International Journal of Sustainable Development and Planning
Vol. 20, No. 12, December, 2025, pp. 5227-5246
Journal homepage: http://iieta.org/journals/ijsdp

Air Pollution Forecasting in Almaty Based on Meteorological Data Using Machine Learning for Sustainable Environmental Management

Yerzhan Domalatov[1,2], Katipa Chezhimbayeva[2,3], Bekzhan Issakov[2,4*], Aizhan Amirgalina[2,4], Makpal Zharkymbekova[2,5], Markhabat Sakitzhanov[2,5], Dinara Tokseit[2,6]

[1] Department of Economics, Management and Finance, Sarsen Amanzholov East Kazakhstan University, Ust-Kamenogorsk 070000, Kazakhstan
[2] Humboldt Innovation GmbH, Humboldt University of Berlin, Berlin 10099, Germany
[3] Department of Telecommunications Engineering, Almaty University of Power Engineering and Telecommunications named Gumarbek Daukeyev, Almaty 050013, Kazakhstan
[4] Department of Industrial Safety and Ecology, Abylkas Saginov Karaganda Technical University, Karaganda 100027, Kazakhstan
[5] Department of Electrical Power Engineering, Almaty University of Power Engineering and Telecommunications named Gumarbek Daukeyev, Almaty 050013, Kazakhstan
[6] Department of Information Security, L.N. Gumilyov Eurasian National University, Astana 010008, Kazakhstan

Corresponding Author Email: b.isakov@ktu.edu.kz

https://doi.org/10.18280/ijsdp.201217

Received: 3 October 2025
Revised: 20 December 2025
Accepted: 25 December 2025
Available online: 31 December 2025

Keywords:
air quality management, LSTM, machine learning, $PM_{2.5}$, PM_{10}, pollution forecasting, Random Forest, Smart City

ABSTRACT

Sustainable air quality management in large and rapidly growing megacities requires the implementation of forecasting systems capable of accounting for nonlinear interactions between meteorological conditions and the dynamics of suspended particles. Almaty, characterized by pronounced mountain-valley circulation and frequent winter inversions, is one of the cities in Central Asia where $PM_{2.5}$ and PM_{10} concentrations regularly exceed WHO recommendations. As part of the study, an interpretable model for short-term and conditional medium-term air pollution forecasting was developed based on Random Forest and LSTM algorithms using data from AQICN, AirKaz, Dashboard.air.org.kz, Ogimet and ERA5 for 2020–2024. Modelling was performed in two scenarios: (A) using only pollutant concentration lags and (B) adding a complete set of meteorological parameters, including temperature, relative humidity, wind speed, boundary layer height (BLH), surface pressure and cloud cover. Accuracy assessment at 7- and 30-day horizons showed that the inclusion of meteorological data significantly improves forecast quality, especially for $PM_{2.5}$, with Random Forest providing the most stable RMSE and MAE values. The LSTM model demonstrates high sensitivity to short-term peak values, more accurately reflecting the dynamics of pollution episodes. Feature importance analysis shows the key role of atmospheric stability (BLH), wind regime, and autocorrelation structure in the formation of winter smog situations. Compared to the baseline methods (Persistence and Seasonal Naïve), the forecast accuracy over a 7-day horizon shows poor performance and in some cases, is inferior to the "persistence" method, while over a 30-day horizon, it improved to 40% for $PM_{2.5}$ and to 15% for PM_{10}. The developed system has high potential for integration into digital monitoring platforms, early warning services, and Smart City solutions. The study fills an existing scientific gap in the field of interpretable weather-dependent air quality forecasting for cities with mountain-valley circulation in Central Asia and strengthens the analytical basis for sustainable environmental management.

1. INTRODUCTION

The problem of atmospheric pollution in large cities has once again become the focus of attention for researchers and urban policymakers in recent years, as growing urbanization, changing climatic conditions and increasing traffic loads pose serious challenges to the sustainable development of the urban environment. The scientific agenda is increasingly focused on the development of effective environmental monitoring and forecasting systems that can support urban planning, health risk management and the development of long-term environmental strategies. Almaty, Kazakhstan's largest metropolis, is one of the most characteristic examples of a city where the environmental situation creates significant barriers to achieving sustainable development goals.

According to data from the National Statistical Service of Kazakhstan [1], between 2020 and 2024, annual pollutant emissions in Almaty remained stable at 40–44 thousand tones, while winter concentrations of $PM_{2.5}$ exceeded WHO recommendations by 5–8 times [2]. The highest contribution comes from the Alatau, Zhetysu and Turksibsky districts, where large man-made sources are located: industrial hubs, thermal power plants, logistics and transport infrastructure. Per capita emissions in some districts reach 173 kg per year, which is significantly higher than in the central districts (Medeu, Almaly), reflecting the spatial asymmetry of the environmental burden. The factors driving this trend are well known: population growth (2.337 million people as of October 2025), building density (3,419 people/km^2), the characteristics of the heating season, and the city's unique foothill topography, which is prone to temperature inversions and the retention of polluted air masses.

Despite the fact that the city's current environmental protection expenditures have increased from USD 12.06 million in 2020 to USD 20.14 million in 2024, there has been no systematic improvement in air quality. The increase in the number of vehicles, the expansion of private heating, and the growth of local industrial emissions are exacerbating the concentration of fine particulate matter $PM_{2.5}$ and PM_{10}. This trend is consistent with global urbanization trends identified by Burnett et al. [3] and is observed in megacities with similar mountain-valley morphology, including Tehran, Ulaanbaatar and Santiago [4, 5]. Recent data from AQICN, AirKaz, Dashboard.air.org.kz and research by Kerimray et al. [6] confirm the operation of a similar mechanism for the formation of a ‘winter bowl’ in Almaty, making the city a key model site for studying weather-dependent smog episodes and developing tools for sustainable environmental management.

The relevance of the study is reinforced by the lack of comprehensive pollution forecasting systems based on modern machine learning (ML) and deep learning (DL) methods that are capable of accounting for nonlinear interactions between $PM_{2.5}/PM_{10}$ concentrations, meteorological parameters (temperature, humidity, wind speed, boundary layer height - BLH, surface pressure, cloud cover) and the topographical heterogeneity of the urban environment. In addition, expanding the set of meteorological parameters to include atmospheric stability characteristics (BLH) and wind conditions significantly enhances the physical validity of forecast models. Despite the expansion of the monitoring network to 71 stations (2024), most studies for Almaty are descriptive in nature, and there are no tools for short- or medium-term pollution forecasting. This creates a critical gap between scientific knowledge and the practical needs of urban policy: without high-precision forecasting systems, it is impossible to develop early warning services, adaptive transport management, environmental health risk assessment, and sustainable urban infrastructure planning.

In these circumstances, the aim of the study is to develop and empirically evaluate a weather-dependent system for short-term and conditional medium-term forecasting of $PM_{2.5}$ and PM_{10} concentrations for Almaty using Random Forest and LSTM algorithms and integrating data from three monitoring networks and key meteorological characteristics. This approach allows us to combine intelligent data analysis methods with sustainable urban planning tasks, improving the basis for decision-making in the field of air quality management.

To achieve this goal, the following research tasks were set:

1. analyze natural, climatic and anthropogenic determinants of pollution;
2. create a unified database based on multiple network sources;
3. build two types of models (RF and LSTM) for 7- and 30-day horizons and assess the contribution of meteorological factors;
4. identify spatial heterogeneity of pollution and conduct a comparative analysis of models;
5. develop recommendations for integrating the results into urban monitoring systems and Smart City infrastructure.

Based on the research tasks, the following hypotheses were formulated:

H1: *Meteorological factors have a statistically significant impact on the short-term dynamics of $PM_{2.5}/PM_{10}$.*
H2: *The inclusion of meteorological parameters significantly improves the accuracy of ML models.*
H3: *The topography of Almaty creates a pronounced spatial asymmetry of pollution.*
H4: *LSTM more accurately predicts peak pollution, while Random Forest provides stability over medium horizons.*
H5: *Combining multi-network monitoring data improves the predictive power of models.*

The scientific novelty of the study includes several points. First, an integrated weather-dependent ML model for short-term forecasting of $PM_{2.5}$ and PM_{10} has been developed, combining data from public and private stations with meteorological parameters for the city of Almaty. Second, a comparative study of RF and LSTM models has been carried out under conditions of mountain-valley circulation and winter inversions. Thirdly, an original conceptual scheme of a ‘weather-dependent forecasting ecosystem’ was proposed, linking topography, climatic factors, the distribution of emission sources and the location of stations with pollution forecasts. Fourth, a quantitative analysis of the impact of the regional emission structure on $PM_{2.5}/PM_{10}$ concentrations in the short term has been carried out. Fifth, recommendations have been formulated for the integration of models into urban digital platforms, Smart City and PPP mechanisms in the field of environmental monitoring.

Thus, the study fills a key scientific and practical gap in the field of sustainable air quality management in Central Asia. The results provide an analytical and technological basis for the implementation of early warning systems based on interpretable ML/DL algorithms and can serve as a model for the development of a digital ecosystem for environmental monitoring in cities in Kazakhstan and the region.

2. LITERATURE REVIEW

The environmental situation in Almaty has been the subject of growing interest among the scientific community in recent decades due to consistently high concentrations of $PM_{2.5}$ and PM_{10} particulate matter, their seasonal variability, and their pronounced dependence on meteorological conditions. Existing studies on air quality in Almaty provide important retrospective observations, but they lack modern approaches to pollution forecasting based on ML methods. To justify the need for such an approach, this section reviews key scientific areas in the field of air quality forecasting, analyses modern ML/DL methods, assesses the role of meteorological factors, topography and anthropogenic sources, and identifies gaps that remain significant for Almaty.

2.1 Air quality studies in Almaty and Central Asia

One of the most significant studies of air quality in Almaty is the work of Kerimray et al. [6], based on data from Airkaz.org. The authors established a clear vertical stratification of pollution, significant variability of $PM_{2.5}$ within the city, and the key role of domestic heating during periods of temperature inversions. However, the study did not go beyond a statistical description of PM dynamics and did not use predictive models.

A comprehensive regional analysis conducted by Kozhagulov et al. [7] shows that over the past three decades, Central Asian countries have remained structurally dependent on fossil fuels: up to 78% of CO_2 emissions are associated with the combustion of energy carriers, and the main exports of Kazakhstan and Turkmenistan consist of raw materials. At the same time, limited financial resources, fragmented regional cooperation and the lack of a modern air quality monitoring system mean that the climate and environmental measures taken so far have not led to a significant reduction in atmospheric emissions. At the city level, the findings of Tursumbayeva et al. [8] demonstrate that cities in Central Asia, including Almaty, are forming a new global 'hot belt' of pollution: average annual $PM_{2.5}$ concentrations exceed WHO recommendations by 4.3–12.6 times, winter peaks are associated with stagnation and slow air mass transport, and coal combustion remains the dominant source of $PM_{2.5}$ in most of the cities studied. The authors also emphasize that official emissions inventories are often based on outdated methodologies, which hinders the development of scientifically sound air quality management strategies and further highlights the need to create modern, weather-dependent predictive models for urbanized areas in the region.

Satellite analysis of NO_2 over Kazakhstan [9] revealed spatial heterogeneity of pollution, but did not address forecasting issues and was limited to retrospective assessment.

Thus, existing studies on Almaty form an important empirical basis, but do not offer predictive ML/DL models, confirming the existence of a significant scientific gap.

2.2 Global studies of $PM_{2.5}$/PM_{10} dynamics under inversion conditions and complex terrain

Almaty, located in a mountain basin, has climatic characteristics similar to cities such as Tehran, Ulaanbaatar, Santiago de Chile, and Lahore. All of these cities are dominated by the effect of winter temperature inversions, which prevent vertical air mixing. Alizadeh-Choobari et al. [4] showed that inversions account for up to 70% of extreme pollution episodes in Tehran.

Wang et al. [5] demonstrated similar dynamics for Ulaanbaatar, where $PM_{2.5}$ can increase 8–10 times during periods of persistent anticyclones and night-time inversions.

Unlike many mountain cities, where increases in $PM_{2.5}$ concentrations are directly linked to air stagnation and increased temperature inversions, Santiago's experience shows more complex dynamics, in which not only the terrain but also the evolution of urban mobility and transport structure play a key role. According to Gallardo et al. [10], over the past three decades, concentrations of coarse PM_{10} particles in Santiago have decreased significantly thanks to the introduction of technological measures – improved fuel quality, the use of catalytic converters and diesel particulate filters. However, $PM_{2.5}$ concentrations remain high due to the rapid growth of the vehicle fleet and the transition of the city's atmosphere to a more oxidative regime, which enhances the formation of secondary aerosols. The authors emphasize that without changes in the transport behavior of the population—in particular, without an increase in the share of public transport—technological measures are insufficient to improve air quality.

