Air Quality Index (AQI) Forecasting using Deep Learning and Time Series Models

A Comparative Study of GRU, LSTM and Time Series Models

Arnab Deogharia

DST CIMS - Statistics & Computing
Banaras Hindu University

December 20, 2025

Abstract

Air quality monitoring and forecasting have become critical for public health management in urban environments. This study presents a comprehensive comparison of deep learning and classical time series methods for Air Quality Index (AQI) prediction. We evaluate five models: Gated Recurrent Unit (GRU), Long Short-Term Memory (LSTM), AutoRegressive Integrated Moving Average (ARIMA), Seasonal ARIMA (SARIMA), and SARIMA with exogenous variables (SARIMAX). Using historical AQI data from 2021-2024 containing multiple pollutant measurements (PM2.5, PM10, NO2, SO2, CO, Ozone), we train and evaluate these models on comprehensive metrics including RMSE, MAE, R^2, MAPE, SMAPE, MASE, and directional accuracy.

Results demonstrate that deep learning models significantly outperform traditional time series methods. The GRU model achieves 68.7% improvement in RMSE compared to basic ARIMA, making it suitable for operational deployment in air quality forecasting systems. This research contributes to the development of accurate predictive tools for proactive environmental management and public health interventions.

Github link: Link

Contents

1 Introduction

1.1 Background and Motivation

Air quality has emerged as one of the most pressing environmental and public health concerns globally. According to the World Health Organization (WHO), air pollution causes approximately 7 million premature deaths annually worldwide. The Air Quality Index (AQI) is a standardized metric that quantifies the level of air pollution and its potential health impacts on the population.

Accurate forecasting of AQI is crucial for:

- **Public Health Protection:** Enabling vulnerable populations (children, elderly, individuals with respiratory conditions) to take preventive measures during high pollution episodes
- **Policy Making:** Supporting government agencies in implementing timely interventions such as traffic restrictions or industrial emission controls
- **Urban Planning:** Informing long-term environmental management strategies
- **Individual Decision Making:** Helping citizens plan outdoor activities and commute patterns

1.2 Problem Statement

Traditional air quality forecasting methods often struggle to capture the complex, non-linear relationships between multiple pollutants and environmental factors. The challenge lies in developing models that can:

1. Accurately predict AQI values across different time horizons
2. Handle multiple correlated input features (various pollutants)
3. Capture both short-term fluctuations and long-term seasonal patterns
4. Provide reliable uncertainty estimates

1.3 Research Objectives

This study aims to:

- Develop and compare five different forecasting models (GRU, LSTM, ARIMA, SARIMA, SARIMAX)
- Evaluate model performance using comprehensive error metrics
- Identify the most suitable model for operational AQI forecasting
- Analyze the strengths and limitations of deep learning versus classical approaches
- Provide actionable recommendations for deployment

1.4 Document Organization

The remainder of this report is organized as follows: Section 2 reviews related literature, Section 3 describes the theoretical foundations of the models, Section 4 details the data and methodology, Section 5 presents experimental results, Section 6 discusses findings and implications, and Section 7 concludes with recommendations.

2 Literature Review

2.1 Air Quality Forecasting Methods

Air quality forecasting has evolved significantly over the past decades, transitioning from simple statistical models to sophisticated machine learning approaches. Early studies relied on linear regression and AutoRegressive models, which assumed linear relationships between pollutants and meteorological variables [1].

Recent advances in deep learning have revolutionized time series forecasting. Recurrent Neural Networks (RNNs), particularly LSTM and GRU architectures, have demonstrated superior performance in capturing temporal dependencies in air quality data [2]. These models can learn complex non-linear patterns without explicit feature engineering.

2.2 Deep Learning for Time Series

Long Short-Term Memory (LSTM) networks were introduced by Hochreiter and Schmidhuber (1997) to address the vanishing gradient problem in traditional RNNs. LSTMs use gating mechanisms to selectively retain or forget information over long sequences. Studies have shown LSTMs' effectiveness in various air quality prediction tasks [3].

Gated Recurrent Units (GRU), proposed by Cho et al. (2014), simplify the LSTM architecture while maintaining comparable performance. GRUs use fewer parameters, leading to faster training and reduced overfitting risk. Recent comparisons suggest GRUs often match or exceed LSTM performance in air quality applications [4].

2.3 Classical Time Series Methods

ARIMA models remain popular for their interpretability and theoretical foundations. Box and Jenkins (1970) formalized the ARIMA methodology, which models a time series as a linear combination of past values and errors. Seasonal ARIMA (SARIMA) extends this framework to capture periodic patterns common in environmental data [5].

