Article

Performance Comparison of Statistical Models in PM2.5 Forecasting: A Case Study of Almaty

Nuray Dauletkhan [1*] and Khaled Mohamad [2]

[1]Department of Computer Science, SDU University, Almaty, Kazakhstan
[2]Department of Computer Science, SDU University, Almaty, Kazakhstan

DOI: 10.47344/b3exq459

Abstract

Air pollution, particularly fine particulate matter (PM2.5), poses a significant threat to public health in urban areas. In Almaty, Kazakhstan, high PM2.5 concentrations require effective forecasting methods to support timely intervention and policy planning. This study aims to evaluate and compare the performance of traditional statistical models and their hybrid counterparts for PM2.5 prediction. Multiple Linear Regression (MLR), Autoregressive Integrated Moving Average (ARIMA), Seasonal ARIMA (SARIMA), Generalized Additive Models (GAM), and several hybrid combinations (e.g., MLR + GAM) were applied to daily air quality and meteorological data from February 2020 to May 2024. Missing values were imputed using Multiple Imputation by Chained Equations (MICE), and model performance was assessed using Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R^2 score. The results show that MLR provided the best explanatory power (R^2 = 0.7160), while SARIMA achieved the lowest RMSE (0.2719), indicating strong short-term predictive accuracy. Among hybrid models, MLR + GAM delivered the most promising results (R2 = 0.6124), although improvements over standalone models were limited. These findings demonstrate the robustness of traditional statistical approaches for air quality forecasting and provide a benchmark for future studies incorporating machine learning techniques. The study offers practical value for environmental monitoring and air quality management in Almaty, and similar urban regions.

Keywords: $PM_{2.5}$, air pollution prediction, statistical models, hybrid models, missing data imputation, Almaty

I. Introduction

Air pollution is still a large issue globally, most notably in cities where industry, cars, as well as weather patterns affect worsening air quality [1], [2]. Particulate PM2.5 is extremely harmful to overall human health as well as to the entire environment. For its

*Corresponding author: nuray.dauletkhan@sdu.edu.kz
Email: nuray.dauletkhan@sdu.edu.kz ORCID: 0009-0001-0773-7203
Email: khaled.mohamad@sdu.edu.kz ORCID: 0000-0002-5980-0147

Received: March 12, 2025. Reviewed: April 12, 2025. Accepted: April 16, 2025.

ability for penetration deep into the respiratory system, long-term exposure with elevated PM2.5 concentrations is associated with respiratory illnesses. It is associated with cardiovascular diseases, and increased mortality rates [3].

Almaty, the largest city in Kazakhstan, frequently experiences high levels of $PM_{2.5}$, posing serious health risks to its residents. These concentrations are influenced not only by meteorological conditions but also by external factors such as industrial activity, traffic emissions, and seasonal variations. In urban settings like Almaty, traffic congestion and emissions from nearby industries are major sources of air pollution. These factors can lead to sharp increases in pollutant levels, particularly during peak traffic hours or in colder seasons when heating demand rises.

Factors like temperature, humidity, wind speed, and atmospheric pressure significantly influence how pollutants spread and settle in the environment. With more environmental data at our fingertips and by statistical models has become crucial for predicting $PM_{2.5}$ levels. This helps authorities jump in early and tackle potential health risks for the public.

Exploring historical data to capture useful patterns has been used for many years in air pollution forecasting using statistical models. Typical time-series models (ARIMA, SARIMA) use past observations to predict future ones, taking into account the temporal dependencies and seasonal trends in pollution levels [4]. GAM allows for non-linear relationships; however, MLR is a commonly used regression approach that measures $PM_{2.5}$ as a function of meteorology and pollutants [4].

Hybrid models combine the strengths of the different statistical methods to gain the most predictive power. Also, hybrids based on MLR (MLR + ARIMA, MLR + SARIMA, and MLR + GAM) mix regression methods with time-series or nonlinear modeling for greater performance. For example, GAM + ARIMA, uses the flexibility of GAM to model complex relationships combined with the time-based forecasting capabilities of ARIMA. Given the rapid advancements in these models, we argue that a comparative assessment of these is needed both to assess their performance at predicting $PM_{2.5}$ as well as to inform future research and policy.

II. Literature Review

Accurate $PM_{2.5}$ prediction is required to mitigate air pollution's impact on the environment and public health. $PM_{2.5}$ is a fine particulate matter with a diameter of 2.5 micrometers or less, making it small enough to penetrate the respiratory system and induces serious health issues [5]. Due to the adverse effects, a lot of models have been developed to forecast the air quality, from traditional to advanced machine learning methods.

One of the most frequently used techniques is Multiple Linear Regression (MLR), in which the effect of meteorological parameters like temperature, humidity, and wind speed on the $PM_{2.5}$ concentrations is identified. Research shows that MLR captures strong patterns of concentration of pollutants, and is a tool appropriate for application in air quality [6].

Another widely used model for time series is Autoregressive Integrated Moving Average (ARIMA) and seasonal version SARIMA. They make predictions for the $PM_{2.5}$ values based on the previous values and tend to capture the short-term variation and trend expected [7]. An extension of ARIMA with an extra part of seasonal variation is a model called SARIMA, particularly in urban regions where pollution varies seasonally due to weather and human activities SARIMA model holds the best accuracy out of all [8].

In addition to ARIMA-based methods, generalized additive models (GAMs) have emerged as a flexible alternative. Generalized additive models (GAMs) are generally able to model nonlinear relationships between variables, so they are well-suited to identify air pollution patterns [9]. Their performance is highly sensitive to data both in terms of shape and variability and should be carefully selected and tuned. Recent research on the use of machine learning and deep learning algorithms to predict air pollution have emerged. An example of the above approach is the use of a Hybrid ARIMA-LSTM where time-series data modeling via ARIMA is combined with the Long Short- Term Memory (LSTM) network, which is very effective in learning complex temporal patterns [10]. Hybrid models that combine AI-driven techniques with classical statistics outperform traditional statistics in revealing the complex patterns and associations around air pollution.

