.

Cobourn, W. G.: An enhanced $PM_{2.5}$ air quality forecast model based on nonlinear regression and back-trajectory concentrations, Atmos. Environ., 44, 3015–3023, 2010.

Cutler, A., Cutler, D. R., and Stevens, J. R.: Random Forests, 157–175, Springer US, Boston, MA, ISBN 978-1-4419-9326-7, https://doi.org/10.1007/978-1-4419-9326-7_5, 2012.

Dubovik, O., Holben, B., Eck, T. F., Smirnov, A., Kaufman, Y. J., King, M. D., Tanré, D., and Slutsker, I.: Variability of absorption and optical properties of key aerosol types observed in worldwide locations, J. Atmos. Sci., 59, 590–608, 2002.

Friedman, J. H.: Greedy function approximation: a gradient boosting machine, Ann. Stat., 29, 1189–1232, 2001.

Giglio, L., Descloitres, J., Justice, C. O., and Kaufman, Y. J.: An enhanced contextual fire detection algorithm for MODIS, Remote Sens. Environ., 87, 273–282, 2003.

Giglio, L., Schroeder, W., and Justice, C. O.: The collection 6 MODIS active fire detection algorithm and fire products, Remote Sens. Environ., 178, 31–41, https://doi.org/10.1016/j.rse.2016.02.054, 2016.

Gregorich, M., Strohmaier, S., Dunkler, D., and Heinze, G.: Regression with highly correlated predictors: variable omission is not the solution, Int. J. Env. Res. Pub. He., 18, 4259, https://doi.org/10.3390/ijerph18084259, 2021.

Guo, L.-C., Bao, L.-J., She, J.-W., and Zeng, E. Y.: Significance of wet deposition to removal of atmospheric particulate matter and polycyclic aromatic hydrocarbons: A case study in Guangzhou, China, Atmos. Environ., 83, 136–144, 2014.

Guo, W., Zhang, B., Wei, Q., Guo, Y., Yin, X., Li, F., Wang, L., and Wang, W.: Estimating ground-level $PM_{2.5}$ concentrations using two-stage model in Beijing-Tianjin-Hebei, China, Atmos. Pollut. Res., 12, 101154, https://doi.org/10.1016/j.apr.2021.101154, 2021.

Gutowski, W. J., Ullrich, P. A., Hall, A., Leung, L. R., O'Brien, T. A., Patricola, C. M., Arritt, R., Bukovsky, M., Calvin, K. V., Feng, Z., Jones, A. D., Kooperman, G. J., Monier, E., Pritchard, M. S., Pryor, S. C., Qian, Y., Rhoades, A. M., Roberts, A. F., Sakaguchi, K., Urban, N., and Zarzycki, C.: The ongoing need for high-resolution regional climate models: Process understanding and stakeholder information, B. Am. Meteorol. Soc., 101, E664–E683, 2020.

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M. H., Brett, M., Haldane, A., del Río, J. F., Wiebe, M., Peterson, P., Gérard-Marchant, P., Sheppard, K., Reddy, T., Weckesser, W., Abbasi, H., Gohlke, C., and Oliphant, T. E.: Array programming with NumPy, Nature, 585, 357–362, https://doi.org/10.1038/s41586-020-2649-2, 2020.

Henao, J. J., Mejía, J. F., Rendón, A. M., and Salazar, J. F.: Subkilometer dispersion simulation of a CO tracer for an inter-Andean urban valley, Atmos. Pollut. Res., 11, 928–945, 2020.

Hernandez, A. J., Morales-Rincon, L. A., Wu, D., Mallia, D., Lin, J. C., and Jimenez, R.: Transboundary transport of biomass burning aerosols and photochemical pollution in the Orinoco River Basin, Atmos. Environ., 205, 1–8, https://doi.org/10.1016/j.atmosenv.2019.01.051, 2019.

Hernández, K. S., Henao, J. J., and Rendón, A. M.: Dispersion simulations in an Andean city: Role of continuous traffic data in the spatio-temporal distribution of traffic emissions, Atmos. Pollut. Res., 13, 101361, https://doi.org/10.1016/j.apr.2022.101361, 2022.

Herrera-Mejía, L. and Hoyos, C. D.: Characterization of the atmospheric boundary layer in a narrow tropical valley using remote-sensing and radiosonde observations and the WRF model: the Aburrá Valley case-study, Q. J. Roy. Meteor. Soc., 145, 2641–2665, https://doi.org/10.1002/qj.3583, 2019.