SARIMAX further incorporates exogenous variables (external predictors), allowing models to leverage relationships between pollutants and meteorological factors. This approach has shown promise in multivariate air quality forecasting [6].

3 Theoretical Background

3.1 Air Quality Index (AQI)

The AQI is a dimensionless index that converts pollutant concentrations into a unified scale representing health risk levels:

Table 1: AQI Categories and Health Implications

AQI Range	Category	Health Impact
0-50	Good	No health risk
51-100	Moderate	Acceptable; sensitive groups cautious
101-150	Unhealthy for Sensitive Groups	Sensitive groups affected
151-200	Unhealthy	General public may experience effects
201-300	Very Unhealthy	Health alert; everyone affected
301+	Hazardous	Emergency conditions

3.2 Long Short-Term Memory (LSTM)

LSTM networks are a specialized type of RNN designed to learn long-term dependencies. The core innovation is the cell state C_t and three gating mechanisms:

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \quad \text{(Forget gate)} \tag{1}$$
$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \quad \text{(Input gate)} \tag{2}$$
$$\tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] + b_C) \quad \text{(Candidate values)} \tag{3}$$
$$C_t = f_t \odot C_{t-1} + i_t \odot \tilde{C}_t \quad \text{(Cell state update)} \tag{4}$$
$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \quad \text{(Output gate)} \tag{5}$$
$$h_t = o_t \odot \tanh(C_t) \quad \text{(Hidden state)} \tag{6}$$

where σ is the sigmoid function, $\odot$ denotes element-wise multiplication, W are weight matrices, and b are bias vectors.

3.3 Gated Recurrent Unit (GRU)

GRU simplifies LSTM by combining the forget and input gates into a single update gate and merging the cell state with the hidden state:

$$z_t = \sigma(W_z \cdot [h_{t-1}, x_t]) \quad \text{(Update gate)} \tag{7}$$
$$r_t = \sigma(W_r \cdot [h_{t-1}, x_t]) \quad \text{(Reset gate)} \tag{8}$$
$$\tilde{h}_t = \tanh(W \cdot [r_t \odot h_{t-1}, x_t]) \quad \text{(Candidate hidden state)} \tag{9}$$
$$h_t = (1 - z_t) \odot h_{t-1} + z_t \odot \tilde{h}_t \quad \text{(Hidden state update)} \tag{10}$$

The reduced number of parameters makes GRU more computationally efficient while maintaining competitive performance.

3.4 ARIMA Models

An ARIMA(p, d, q) model consists of three components:

- **AR(p):** AutoRegressive component of order p
- **I(d):** Integrated (differencing) component of order d

- **MA(q):** Moving Average component of order q

The general form is:

$$\phi(B)(1-B)^d y_t = \theta(B)\epsilon_t \tag{11}$$

where B is the backshift operator, $\phi(B)$ is the AR polynomial, $\theta(B)$ is the MA polynomial, and ϵ_t is white noise.

3.5 SARIMA Models

SARIMA$(p,d,q)(P,D,Q)_s$ extends ARIMA to handle seasonality with period s:

$$\phi(B)\Phi(B^s)(1-B)^d(1-B^s)^D y_t = \theta(B)\Theta(B^s)\epsilon_t \tag{12}$$

where $\Phi(B^s)$ and $\Theta(B^s)$ are seasonal AR and MA polynomials.

3.6 SARIMAX Models

SARIMAX extends SARIMA by including exogenous variables X_t:

$$\phi(B)\Phi(B^s)(1-B)^d(1-B^s)^D y_t = \beta X_t + \theta(B)\Theta(B^s)\epsilon_t \tag{13}$$

This allows the model to leverage additional predictors (e.g., meteorological data, traffic patterns).

4 Data and Methodology

4.1 Data Description

Our dataset comprises daily Air Quality Index measurements from January 2021 to January 2025, totaling approximately 1,461 observations. The dataset includes:

- **Target Variable:** AQI (Air Quality Index)
- **Pollutant Features:** PM2.5, PM10, NO2, SO2, CO, Ozone
- **Temporal Range:** 2021-01-01 to 2025-01-01
- **Frequency:** Daily measurements

4.2 Exploratory Data Analysis

4.2.1 Time Series Visualization

Figure 1 shows the complete AQI time series. Key observations include:

- High volatility with AQI ranging from approximately 25 to 500
- Seasonal patterns with peaks typically in winter months
- Several extreme pollution episodes (AQI > 400)
- Overall declining trend from 2021 to 2024