Table I provides a comparative summary of the four statistical models used in this study - MLR, ARIMA, SARIMA, and GAM, highlighting their applications in previous literature, key strengths, and known limitations for $PM_{2.5}$ forecasting tasks.

As shown in the table, while MLR offers interpretability and simplicity, models like SARIMA and GAM provide enhanced capabilities for capturing seasonal and nonlinear patterns, respectively. This comparison justifies their inclusion in our modeling framework for urban air quality prediction.

From these advances, we build on this subsequent work and compare a range of machine learning and deep learning algorithms, including Random Forest, XGBoost, LSTM, and CNN. All of these algorithms have been demonstrated to produce robust results in the literature. For example, Random Forest and XGBoost have outperformed simpler algorithms on R^2 and RMSE in many regions [11], [12]. LSTM based deep architectures have emerged as powerful scheme in modeling temporal relation as in case of Ulaanbaatar [13], whereas CNNs have been exploited to extract spatial features from pollution data [14]. The purpose of the subsequent work

Received: March 12, 2025. Reviewed: April 12, 2025. Accepted: April 16, 2025.

TABLE I
COMPARATIVE SUMMARY OF STATISTICAL MODELS USED FOR $PM_{2.5}$ FORECASTING

Model	Application in Literature	Strengths	Limitations
Multiple Linear Regression (MLR)	Used to model $PM_{2.5}$ as a linear function of meteorological variables such as temperature, humidity, and wind speed [6]	Easy to interpret, computationally efficient, performs well when linear assumptions hold	Struggles with nonlinear relationships and underperforms during extreme pollution spikes
Autoregressive Integrated Moving Average (ARIMA)	Forecasts $PM_{2.5}$ using past pollution values, capturing temporal correlations [7]	Strong for short-term forecasting, effective with stationary time series data	Limited in capturing seasonal patterns and exogenous variables
Seasonal ARIMA (SARIMA)	Extension of ARIMA that incorporates seasonality, suitable for periodic $PM_{2.5}$ trends [8]	Captures seasonal and cyclical pollution behaviors, lower RMSE in time series	Sensitive to parameter tuning; not suited for modeling external variables
Generalized Additive Models (GAM)	Models nonlinear relationships between $PM_{2.5}$ and predictors using smooth functions [9]	Flexible, handles nonlinearity effectively, adaptable to diverse datasets	Sensitive to noise and data distribution; requires careful smoothing parameter selection

will be to test if such advanced algorithms are able to represent an improvement on the statistical baselines in the current study. Analysis will also explore the influence of model structure, feature extraction, and data preprocessing, including imputation, on predictive ability. Inspired by novel advances in ensemble and hybrid learning strategies [15], [16], research will explore composite methods to develop stronger, scalable, and adaptive air quality forecasting systems for the conditions in Almaty.

Air quality forecasting, from simple regression to complex hybrid and deep learning algorithms, has evolved with time. Statistical models, external factors, and strong algorithms have improved predictability tremendously and provided valuable inputs to the policymakers to frame pollution control strategies. In the current research, we apply forecasting techniques on the air quality data in Almaty to find the most suitable statistical model to forecast the levels of $PM_{2.5}$ and assist in enhancing air quality control.

III. METHODS

In this study, we develop a structured approach to assess statistical and hybrid models for $PM_{2.5}$ forecasting in Almaty. The development process consists of four essential stages: (A) Data collection; (B) Data preprocessing; (C) Training and testing of models; (D) Evaluation metrics.

We started by obtaining daily air quality data and meteorological data from government sources. Data Preprocessing included missing values imputation using Multiple Imputation by Chained Equations (MICE), selection of relevant features based on correlation analysis and scaling of the variables.

We used combined statistical models (MLR, ARIMA, SARIMA, GAM) in simple as well as hybrid combinations to capture linear, non-linear, and temporal patterns. Model performance was evaluated employing Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R^2 to characterize accuracy and explanatory strength.

A. Data Collection

This study uses a dataset comprising 1,558 daily observations from February 2020 to May 2024, collected to support PM2.5 forecasting in Almaty, Kazakhstan. The period captures seasonal variability, pollution episodes, and changes in emission patterns, including those during the COVID-19 lockdown.

Meteorological data included the following: temperature, humidity, wind speed, atmospheric pressure (at station and sea level), and precipitation. These were obtained from Kazhydromet.kz, the official hydrometeorological service of Kazakhstan. These variables influence pollutant dispersion, chemical transformation, and removal through wet deposition.

Air quality data, including $PM_{2.5}$, PM_{10} , NO_2, SO_2 and CO, were retrieved from aqicn.org, which aggregates data from government-certified monitoring stations. $PM_{2.5}$ serves as the primary target due to its high health risk, while co-pollutants aid in capturing complex interactions affecting air quality.

Received: March 12, 2025. Reviewed: April 12, 2025. Accepted: April 16, 2025.

The meteorological and pollutant datasets were aligned in time to ensure that each record represents environmental conditions for a single day. This synchronization is crucial for accurate time series modeling. An initial inspection revealed missing values, which are common in environmental monitoring. These were handled during preprocessing using appropriate imputation methods.