These studies demonstrate that topography and meteorology create nonlinear pollution dynamics that cannot be adequately described by classical analysis methods. That is why, in global scientific practice, considerable attention is paid to the integration of meteorological parameters into ML forecasting models.

2.3 Meteorological determinants of $PM_{2.5}$ and PM_{10} concentrations

Numerous studies confirm that short-term fluctuations in $PM_{2.5}$ and PM_{10} are determined by a combination of temperature, relative humidity, wind activity and the height of the planetary boundary layer (PBL). These factors influence the processes of accumulation, dispersion, and secondary formation of aerosols, shaping the nature of pollution in urban environments.

Atmospheric air temperature is a key regulator of both particle formation processes and their temporal dynamics. According to Bai et al. [11], a decrease in perceived temperature is associated with an increase in the impact of $PM_{2.5}$ and NO_2 on the body, reflecting a general mechanism: at low temperatures, air stagnation increases, ventilation of the surface layer deteriorates, and the likelihood of solid particle accumulation increases. Similar physical processes are described in studies on COPD [12], which reveal U-shaped relationships between temperature and $PM_{2.5}$ exposure, particularly pronounced during prolonged periods of exposure (7–30 days). These results emphasize that extremely low and high temperatures exacerbate the negative effects of $PM_{2.5}$, indicating the importance of temperature as a pollution factor.

Air humidity has a significant effect on the hygroscopic growth of aerosols and the intensity of chemical reactions leading to the formation of secondary particles. A study by Niu et al. [13] showed that high humidity enhances the impact of the main components of $PM_{2.5}$ - nitrates, ammonium, black carbon and organic aerosols. Zender-Świercz et al. [14] found a consistent positive correlation between humidity and $PM_{2.5}$/PM_{10} concentrations in areas with 'fair & moderate' air quality, where hygroscopic growth of particles is most noticeable. These data confirm that high humidity contributes to an increase in particle mass, especially at low temperatures.

Wind activity determines the degree of dispersion of suspended particles. In light winds, $PM_{2.5}$ and PM_{10} concentrations increase due to limited horizontal transport. According to Purnomo et al. [15], an increase in wind speed leads to a decrease in the measured concentration of $PM_{2.5}$ from 25.2 to 16.4 μg/m^3 when the wind speed increases from 0.86 to 2.79 m/s. Although the study was conducted on sensors, it confirms the general aerodynamic principle: wind is the main mechanism of natural aerosol dilution.

The height of the planetary boundary layer (PBL) determines the vertical volume available for mixing pollutants. A lower PBL leads to a sharp increase in ground-level concentrations. Long-term studies in São Paulo show a consistent relationship between low PBL, temperature inversions, and increases in $PM_{2.5}$ [16]. The work of Han et al.

[17] confirms that the influence of the PBL is particularly pronounced for primary aerosols, and during periods of high pollution, the PBL exhibits different behavior depending on the measurement method, which is important for data interpretation.

It is important to note that in megacities with mountain-valley circulation, such as Almaty, meteorological factors are amplified by topography. A study by Kerimray et al. [6] showed that low temperatures, high humidity, and weak winds, combined with a reduced boundary layer height, lead to pronounced winter pollution peaks – an effect that had not previously been quantified using ML.

2.4 Machine learning and deep learning methods in air quality forecasting

In recent years, a separate body of work has emerged in the literature devoted to the use of ML and DL methods for short-term forecasting of $PM_{2.5}$ and PM_{10}. A review by Wu et al. [18] shows exponential growth in the number of such studies after 2015: ensemble trees (Random Forest, gradient boosting) and recurrent neural networks (LSTM/GRU), as well as their hybrids with spatial models, dominate.

2.4.1 Random Forest and gradient boosting

Ensemble decision trees remain one of the basic tools for predicting $PM_{2.5}$ concentrations, especially when mixed (meteorological and emission) predictors are available. Pan et al. [19] proposed a $PM_{2.5}$ prediction model based on Random Forest with subsequent interpretation using SHAP: it was shown that such models not only provide high accuracy ($R^2 > 0.9$ on validation), but also allow ranking the contribution of temperature, humidity, wind speed, and background pollution levels by importance. A similar conclusion is made by Alrashidi et al. [20] for monitoring stations in Kuwait, where ensemble methods (Random Forest, XGBoost) showed an advantage over classical regression approaches for predicting the air quality index based on $PM_{2.5}$.

2.4.2 LSTM and hybrid CNN-LSTM

Recurrent LSTM networks are used to model the temporal structure of pollution, taking into account the inertia of processes and the delayed effects of meteorological factors. Chang et al. [21] showed that the LSTM model provides a significant RMSE advantage over classical statistical models and simple neural networks when forecasting $PM_{2.5}$ and other pollutants in a metropolis, especially over a 24–48 hour horizon. Hybrid architectures have been further developed: Bai et al. [22] proposed a CNN-LSTM model in which the convolutional block extracts local spatio-temporal patterns between stations, and LSTM is responsible for dynamics over time; this scheme improved the accuracy of $PM_{2.5}$ forecasting and better reproduced episodes of high pollution.

There are still few direct DL studies for Kazakhstan. The closest to our work is the article by Yedilkhan et al. [23], which compares LightGBM and LSTM with an attention mechanism for $PM_{2.5}$ and PM_{10} forecasting based on meteorological data for the city of Almaty; LSTM with attention demonstrates the best RMSE values and better captures daily and seasonal variations in pollution. However, this work did not provide a detailed interpretation of the influence of individual meteorological factors and did not analyze spatial heterogeneity within the city, which leaves a methodological gap.

2.4.3 Interpretability of models and SHAP

A key focus in recent years has been the interpretation of ML model 'black boxes.' Wu et al. [18] emphasized that without explaining the contribution of individual features, DL predictions are difficult to use in environmental policy and urban management. Pan et al. [19] demonstrated that the use of SHAP makes it possible to quantitatively assess how changes in temperature, humidity, wind speed, and other variables shift the predicted $PM_{2.5}$ concentrations, and that such assessments are consistent with physical concepts of dispersion and accumulation of pollutants. A systematic review by Houdou et al. [24] shows that the combination of ensemble/neural network models with SHAP analysis is becoming the de facto standard in interpretable air quality forecasting, but there are still virtually no examples of its application for cities in Central Asia.

Thus, although global literature demonstrates a mature set of ML/DL tools for forecasting $PM_{2.5}/PM_{10}$, there are still no studies for Almaty and comparable mountain-valley megacities that simultaneously: (1) use an extensive network of stations, (2) explicitly take into account the weather-dependent nature of smog, and (3) apply interpretable models (e.g., Random Forest / gradient boosting + SHAP) to quantitatively assess the role of individual meteorological factors. This study fills this gap.

2.5 Research gap

Despite a significant increase in the number of studies on air quality in Central Asia and Almaty, the existing scientific literature remains fragmented and limited mainly to retrospective analysis of pollution. The work of Kerimray et al. [6] provides important insights into the spatiotemporal structure of $PM_{2.5}$ and PM_{10} in Almaty, including the influence of the heating season, inversions, and local sources, but there is a complete lack of short-term weather-dependent forecasting models. Similarly, previous studies show systemic features of pollution in Central Asia – dependence on fossil fuels, strong winter peaks, low boundary layer height and weak air ventilation – but do not contain predictive digital models that take into account the nonlinear effects of meteorological factors [7, 8].

Global studies on inversions and complex topography [4, 5, 10] emphasize that relief and temperature inversions form nonlinear $PM_{2.5}$ retention regimes that require the use of ML methods to adequately describe the dynamics. However, none of these studies apply to Almaty, despite the similarity of climatic conditions.

In the field of studying the influence of meteorological factors, contemporary literature demonstrates that temperature, humidity, wind speed, and the height of the PBL have a decisive influence on daily changes in $PM_{2.5}/PM_{10}$ concentrations. These factors in different climatic zones, but they do not focus on the mountain-valley conditions of Almaty and do not attempt to quantitatively integrate meteorological factors into ML models specifically for this city [11-13, 15-17].

At the same time, global experience in using ML/DL for pollution forecasting is growing rapidly: Random Forest, XGBoost, LSTM and CNN-LSTM demonstrate high prediction accuracy in various cities. However, in studies related to Kazakhstan, predictive models have been used to a limited extent: for example, Yedilkhan et al. [23] applied LSTM to Almaty, but the model did not take into account the extensive network of monitoring stations (71 points), did not analyze the meteorological dependence of smog, and did not interpret the factors using SHAP.

Thus, the scientific gap consists in the absence of a comprehensive weather-dependent ML model for short-term forecasting of $PM_{2.5}$ and PM_{10} for Almaty, which simultaneously:

- uses multi-network monitoring data (AQICN, AirKaz, Dashboard.air.org.kz);
- takes into account key meteorological factors (temperature, humidity, wind, PBL);
- reflects the characteristics of mountain-valley circulation;
- applies modern algorithms (Random Forest, LSTM) in comparative analysis;
- provides interpretation of the influence of factors on the forecast (SHAP).

It is this scientific gap that this study fills.

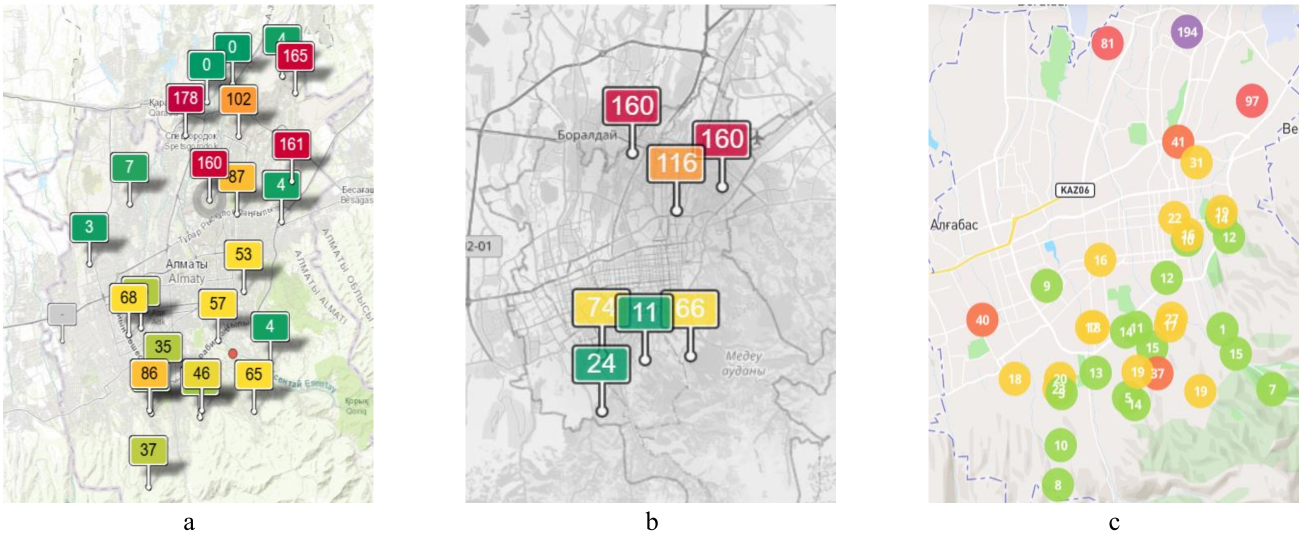
3. MATERIALS AND METHODS

3.1 Study area and observation network

The object of the study is the city of Almaty, Kazakhstan's largest metropolis, located in a foothill basin at the northern foot of the Trans-Ili Alatau. The city is characterized by pronounced mountain-valley circulation, frequent winter inversions and seasonal episodes of smog, making it an ideal testing ground for the development of weather-dependent $PM_{2.5}$ and PM_{10} forecasting models.

The monitoring network includes 71 air quality observation stations within the Almaty urban agglomeration. It consists of national government network stations, low-cost sensors integrated into the AirKz/Airkaz and Dashboard.air.org.kz mobile applications, Kazhydromet equipment, and stations of the World Air Quality Index (WAQI, AQICN) global network. The spatial distribution of monitoring points is shown in Figure 1: the central and northern parts of the city are characterized by a high density of residential and transport development and, accordingly, contain a group of stations that record conditions of increased anthropogenic load; the southern foothill zone contains stations at higher elevations, reflecting the influence of mountain-valley circulation and relatively better ventilation; the eastern and western areas are represented by a combination of residential areas and local industrial sites, providing representative coverage of different types of urban environments.