Hoyos, C. D., Herrera-Mejía, L., Roldán-Henao, N., and Isaza, A.: Effects of fireworks on particulate matter concentration in a narrow valley: the case of the Medellín metropolitan area, Environ. Monit. Assess., 192, 6, https://doi.org/10.1007/s10661-019-7838-9, 2020.

Hunter, J. D.: Matplotlib: A 2D graphics environment, Comput. Sci. Eng., 9, 90–95, https://doi.org/10.1109/MCSE.2007.55, 2007.

Inness, A., Ades, M., Agustí-Panareda, A., Barré, J., Benedictow, A., Blechschmidt, A.-M., Dominguez, J. J., Engelen, R., Eskes, H., Flemming, J., Huijnen, V., Jones, L., Kipling, Z., Massart, S., Parrington, M., Peuch, V.-H., Razinger, M., Remy, S., Schulz, M., and Suttie, M.: The CAMS reanalysis of atmospheric composition, Atmos. Chem. Phys., 19, 3515–3556, https://doi.org/10.5194/acp-19-3515-2019, 2019.

Isaza Uribe, A.: Evaluación de la variabilidad temporal de la estructura termodinámica de la atmósfera y su influencia en las concentraciones de material particulado dentro del Valle de Aburrá, Escuela de Geociencias y Medio Ambiente, Master's thesis, Collections: Maestría en Ingeniería – Recursos Hidráulicos [171], Universidad Nacional de Colombia, Medellín, https://repositorio.unal.edu.co/handle/unal/69429 (last access: 19 December 2023), 2020.

Pérez-Carrasquilla, J. S.: jhayron-perez/ForecastPM2.5-SIATA: ForecastPM2.5-SIATA (v1.0.0), Zenodo [code], https://doi.org/10.5281/zenodo.10383573, 2023.

Justice, C., Giglio, L., Korontzi, S., Owens, J., Morisette, J., Roy, D., Descloitres, J., Alleaume, S., Petitcolin, F., and Kaufman, Y.: The MODIS fire products, Remote Sens. Environ., 83, 244–262, 2002.

Kang, G. K., Gao, J. Z., Chiao, S., Lu, S., and Xie, G.: Air quality prediction: Big data and machine learning approaches, Int. J. Environ. Sci. Dev, 9, 8–16, 2018.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., and Liu, T.-Y.: Lightgbm: A highly efficient gradient boosting decision tree, Adv. Neur. In., 30, 2017.

Ke, H., Gong, S., He, J., Zhang, L., Cui, B., Wang, Y., Mo, J., Zhou, Y., and Zhang, H.: Development and application of an automated air quality forecasting system based on machine learning, Sci. Total Environ., 806, 151204, https://doi.org/10.1016/j.scitotenv.2021.151204, 2022.

Lee, M., Lin, L., Chen, C.-Y., Tsao, Y., Yao, T.-H., Fei, M.-H., and Fang, S.-H.: Forecasting air quality in Taiwan by using machine learning, Scientific Reports, 10, 1–13, https://doi.org/10.1038/s41598-020-61151-7, 2020.

Lelieveld, J., Evans, J. S., Fnais, M., Giannadaki, D., and Pozzer, A.: The contribution of outdoor air pollution sources to premature mortality on a global scale, Nature, 525, 367–371, 2015.

Lepeule, J., Laden, F., Dockery, D., and Schwartz, J.: Chronic exposure to fine particles and mortality: an extended follow-up of the Harvard Six Cities study from 1974 to 2009, Environ. Health Persp., 120, 965–970, 2012.

Lewis, T. C., Robins, T. G., Dvonch, J. T., Keeler, G. J., Yip, F. Y., Mentz, G. B., Lin, X., Parker, E. A., Israel, B. A., Gonzalez, L., and Hill, Y.: Air pollution–associated changes in lung function among asthmatic children in Detroit, Environ. Health Persp., 113, 1068–1075, 2005.

Liao, T., Wang, S., Ai, J., Gui, K., Duan, B., Zhao, Q., Zhang, X., Jiang, W., and Sun, Y.: Heavy pollution episodes, transport pathways and potential sources of $PM_{2.5}$ during the winter of 2013 in Chengdu (China), Sci. Total Environ., 584, 1056–1065, 2017.

Lin, C.-Y., Chang, Y.-S., and Abimannan, S.: Ensemble multifeatured deep learning models for air quality forecasting, Atmos. Pollut. Res., 12, 101045, https://doi.org/10.1016/j.apr.2021.03.008, 2021.

Loecher, M.: Unbiased variable importance for random forests, Communications in Statistics – Theory and Methods, 51, 1413–1425, 2022.