TABLE II
SUMMARY OF METEOROLOGICAL AND AIR QUALITY DATA IN ALMATY FROM 2020 TO 2024

Parameter	Lower Limit	Average	Upper Limit
Temperature, °C	-19.3	11.4	33.5
Wind speed, m/s	0.0	0.55	2.0
Humidity, %	19.0	58.9	98.0
Precipitation, mm	0.0	1.69	49.0
Atmospheric Pressure (Sea Level), hPa	994.9	1017.8	1039.4
$PM_{2.5}$, μg/m³	14.0	74.7	160.0
NO_2, μg/m³	0.0	11.4	45.2
SO_2, μg/m³	0.0	1.21	5.6
CO, mg/m³	0.0	6.56	18.3

Table II presents a statistical summary of the meteorological and air quality variables used in this study, based on data collected in Almaty from 2020 to 2024. The temperature ranged from $-19.3\,^{\circ}\mathrm{C}$ to $33.5\,^{\circ}\mathrm{C}$, with an average of $11.4\,^{\circ}\mathrm{C}$, reflecting the city's continental climate. Wind speed showed low variability, averaging $0.55\,\mathrm{m/s}$, which may contribute to pollutant accumulation due to limited atmospheric dispersion. Humidity levels varied widely (from $19.0\,\%$ to $98.0\,\%$), while precipitation ranged from 0 to $49.0\,\mathrm{mm}$, with an average of $1.69\,\mathrm{mm}$, indicating mostly dry conditions. Atmospheric pressure at sea level remained relatively stable, averaging $1017.8\,\mathrm{hPa}$.

Regarding air quality, the mean $PM_{2.5}$ concentration was $74.7\,\mu\mathrm{g/m^3}$, significantly exceeding WHO air quality guidelines, with daily values reaching up to $160.0\,\mu\mathrm{g/m^3}$. NO_2 and SO_2 levels were moderate, with averages of $11.4\,\mu\mathrm{g/m^3}$ and $1.21\,\mu\mathrm{g/m^3}$, respectively. CO concentrations averaged $6.56\,\mathrm{mg/m^3}$, with a peak of $18.3\,\mathrm{mg/m^3}$. These statistics highlight the persistent air pollution challenges in Almaty and provide the basis for model development and evaluation in the study.

B. Data Preprocessing

Several preprocessing steps were applied to enhance data quality. The dataset included numerous missing values, particularly for pollutant variables. The longest gap occurred between May 2, 2022, and September 5, 2022.

To address this, missing values were imputed using Multiple Imputation by Chained Equations (MICE). It generates a series of complete data sets by modeling one incomplete variable conditionally on others. MICE is standardly applied to continuous and dependent environmental data, as it maintains the structure and the variations present in the data [5]. It is more robust compared to simple methods like mean imputation or K-Nearest Neighbors (KNN), which are not always effective when dealing with complex multivariate dependencies.

In order to facilitate the choice of predictors in $PM_{2.5}$ prediction, it is necessary to know how $PM_{2.5}$ relates to meteorological and pollutant variables. Figure 1 shows a correlation matrix, a visual representation of the direction and magnitude of the correlations of the variables in the dataset.

As shown in Figure 1, $PM_{2.5}$ exhibits strong positive correlations with variables such as carbon monoxide (CO) and nitrogen dioxide (NO_2), indicating that emissions from traffic and combustion processes are key contributors. Moderate correlations with humidity and temperature also suggest that weather conditions influence pollution levels, likely through effects on pollutant dispersion and atmospheric stability. These insights were used to guide feature selection for the models, ensuring only variables with significant relationships were included.

Standardization was subsequently performed on all the continuous variables to achieve uniform scaling. This step results in improved model performances and stability, especially of those models that are sensitive to the feature magnitude.

Figure 2 presents the temporal profile of standardized $PM_{2.5}$ concentration during the study period spanning February 2020 to May 2024, subsequent to using Multiple Imputation by Chained Equations (MICE) to handle the gaps in the data. The time series exhibit strong seasonal trends, with concentration spikes occurring repeatedly in the winter months. These spikes are due to high

Received: March 12, 2025. Reviewed: April 12, 2025. Accepted: April 16, 2025.

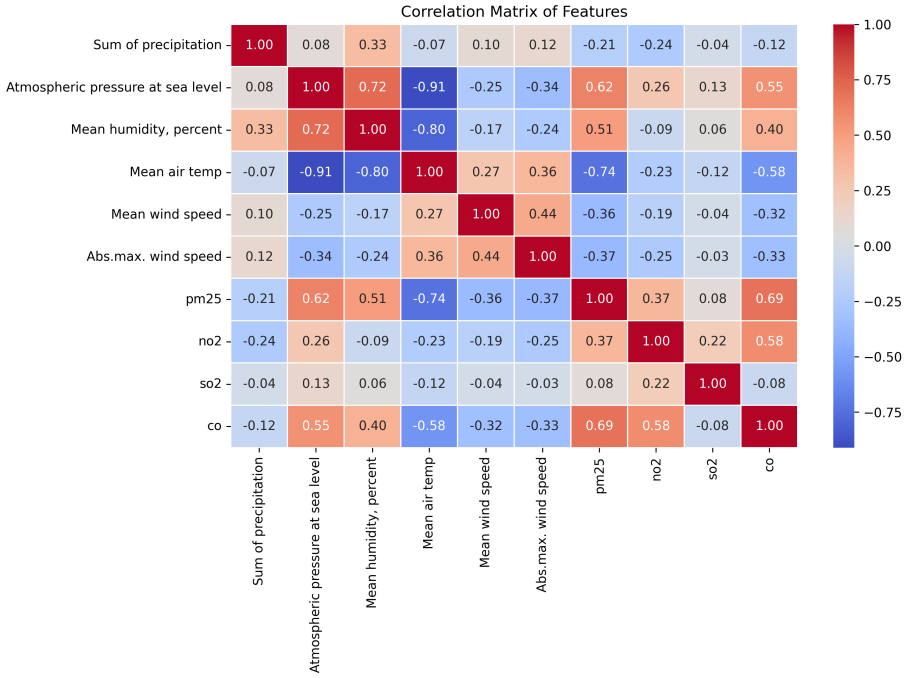

Fig. 1. Correlation matrix showing relationships between $PM_{2.5}$, meteorological variables and other pollutants. Strong correlations help guide feature selection for modeling.