The spatial distribution of stations reflects the marked heterogeneity of air pollution in the city. Central areas (Almaly, Auezovsky) with high traffic loads show elevated background levels of $PM_{2.5}$ and PM_{10}, while in the Turksibsky and Zhetysu districts, individual industrial sites form local peaks. The southern foothill areas (Bostandyk, Medeu) have lower concentrations due to better ventilation, but are prone to pollution accumulation during winter inversions.



a b c

Figure 1. Geographical location of 71 pollutant monitoring stations in Almaty included in the analytical dataset (a: AQICN, b: AirKaz, c: Dashboard.air.org.kz)

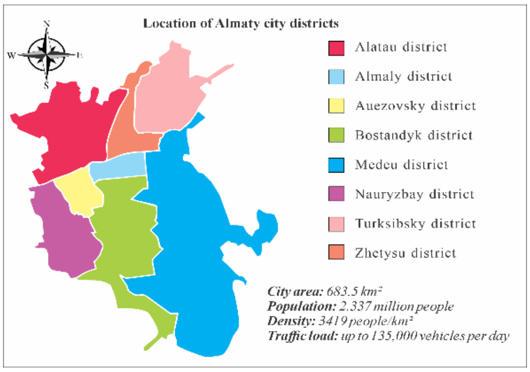


Figure 2. Location of Almaty City districts

Given this territorial heterogeneity, the study uses the median daily concentration as a stable urban integral indicator, reducing the impact of local emissions and data gaps. If necessary, the method can be extended to station-specific models. The cartographic location and names of the districts of Almaty are shown in Figure 2.

3.2 Data sources and observation period

The following open data sources were used to build predictive models:

- $PM_{2.5}$ and PM_{10} concentrations—hourly measurements from the city's air quality monitoring network, aggregated by the AQICN, AirKaz and Dashboard.air.org.kz platforms

(2020–2024 period);

- meteorological parameters—average daily relative humidity (Ogimet portal), as well as ERA5 reanalysis parameters, including air temperature wind speed, BLH, surface pressure and cloud cover (2020–2024).

Hourly PM values were converted to daily averages to align with WHO recommendations on daily limits and to reduce the impact of short-term emissions. The Ogimet and ERA5 meteorological series initially have a daily time step.

Only the primary time series AQICN / AirKaz / Dashboard.air.org.kz and Ogimet / ERA5 were used in the modelling. BNS statistical materials and WHO recommendations were used exclusively to describe the environmental situation in the city and were not included in the training sample.

3.3 Data pre-processing and quality control

Pre-processing consisted of three consecutive steps.

Completeness check. Only days for which at least 75% of valid hourly observations for the relevant indicator were available were included in the daily calculation. This filter is in line with international practice for ensuring the representativeness of daily air quality values [25, 26].

Emissions filtering. Unrealistic values (negative concentrations, extreme peaks associated with technical failures) were removed using range rules and subsequent time series analysis (an approach similar to that implemented in the open-air package [27].

Synchronization and interpolation. The median was calculated for each day based on the available stations. Single-day gaps in the final city series were filled using linear interpolation, provided that the length of the continuous gap did not exceed three days; longer intervals were marked as missing and were not used in model training.

At this stage, visual inspection of time series (graphs, swings, seasonality) was also performed, allowing for additional identification of anomalous areas and verification of PM consistency with meteorological data (increase in concentrations during periods of low temperatures and weak winds, etc.).

3.4 Regulatory thresholds and setting forecasting targets

In accordance with the WHO Air Quality Guidelines (2021), the study used the recommended daily air quality guidelines (AQG levels):

- for $PM_{2.5}$ – 15 μg/m^3;
- for PM_{10} – 45 μg/m^3.

Based on these, two interrelated tasks were formulated.

3.4.1 Regression (concentration forecast)

The regression model estimates the expected concentration of the pollutant *h* days ahead:

$$\widehat{y_t} + h = f(\mathrm{y_{t:t-L}}, \mathrm{m_{t:t-L}}, \mathrm{x_{t:t-L}}), \tag{1}$$

where, *y* is the target concentration of the pollutant ($PM_{2.5}$ or PM_{10}), *m* is meteorological variables, *x* is additional predictors (e.g., calendar features), *L* is the length of the historical window.

3.4.2 Classification of exceedances (early warning)

A binary variable was formed based on a regression forecast:

$$\begin{gathered}\mathrm{Z}_{\mathrm{t}+h} = \upuparrows (\mathrm{y}_{\mathrm{t}+h} > \mathrm{MPC}\,), \\ \hat{\mathrm{z}}_{\mathrm{t}+h} = \mathrm{g}(\mathrm{y_{t:t-L}}, \mathrm{m_{t:t-L}}, \mathrm{x_{t:t-L}}),\end{gathered} \tag{2}$$

where, *MPC* is the threshold value specified above.

This formulation corresponds to the applied task of urban environmental services—to issue a signal about the risk of exceeding the standard several days before the event.

3.5 Scenarios of signs and forecasting horizons

To assess the impact of meteorological conditions, two scenarios for the formation of input characteristics were considered:

- Scenario A (without meteorological parameters)—only $PM_{2.5}$/PM_{10} concentration lags for the previous day (up to 24 lags) are transferred to the model.
- Scenario B (with meteorological parameters)—in addition to concentration lags, lags of meteorological variables are included: relative humidity (Ogimet), as well as temperature, wind speed, BLH, surface pressure and cloud cover (ERA5). Values are generated for each parameter for the current and previous 1–7 days. The use of ERA5 data ensures statistical continuity of meteorological series for the entire period from 2020 to 2024.

Two forecast horizons were tested in both scenarios:

- 7 days—short-term operational forecast;
- 30 days—a conditional medium-term forecast, allowing the stability of models to be assessed over an extended time interval.

Before training, all input variables were scaled using the Min-Max method to the range [0;1] according to the parameters of the training sample, which eliminates information leakage between train and test [28].

3.6 Forecasting models

Two classes of models were used to construct forecasts [29, 30].

3.6.1 Random Forest (RF)

Random Forest is used in the study as an interpretable and robust ML algorithm capable of identifying nonlinear relationships between pollutant concentrations and meteorological factors.

For each scenario, the model was trained on lagged PM concentration values (24 previous steps) and, depending on the scenario, on lagged meteorological parameter values.

Scenario A (without meteorological data):

- only 24 lagged PM values are used.

Scenario B (with meteorological data):

- lags of the following meteorological variables are added to PM lags;
- air temperature;
- relative humidity;
- wind speed;
- BLH;
- surface pressure;
- cloud cover.

For each meteorological parameter, lag 0 denotes the most recent available historical observation at the forecast origin, while lags 1–7 correspond to preceding days; no future information was used in model training or forecasting. This feature set reflects the physical mechanisms of pollutant dispersion and inversions in the mountain-valley conditions of

Almaty.

Main RF hyperparameters:

-number of trees: 400 (scenario A) and 500 (scenario B);

-maximum depth: selected by cross-validation;

-random_state = 42 for reproducibility.

Feature importance indicators were also calculated, which made it possible to quantitatively assess the contribution of each meteorological factor to improving forecast accuracy.

3.6.2 LSTM (Long Short-Term Memory)

An LSTM recurrent neural network was used to account for the temporal structure of the data and the inertia of pollution accumulation processes. The input data for the network consisted of sequences with a length of 24-time steps (approximately 24 previous days of observations).

The output of the model was the PM concentration at a forecast horizon of 7 or 30 days.

LSTM architecture:

- one LSTM layer with 64 neurons;
- fully connected Dense layer with 32 neurons and ReLU activation function;
- one output neuron with linear activation (regression).

Training parameters:

- Adam optimizer;
- MSE loss function;
- 40 training epochs;
- batch size – 32;
- early stopping mechanism (patience = 5) to prevent overfitting.

In scenario B, the model input data included not only PM lags, but also the lagged meteorological parameters listed above.

All algorithms were implemented in Python using the NumPy, pandas, scikit-learn, and TensorFlow/Keras libraries.

3.7 Accuracy assessment and validation scheme

The division into training and test samples was performed strictly in chronological order without mixing. The final version of the study used a horizon-based hold-out validation scheme, which is methodologically consistent with the general principles of time samples described in the works [31, 32].

Separate test samples were formed for each modelling horizon:

- for a short-term forecast of 7 days, the test sample included the last 7 days of the time series;
- for a conditional medium-term forecast of 30 days, the test sample included the last 30 days of the series;

The quality of regression forecasts was assessed using metrics where y_t is the true value, $\hat{y}_t$ is the forecast, and N is the number of test points:

MAE (mean absolute error):

$$\mathrm{MAE} = \frac{1}{\mathrm{N}}\sum_{t=1}^{\mathrm{N}} |y_t - \hat{y}_t|. \tag{3}$$

RMSE (mean square error):

$$\mathrm{RMSE} = \sqrt{\frac{1}{\mathrm{N}}\sum_{t=1}^{\mathrm{N}} (y_t - \hat{y}_t)^2}. \tag{4}$$

MAPE (mean absolute percentage error):

$$\mathrm{MAPE} = \frac{100\%}{\mathrm{N}}\sum_{t=1}^{\mathrm{N}} \left|\frac{y_t - \hat{y}_t}{y_t}\right|. \tag{5}$$

MAPE is undefined at $y_t = 0$ and overestimates errors at low concentrations, so a safe value of the denominator $y_t + \varepsilon$ was used in the calculations. Comparing MAE and RMSE allows us to assess the sensitivity of the model to outliers: if RMSE is significantly higher than MAE, the model tends to be penalized for large errors at peak values.

For binary classification tasks, the following were calculated:

$$\mathrm{Accuracy} = \frac{\mathrm{TP+TN}}{\mathrm{TP+FP+TN+FN}}, \mathrm{Precision} = \frac{\mathrm{TP}}{\mathrm{TP+FP}}, \tag{6}$$

$$\mathrm{Recall} = \frac{\mathrm{TP}}{\mathrm{TP+FN}}, \quad \mathrm{F1} = 2 \cdot \frac{\mathrm{Precision \cdot Recall}}{\mathrm{Precision+Recall}}. \tag{7}$$

where, *TP* is correctly predicted exceedances, *FP* is false alarms, *TN* is correctly predicted ‘norms,’ and *FN* is missed exceedances. Particular attention was paid to the Recall (completeness) indicator, since it is critical for early warning systems to minimize the omission of dangerous episodes (FN).

The configurations and key characteristics of the models used are presented in Table 1, which summarizes the scenarios applied, input features, hyperparameters, and the main advantages and limitations of each approach.

Table 1. Parameters and characteristics of forecast models

Parameter	Random Forest	LSTM (Long Short-Term Memory)
Scenarios	A (without weather), B (with weather)	A (without weather), B (with weather)
Input Features	24 pollutant concentration lags; Scenario B additionally uses lagged meteorological variables (tmean, rh, wind speed, BLH, pressure, cloud cover)	24-step input sequences (24 previous observations); Scenario B additionally uses lagged meteorological variables (tmean, rh, wind speed, BLH, pressure, cloud cover)
Main Hyperparameters	n_estimators = 400 – 500; max_depth – auto; random_state = 42	64 LSTM neurons; 32 Dense (ReLU); Adam; 40 epochs; batch size = 32; early stopping (patience = 5)
Normalization Type	Min – Max scaling (based on training set)	Min-Max scaling (by training set)
Advantages	Robustness, interpretability, handling heterogeneous features	Captivates long-term dependencies and nonlinearities in time series
Limitations	Smoothing of extremes, limited adaptability to sudden changes	Requires normalization; sensitive to volume

The approach combines the interpretability of ensemble models with the adaptability of recurrent networks, which improves the reliability of short-term pollution forecasts and the applicability of results for environmental monitoring and air quality management systems.