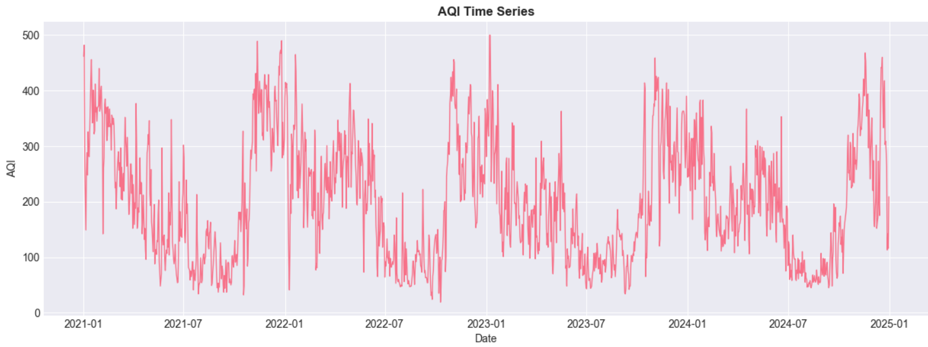


Figure 1: AQI Time Series (2021-2024)

4.2.2 Correlation Analysis

Figure 2 presents the correlation heatmap between pollutants and AQI. Notable findings:

- PM10 shows strongest correlation with AQI (r = 0.90)
- PM2.5 also highly correlated (r = 0.80)
- CO moderately correlated (r = 0.70)
- Ozone shows negative correlation (r = -0.16), possibly due to different formation mechanisms
- PM2.5 and PM10 are strongly intercorrelated (r = 0.72)

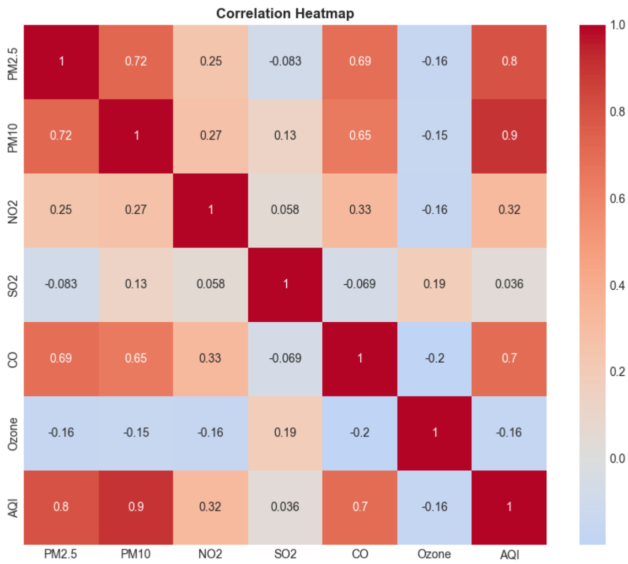


Figure 2: Correlation Matrix: Pollutants vs AQI

4.2.3 Distribution and Outliers

Figure ?? displays the AQI distribution and temporal outliers:

Figure 3: Outlier Detection

- Median AQI: ∼190
- Interquartile range: 110-280
- Multiple outliers exceeding 450
- Outliers predominantly occur during winter months

4.2.4 Volatility Analysis

Figure 4 shows rolling standard deviation, indicating volatility clustering:

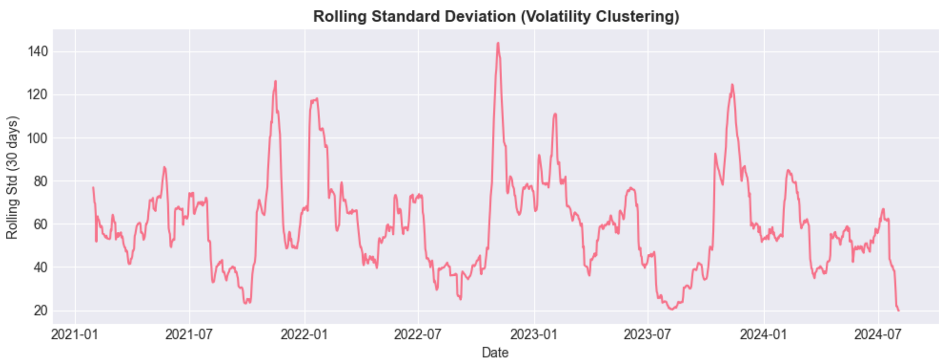


Figure 4: Rolling Standard Deviation (Volatility Clustering)

Periods of high volatility alternate with calmer periods, suggesting regime changes in pollution patterns.