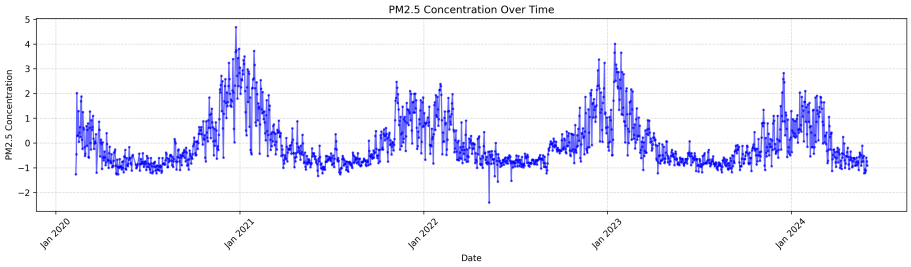

Fig. 2. Multiple Imputation by Chained Equations (MICE)

coal burning for household heating, atmospheric inversion, and low dispersion caused by low wind velocities. Valleys in the series in the summer months are indicative of better air quality due to good meteorological conditions, and low heating needs.

Application of MICE permitted the interpolation of missing values without interrupting these seasonal and long-term trends. Imputed values keep the same structure and volatility of the original data, preserving the continuity fit for time series modeling purposes. In particular, no artificial discontinuities or flattening effects were seen after imputation, indicating the method preserved the natural trend of $PM_{2.5}$ levels over time.

This visualization confirms the suitability of the dataset for forecasting tasks and highlights the importance of seasonal modeling approaches. In particular, models that can account for periodic patterns, such as SARIMA, are expected to perform well in capturing the dynamics shown in the figure. The successful application of MICE in this context enhances the reliability of subsequent analyses

Received: March 12, 2025. Reviewed: April 12, 2025. Accepted: April 16, 2025.

and strengthens the statistical foundation of the study.

C. Models

1) Multiple Linear Regression (MLR): MLR was used to model the linear relationship between $PM_{2.5}$ and meteorological variables. No regularization was applied to retain interpretability. Research has shown that MLR can use weather data to predict $PM_{2.5}$ with the same level of success as other statistical models like random forests [17], [18]. When used in a study to predict indoor $PM_{2.5}$, MLR performed well with a cross-validation R^2 of 60.48%, showing that it is reliable for such an application [17].

2) ARIMA: ARIMA is a regularly applied model in air quality research because it gives stable predictions. Ramadan et al., for example, designed customized ARIMA models to enhance the accuracy in the forecast of pollutants and guide air quality policy in urban areas like Abu Dhabi [7]. Koleva et al. also applied ARIMA in daily pollution data and proved the ability in tracing the trend in $PM_{2.5}$ [19]. Muzakki et al. also vouched for the fact that ARIMA is able to describe the manner in which air pollutants are sustained in the long term and therefore apt in forecasting future concentration [20]. In this study ARIMA model was trained using Statsmodels. Initial stationarity was tested using the Augmented Dickey-Fuller (ADF) test. To identify optimal parameters (p, d, q), the auto_arima() function from the pmdarima package was used with stepwise selection and AIC minimization.

3) SARIMA: SARIMA is a more sophisticated version of ARIMA, considering seasonal patterns, e.g., daily, monthly, or yearly cycles. SARIMA is found to be more precise compared to ARIMA in accommodating these seasonal patterns in $PM_{2.5}$ data [21]. It is important because $PM_{2.5}$ does not remain constant throughout the year. Recent studies show that SARIMA is better in accuracy measures (RMSE and MAE), reflecting the capability of SARIMA in accommodating seasonal and long-run patterns in air pollution data [21], [22]. We used seasonal order (P, D, Q, s). It was manually tuned based on prior decomposition and AIC minimization. The chosen model was SARIMA(1, 0, 1)(1, 0, 1, 12), assuming monthly seasonality.

4) Generalized Additive Model (GAM): GAM is a good method in predicting PM2.5 levels since this method can deal with multiple forms of environmental information. The adoption of ground measurements and satellite data is shown as contributing to the ability of GAM to estimate levels of PM2.5 over a global scale [23]. GAM is used with $PM_{2.5}$ pollution to analyze the association with Kawasaki disease and we show that it can capture complex mapping between environmental and health data [24]. GAM was trained using the pyGAM package. Spline smoothers were applied to the most relevant features. The smoothing parameter (lambda) was selected using a grid search over the range 10^{-3} to 10^{3}. This allowed the model to adaptively fit non-linear relationships.

5) Hybrid Statistical Models: To further enhance the predictability of $PM_{2.5}$, the current research utilizes the strengths of different statistical techniques in hybrid models. The objective is to capture both the linear and non-linear relations and time patterns in the data more effectively.

- MLR + ARIMA: MLR models the relationship between PM2.5 and meteorological variables, while ARIMA models the residuals to capture time-based trends.
- MLR + SARIMA: Similar to the above, but SARIMA accounts for seasonality in the residuals, improving performance in seasonal patterns.
- MLR + GAM: MLR handles linear effects; GAM models the non-linear patterns left in the residuals, enhancing flexibility.
- GAM + ARIMA: GAM captures complex non-linear relationships, and ARIMA handles the remaining temporal structure.

While ML models have gained extensive use in air quality forecasting, statistical models are discussed here to first assess their performance on air pollution data in the city of Almaty. Statistical models are easier to interpret and are better suited for the analysis of relationships between variables and temporal-based patterns. The initial analysis here sets a strong benchmark and provides a basis for understanding the structure in the data. Future studies will follow on from the current research through comparisons with the performance of advanced ML and deep learning (DL) models in order to determine the value added in forecasting levels of $PM_{2.5}$.

D. Evaluation Metrics

The $PM_{2.5}$ prediction model accuracy was measured using Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R^2 Score. These metrics provide a general measure of model accuracy and reliability.

Received: March 12, 2025. Reviewed: April 12, 2025. Accepted: April 16, 2025.

1) Mean Absolute Error (MAE): MAE measures the average absolute difference between predicted ($\hat{y}_i$) and actual (y_i) values:

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{1}$$

A lower MAE indicates better predictive accuracy, as it represents the average error magnitude.