3.8 Structural diagram of the methodology

The general sequence of modelling stages is shown in Figure 3. The diagram includes the following blocks:

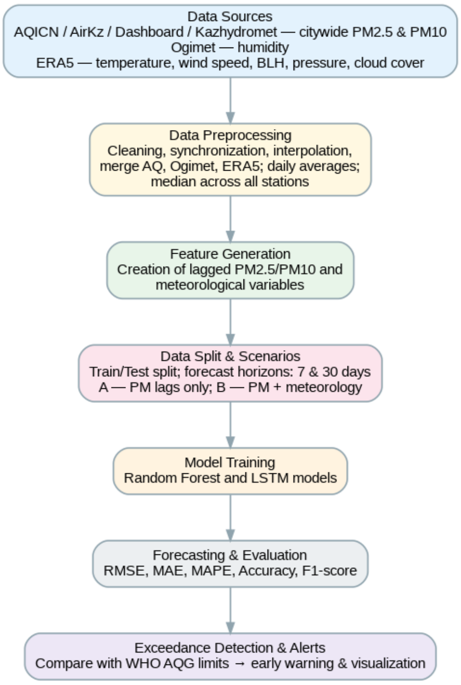

Data Sources
AQICN / AirKz / Dashboard / Kazhydromet — citywide PM2.5 & PM10
Ogimet — humidity
ERA5 — temperature, wind speed, BLH, pressure, cloud cover
Data Preprocessing
Cleaning, synchronization, interpolation,
merge AQ, Ogimet, ERA5; daily averages;
median across all stations
Feature Generation
Creation of lagged PM2.5/PM10 and
meteorological variables
Data Split & Scenarios
Train/Test split; forecast horizons: 7 & 30 days
A — PM lags only; B — PM + meteorology
Model Training
Random Forest and LSTM models
Forecasting & Evaluation
RMSE, MAE, MAPE, Accuracy, F1-score
Exceedance Detection & Alerts
Compare with WHO AQG limits → early warning & visualization

Figure 3. Schematic diagram of the forecasting methodology for pollutant concentration and exceedance detection

This methodological framework ensures transparency, reproducibility, and scalability of results: if additional stations or meteorological parameters become available, the algorithm can be easily scaled to new data sources and other cities.

4. RESULTS

4.1 Time series dynamics of pollutants and meteorological parameters

Visualization of time series of average daily concentrations of $PM_{2.5}$ and PM_{10}, as well as air temperature and relative humidity for the period 2020-2024, reveals the structure of seasonal and interannual fluctuations in atmospheric pollution in Almaty (see Figure 4).

A characteristic feature of the dynamics is stable winter pollution peaks. During the cold season, $PM_{2.5}$ and PM_{10} concentrations increase by 2-4 times compared to summer levels. These seasonal peaks are explained by:

(1) the active phase of the heating season, accompanied by an increase in emissions from coal and mixed heating systems;

(2) recurring episodes of temperature inversions;

(3) weakening of mountain-valley circulation and a decrease in wind speed.

Higher amplitudes of fluctuations are recorded for PM_{10}, reflecting the influence of dust emissions and mechanical resuspension. $PM_{2.5}$ shows a stable baseline level, and its increase in winter indicates the predominance of fine particles of anthropogenic origin. Temperature and humidity show typical climatic seasonality, confirming the correctness of data processing and the suitability of meteorological parameters for inclusion in forecasting models. Additional ERA5 parameters (wind speed, BLH, etc.) also agree with the identified seasonal phases, but are not included in the visualization to maintain the readability of the graph.

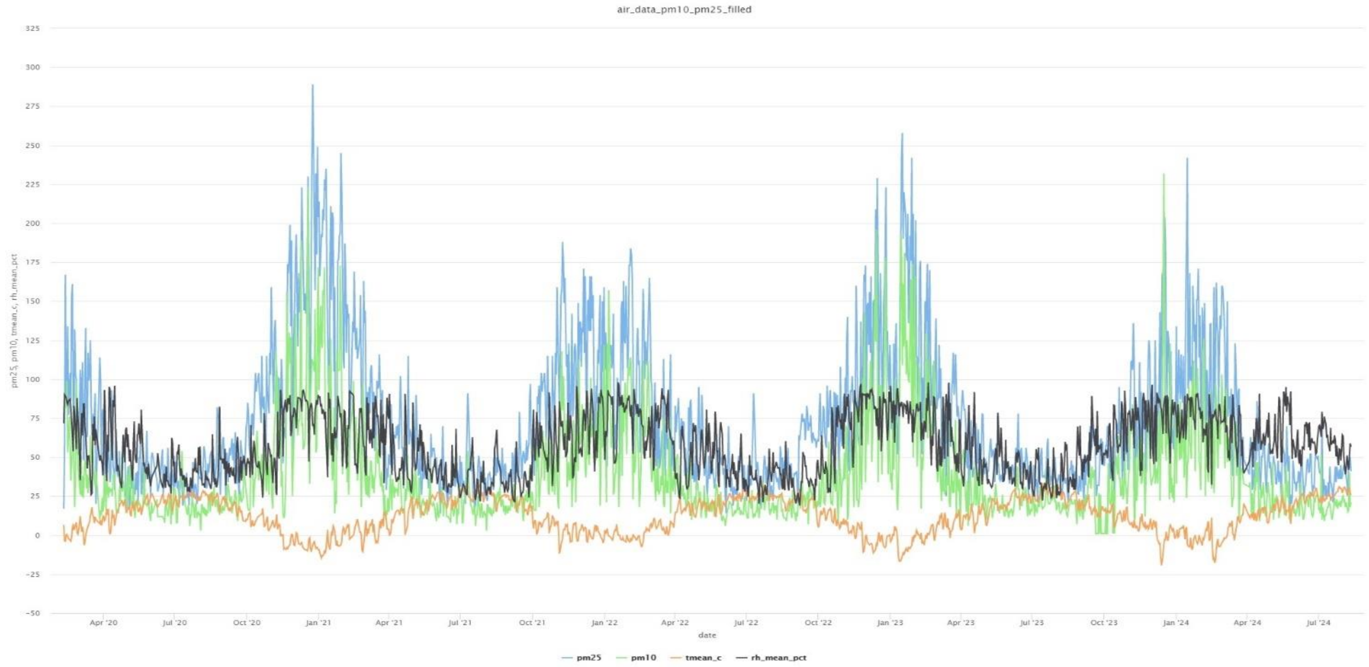


Figure 4. Time series of daily mean concentrations of $PM_{2.5}$ and PM_{10}, air temperature (°C), and relative humidity (%) during the period 2020-2024

4.2 Correlations between $PM_{2.5}$/PM_{10} concentrations and meteorological factors

Correlation analysis revealed statistically significant relationships between pollutants and weather variables (see Figure 5 and Table 2).

Key findings:

(1) air temperature

$PM_{2.5}$ and PM_{10} show a consistent negative correlation with temperature (r = –0.739 and r = –0.708, respectively), confirming the formation of winter smog under conditions of cooling of the surface layer, temperature inversions, and limited vertical turbulence.

(2) relative humidity

The correlation is moderately positive ($PM_{2.5}$: r = +0.483; PM_{10}: r = +0.504), reflecting the hygroscopic growth of particles and the intensification of secondary aerosol formation processes at elevated humidity.

(3) boundary layer height (BLH)

BLH has a pronounced negative correlation with pollutant concentrations ($PM_{2.5}$: r = –0.656; PM_{10}: r = –0.580), which is consistent with the mechanism of PM accumulation at reduced mixing layer heights characteristic of stagnant cold periods.

(4) wind speed

A weak negative correlation is observed ($PM_{2.5}$: r = –0.216; PM_{10}: r = –0.289), reflecting the role of wind activity in the dispersion and transport of pollutants.

(5) surface pressure

A positive correlation ($PM_{2.5}$: r = +0.461; PM_{10}: r = +0.516) indicates the influence of anticyclonic regimes, which contribute to stagnant conditions and increased PM concentrations.

(6) cloud cover

The correlation is weak ($PM_{2.5}$: r = –0.046; PM_{10}: r = –0.063), the influence is indirect and does not have a key effect on particle dynamics.

(7) relationship between $PM_{2.5}$ and PM_{10}

There is a strong correlation between the two fractions (r = +0.734), indicating common anthropogenic sources of pollution.

These results confirm the validity of including an extended set of meteorological parameters (ERA5 + Ogimet) in the ML model.

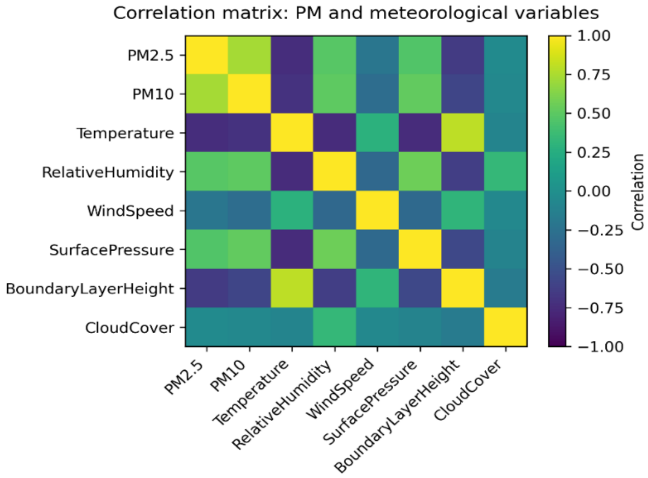


Figure 5. Correlation matrix between pollutant concentrations ($PM_{2.5}$, PM_{10}) and meteorological parameters

Table 2. Correlations between pollutants and meteorological parameters

Indicator	$PM_{2.5}$	PM_{10}	Temperature	Relative Humidity	Wind Speed	Surface Pressure	BLH	Cloud Cover
$PM_{2.5}$	1.000	0.734	–0.739	0.483	–0.216	0.461	–0.656	–0.046
PM_{10}	0.734	1.000	–0.708	0.504	–0.289	0.516	–0.580	–0.063
Temperature	–0.739	–0.708	1.000	–0.746	0.281	–0.747	0.802	–0.094
Relative Humidity	0.483	0.504	–0.746	1.000	–0.325	0.560	–0.630	0.331
Wind Speed	–0.216	–0.289	0.281	–0.325	1.000	–0.313	0.307	–0.067
Surface Pressure	0.461	0.516	–0.747	0.560	–0.313	1.000	–0.568	–0.106
BLH	–0.656	–0.580	0.802	–0.630	0.307	–0.568	1.000	–0.164
Cloud Cover	–0.046	–0.063	–0.094	0.331	–0.067	–0.106	–0.164	1.000

4.3 Results of modelling and forecasting pollutant concentrations

Two models were used to evaluate predictive capabilities:

- Random Forest – an interpretable ensemble model that is robust to nonlinearities;
- LSTM – a recurrent neural network focused on temporal dependencies and smoothing short-term fluctuations.

Both models were tested in two scenarios:

- A – pollutant lags only;
- B – pollutant lags + meteorological parameters (temperature, humidity, wind speed, BLH, pressure, cloud cover).

Forecasts were made for 7- and 30-day horizons.

Forecast accuracy assessment.

7-day horizon (see Figure 6 – ‘observed vs. predicted’ correspondence diagrams and Figure 7 – time series).

For $PM_{2.5}$, both models show satisfactory correspondence with observations, however:

- RF gives more stable predictions that are closer to the diagonal,
- LSTM better captures local variations but tends to smooth

them out.

For PM_{10}, the quality is noticeably lower:

- RF and LSTM predictions show dispersion and underestimation of sharp jumps,
- This reflects the higher variability of PM_{10} and its dependence on wind and dust processes.

The addition of meteorological data (scenario B) improves accuracy, especially for $PM_{2.5}$, as evidenced by the reduction in deviations from the diagonal.

30-day horizon (see Figure 8 – time series).

Over a long horizon, both models show a regular decrease in accuracy relative to the short-term forecast.

RF demonstrates more consistent reproduction of the overall dynamics of $PM_{2.5}$ and PM_{10}.

LSTM better follows the structure of the series, but tends to underestimate high values and smooth out peaks.

The inclusion of meteorological factors (scenario B) improves predictability, especially for PM_{10}, which is consistent with the findings of international studies [33, 34].