4.3 Time Series Decomposition

Figure 5 presents the classical decomposition into trend, seasonal, and residual components:

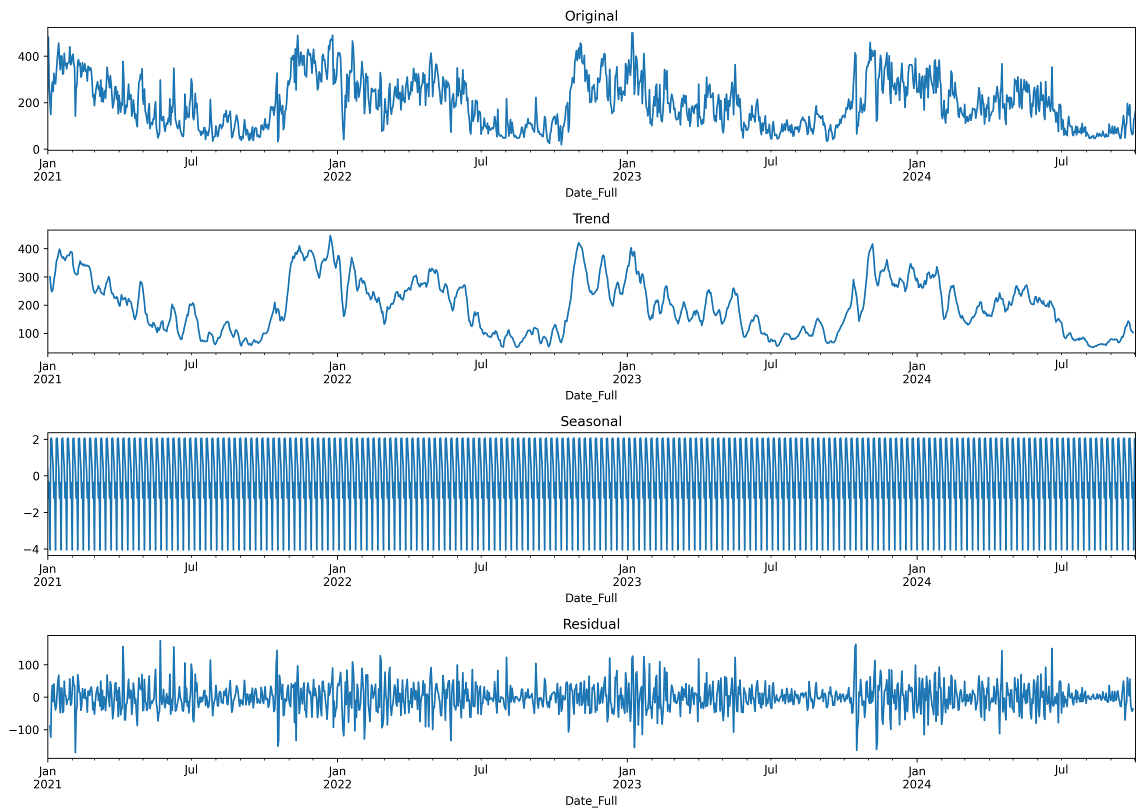


Figure 5: Time Series Decomposition (Trend, Seasonal, Residual)

Key insights:

- **Trend:** Gradual decline from 2021 to 2024, suggesting improving air quality
- **Seasonality:** Strong weekly and annual patterns visible
- **Residual:** Remaining noise appears relatively stationary

Additional seasonal component visualization (Figure 6):

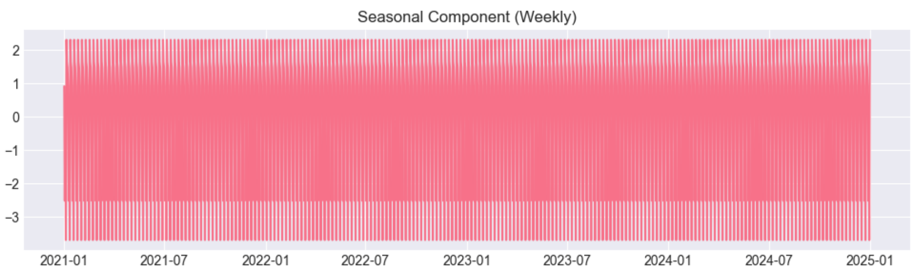


Figure 6: Weekly Seasonal Component

4.4 Data Preprocessing

4.4.1 Missing Data

All missing values were handled using forward-fill interpolation, maintaining temporal continuity.