2) Root Mean Squared Error (RMSE): RMSE evaluates the standard deviation of prediction errors, penalizing larger deviations more heavily:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{2}$$

Lower RMSE values indicate better model performance, particularly in handling variations and extreme fluctuations in $PM_{2.5}$ levels.

3) R^2 Score (Coefficient of Determination): The R^2 score measures how well the model explains variance in $PM_{2.5}$ concentrations:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{3}$$

where $\bar{y}$ is the mean of actual values. An R^2 score closer to 1 suggests a stronger fit between predictions and observations.

These steps collectively give an assessment of model accuracy, guiding the selection of optimal prediction strategy.

IV. Results and Discussion

Comparison of statistical and hybrid models for $PM_{2.5}$ prediction in Almaty provides valuable information on their ability to capture patterns and trends in air pollution data. The evaluation metrics used include MAE, RMSE, and R^2 score. Visualizations further illustrate how well each model tracks changes in $PM_{2.5}$ levels over time. Table III summarizes the results across all standalone and hybrid models. These metrics enable a comprehensive assessment of both error magnitude and explanatory power, allowing for fair comparison between models of varying complexity.

TABLE III
Performance Comparison Across Different Models

Statistical Models			
	MAE	RMSE	R^2
Multiple Linear Regression (MLR)	0.3831	0.5268	0.7160
ARIMA	0.4235	0.5224	0.6056
SARIMA	0.4156	0.2719	0.6058
Generalized Additive Model (GAM)	0.4415	0.5701	0.5357
Hybrid Models			
	MAE	RMSE	R^2
MLR + ARIMA	0.4248	0.5273	0.6027
MLR + SARIMA	0.4258	0.5281	0.6015
MLR + GAM	0.3944	0.5209	0.6124
GAM + ARIMA	0.4052	0.5772	0.5240

Received: March 12, 2025. Reviewed: April 12, 2025. Accepted: April 16, 2025.

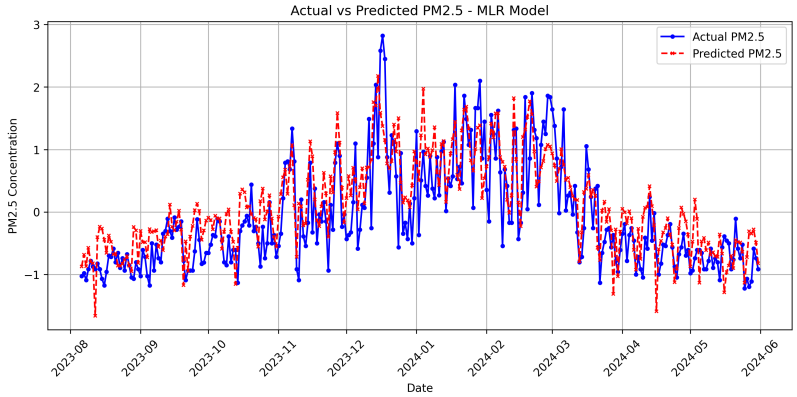

Fig. 3. Forecasted vs. actual $PM_{2.5}$ levels for MLR model

1) Statistical Models: Statistical models offer a strong baseline for time series forecasting by modeling linear relationships and temporal dependencies. Multiple Linear Regression (MLR) performed well, achieving an MAE of 0.3831, an RMSE of 0.5268, and an R^2 of 0.7160. Despite its simplicity and assumption of linearity, MLR was the best performing model in terms of R^2, indicating its strength to capture the relationship between meteorological variables and $PM_{2.5}$ levels. However, the relatively high RMSE shows that it may not fully capture more complex variations. To visualize how well the Multiple Linear Regression model tracks $PM_{2.5}$ levels over time, Figure 3 shows a comparison between actual and predicted values on the test set.

Figure 3 demonstrates that the MLR model captures the general trends and seasonal patterns in $PM_{2.5}$ levels, with predicted values (red dashed line) closely tracking actual observations (blue solid line) throughout the test period. The model performs well during periods of moderate pollution and maintains a consistent alignment between predicted and actual values.

However, deviations become more noticeable during peak pollution events, particularly in winter months. The model tends to underpredict extreme spikes and overpredict during sudden drops. This is a known limitation of linear models. They may struggle to fully capture nonlinear interactions between meteorological variables and pollutant concentrations. Despite this, the overall alignment between the two series is satisfactory, reflecting the strength of MLR in modeling long-term pollution behavior driven by dominant weather patterns.

The analysis confirms the utility of MLR as a baseline statistical model for $PM_{2.5}$ forecasting. Its simplicity, interpretability, and strong explanatory power make it a reliable first step in air quality modeling. Nonetheless, more advanced or hybrid approaches may be required to improve performance during extreme events and capture complex dependencies in the data.

ARIMA achieved a slightly better RMSE (0.5224) than MLR but a lower R^2 score (0.6056) and higher MAE (0.4235), suggesting that while ARIMA is effective in modeling temporal patterns, it may miss important external influences. To assess how well the ARIMA model captures temporal patterns in $PM_{2.5}$ levels, Figure 4 compares the model's predictions against the actual observations. As a time-series model, ARIMA is expected to track short-term dependencies, though it does not explicitly account for seasonality.

SARIMA showed a much lower RMSE (0.2719), reflecting strong short-term predictive accuracy and the ability to model seasonal fluctuations. However, its R^2 (0.6058) was similar to ARIMA, indicating that its overall explanatory power was not significantly higher. Figure 5 presents the SARIMA model's forecasts compared to actual $PM_{2.5}$ values.

GAM, which allows for non-linear relationships, had the lowest R^2 (0.5357), with an MAE of 0.4415 and RMSE of 0.5701. This suggests that despite its flexibility, GAM alone did not provide significant gains in this context, possibly due to the nature of the data or interactions between variables. To explore the performance of a non-linear model, Figure 6 shows the results of the Generalized Additive Model (GAM).