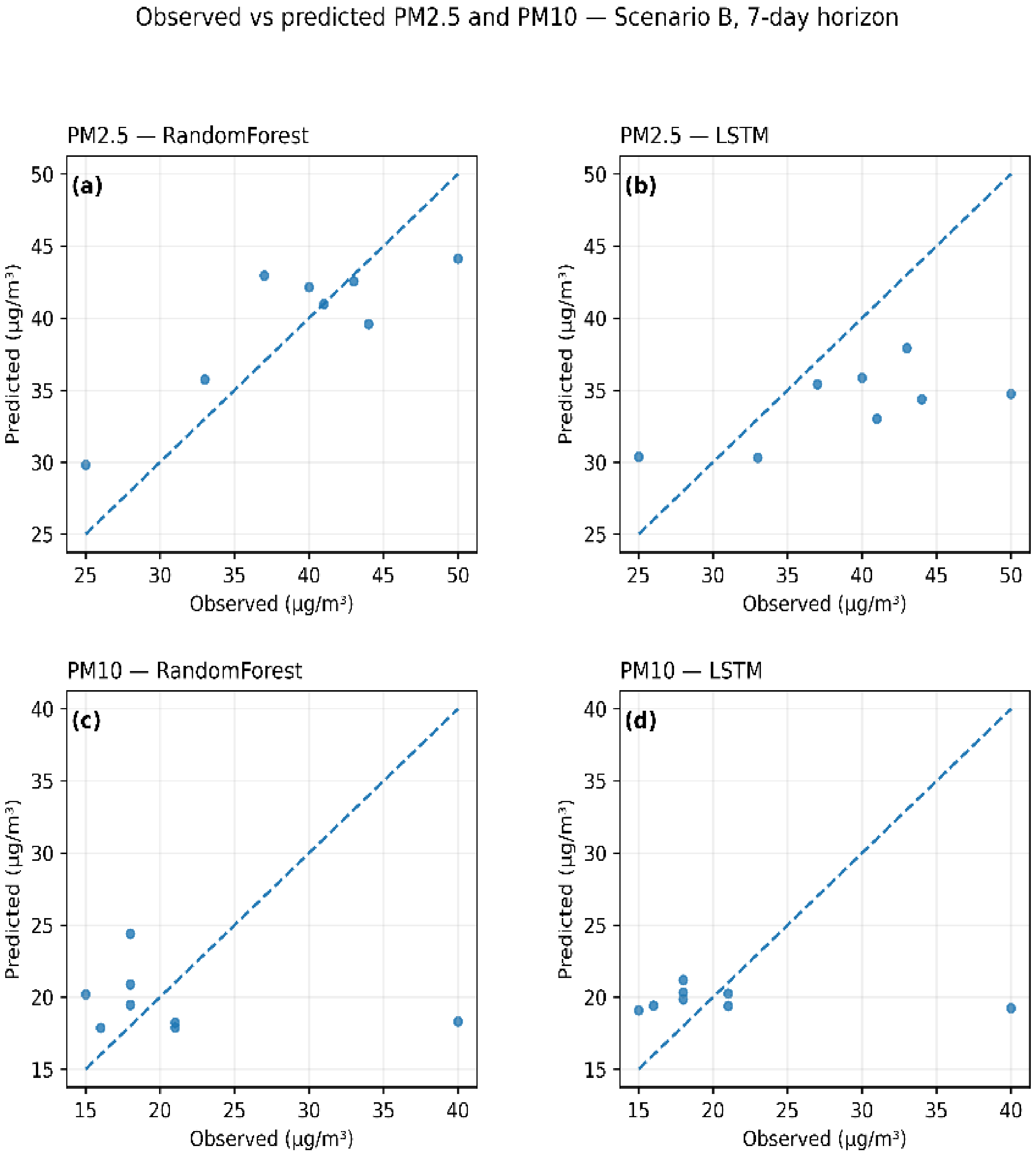


Figure 6. Comparison of actual and forecast pollutant concentration values over a 7-day period: (a) $PM_{2.5}$ – Random Forest, (b) $PM_{2.5}$ – LSTM, (c) PM_{10} – Random Forest, (d) PM_{10} – LSTM

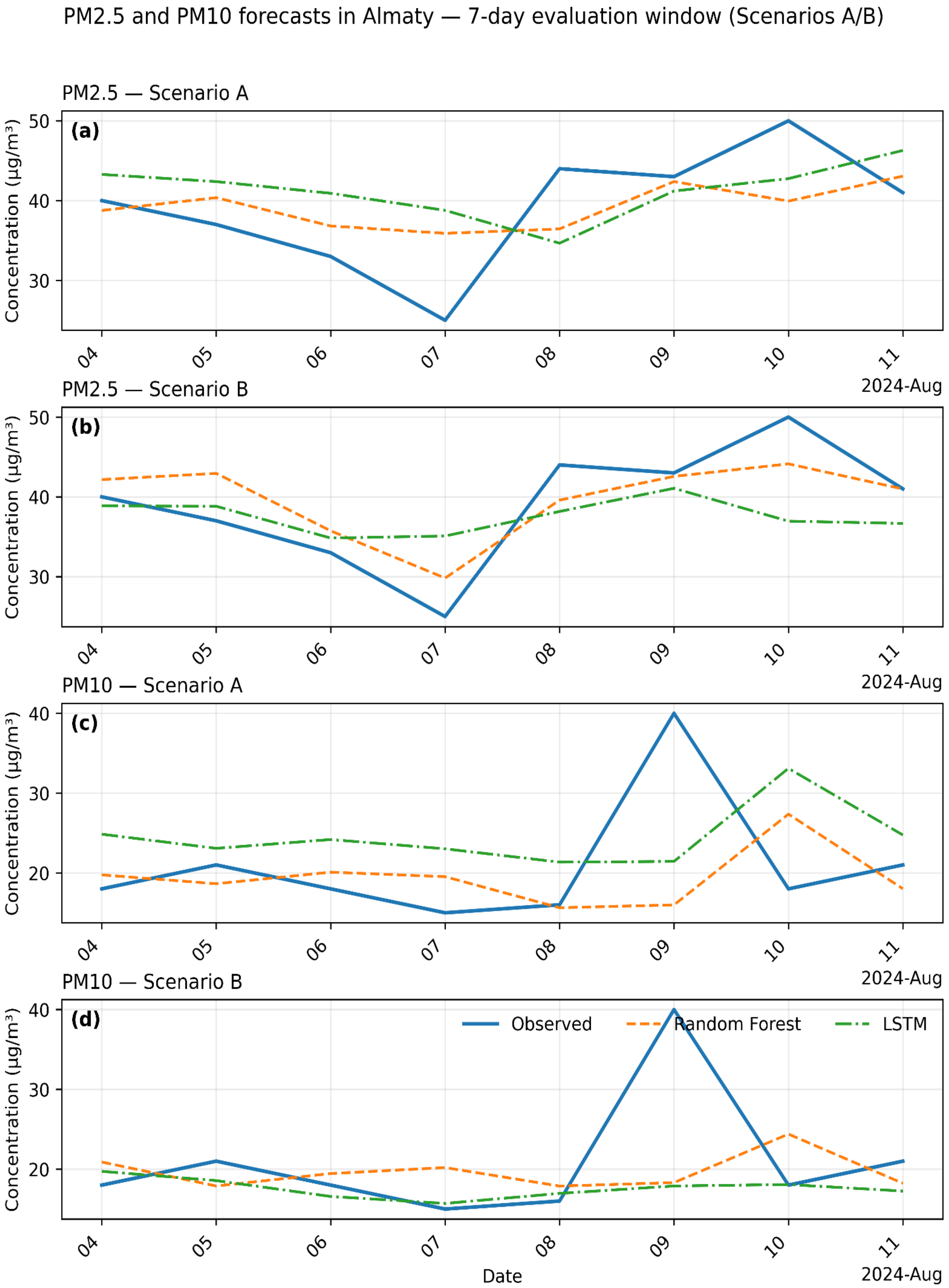


Figure 7. Results of the short-term forecast (7-day horizon). (a) $PM_{2.5}$ – scenario A; (b) $PM_{2.5}$ – scenario B; (c) PM_{10} – scenario A; (d) PM_{10} – scenario B. Each panel shows observations and forecasts of the Random Forest and LSTM models

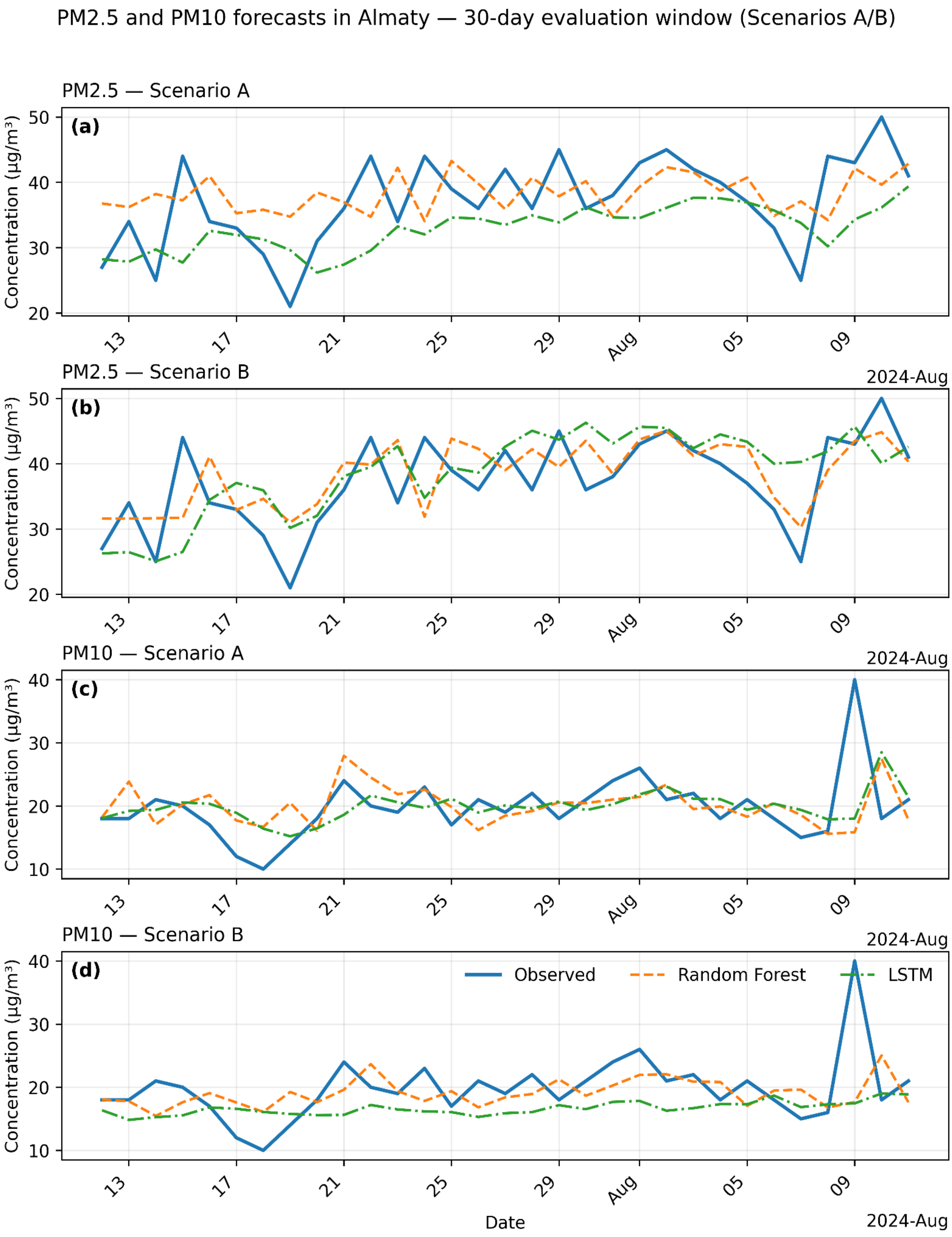


Figure 8. 30-day forecast results. (a) $PM_{2.5}$ – Scenario A; (b) $PM_{2.5}$ – Scenario B; (c) PM_{10} – Scenario A; (d) PM_{10} – Scenario B. Each panel shows the observations and forecasts from the Random Forest and LSTM models

4.4 Quantitative assessment of the accuracy of Random Forest and LSTM models

Table 3 reflects the values of RMSE, MAE, MAPE, Accuracy, and F1-score metrics for all combinations of 'model × scenario × forecast horizon.' The results demonstrate consistent differences in model behavior and the influence of meteorological factors on forecast accuracy.

$PM_{2.5}$

Random Forest shows the lowest errors at both horizons (RMSE 3.96–6.91 µg/m³).

LSTM has higher RMSE (6.49–7.64 µg/m³), but better reproduces short-term peaks.

Scenario B leads to a significant reduction in errors, especially at the 7-day horizon (36–43% improvement in RMSE).

The improvement is directly related to the addition of key meteorological parameters (temperature, humidity, wind speed, boundary layer height), which explain the dynamics of winter episodes of air stagnation.

PM_{10}

Over a 7-day horizon, Random Forest shows more stable errors (RMSE 8.44–9.40 µg/m³), while LSTM sometimes outperforms RF on a 30-day forecast.

Meteorological variables have a more noticeable effect on PM_{10} accuracy, reflecting the high sensitivity of coarse particles to wind conditions and atmospheric stratification.

It should be noted that the values Accuracy = 1.00 and F1-score = 0.00 for PM_{10} are due to the absence of threshold exceedances in the test dataset, which makes the classification metrics non-functional but does not affect the interpretation of the regression results.

The inclusion of an extended set of meteorological parameters (Ogimet + ERA5) significantly improves the accuracy of short-term and medium-term forecasting. Random Forest demonstrates the greatest stability, while LSTM better models short-term emissions. This combination of models allows for the creation of a more reliable operational air quality forecasting system.

