AQI Forecast Project
10pearls Shine Internship

Real Time AQI Monitoring & 72 Hour Forecasting for Karachi

Introduction:

AQI SeekAI is an end to end machine learning system that monitors and forecasts Air Quality Index (AQI) for Karachi, Pakistan. The system automatically collects hourly environmental data, engineers 186+ predictive features, trains LightGBM models, and serves a live 72 hour AQI forecast through an interactive Streamlit dashboard all deployed using free tier infrastructure. The AQI standard follows the methodology of the United States Environmental Protection Agency.

Problem Statement

Karachi lacks accessible real time air quality forecasting. Existing systems:

- Report only current AQI (no prediction)
- Do not provide public dashboards
- Require paid infrastructure for automation
- Offer limited historical analytics

This project addresses these limitations using a fully serverless ML pipeline.

Tools & Technologies

- **Language:** Python 3.11/3.12
- **ML Model:** LightGBM (Gradient Boosted Trees)
- **Database:** MongoDB Atlas (Free Tier)
- **Data Source:** Open Meteo Weather + Air Quality APIs
- **Frontend:** Streamlit + Plotly
- **Automation:** GitHub Actions (cron workflows)
- **Deployment:** Streamlit Community Cloud

System Architecture

The system uses a 3 layer architecture:

1. **Hourly Pipeline (Every Hour)**
    - Fetches weather + air quality data
    - Engineers 186+ features
    - Stores results in MongoDB
    - Includes deduplication + retry logic

2. **Retrain Pipeline (Every 12 Hours)**
    - Pulls full dataset
    - Applies temporal 80/20 split
    - Trains 3 LightGBM models
    - Saves models back to MongoDB
3. **Streamlit Dashboard**
    - Displays live AQI
    - Shows 72 hour forecast
    - Provides historical analytics & EDA
    - Displays model performance metrics

Feature Engineering

The model uses 186+ engineered features across:

- Weather rolling statistics (6h/12h/24h)
- Atmospheric trends
- EPA sub AQI indices
- Autoregressive AQI lags (1h–24h)

- Cyclical time encodings (hour/day/month)
- Interaction features (e.g., humidity × temperature)
- Wind vector decomposition (u/v components)

Autoregressive AQI features dominate short term accuracy, while rolling weather features support medium and long term forecasts.

Model Design

3 Band Forecast Architecture

Instead of training 72 separate models, the system uses 3 specialized models:

Band	Horizons	Purpose
Short	1–8h	Persistence-dominated
Medium	9–24h	Diurnal pattern learning
Long	25–72h	Weather regime modeling

Each prediction uses:

- 186 base features
- 5 autoregressive AQI features
- 1 normalized horizon encoding

Total input size: **192 features**

Model Performance

Horizon	R^2	Interpretation
t+1h	0.98	Near-perfect short-term prediction
t+6h	0.89	Strong autoregressive accuracy
t+24h	~0.05	Difficult 1 day forecasting