2) Hybrid Models: Hybrid models were applied to combine the strengths of individual approaches. MLR + ARIMA produced an R^2 of 0.6027, MAE of 0.4248, and RMSE of 0.5273 - very similar to standalone ARIMA, indicating little added benefit from combining the two. MLR + SARIMA followed a similar pattern, with an RMSE of 0.5281 and R^2 of 0.6015.

MLR + GAM showed the best performance among hybrid models, with an R^2 of 0.6124, an RMSE of 0.5209, and an MAE of 0.3944. This indicates a slight improvement, likely due to the combination of MLR's structure and GAM's ability to model

Received: March 12, 2025. Reviewed: April 12, 2025. Accepted: April 16, 2025.

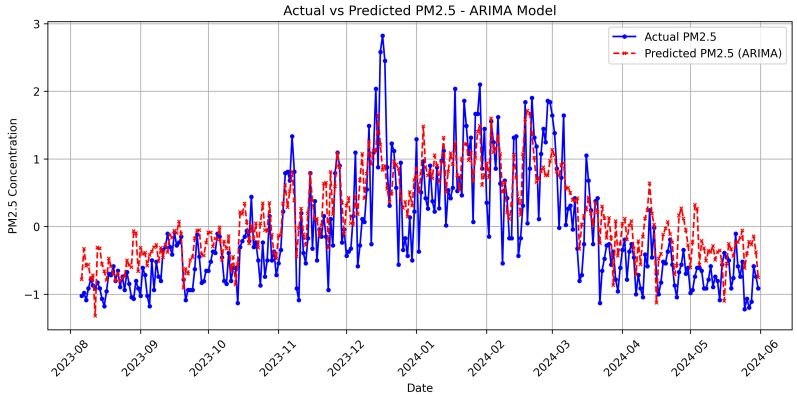

Fig. 4. Forecasted vs. actual $PM_{2.5}$ levels for ARIMA model

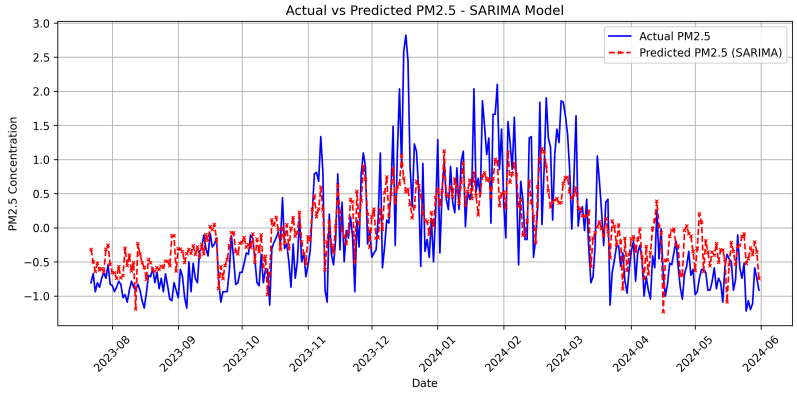

Fig. 5. Forecasted vs. actual $PM_{2.5}$ levels for SARIMA model

non-linear effects. Figure 7 illustrates the predictive performance of the hybrid MLR + GAM model.

On the other hand, GAM + ARIMA performed the worst among hybrid models, with an R^2 of 0.5240, RMSE of 0.5772, and MAE of 0.4052. This suggests that combining two flexible but complex models does not necessarily lead to better results and may introduce redundancy or overfitting.

Although hybrid models were expected to outperform individual models, improvements were minimal. One reason is that models like GAM already captured much of the variation in the data, leaving little structure for ARIMA or SARIMA to model further. Furthermore, long periods of missing values were imputed using MICE. It may have smoothed out key time series patterns. The relatively small dataset (1,558 records) may also have limited the effectiveness of more complex, multistage models.

3) Key Findings and Implications: Overall, statistical models offered a solid baseline for $PM_{2.5}$ forecasting in Almaty. MLR performed best in explaining variance, while SARIMA achieved the lowest RMSE, highlighting its strength in short-term and seasonal forecasting. Among hybrid models, MLR + GAM was the most promising, and showed a modest gain.

These results suggest that while combining models can add flexibility, it does not guarantee better generalization. The study also highlights the limitations of current approaches and the potential benefit of exploring machine learning or deep learning techniques in future research. Ensemble learning and enhanced feature engineering, especially incorporating real-time traffic, industrial activity, and emission data, could significantly improve predictive performance.

Received: March 12, 2025. Reviewed: April 12, 2025. Accepted: April 16, 2025.