Table 3. Forecasting results of Random Forest and LSTM models for short-term (7-day) and medium-term (30-day) horizons under two scenarios: A (without weather data) and B (with meteorological variables)

Pollutant	Scenario	Horizon (days)	Model	RMSE (µg/m³)	MAE (µg/m³)	MAPE (%)	Accuracy	F1
$PM_{2.5}$	A (no weather)	7	Random Forest	6.22	4.95	13.89	1.00	1.00
			LSTM	7.61	6.75	19.34	1.00	1.00
	B (with weather)	7	Random Forest	3.96	3.30	8.98	1.00	1.00
			LSTM	6.49	5.00	13.51	1.00	1.00
	A (no weather)	30	Random Forest	6.91	5.79	17.39	1.00	1.00
			LSTM	7.64	6.00	15.86	1.00	1.00
	B (with weather)	30	Random Forest	5.70	4.64	13.50	1.00	1.00
			LSTM	6.68	4.97	14.36	1.00	1.00
PM_{10}	A (no weather)	7	Random Forest	9.40	5.93	23.92	1.00	0.00
			LSTM	9.80	8.24	39.66	1.00	0.00
	B (with weather)	7	Random Forest	8.44	5.67	23.53	1.00	0.00
			LSTM	8.03	4.15	14.16	1.00	0.00
	A (no weather)	30	Random Forest	5.79	3.93	20.18	1.00	0.00
			LSTM	5.27	3.41	17.50	1.00	0.00
	B (with weather)	30	Random Forest	5.35	3.66	18.42	1.00	0.00
			LSTM	5.83	4.19	19.95	1.00	0.00

4.5 The importance of signs and the interpretability of models

Figure 9 shows the ranking of predictors in the Random Forest model for scenario B (taking into account meteorological data) over a 7-day horizon. The results obtained are consistent with the physical mechanisms of pollution formation in the Almaty Basin.

The boundary layer height (BLH) is the absolute leading predictor for both pollutants (blh_0).

This reflects the key role of vertical air mixing: the lower the BLH in winter, the greater the accumulation of $PM_{2.5}$ and PM_{10}.

Pollutant lags are among the most significant features (lag_1, lag_2, lag_24), confirming the strong autocorrelation and inertia of pollution.

Temperature ranks second in influence after BLH (temp_0, temp_1, temp_2).

Low temperatures intensify inversions and reduce BLH → this increases PM concentrations.

Surface pressure (sp_0–sp_3) shows a noticeable contribution, which corresponds to anticyclonic, stagnant winter conditions.

Wind speed and cloud cover have a small but interpretable contribution:

- wind_0 and wind_1 weaken pollution (dispersion),
- cloud_0 reflects changes in radiative cooling and layer stability.

The resulting importance structure demonstrates the physical consistency of the model and justifies the inclusion of an extended set of meteorological parameters.

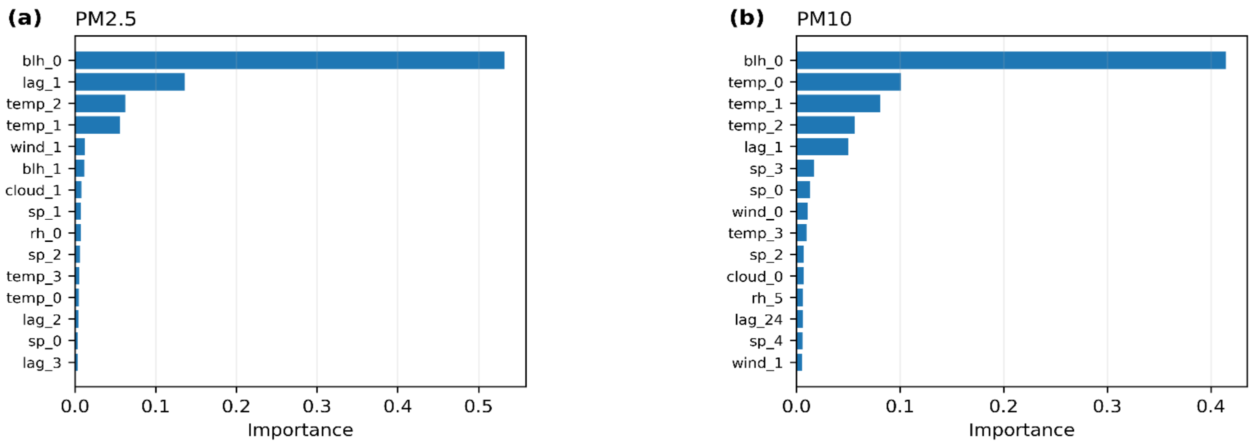


Figure 9. Importance of features in the Random Forest model for predicting concentrations of: (a) $PM_{2.5}$ and (b) PM_{10} in scenario B (taking meteorological factors into account), forecast horizon – 7 days

4.6 Comparison of models with basic approaches (Persistence and Seasonal Naïve)

The comparison results (Table 4) show that the use of ML models provides significant advantages in terms of RMSE metrics compared to standard statistical approaches. The improvements are particularly pronounced when using an extended set of meteorological variables.

$PM_{2.5}$

Over a 30-day horizon, Random Forest demonstrates the most significant increase in accuracy: the RMSE improvement is +38.6% in scenario A and up to +49.4% in scenario B.

LSTM also consistently outperforms baselines (+32.2% in A, +40.7% in B).

Over a 7-day horizon, the effect is moderate:

- RF improves RMSE by +2.9% (A) and +38.3% (B);
- LSTM is almost comparable to Persistence, which is explained by the high autocorrelation of the series.

Including temperature, humidity, wind speed, boundary layer height, and pressure significantly improves the accuracy of the short-term forecast.

PM_{10}

A computational modeling experiment revealed that, over a 7-day forecast horizon, the quality of results obtained using the constructed models is, in some cases, inferior to the "robustness" method (e.g., Δ = –24.1%). This effect is not due to errors or incorrect model specifications, but to the structural features of the test time series. Under conditions of smooth short-term dynamics and high autocorrelation of observations, the "robustness" method serves as a statistically optimal benchmark strategy, minimizing forecast error within the experimental problem formulation.

At a 30-day horizon, both models confidently outperform the baseline approaches:

- RF: +7.3% (A) and +14.5% (B)
- LSTM: +15.7% (A) and +6.8% (B)

Weather parameters have a noticeable effect due to the sensitivity of PM_{10} to BLH, humidity and wind conditions.

The Random Forest and LSTM models demonstrate stable advantages over the baseline models, especially in medium-term forecasts and when meteorological factors are included. Negative Δ values in some short-term scenarios are explained by the properties of the time series, rather than shortcomings of the ML models, which confirms the correctness and interpretability of the comparisons.

Table 4. Comparison of Random Forest and LSTM model performance with baseline forecasting methods (Persistence and Seasonal Naïve)

Pollutant	Scenario	Horizon (days)	Model	RMSE (μg/m³)	Baseline RMSE (μg/m³)	Δ vs Baseline (%)
$PM_{2.5}$	A (no weather)	7	Random Forest	6.22	6.40	+2.92
			LSTM	7.61	6.40	–18.82
	A (no weather)	30	Random Forest	6.91	11.27	+38.64
			LSTM	7.64	11.27	+32.20
	B (with weather)	7	Random Forest	3.95	6.40	+38.30
			LSTM	6.49	6.40	–1.29
	B (with weather)	30	Random Forest	6.91	11.27	+49.39
			LSTM	7.64	11.27	+40.70
PM_{10}	A (no weather)	7	Random Forest	9.40	7.57	–24.10
			LSTM	9.80	7.57	–29.43
	A (no weather)	30	Random Forest	5.79	6.25	+7.34
			LSTM	5.27	6.25	+15.66
	B (with weather)	7	Random Forest	8.44	7.57	–11.39
			LSTM	8.03	7.57	–5.95
	B (with weather)	30	Random Forest	5.35	6.25	+14.52
			LSTM	5.83	6.25	+6.78

Note: Positive values of Δ indicate an improvement in accuracy (lower RMSE) compared to baseline models. Negative values indicate cases where the selected model performed worse than the baseline.

4.7 Practical interpretation and significance of results

The results demonstrate that the combination of the Random Forest ensemble algorithm and the LSTM recurrent neural network provides high accuracy in short-term forecasting of $PM_{2.5}$ and PM_{10} concentrations in the complex mountain-valley circulation conditions of Almaty. The models are resistant to data noise, correctly capture the inertia of time series, and demonstrate sensitivity to key meteorological parameters, making them applicable to practical tasks of environmental monitoring and air quality management.

4.7.1 Practical significance of the results

1) The basis for an early warning system for pollution. The models obtained can be integrated into automated monitoring and early warning platforms, providing daily and weekly forecasts of likely exceedances of air quality standards. This is particularly important for cities in Kazakhstan, where frequent winter smog and poor air ventilation require rapid response measures.

2) Use in mobile applications and services for the general public. The models can be implemented in existing mobile applications, such as AirKz / Airkaz, in the form of a '7-day $PM_{2.5}$/PM_{10} forecast' module. Users will be able to plan outdoor physical activity in advance, which is particularly relevant for vulnerable groups (children, the elderly, patients with respiratory diseases).

3) Scalability and transferability. The proposed methodology could potentially be adapted for other cities in Kazakhstan and Central Asia, but it needs to be calibrated to local emission structures, heating practices, and meteorological conditions. With low input data requirements (PM and basic meteorological parameters), the approach remains applicable in conditions of limited data availability, but additional verification in other cities is necessary to confirm the generalizability of the results.

4) Potential for environmental policy and city management. Forecast smog maps can be used to:

- optimize public transport route networks;
- introducing temporary environmental restrictions (low-emission days);
- planning utility service schedules;
- adjusting heating regimes in the private sector.

Thus, forecast models become a tool to support decision-making at the level of government authorities.

4.7.2 Integration of results into public-private partnerships (PPPs) in the field of ecology

Digital environmental monitoring and forecasting technologies are increasingly becoming the functional foundation of next-generation public-private partnerships (PPPs) focused on decarbonization, sustainable infrastructure, and the mitigation of environmental risks in cities. Recent research emphasizes that the effectiveness of environmental PPPs increases significantly with the presence of quantifiable impact indicators and digital management tools [35]. The results of forecasting $PM_{2.5}$ and PM_{10} concentrations obtained in this study can be directly integrated into the architecture of such partnerships, complementing the investment and institutional logic of "green" PPPs previously substantiated by the authors [36].

1) Digital air monitoring services within PPPs

$PM_{2.5}$ and PM_{10} forecasting models can serve as a key analytical module for digital air quality platforms created through PPPs between the government, IT companies, and sensor infrastructure operators. International practice confirms the viability of such solutions: in Singapore, the National Environment Agency (NEA), together with private technology partners, is using ML-based pollution forecasts as part of the Smart Environment Platform for operational environmental management [37].

In the context of Kazakhstan, a similar model could be implemented as a concession or service PPP project, including:

- a network of low-cost air quality sensors maintained by a private operator;
- a pollution forecasting module as a B2G service for the city administration;
- integration of forecasts into the Smart City ecosystem.

This approach is consistent with the characteristics of effective PPP projects in “green” sectors identified by the authors earlier [36], where the key role is played by the technological structure of the project and the participation of the private partner in the management of the innovative component.

2) Justification of investments in environmental infrastructure

Forecasting MAC exceedances for $PM_{2.5}$ and PM_{10} allows us to move from declarative environmental impacts to quantifiable justifications for investments in PPPs. According to World Bank and WHO estimates, an increase in $PM_{2.5}$ concentrations by 10–15 μg/m³ is accompanied by a 1–3% increase in the burden on the healthcare system, making it possible to use Cost of Illness (COI) methods to calculate the socioeconomic benefits of infrastructure projects [38, 39].

In practical terms, this means that ML forecasts can be used:

- to model the benefits of replacing coal-fired heating systems with gas or electric ones;
- to assess the effectiveness of clean transport support programs;
- to calculate the environmental impacts of concession projects for the modernization of combined heat and power plants and distributed energy systems.

This logic is fully consistent with the findings of Casady et al. [35], which highlight that low-carbon and sustainable PPP environments require comprehensive analytical tools that link environmental performance to project patterns and institutional parameters.

3) Improving the transparency and manageability of environmental data

The use of predictive models in PPPs increases the transparency of environmental data and creates the basis for objective monitoring of the private partner's performance. Public air quality forecasts enable the development of KPIs based not only on actual measurements but also on the operator's ability to prevent projected exceedances of maximum permissible concentrations.