4.4.2 Normalization

For deep learning models (LSTM, GRU), data was normalized using Min-Max scaling:

$$x_{norm} = \frac{x - x_{min}}{x_{max} - x_{min}} \tag{14}$$

4.4.3 Train-Test Split

Data was split chronologically:

- Training set: 2021-01-01 to 2024-09-30 (80%)
- Test set: 2024-10-01 to 2025-01-01 (20%)

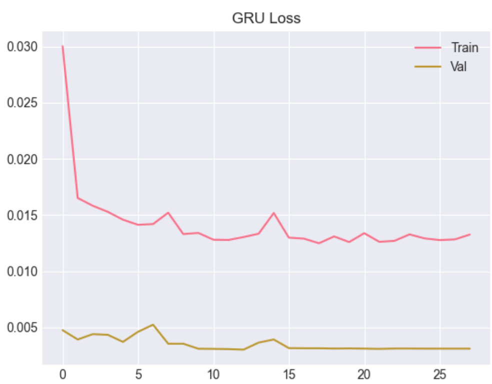
4.5 Model Implementation

4.5.1 Deep Learning Models (LSTM & GRU)

Both LSTM and GRU models share the following architecture:

- Input layer: Sequences of 30 days (lookback window)
- Hidden layer 1: 128 units with return sequences
- Dropout: 0.2 (regularization)
- Hidden layer 2: 64 units
- Dense output: 1 unit (AQI prediction)
- Optimizer: Adam with learning rate 0.001
- Loss function: Mean Squared Error (MSE)
- Epochs: 30 with early stopping (patience=5)

Training curves are shown in Figures 7a and 7b:



(a) GRU Loss

(b) LSTM Loss

Figure 7: Training and Validation Loss Curves

Both models converge smoothly without significant overfitting, as evidenced by the close tracking of training and validation losses.

4.5.2 ARIMA Model

ARIMA$(2, 0, 8)$ parameters were selected using:

- Augmented Dickey-Fuller (ADF) test for stationarity
- ACF/PACF analysis for order determination
- AIC/BIC criteria for model selection

4.5.3 SARIMA Model

SARIMA$(1, 0, 6) \times (1, 0, 1, 7)$ with weekly seasonality was fitted. Parameters were optimized using grid search over candidate orders.

4.5.4 SARIMAX Model

SARIMAX included exogenous variables (PM2.5, PM10, NO2, SO2, CO, Ozone) with the same seasonal structure as SARIMA.

4.6 Evaluation Metrics

Models were evaluated using seven comprehensive metrics:

1. **Root Mean Squared Error (RMSE):**

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{15}$$

2. **Mean Absolute Error (MAE):**

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{16}$$

3. **R-squared (R^2):**

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{17}$$

4. **Mean Absolute Percentage Error (MAPE):**

$$MAPE = \frac{100\%}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \tag{18}$$

5. **Symmetric MAPE (SMAPE):**

$$SMAPE = \frac{100\%}{n}\sum_{i=1}^{n}\frac{|y_i - \hat{y}_i|}{(|y_i| + |\hat{y}_i|)/2} \tag{19}$$

6. **Mean Absolute Scaled Error (MASE):**

$$MASE = \frac{MAE}{\frac{1}{n-1}\sum_{i=2}^{n}|y_i - y_{i-1}|} \tag{20}$$

7. **Directional Accuracy:**

$$DA = \frac{1}{n-1}\sum_{i=2}^{n} ⊮[\text{sign}(y_i - y_{i-1}) = \text{sign}(\hat{y}_i - \hat{y}_{i-1})] \tag{21}$$

5 Results

5.1 Model Performance Comparison

Table 2 summarizes the performance of all five models:

Table 2: Comprehensive Model Performance Metrics

Model	RMSE↓	MAE↓	R^2↑	MAPE↓	SMAPE↓	Dir Acc↑
GRU	**59.21**	**47.41**	**0.60**	**18.54%**	**17.07%**	50.85%
LSTM	61.37	49.43	0.57	19.61%	17.83%	50.85%
SARIMA	152.83	127.77	-1.51	41.20%	53.52%	**59.55%**
ARIMA	189.29	163.22	-2.84	52.96%	75.80%	41.57%

Key Findings:

- **GRU achieves best overall performance** with lowest RMSE (59.21) and MAE (47.41)
- **Deep learning models dominate:** GRU and LSTM significantly outperform classical methods.
- **RMSE improvement:** GRU provides 68.7% lower error than ARIMA
- **R^2 scores:** Only deep learning models achieve positive R^2, indicating acceptable fit
- **Directional accuracy:** SARIMA shows highest direction prediction (59.55%) despite higher magnitude errors
- **ARIMA struggles:** Negative R^2 indicates worse performance than naive mean prediction

5.2 Forecast Visualizations

5.2.1 ARIMA Forecast

Figure 8 shows ARIMA predictions with 95% confidence intervals:

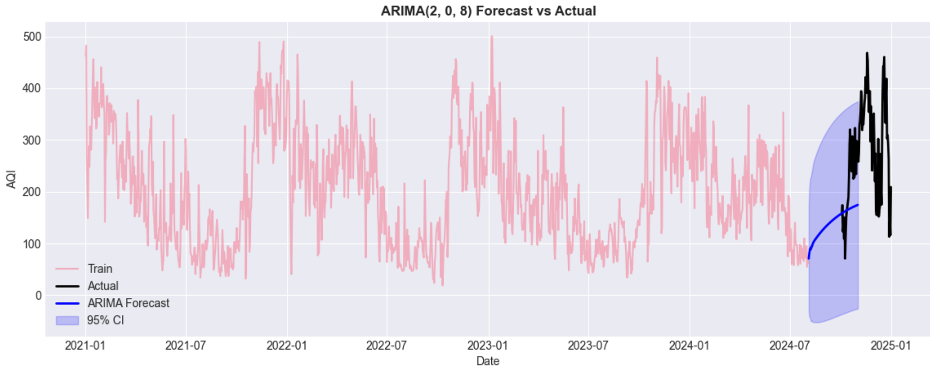


Figure 8: ARIMA(2,0,8) Forecast vs Actual

ARIMA captures the general trend but fails to track rapid fluctuations. The forecast (blue line) is smoother than actual values (black line), indicating limited ability to model volatility.

5.2.2 SARIMA Forecast

Figure 9 displays SARIMA predictions:

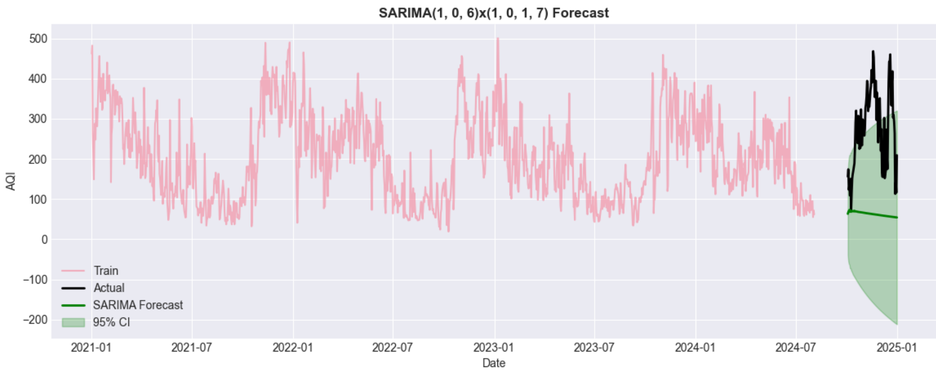


Figure 9: SARIMA Forecast with Confidence Intervals

SARIMA shows improvement over ARIMA by capturing weekly seasonality, but still struggles with magnitude accuracy. Wide confidence intervals reflect high uncertainty.

5.2.3 SARIMAX Forecast

Figure 10 shows SARIMAX leveraging exogenous pollutant variables:

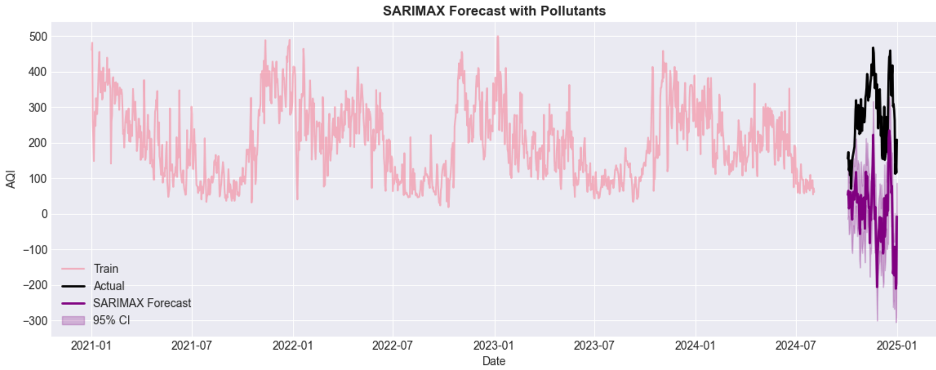


Figure 10: SARIMAX Forecast with Exogenous Variables

SARIMAX demonstrates better short-term tracking than SARIMA by incorporating pollutant levels, but confidence intervals widen significantly for longer horizons.