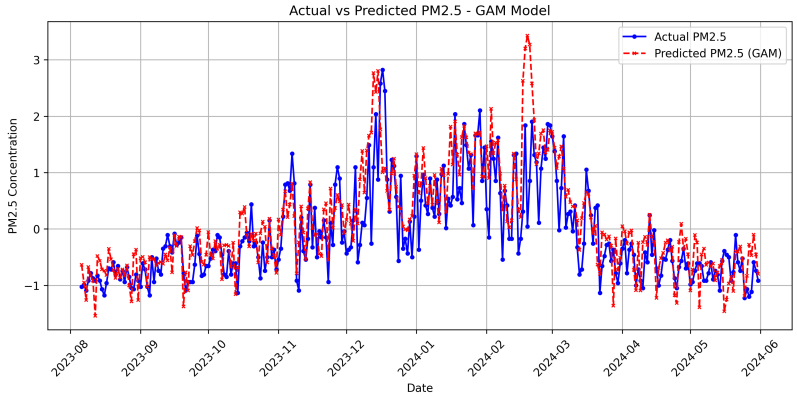

Fig. 6. Forecasted vs. actual $PM_{2.5}$ levels for GAM model

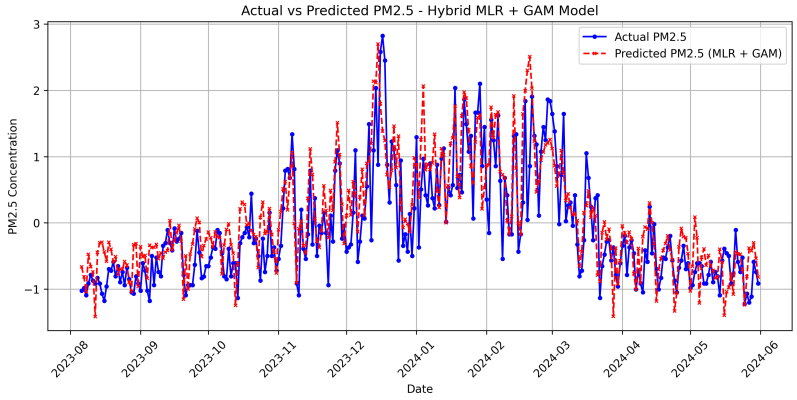

Fig. 7. Forecasted vs. actual $PM_{2.5}$ levels for the hybrid MLR GAM model

4) Practical Relevance: These findings offer practical value for city officials and environmental agencies in Almaty. By using these models to forecast $PM_{2.5}$ levels, they can take earlier action, such as issuing health warnings, managing traffic, or limiting industrial operations on high pollution days. Since the models rely on data that is already being collected, they offer a cost-effective tool for real-time air quality management, ultimately helping protect public health, especially for vulnerable groups.

V. Conclusion and Future Work

Among the compared models, Multiple Linear Regression (MLR) was the most competent in describing the relationship between weather conditions and $PM_{2.5}$ concentrations. Having an R^2 of 0.7160, it successfully captured pollution trends with reasonable accuracy. Despite the assumption of linearity in MLR, its predictions were robust, as shown by the Mean Absolute Error (MAE) of 0.3831 and Root Mean Squared Error (RMSE) of 0.5268.

For seasonal trends, the SARIMA model provided the lowest RMSE (0.2719), it was the best performing for short-term prediction. Its R^2 value (0.6058) was, however, less than that of MLR, suggesting that while SARIMA is effective at modeling seasonal variation, it may not capture long-term pollution trends.

Received: March 12, 2025. Reviewed: April 12, 2025. Accepted: April 16, 2025.

Among the hybrid models, the best performing one was MLR combined with the Generalized Additive Model (MLR + GAM). With an R^2 of 0.6124 and an RMSE of 0.5209, this model demonstrated that combining MLR's structured approach with GAM's ability to capture nonlinear trends led to moderate improvements over traditional statistical methods.

In general, MLR was the best model for explaining $PM_{2.5}$ variations, and SARIMA was the most accurate for short-term forecasting. These findings can help in the development of more efficient air pollution control strategies for Almaty and other cities.

This study is designed as the first phase of a broader investigation of $PM_{2.5}$ forecasting in Almaty. In this phase here, the key focus is right on statistical as well as hybrid statistical models in order to establish a solid baseline to understand the full structure and behavior of local air quality data. For future research work, by integrating machine learning (ML) along with deep learning (DL) models – like Random Forests, Gradient Boosting, and LSTM networks – may very well further improve overall forecasting accuracy, particularly in the handling of nonlinear patterns, interactions, and much longer time dependencies. A comparative analysis between statistical, machine learning (ML), and deep learning (DL) approaches on the same dataset can provide a broader understanding of their respective strengths and help identify the most effective tools to support air quality management in Almaty and similar cities. Future research can focus on developing more adaptive air quality forecasting systems, as well as truly scalable, real-time solutions that directly inform public health strategies and environmental policy.

REFERENCES

[1] A. Bekbossynova, D. Duvanova, N. Jones, K. Lyden, T. McGinley, and H. Moss, "How attitudes towards air pollution may impact public health: a case study of almaty, kazakhstan," *Journal of Environmental Protection*, vol. 14, no. 07, pp. 583–601, 2023. [Online]. Available: https://doi.org/10.4236/jep.2023.147034

[2] T. B. Ogbuabia, M. Guney, N. Baimatova, I. Ulusoy, and F. Karaca, "Assessing the impact of combined heat and power plants (chpps) in central asia: a case study in almaty for pm2.5 simulations using wrf-aermod and ground level verification," *Atmosphere*, vol. 14, no. 10, p. 1554, 2023. [Online]. Available: https://doi.org/10.3390/atmos14101554

[3] S. Jalali, M. Karbakhsh, M. Momeni, M. Taheri, S. B. Amini, M. Mansourian, and N. Sarrafzadegan, "Long-term exposure to pm2.5 and cardiovascular disease incidence and mortality in an eastern mediterranean country: findings based on a 15-year cohort study," 2021, preprint. [Online]. Available: https://doi.org/10.21203/rs.3.rs-142122/v1

[4] P. Nath, P. Saha, A. I. Middya, and S. Roy, "Long-term time-series pollution forecast using statistical and deep learning methods," *Neural Computing and Applications*, vol. 33, no. 19, pp. 12 551–12 570, 2021. [Online]. Available: https://doi.org/10.1007/s00521-021-05901-2

[5] B. Liu, Y. Jin, X. De-zhi, Y. Wang, and C. Li, "A data calibration method for micro air quality detectors based on a lasso regression and narx neural network combined model," *Scientific Reports*, vol. 11, no. 1, 2021. [Online]. Available: https://doi.org/10.1038/s41598-021-00804-7

[6] X. Qu and Y. Cao, "Empirical analysis of air quality in china based on multiple linear regression analysis," in *Second International Conference on Statistics, Applied Mathematics, and Computing Science (CSAMCS 2022)*, 2023. [Online]. Available: https://doi.org/10.1117/12.2671882

[7] M. S. Ramadan, A. Abuelgasim, and N. A. Hosani, "Advancing air quality forecasting in abu dhabi, uae using time series models," *Frontiers in Environmental Science*, vol. 12, 2024. [Online]. Available: https://doi.org/10.3389/fenvs.2024.1393878