The practice of OECD countries shows that the inclusion of predictive indicators in PPP monitoring systems contributes to increased public confidence, reduced social risks, and increased accountability of environmental projects [40]. As a result, forecasting becomes not just an auxiliary tool, but an element of the institutional design of environmental PPPs. Thus, the developed forecasting system for $PM_{2.5}$ and PM_{10} can be considered a technological component of next-generation environmental PPPs, providing quantitative justification for investments, operationalizing environmental effects, and increasing management transparency. The integration of forecasting models into PPPs is consistent with modern international approaches to green infrastructure development and enhances the practical applicability of the study's results, complementing previously obtained conclusions on the structure and effectiveness of PPP projects in the green economy.

4.7.3 Scientific contribution of the research

The study has several significant scientific results:

1) Weather-dependent ML model for a mountain-valley metropolis. An interpretable air pollution prediction model has been developed for the city of Almaty, taking into account topographical specifics, pronounced temperature inversions and seasonal features.

2) Comparison of two architectures on a single database. A direct comparative analysis of Random Forest and LSTM, trained on the same sample, was conducted, which made it possible to identify their advantages and limitations in the context of real data from Central Asia.

3) Quantitative assessment of the influence of meteorological factors. The significant role of temperature, humidity, wind speed, atmospheric pressure, cloud cover, and boundary layer height in shaping the short-term dynamics of $PM_{2.5}$ and PM_{10} is demonstrated, which is confirmed by both ML methods (feature importance) and physical mechanisms of winter smog.

4) Basis for an intelligent early warning system. The results from the scientific and technical basis for the creation of a predictive air quality platform applicable to cities in Central Asia within the framework of the Smart City ecosystem and PPP projects.

5. DISCUSSION

This study has shown that the use of ML (Random Forest) and DL (LSTM) methods provides reliable, interpretable, and practically applicable short-term forecasting of $PM_{2.5}$ and

PM_{10} concentrations in the complex orography and pronounced meteorological dependence of the city of Almaty. This section presents the key results of the analysis, their scientific interpretation, comparison with international studies, as well as limitations and prospects for further development of the forecasting system in the context of sustainable air quality management in a metropolis.

5.1 Comparative analysis of Random Forest and LSTM

A comparison of the simulation results shows that Random Forest and LSTM demonstrate different accuracy and stability characteristics at different forecast horizons (see Table 3). For $PM_{2.5}$, the Random Forest model provides the lowest RMSE and MAE values for both the 7-day and 30-day horizons, especially in scenario B, where the inclusion of meteorological parameters leads to the most significant improvement in accuracy (RMSE = 3.96 µg/m^3). This stability is consistent with the findings of international studies, where RF is considered a reliable tool for early warning systems [41]. The LSTM model for $PM_{2.5}$ shows higher errors but remains sensitive to short-term fluctuations in concentrations, which is due to its recurrent architecture. However, over a 30-day horizon, there is an increase in error variability – a limitation characteristic of LSTM with complex seasonality of time series [42]. For PM_{10}, the results are more heterogeneous. At a 7-day horizon in scenario B, LSTM shows a lower RMSE (8.03 µg/m^3) than RF, indicating a better response of the network to short-term changes in coarse particles. However, on a 30-day horizon, Random Forest remains the most stable and shows the lowest errors in both scenarios. Thus, Random Forest is the preferred algorithm for obtaining stable and interpretable forecasts, while LSTM is appropriate for increasing sensitivity to short-term peak episodes, especially when forecasting PM_{10} in the short term.

5.2 Role of meteorological factors and confirmation of seasonal dependence

The results for two scenarios (A – without meteorological parameters, B – with meteorological parameters) show a significant reduction in RMSE and MAE errors when temperature, relative humidity, wind, pressure, and PBL height (BLH) are added. The greatest improvement is achieved for $PM_{2.5}$, where taking BLH and temperature into account significantly enhances the physical explain ability of the model; for PM_{10}, the contribution of meteorological parameters is also noticeable, reflecting the high sensitivity of coarse particles to vertical mixing and wind activity.

Correlation analysis (Figure 4, Table 2) confirms key physical relationships:

- a pronounced negative correlation between $PM_{2.5}$ and PM_{10} and temperature (–0.739 and –0.708), which corresponds to the mechanism of winter inversions and weakening turbulence;
- a moderate positive correlation with humidity (0.483 and 0.504), reflecting hygroscopic particle growth and enhanced secondary aerosol processes;
- a weak negative correlation with wind speed (up to –0.289), indicating the role of horizontal transport;
- positive correlation with surface pressure (0.461 and 0.516), consistent with the formation of stagnant anticyclonic conditions;
- strong negative correlation with boundary layer height (BLH) (–0.656 and –0.580), confirming the key role of vertical mixing volume in aerosol accumulation.

These patterns fully reflect the regional specifics of Almaty: low temperatures, high humidity, frequent anticyclones, and weakened vertical mixing form stable winter pollution peaks. Similar conclusions are presented in studies by Gao et al. [43] and Özüpak et al. [34], where the inclusion of meteorological factors significantly improves the accuracy of short-term forecasts.

5.3 Analysis of the predictive capabilities of models at 7- and 30-day horizons

Graphical visualization of forecasts (Figures 6–8) shows differences in model behavior in the short term and medium term. Over a 7-day interval, both models adequately reproduce the overall dynamics of $PM_{2.5}$ and PM_{10}, but Random Forest generates more stable forecasts that are closer to the observed values, especially in scenario B for $PM_{2.5}$, while LSTM demonstrates a more pronounced sensitivity to local fluctuations and, in some cases, better reflects short-term changes, particularly for PM_{10}. When the horizon is increased to 30 days, the forecasts of both models become smoother, which corresponds to an increase in uncertainty and an increase in RMSE. Nevertheless, Random Forest maintains a more stable correspondence with the observed values, while LSTM shows greater smoothing and slightly underestimates the concentration peaks. In scenario B, the influence of meteorological factors becomes more noticeable, improving the model's fit to the trend, which is consistent with global air quality studies [44], according to which long-term forecasts are more dependent on large-scale atmospheric dynamics, while short-term forecasts are formed mainly due to the autocorrelation structure of time series.

5.4 Comparison with international studies

A comparison of the results obtained with international studies shows that the dynamics of air pollution in Almaty generally correspond to the patterns characteristic of large cities subject to winter temperature inversions. Similar profiles of seasonal peaks in $PM_{2.5}$ and PM_{10} have been described in detail for Ulaanbaatar, Tehran and Tashkent [5, 8], where a combination of low boundary layer height, weak wind activity and intensive use of carbon-containing fuels leads to prolonged periods of aerosol accumulation. However, the regional specifics of Almaty are more pronounced and manifest themselves in a combination of factors: the widespread use of coal heating in the private sector, the peculiarities of mountain-valley circulation that limits vertical ventilation, and the high frequency of calm conditions in winter. This combination forms a unique ‘smog profile’ that significantly increases the meteorological dependence of pollution and requires the use of models capable of accounting for the nonlinear interaction of meteorological parameters and topography. In this context, the scientific contribution of this study is the construction of an interpretable ML/DL model for Almaty based on long-term data from 2020-2024, which fills the identified gap and complements the international literature on air quality forecasting in cities with complex orographic conditions.

5.5 Research limitations

Despite the results obtained, the study has a number of limitations that must be taken into account when interpreting the conclusions. First, despite the expanded set of meteorological characteristics (temperature, humidity, wind speed, atmospheric pressure, cloud cover, and planetary boundary layer height), the model remains deterministic and does not take into account possible variations in emissions, heat energy load and fuel consumption dynamics, which may affect the reproducibility of peak episodes. Secondly, the analysis was performed for only one urban agglomeration – Almaty – which limits the external validity of the results and does not allow the conclusions to be directly extrapolated to other cities in Central Asia without additional adaptation; the transferability of the models may be limited by differences in emission structures, heating types, terrain and circulation patterns. Thirdly, spatial heterogeneity is only partially taken into account: aggregating data into a median urban time series reduces the influence of local emissions, but at the same time limits the ability to identify station-specific patterns. Fourth, the high Accuracy and F1 scores in a number of scenarios are due to the absence of threshold exceedances in the test period, which reflects the specificity of the sample rather than the universal diagnostic capability of the models. Finally, the LSTM model showed increased error variability over long horizons and a high computational load, which is typical for recurrent architectures with limited data volumes. These limitations determine the directions for further development, including expanding the set of factors, integrating spatiotemporal models, and assessing the transferability of algorithms to other cities in the region.

5.6 Prospects for the development and integration of the model into air quality management systems

The prospects for further development of the proposed forecasting system are linked to expanding the set of input factors and improving model architectures. Despite the inclusion of key meteorological parameters (temperature, humidity, wind, pressure, cloud cover, BLH), additional data on emissions, energy load and heating characteristics can improve the accuracy of forecasts during periods of winter inversions. Promising areas also include the use of spatio-temporal architectures (CNN - LSTM, transformers) and the transition to probabilistic forecasting using quantile regression, bootstrap ensembles or Bayesian LSTM, which will allow the formation of confidence intervals and the assessment of model uncertainty.

Integrating forecasting algorithms with low-cost sensor IoT infrastructure will enable real-time data updates and automatic model retraining. Creating dynamic pollution forecast maps will enhance the practical value of the results for city services, researchers, and the public.

The resulting predictive models can be used as analytical and management tools within environmentally oriented public-private partnerships (PPPs). Embedding $PM_{2.5}$ and PM_{10} concentration forecasting modules into Smart City digital dashboards, as well as using the modeling results to justify heating system modernization projects and deploy air quality monitoring networks within PPPs, creates the preconditions for a transition from retrospective monitoring to proactive air quality management. The practical implementation of such solutions requires a coordinated digital architecture and unified performance indicators, which allows forecasting to be considered as an element of the emerging adaptive environmental management system at the municipal level.

6. CONCLUSIONS

This study develops a modern approach to air quality management in large cities in Central Asia and proposes an interpretable weather-dependent system for short-term forecasting of $PM_{2.5}$ and PM_{10} concentrations for the city of Almaty based on ML (Random Forest) and DL (LSTM) methods. The use of long-term data for 2020 - 2024 from the open platforms AQICN, AirKaz and Dashboard.air.org.kz, as well as the construction of two modelling scenarios – with and without meteorological factors – made it possible to comprehensively assess the role of atmospheric conditions in the formation of smog episodes and to improve the accuracy of forecasts for mountainous terrain conditions.

The results show that the inclusion of an extended set of meteorological parameters (temperature, humidity, wind speed, pressure, cloud cover and planetary boundary layer height) significantly improves forecast accuracy, especially for $PM_{2.5}$, highlighting the key role of atmospheric stratification and winter inversions. The Random Forest algorithm demonstrated the most stable RMSE/MAE values at 7- and 30-day horizons, ensuring high interpretability and stability of forecasts. The LSTM model, in turn, showed an advantage in reproducing short-term fluctuations and daily peaks, making it a useful component for operational environmental monitoring. The combination of RF stability and LSTM adaptability increases the applied value of the developed approach for sustainable urban planning tasks.

The data obtained forms the basis for the creation of an intelligent early warning system for risks of exceeding maximum permissible concentrations and can support decision-making in the areas of public health, urban transport and environmental infrastructure management. The integration of predictive models into mobile services (AirKZ/Airkaz), urban visualization dashboards, Smart City systems and public-private partnership projects opens up opportunities for the development of preventive environmental policies, raising public awareness and strengthening the resilience of the urban environment.

The main scientific and practical results of the study are as follows:

1) a fully functional ML/DL-based air quality forecasting system has been developed for the mountain-valley territory of Almaty, taking into account weather-dependent smog mechanisms;

2) the high significance of meteorological factors and the key contribution of temperature inversions and boundary layer structure to the formation of winter pollution episodes have been confirmed;

3) methodological and practical prerequisites have been established for adapting the approach to other cities in Kazakhstan and Central Asia, taking into account differences in emission structures and meteorological conditions;

4) the possibility of integrating models into existing digital monitoring platforms and the infrastructure of environmental PPP projects within the framework of the concept of sustainable urban development has been demonstrated.

Prospects for further research are related to expanding the set of meteorological and emission factors, transitioning to spatiotemporal modelling based on CNN - LSTM and transformers, integrating data from low-cost IoT sensor networks, and creating a unified urban predictive analytics platform. The implementation of these areas will contribute to the formation of scientifically sound solutions in the field of sustainable air quality management and the development of modern environmental planning tools in the context of accelerated urbanization.

ACKNOWLEDGMENT

The authors acknowledge that this research was carried out within the framework of the Bolashak International Scholarship Program (Republic of Kazakhstan) during a joint academic mobility project hosted by Humboldt University of Berlin (Germany).

REFERENCES

[1] Bureau of National Statistics of the Agency for Strategic Planning and Reforms of the Republic of Kazakhstan [BNS]. (2025). Statistics of environment (Almaty city: Dynamic tables). https://stat.gov.kz/en/region/almaty/dynamic-tables/1523/.