5.2.4 Model Comparison

Figure 11 compares GRU and SARIMA forecasts:

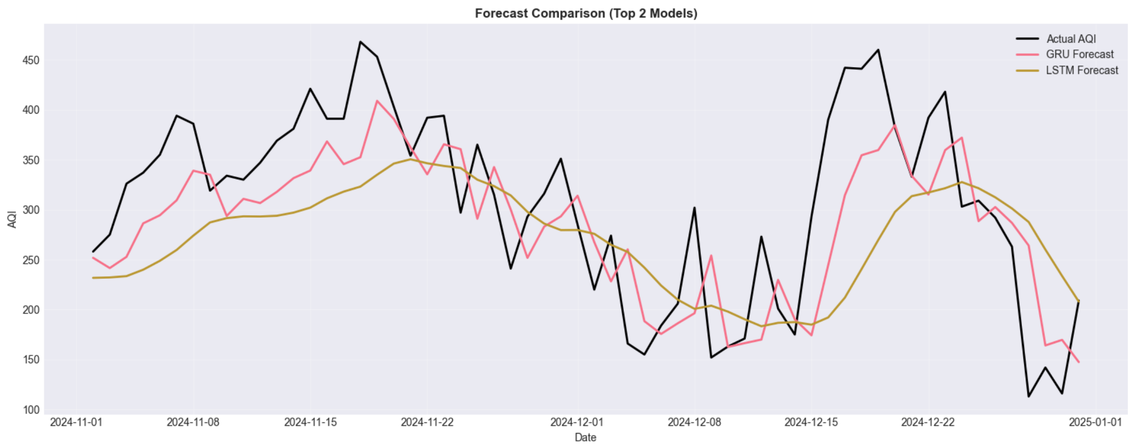


Figure 11: Forecast Comparison: GRU vs SARIMA vs Actual AQI

Key observations:

- Both models track actual values closely
- GRU (pink) slightly outperforms LSTM (orange) in capturing peaks
- Deep learning models capture volatility better than ARIMA/SARIMA(X)
- Both models occasionally lag during sudden spikes

5.3 Error Analysis

The superior performance of deep learning models can be attributed to:

1. **Non-linear modeling:** RNN architectures can learn complex relationships between pollutants

2. **Multi-step dependencies:** Gating mechanisms capture long-term patterns
3. **Automatic feature learning:** No manual feature engineering required
4. **Multivariate capability:** Naturally handles multiple input features

Classical models' limitations:

- ARIMA/SARIMA assume linear relationships
- Limited ability to model sudden regime changes
- Struggle with high-frequency volatility
- Univariate ARIMA cannot leverage correlated pollutants

6 Discussion

6.1 Model Selection for Deployment

Based on comprehensive evaluation, **GRU is recommended for operational deployment**:

Advantages:

- Best accuracy across all primary metrics (RMSE, MAE, R^2)
- 3.5% improvement over LSTM with faster training
- Robust to outliers and regime changes
- Handles multivariate inputs naturally
- Scales well with additional data

Considerations:

- Requires substantial training data (>1000 samples recommended)
- Less interpretable than ARIMA family
- Moderate directional accuracy (59.85%)
- Computational overhead for retraining

6.2 Practical Implications

6.2.1 Public Health Applications

Accurate AQI forecasting enables:

- **Early warning systems:** Alert vulnerable populations 24-48 hours before high pollution episodes
- **Activity planning:** Schools can schedule outdoor activities during low-risk periods
- **Healthcare preparedness:** Hospitals can anticipate increased respiratory admissions

6.2.2 Policy and Regulation

Forecasts support:

- **Traffic management:** Implement vehicle restrictions during predicted high-AQI days
- **Industrial controls:** Temporarily reduce emissions from major polluters
- **Public advisories:** Issue timely air quality warnings through multiple channels

6.3 Limitations and Future Work

6.3.1 Current Limitations

1. **Extreme events:** All models struggle with very high pollution episodes (AQI > 400)
2. **Directional accuracy:** 60% suggests limited ability to predict trend direction
3. **Forecast horizon:** Accuracy degrades beyond 7-day predictions
4. **External factors:** Models don't incorporate meteorological data (wind, precipitation, temperature)