[8] H. Bouzghiba, A. Mendyl, K. Khomsi, and G. Géczi, "Short-term predictions of pm10 and no2 concentrations in urban environments based on arima search grid modeling," *CLEAN – Soil, Air, Water*, vol. 52, no. 6, 2024. [Online]. Available: https://doi.org/10.1002/clen.202300395

[9] L. Zhang, X. Tian, Y. Zhao, L. Liu, Z. Li, L. Tao, and Y. Luo, "Application of nonlinear land use regression models for ambient air pollutants and air quality index," *Atmospheric Pollution Research*, vol. 12, no. 10, p. 101186, 2021. [Online]. Available: https://doi.org/10.1016/j.apr.2021.101186

[10] J. D. Kurniawan, H. A. Parhusip, and S. Trihandaru, "Predictive performance evaluation of arima and hybrid arima-lstm models for particulate matter concentration," *Jurnal Online Informatika*, vol. 9, no. 2, pp. 259–268, 2024. [Online]. Available: https://doi.org/10.15575/join.v9i2.1318

[11] Džaferović and Karauzović-Hadžiabdić, "Air quality prediction using machine learning methods: A case study of bjelave neighborhood, sarajevo, bih," in *Proceedings of the International Conference*. Springer, 2020. [Online]. Available: https://doi.org/10.1007/978-3-030-54765-3_29

[12] M. D. Yazdi *et al.*, "Predicting fine particulate matter (pm2.5) in the greater london area: An ensemble approach using machine learning methods," *Remote Sensing*, vol. 12, no. 6, 2020. [Online]. Available: https://doi.org/10.3390/rs12060914

Received: March 12, 2025. Reviewed: April 12, 2025. Accepted: April 16, 2025.

[13] Badrakh and Choimaa, "Air quality predictions of ulaanbaatar using machine learning approach," in *Proceedings of the 24th International Conference on Computing in High Energy and Nuclear Physics (CHEP 2019)*, 2021. [Online]. Available: https://doi.org/10.22323/1.378.0012

[14] A. Dairi *et al.*, "Integrated multiple directed attention-based deep learning for improved air pollution forecasting," *IEEE Transactions on Instrumentation and Measurement*, 2021. [Online]. Available: https://doi.org/10.1109/tim.2021.3091511

[15] R. Idroes *et al.*, "Urban air quality classification using machine learning approach to enhance environmental monitoring," *Lhokseumawe Journal of Engineering and Science (LJES)*, vol. 1, no. 2, 2023. [Online]. Available: https://doi.org/10.60084/ljes.v1i2.99

[16] T. V. Vu *et al.*, "Assessing the impact of clean air action on air quality trends in beijing using a machine learning technique," *Atmospheric Chemistry and Physics*, vol. 19, pp. 11 303–11 314, 2019. [Online]. Available: https://doi.org/10.5194/acp-19-11303-2019

[17] Y. Shi, Z. Du, J. Zhang, F. Han, F. Chen, D. Wang, and S. Sui, "Construction and evaluation of hourly average indoor pm2.5 concentration prediction models based on multiple types of places," *Frontiers in Public Health*, vol. 11, 2023. [Online]. Available: https://doi.org/10.3389/fpubh.2023.1213453

[18] A. Agibayeva, R. Khalikhan, M. Güney, F. Karaca, A. Torezhan, and E. Avcu, "An air quality modeling and disability-adjusted life years (daly) risk assessment case study: Comparing statistical and machine learning approaches for pm2.5 forecasting," *Sustainability*, vol. 14, no. 24, p. 16641, 2022. [Online]. Available: https://doi.org/10.3390/su142416641

[19] S. Koleva, S. Gocheva-Ilieva, and H. Kulina, "Stochastic modelling of daily air pollution in burgas, bulgaria," *Journal of Physics: Conference Series*, vol. 2675, no. 1, p. 012003, 2023. [Online]. Available: https://doi.org/10.1088/1742-6596/2675/1/012003

[20] N. F. Muzakki, A. Z. Putri, S. Maruli, and F. Kartiasih, "Forecasting the air quality index by utilizing several meteorological factors using the arimax method (case study: Central jakarta city)," *Jurnal JTIK (Jurnal Teknologi Informasi dan Komunikasi)*, vol. 8, no. 3, pp. 569–586, 2024. [Online]. Available: https://doi.org/10.35870/jtik.v8i3.2012

[21] T. Bunnag, "Forecasting pm10 caused by bangkok's leading greenhouse gas emission using the sarima and sarima-garch model," *International Journal of Energy Economics and Policy*, vol. 14, no. 1, pp. 418–426, 2024. [Online]. Available: https://doi.org/10.32479/ijeep.15275

[22] G. Reddy, M. Manjunath, R. Patil, and P. Kulkarni, "Predicting potential evapotranspiration for kalaburagi district using a seasonal arima model," *International Journal of Environment and Climate Change*, vol. 13, no. 11, pp. 2073–2082, 2023. [Online]. Available: https://doi.org/10.9734/ijecc/2023/v13i113367

[23] M. S. Hammer, A. v. Donkelaar, C. Li, A. Lyapustin, A. M. Sayer, N. C. Hsu, and R. V. Martin, "Global estimates and long-term trends of fine particulate matter concentrations (1998–2018)," *Environmental Science & Technology*, vol. 54, no. 13, pp. 7879–7890, 2020. [Online]. Available: https://doi.org/10.1021/acs.est.0c01764

[24] F. Si, C. Zhou, Y. Yang, and L. Huang, "Study of the relationship between occurrence of kawasaki disease and air pollution in chengdu by parametric and semi-parametric models," *Environmental Science and Pollution Research*, vol. 30, no. 55, pp. 117 706–117 714, 2023. [Online]. Available: https://doi.org/10.1007/s11356-023-30533-5

Received: March 12, 2025. Reviewed: April 12, 2025. Accepted: April 16, 2025.