[2] WHO. (2021). WHO global air quality guidelines. https://www.who.int/news-room/questions-and-answers/item/who-global-air-quality-guidelines.

[3] Burnett, R., Chen, H., Szyszkowicz, M., Fann, N., et al. (2018). Global estimates of mortality associated with long-term exposure to outdoor fine particulate matter. Proceedings of the National Academy of Sciences, 115(38): 9592-9597. https://doi.org/10.1073/pnas.1803222115

[4] Alizadeh-Choobari, O., Bidokhti, A.A., Ghafarian, P., Najafi, M.S. (2016). Temporal and spatial variations of particulate matter and gaseous pollutants in the urban area of Tehran. Atmospheric Environment, 141: 443-453. https://doi.org/10.1016/j.atmosenv.2016.07.003

[5] Wang, M., Kai, K., Jin, Y., Sugimoto, N., Dashdondog, B. (2017). Air particulate pollution in Ulaanbaatar, Mongolia: Variation in atmospheric conditions from autumn to winter. SOLA, 13: 90-95. https://doi.org/10.2151/sola.2017-017

[6] Kerimray, A., Azbanbayev, E., Kenessov, B., Plotitsyn, P., Alimbayeva, D., Karaca, F. (2020). Spatiotemporal variations and contributing factors of air pollutants in Almaty, Kazakhstan. Aerosol and Air Quality Research, 20: 1340-1352. https://doi.org/10.4209/aaqr.2019.09.0464

[7] Kozhagulov, S., Adambekova, A., Quadrado, J.C., Salnikov, V., Rysmagambetova, A., Tanybayeva, A. (2025). Trends in atmospheric emissions in Central Asian countries since 1990 in the context of regional development. Climate, 13(9): 176. https://doi.org/10.3390/cli13090176

[8] Tursumbayeva, M., Muratuly, A., Baimatova, N., Karaca, F., Kerimray, A. (2023). Cities of Central Asia: New hotspots of air pollution in the world. Atmospheric Environment, 309: 119901. https://doi.org/10.1016/j.atmosenv.2023.119901

[9] Darynova, Z., Maksot, A., Kulmukanova, L., Malekipirbazari, M., Sharifi, H., Amouei Torkmahalleh, M., Holloway, T. (2018). Evaluation of NO_2 column variations over the atmosphere of Kazakhstan using satellite data. Journal of Applied Remote Sensing, 12(4): 042610. https://doi.org/10.1117/1.JRS.12.042610

[10] Gallardo, L., Barraza, F., Ceballos, A., Galleguillos, M., et al. (2018). Evolution of air quality in Santiago: The role of mobility and lessons from the science-policy interface. Elementa: Science of the Anthropocene, 6(1): 38. https://doi.org/10.1525/elementa.293

[11] Bai, X., Ming, X., Zhao, M., Zhou, L. (2024). Explore the effect of apparent temperature and air pollutants on the admission rate of acute myocardial infarction in Chongqing, China: A time-series study. BMJ Open, 14(4): e084376. https://doi.org/10.1136/bmjopen-2024-084376

[12] Tran, H.M., Tsai, F.J., Wang, Y.H., et al. (2025). Joint effects of temperature and humidity with $PM_{2.5}$ on COPD. BMC Public Health, 25(1): 424. https://doi.org/10.1186/s12889-025-21564-3

[13] Niu, Y., Yuan, M., Jiang, F., Yang, Y., Jia, X., Yang, C., Bao, J., Shi, X. (2025). Modification effects of ambient temperature and relative humidity on acute upper respiratory infection morbidity by $PM_{2.5}$ components in university students. Atmospheric Pollution Research, 16(4): 102430. https://doi.org/10.1016/j.apr.2025.102430

[14] Zender-Świercz, E., Galiszewska, B., Telejko, M., Starzomska, M. (2024). The effect of temperature and humidity of air on the concentration of particulate matter – $PM_{2.5}$ and PM_{10}. Atmospheric Research, 312: 107733. https://doi.org/10.1016/j.atmosres.2024.107733

[15] Purnomo, A., Andang, A., Badriah, S., Paryono, E., Sambas, A., Umar, R. (2024). Influence of wind speed and direction on the performance of low-cost particulate matter sensors. Environment and Ecology Research, 12(4): 446-455. https://doi.org/10.13189/eer.2024.120409

[16] de Arruda Moreira, G., Marques, M.T. A., da Silva Lopes, F.J., de Fátima Andrade, M., Landulfo, E. (2024). Analyzing the influence of the planetary boundary layer height, ventilation coefficient, thermal inversions, and aerosol optical depth on the concentration of $PM_{2.5}$ in the city of São Paulo: A long-term study. Atmospheric Pollution Research, 15(8): 102179. https://doi.org/10.1016/j.apr.2024.102179

[17] Han, Z., Wang, Y., Xu, J., Shang, Y., et al. (2024). Assessment of multiple planetary boundary layer height retrieval methods and their impact on $PM_{2.5}$ and its chemical compositions throughout a year in Nanjing. Remote Sensing, 16(18): 3464. https://doi.org/10.3390/rs16183464

[18] Wu, C., Wang, R., Lu, S., Tian, J., Yin, L., Wang, L., Zheng, W. (2025). Time-series data-driven $PM_{2.5}$ forecasting: From theoretical framework to empirical analysis. Atmosphere, 16(3): 292. https://doi.org/10.3390/atmos16030292

[19] Pan, M., Xia, B., Huang, W., Ren, Y., Wang, S. (2024). $PM_{2.5}$ concentration prediction model based on random forest and SHAP. International Journal of Pattern Recognition and Artificial Intelligence, 38(5): 2452012. https://doi.org/10.1142/S0218001424520128

[20] Alrashidi, H., Sibai, F.N., Abonamah, A., Alrashidi, M.,

Alsaber, A. (2025). $PM_{2.5}$: Air quality index prediction using machine learning: Evidence from Kuwait's air quality monitoring stations. Sustainability, 17(20): 9136. https://doi.org/10.3390/su17209136

[21] Chang, Y.S., Chiao, H.T., Abimannan, S., Huang, Y.P., Tsai, Y.T., Lin, K.M. (2020). An LSTM-based aggregated model for air pollution forecasting. Atmospheric Pollution Research, 11(8): 1451-1463. https://doi.org/10.1016/j.apr.2020.05.015

[22] Bai, X., Zhang, N., Cao, X., Chen, W. (2024). Prediction of $PM_{2.5}$ concentration based on a CNN–LSTM neural network algorithm. PeerJ, 12: e17811. https://doi.org/10.7717/peerj.17811

[23] Yedilkhan, M., Berdyshev, A., Galiyev, M., Merembayev, T. (2025). Air quality prediction based on the LSTM with attention using meteorological data in urban area in Kazakhstan. Journal of Problems in Computer Science and Information Technologies, 3(1): 3-12. https://doi.org/10.26577/jpcsit20253101

[24] Houdou, A., El Badisy, I., Khomsi, K., Abdala, S.A., Abdulla, F., Najmi, H., Obtel, M., Belyamani, L., Ibrahimi, A., Khalis, M. (2024). Interpretable machine learning approaches for forecasting and predicting air pollution: A systematic review. Aerosol and Air Quality Research, 24: 230151. https://doi.org/10.4209/aaqr.230151

[25] U.S. Environmental Protection Agency (EPA). (1999). Guideline on data handling conventions for the PM NAAQS (EPA-454/R-99-008). Office of Air Quality Planning and Standards. Research Triangle Park, NC. https://nepis.epa.gov/Exe/ZyPDF.cgi/2000D6J7.PDF?Dockey=2000D6J7.PDF.

[26] Grange, S.K., Carslaw, D.C. (2019). Using meteorological normalisation to detect interventions in air quality time series. Science of the Total Environment, 653: 578-588. https://doi.org/10.1016/j.scitotenv.2018.10.344

[27] Carslaw, D.C., Ropkins, K. (2012). Openair – An R package for air quality data analysis. Environmental Modelling & Software, 27-28: 52-61. https://doi.org/10.1016/j.envsoft.2011.09.008

[28] Hyndman, R.J., Athanasopoulos, G. (2021). Forecasting: Principles and Practice (3rd ed.). OTexts.

[29] Breiman, L. (2001). Random forests. Machine Learning, 45: 5-32. https://doi.org/10.1023/A:1010933404324

[30] Hochreiter, S., Schmidhuber, J. (1997). Long short-term memory. Neural Computation, 9(8): 1735-1780. https://doi.org/10.1162/neco.1997.9.8.1735

[31] Tashman, L.J. (2000). Out-of-sample tests of forecasting accuracy: An analysis and review. International Journal of Forecasting, 16(4): 437-450. https://doi.org/10.1016/S0169-2070(00)00065-0

[32] Bergmeir, C., Benítez, J.M. (2012). On the use of cross-validation for time series predictor evaluation. Information Sciences, 191: 192-213. https://doi.org/10.1016/j.ins.2011.12.028

[33] Zhou, S., Wang, W., Zhu, L., Qiao, Q., Kang, Y. (2024). Deep-learning architecture for $PM_{2.5}$ concentration prediction: A review. Environmental Science and Ecotechnology, 21: 100400. https://doi.org/10.1016/j.ese.2024.100400

[34] Özüpak, Y., Alpsalaz, F., Aslan, E. (2025). Air quality forecasting using machine learning: Comparative analysis and ensemble strategies for enhanced prediction. Water, Air, and Soil Pollution, 236: 464. https://doi.org/10.1007/s11270-025-08122-8

[35] Casady, C.B., Cepparulo, A., Giuriato, L. (2024). Public-private partnerships for low-carbon, climate-resilient infrastructure: Insights from the literature. Journal of Cleaner Production, 470: 143338. https://doi.org/10.1016/j.jclepro.2024.143338

[36] Domalatov, Y., Turginbayeva, A., Apysheva, A., Azimkhan, A., Kamali, K., Kuangaliyeva, T., Kenzhin, Z., Aidaraliyeva, A. (2024). Identifying the characteristics of public-private partnership projects on green energy in developing countries with different incomes. Eastern-European Journal of Enterprise Technologies, 131(13): 14-21. https://doi.org/10.15587/1729-4061.2024.311836

[37] National Environment Agency. (2019). Air and coastal water quality monitoring. https://www.nea.gov.sg/our-services/pollution-control/air-and-coastal-water-quality-monitoring.

[38] World Bank. (2022). The Global Health Cost of PM2.5 Air Pollution: A Case for Action Beyond 2021. World Bank Group. https://openknowledge.worldbank.org/entities/publication/c96ee144-4a4b-5164-ad79-74c051179eee.

[39] United Nations Environment Programme. (2024). Actions on Air Quality Report Update. UNEP. https://www.unep.org/topics/air/multi-level-air-quality-management/actions-air-quality-report-update.

[40] OECD. (2023). Improving the Landscape for Sustainable Infrastructure Financing. https://www.oecd.org/content/dam/oecd/en/publications/reports/2023/01/improving-the-landscape-for-sustainable-infrastructure-financing_637bd452/bc2757cd-en.pdf.

[41] Samal, K.K.R., Panda, A.K., Babu, K.S., Das, S.K. (2021). An improved pollution forecasting model with meteorological impact using multiple imputation and fine-tuning approach. Sustainable Cities and Society, 70: 102923. https://doi.org/10.1016/j.scs.2021.102923

[42] Li, X., Peng, L., Hu, Y., Shao, J., Chi, T. (2016). Deep learning architecture for air quality predictions. Environmental Science and Pollution Research, 23(22): 22408-22417. https://doi.org/10.1007/s11356-016-7812-9

[43] Gao, Z., Do, K., Li, Z., Jiang, X., Maji, K.J., Ivey, C.E., Russell, A.G. (2024). Predicting $PM_{2.5}$ levels and exceedance days using machine learning methods. Atmospheric Environment, 323: 120396. https://doi.org/10.1016/j.atmosenv.2024.120396

[44] Garbagna, L., Saheer, L.B., Oghaz, M.M.D. (2025). AI-driven approaches for air pollution modelling: A comprehensive systematic review. Environmental Pollution, 373: 125937. https://doi.org/10.1016/j.envpol.2025.125937

NOMENCLATURE

WHO	World Health Organization
$PM_{2.5}$	Particulate Matter ≤ 2.5 μm
PM_{10}	Particulate Matter ≤ 10 μm
RF	Random Forest
LSTM	Long Short-Term Memory
ML	Machine Learning

DL	Deep Learning
IoT	Internet of Things
LV	Limit Value
RMSE	Root Mean Square Error
MAE	Mean Absolute Error
MAPE	Mean Absolute Percentage Error
AQG	Air Quality Guidelines
AQI	Air Quality Index