Lorenz, E. N.: Three approaches to atmospheric predictability, B. Am. Meteorol. Soc, 50, 345–349, 1969.

Louie, P. K., Watson, J. G., Chow, J. C., Chen, A., Sin, D. W., and Lau, A. K.: Seasonal characteristics and regional transport of $PM_{2.5}$ in Hong Kong, Atmos. Environ., 39, 1695–1710, 2005.

Lundberg, S. M., Erion, G. G., and Lee, S.-I.: Consistent individualized feature attribution for tree ensembles, arXiv [preprint], https://doi.org/10.48550/arXiv.1802.03888, 2018.

Lv, L., Wei, P., Li, J., and Hu, J.: Application of machine learning algorithms to improve numerical simulation prediction of $PM_{2.5}$ and chemical components, Atmos. Pollut. Res., 12, 101211, https://doi.org/10.1016/j.apr.2021.101211, 2021.

Mabahwi, N. A. B., Leh, O. L. H., and Omar, D.: Human health and wellbeing: Human health effect of air pollution, Procedia – Social and Behavioral Sciences, 153, 221–229, 2014.

Mao, X., Shen, T., and Feng, X.: Prediction of hourly ground-level $PM_{2.5}$ concentrations 3 days in advance using neural networks with satellite data in eastern China, Atmos. Pollut. Res., 8, 1005–1015, 2017.

McDonald, G. C.: Ridge regression, Wiley Interdisciplinary Reviews: Computational Statistics, 1, 93–100, 2009.

Mendez-Espinosa, J., Belalcazar, L., and Betancourt, R. M.: Regional air quality impact of northern South America biomass burning emissions, Atmos. Environ., 203, 131–140, 2019.

Meyer, H. and Pebesma, E.: Predicting into unknown space? Estimating the area of applicability of spatial prediction models, Methods Ecol. Evol., 12, 1620–1633, 2021.

National Centers for Environmental Prediction, National Weather Service, NOAA, U.S. Department of Commerce: NCEP GFS 0.25 Degree Global Forecast Grids Historical Archive, NCAR [data set], https://doi.org/10.5065/D65D8PWK, 2015.

Orru, H., Maasikmets, M., Lai, T., Tamm, T., Kaasik, M., Kimmel, V., Orru, K., Merisalu, E., and Forsberg, B.: Health impacts of particulate matter in five major Estonian towns: main sources of exposure and local differences, Air Quality, Atmosphere & Health, 4, 247–258, 2011.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E.: Scikit-learn: Machine Learning in Python, J. Mach. Learn. Res., 12, 2825–2830, 2011.

Perez, P. and Gramsch, E.: Forecasting hourly $PM_{2.5}$ in Santiago de Chile with emphasis on night episodes, Atmos. Environ., 124, 22–27, 2016.

Pérez-Carrasquilla, J. S.: Forecasting 24-hour-averaged $PM_{2.5}$ concentration in the Aburrá Valley using tree-based ML models, global forecasts, and satellite information: Dataset, Zenodo [data set], https://doi.org/10.5281/zenodo.7091239, 2022.

Perišić, M., Maletić, D., Stojić, S. S., Rajšić, S., and Stojić, A.: Forecasting hourly particulate matter concentrations based on the advanced multivariate methods, Int. J. Environ. Sci. Te., 14, 1047–1054, 2017.

Posada-Marín, J. A., Rendón, A. M., Salazar, J. F., Mejía, J. F., and Villegas, J. C.: WRF downscaling improves ERA-Interim representation of precipitation around a tropical Andean valley during El Niño: implications for GCM-scale simulation of precipitation over complex terrain, Clim. Dynam., 52, 3609–3629, 2019.

Quinlan, J. R.: Induction of decision trees, Machine Learning, 1, 81–106, 1986.

Rincón-Riveros, J. M., Rincón-Caro, M. A., Sullivan, A. P., Mendez-Espinosa, J. F., Belalcazar, L. C., Quirama Aguilar, M., and Morales Betancourt, R.: Long-term brown carbon and smoke tracer observations in Bogotá, Colombia: association with medium-range transport of biomass burning plumes, Atmos. Chem. Phys., 20, 7459–7472, https://doi.org/10.5194/acp-20-7459-2020, 2020.

Rodriguez-Gomez, C., Echeverry, G., Jaramillo, A., and Ladino, L. A.: The negative impact of biomass burning and the Orinoco low-level jet on the air quality of the Orinoco River basin, edited by: Grutter, M., Atmósfera, 35, 497–520, https://doi.org/10.20937/atm.52979, 2022.