t+72h	Negative	High atmospheric uncertainty

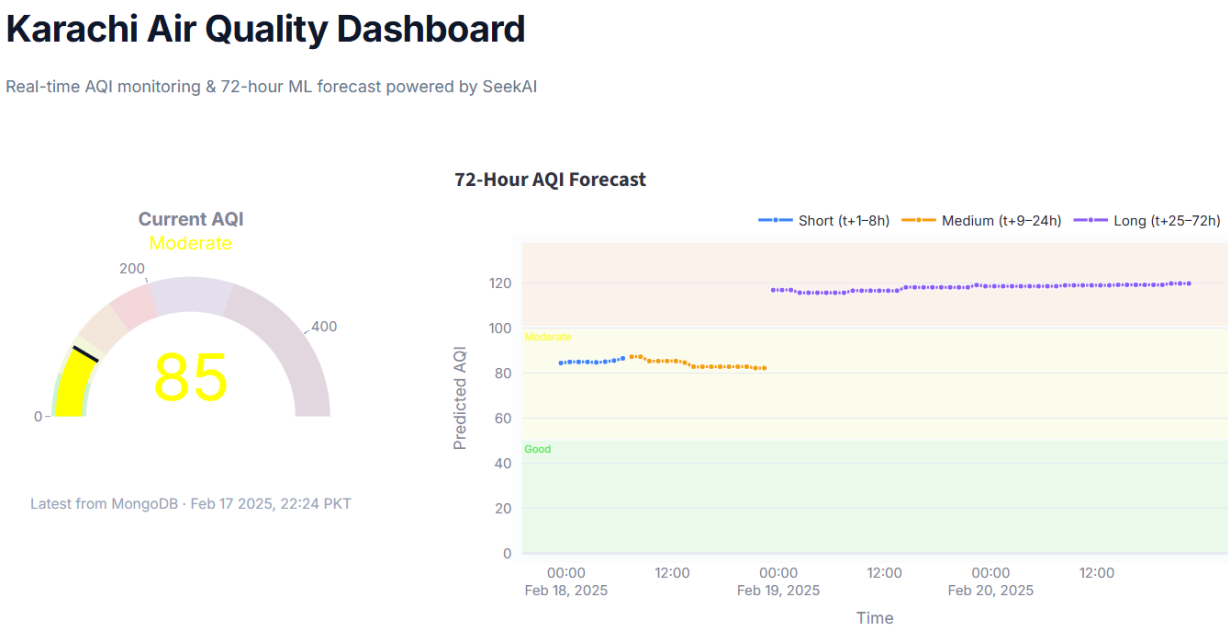
Key insight:
Short term AQI is highly predictable due to autocorrelation. Long term AQI is inherently chaotic due to weather shifts and emission variability.

Dashboard Features

The deployed dashboard includes:

- Live AQI gauge with EPA color bands
- 72 hour forecast visualization
- Forecast milestones (t+1h, t+6h, etc.)
- Historical AQI analysis
- Diurnal heatmaps
- Pollutant breakdown charts
- Model performance metrics

All timestamps displayed in PKT (UTC+5).



Automation & CI/CD

- Hourly pipeline: runs at minute :10 every hour
- Retraining pipeline: runs every 12 hours at :30
- Automatic data freshness checks
- MongoDB model version updates
- Fully automated via GitHub Actions

Challenges & Solutions

1. 72 Model Storage Explosion

Challenge:
The initial design trained one LightGBM model for each of the 72 forecast horizons (t+1 to t+72). After serialization, total size exceeded ~500MB — impossible to store within MongoDB Atlas free-tier limits (512MB total + 16MB per document).

Root Cause:
Separate models duplicated structure and feature space across horizons, creating unnecessary redundancy.

Solution:
Redesigned to a 3 band architecture (Short / Medium / Long) with horizon encoding (h/72).

- Reduced model count from 72 → 3
- Storage reduced to ~15MB total
- Improved generalization by pooling training samples within bands
- Easier maintainability and faster retraining

2. Long Horizon Accuracy Degradation

Challenge:
Prediction quality dropped significantly beyond 24 hours. At 72 hours, R^2 became negative.

Root Cause:
Urban AQI is chaotic at multi-day horizons due to unpredictable emissions and atmospheric regime changes. Autoregressive features lose predictive strength over time.

Solution:

- Specialized model bands for different horizon behaviors
- Reduced performance expectations for long band
- Visually separated forecast bands in dashboard
- Communicated uncertainty transparently instead of hiding degradation

This improves interpretability and user trust.

3. Forecast Collapse (Flat Line Predictions)

Challenge:
Earlier recursive models produced flat forecasts (no variance), even when RMSE appeared reasonable.

Root Cause:
Recursive forecasting fed model predictions back as inputs, amplifying smoothing bias.

Solution:

- Switched to direct multi-horizon prediction
- Avoided recursive dependency loops
- Used horizon encoding instead of iterative roll-forward

This restored natural forecast variability.

Conclusion

AQI SeekAI demonstrates that a fully automated, real-time environmental forecasting system can be built using:

- Advanced feature engineering
- Multi horizon ML modeling
- Serverless cloud infrastructure
- Zero operational cost

The system achieves high short term predictive accuracy while realistically acknowledging long-term atmospheric uncertainty.

Project Repository: github.com/SeekAI-786/AQI_Predictor
Live Dashboard: aqi-seek786.streamlit.app