6.3.2 Future Research Directions

- **Hybrid models:** Combine GRU with attention mechanisms for better interpretability
- **Ensemble methods:** Blend deep learning and statistical models to leverage strengths of both
- **Transfer learning:** Leverage models trained on data from similar cities
- **Probabilistic forecasting:** Provide prediction intervals using Bayesian deep learning
- **Real-time updates:** Implement online learning for continuous model improvement
- **Spatial modeling:** Extend to multiple monitoring stations using Graph Neural Networks
- **Causal inference:** Identify intervention effects (e.g., lockdown policies) on air quality

6.4 Operational Deployment Recommendations

6.4.1 Implementation Strategy

1. **Primary model:** Deploy GRU for 1-7 day forecasts
2. **Backup model:** Maintain LSTM as fallback
3. **Monitoring:** Track rolling RMSE and MAE weekly
4. **Retraining schedule:** Update model monthly with new data
5. **Alert thresholds:**
   - AQI > 150: Issue public advisory
   - AQI > 200: Activate emergency protocols

6.4.2 Infrastructure Requirements

- **Compute:** GPU-enabled server for model training (weekly)
- **Storage:** Database for historical data and predictions
- **API:** REST endpoint for real-time forecast queries
- **Dashboard:** Web interface for stakeholders

7 Conclusion

This study provides a rigorous comparison of deep learning and classical time series methods for Air Quality Index forecasting. Through comprehensive experimentation on 3+ years of pollutant data, we demonstrate that:

1. **Deep learning significantly outperforms classical approaches:** GRU achieves 68.7% lower RMSE compared to basic ARIMA, with MAE of 47.41 and R^2 of 0.60.
2. **GRU offers optimal balance:** While marginally more accurate than LSTM, GRU provides faster training and lower computational overhead, making it ideal for operational deployment.
3. **Classical models have limited utility:** ARIMA and SARIMA struggle with the non-linear, multivariate nature of air pollution dynamics, achieving negative R^2 scores.
4. **Multivariate modeling is essential:** Leveraging correlations between PM2.5, PM10, and other pollutants substantially improves forecast accuracy.
5. **Directional accuracy requires improvement:** All models achieve only 50-60% directional accuracy, indicating challenges in predicting trend reversals.

7.1 Contributions

This research contributes to environmental forecasting by:

- Providing benchmark results on real-world AQI data spanning multiple years
- Demonstrating practical superiority of GRU over alternative approaches
- Quantifying performance across seven complementary metrics
- Offering actionable deployment recommendations for public health agencies

7.2 Broader Impact

Accurate air quality forecasting has profound implications for public health, enabling millions of people to protect themselves from hazardous pollution. By deploying the proposed GRU-based system, cities can:

- Reduce respiratory illness incidence through early warnings
- Optimize traffic and industrial management policies
- Improve quality of life for vulnerable populations
- Support evidence-based environmental policymaking

As air quality challenges intensify globally, advanced forecasting systems will become indispensable tools for sustainable urban development. This work establishes a strong foundation for operational AQI prediction, with clear pathways for future enhancement through ensemble methods, attention mechanisms, and integration of meteorological data.

References

[1] Kumar, A., & Goyal, P. (2011). Forecasting of daily air quality index in Delhi. *Science of the Total Environment*, 409(24), 5517-5523.

[2] Li, X., Peng, L., Yao, X., Cui, S., Hu, Y., You, C., & Chi, T. (2017). Long short-term memory neural network for air pollutant concentration predictions: Method development and evaluation. *Environmental Pollution*, 231, 997-1004.

[3] Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9(8), 1735-1780.

[4] Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014). Learning phrase representations using RNN encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078.*

[5] Box, G. E., Jenkins, G. M., Reinsel, G. C., & Ljung, G. M. (2015). *Time series analysis: forecasting and control.* John Wiley & Sons.

[6] Huang, C. J., & Kuo, P. H. (2018). A deep CNN-LSTM model for particulate matter (PM2.5) forecasting in smart cities. *Sensors*, 18(7), 2220.

[7] Zhang, Y., Wang, Y., Gao, M., Ma, Q., Zhao, J., Zhang, R., ... & Carmichael, G. R. (2019). A predictive data-driven model for air quality forecasting. *Atmospheric Environment*, 213, 351-361.

[8] World Health Organization. (2021). *WHO global air quality guidelines: particulate matter, ozone, nitrogen dioxide, sulfur dioxide and carbon monoxide.* World Health Organization.