Roldán-Henao, N., Hoyos, C. D., Herrera-Mejía, L., and Isaza, A.: An Investigation of the Precipitation Net Effect on the Particulate Matter Concentration in a Narrow Valley: Role of Lower-Troposphere Stability, J. Appl. Meteorol. Clim., 59, 401–426, 2020.

Samet, J. M., Dominici, F., Curriero, F. C., Coursac, I., and Zeger, S. L.: Fine particulate air pollution and mortality in 20 US cities, 1987–1994, New Engl. J. Med., 343, 1742–1749, 2000.

Schneider, S. R., Lee, K., Santos, G., and Abbatt, J. P.: Air quality data approach for defining wildfire influence: Impacts on $PM_{2.5}$, NO_2, CO, and O_3 in Western Canadian cities, Environ. Sci. Technol., 55, 13709–13717, 2021.

Spyromitros-Xioufis, E., Tsoumakas, G., Groves, W., and Vlahavas, I.: Multi-target regression via input space expansion: treating targets as inputs, Machine Learning, 104, 55–98, 2016.

Steininger, M., Kobs, K., Davidson, P., Krause, A., and Hotho, A.: Density-based weighting for imbalanced regression, Machine Learning, 110, 2187–2211, 2021.

Tao, Q., Li, Z., Xu, J., Xie, N., Wang, S., and Suykens, J. A.: Learning with continuous piecewise linear decision trees, Expert Syst. Appl., 168, 114214, https://doi.org/10.1016/j.eswa.2020.114214, 2021.

Tian, J. and Chen, D.: A semi-empirical model for predicting hourly ground-level fine particulate matter ($PM_{2.5}$) concentration in southern Ontario from satellite remote sensing and ground-based meteorological measurements, Remote Sens. Environ., 114, 221–229, 2010.

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., Carey, C. J., Polat, İ., Feng, Y., Moore, E. W., VanderPlas, J., Laxalde, D., Perktold, J., Cimrman, R., Henriksen, I., Quintero, E. A., Harris, C. R., Archibald, A. M., Ribeiro, A. H., Pedregosa, F., van Mulbregt, P., and SciPy 1.0 Contributors: SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python, Nat. Methods, 17, 261–272, https://doi.org/10.1038/s41592-019-0686-2, 2020.

Wang, P., Zhang, H., Qin, Z., and Zhang, G.: A novel hybrid-Garch model based on ARIMA and SVM for $PM_{2.5}$ concentrations forecasting, Atmos. Pollut. Res., 8, 850–860, 2017.

Wang, X., Wei, W., Cheng, S., Li, J., Zhang, H., and Lv, Z.: Characteristics and classification of $PM_{2.5}$ pollution episodes in Beijing from 2013 to 2015, Sci. Total Environ., 612, 170–179, 2018.

Xing, Y.-F., Xu, Y.-H., Shi, M.-H., and Lian, Y.-X.: The impact of $PM_{2.5}$ on the human respiratory system, J. Thorac. Dis., 8, E69–E74, https://doi.org/10.3978/j.issn.2072-1439.2016.01.19, 2016.

Xu, X., Tong, T., Zhang, W., and Meng, L.: Fine-grained prediction of $PM_{2.5}$ concentration based on multisource data and deep learning, Atmos. Pollut. Res., 11, 1728–1737, 2020.

Yang, G., Lee, H., and Lee, G.: A hybrid deep learning model to forecast particulate matter concentration levels in Seoul, South Korea, Atmosphere, 11, 348, https://doi.org/10.3390/atmos11040348, 2020.

Yang, J., Yan, R., Nong, M., Liao, J., Li, F., and Sun, W.: $PM_{2.5}$ concentrations forecasting in Beijing through deep learning with different inputs, model structures and forecast time, Atmos. Pollut. Res., 12, 101168, https://doi.org/10.1016/j.apr.2021.101168, 2021.

Zhang, T., He, W., Zheng, H., Cui, Y., Song, H., and Fu, S.: Satellite-based ground $PM_{2.5}$ estimation using a gradient boosting decision tree, Chemosphere, 268, 128801, https://doi.org/10.1016/j.chemosphere.2020.128801, 2021.

Zhang, X., Sun, J., Wang, Y., Li, W., Zhang, Q., Wang, W., Quan, J., Cao, G., Wang, J., Yang, Y., and Zhang, Y.: Factors contributing to haze and fog in China, Chinese Sci. Bull., 58, 1178–1187, 2013.