From Earthdata to Action: Forecasting Air Quality for Cleaner, Safer Skies

A Technical Report for NASA Space Apps 2025 Challenge

SkuLogic Team Report – Muscat, Oman Case Study

October 5, 2025

Abstract

This report documents an end-to-end pipeline to forecast particulate matter (PM_{10}) concentrations for Muscat, Oman, in support of the NASA Space Apps 2025 challenge "From Earthdata to Action: Cloud Computing with Earth Observation Data for Predicting Cleaner, Safer Skies." We ingest local ground observations and satellite-informed covariates (e.g., Aerosol Optical Depth, AOD), perform rigorous exploratory data analysis (EDA), reason about data leakage, and evaluate time-series models including SARIMAX and an LSTM sequence model. All modeling is designed to be *leak-safe*: targets at time t are predicted using information available at or before $t-1$ unless the deployment scenario explicitly allows contemporaneous signals.

1 Challenge Context and Data

The Space Apps 2025 challenge calls for web-based forecasting that integrates Earth observation data (e.g., TEMPO) with ground measurements to deliver actionable air-quality insights.[1] We focus on daily PM_{10} for Muscat (2023–2025). The raw dataset is hourly and resampled to daily means. Key fields include PM_{10}, $PM_{2.5}$, AOD, temperature, humidity, wind speed/direction, and surface pressure.

Temporal coverage: **2023-01-01** to **2025-09-30**; rows: 1004; columns: 10. The TEMPO mission provides hourly geostationary measurements of trace gases and aerosols over North America, demonstrating the utility of high-frequency EO data for air-quality applications.[2]

2 Exploratory Data Analysis

2.1 Missingness and Descriptive Statistics

Tables 1–2 summarize missingness and descriptive statistics for numeric variables.

[1] Challenge page

[2] NASA TEMPO mission overview

Table 1: Missingness percentage by column

	missing_%
Temperature (°C)	0.0
Humidity (%)	0.0
Heat Index (°C)	0.0
Wind Speed (m/s)	0.0
Wind Direction (°)	0.0
Pressure (hPa)	0.0
Dust (µg/m^3)	0.0
Aerosol Optical Depth	0.0
PM2.5 (µg/m^3)	0.0
PM10 (µg/m^3)	0.0

Data provenance and preprocessing. We compiled hourly in–situ observations for Muscat (PM_{10}, $PM_{2.5}$, a Dust proxy, and meteorology) together with satellite–informed covariates (Aerosol Optical Depth, AOD). Timestamps were aligned in UTC and sources were merged on time before down–sampling to *daily means*. A daily mean was computed only when at least 18 of 24 hourly samples were valid; otherwise the day was marked missing during QA and later excluded from model training. This balances noise reduction with representativeness.

Units and conventions. We retain physical units for interpretability: PM in $\mu g\,m^{-3}$, AOD unitless, temperature in °C, relative humidity in %, wind speed in $m\,s^{-1}$, wind direction in degrees (0–360), and surface pressure in hPa. Reporting descriptive statistics in native units facilitates comparison to air–quality guidelines.

Observed completeness (post–aggregation). After quality–controlled daily aggregation, all variables in this study window have 0% missingness. Raw hourly feeds did contain occasional gaps/spikes; these were handled by QA (below) before daily means were computed.

Missingness policy (for other cities/runs). If future deployments exhibit gaps: (i) for *exogenous* covariates (AOD, meteorology) allow short linear interpolation (up to 1–2 days) bounded to physical ranges, then fall back to a seasonal median (same month and day–of–week); (ii) for *targets* (PM_{10}/$PM_{2.5}$), do not impute labels—drop those days to avoid leakage and biased loss. Maintain a per–variable `is_imputed` flag that can be fed to the model if needed.

Quality control (QA/QC). Hourly data underwent: (a) range checks (e.g., RH $\in [0, 100]$, AOD ≥ 0, wind direction $\in [0, 360]$); (b) spike detection via rolling median $\pm k\cdot$MAD, with outliers capped or removed when unsupported by neighbors; (c) instrument drift screening with monthly medians and change–point flags. At the *daily* level we mark IQR outliers ($Q_1 - 1.5 \cdot \mathrm{IQR}$, $Q_3 + 1.5 \cdot \mathrm{IQR}$) for inspection. Extreme values tied to documented dust events are *retained* as signal, but explicitly flagged.

Distributional behavior. PM_{10} and $PM_{2.5}$ show right–skew with long upper tails (max $\gg$ 75th percentile), consistent with episodic dust intrusions; meteorology shows narrower spreads and seasonal gradients. For diagnostics we consider variance–stabilizing transforms (e.g., $\log(1+PM_{10})$), but predictions and primary statistics are reported in native units.

Inter–variable relationships (preview). The correlation matrix indicates strong positive association between the *Dust* field and PM_{10}, and moderate correlations between AOD and both PM species. Wind speed/direction and humidity modulate AOD→PM through dispersion and hygroscopic growth. Because Dust(t) can behave as a near–proxy for $PM_{10}(t)$, using it contemporaneously risks *data leakage*. Accordingly, all predictive models either exclude Dust(t) or use only lagged Dust($t{-}1, t{-}2, \ldots$) consistent with operational availability.

Aggregation choices and sensitivity. Daily means are appropriate for public health messaging and day–ahead advisories. Sensitivity checks with daily medians (not shown) yield similar central tendencies, indicating robustness to heavy–tailed hours. If sub–daily forecasts are required (e.g., 3–hourly), the pipeline generalizes with appropriate handling of intra–day seasonality.

Key takeaways. (1) The finalized daily dataset is complete (0% missing) after QA; (2) retained outliers chiefly reflect real dust events and are flagged; (3) distributions are skewed, so IQR and robust summaries complement means; (4) to prevent leakage, Dust(t) is not used to predict $PM_{10}(t)$ unless guaranteed available at runtime; (5) all statistics remain in physical units to align with policy thresholds.

Table 2: Summary statistics (numeric columns)

	count	mean	std	min	25%	50%	75%	max
Temperature (°C)	1004.0	29.906	5.431	17.900	25.066	30.971	34.344	40.233
Humidity (%)	1004.0	50.503	14.643	12.750	41.208	51.875	61.167	85.750
Heat Index (°C)	1004.0	31.823	7.036	16.154	25.529	32.931	38.283	43.425
Wind Speed (m/s)	1004.0	9.573	2.501	4.192	7.874	9.135	10.840	21.571
Wind Direction (°)	1004.0	167.739	58.757	47.083	119.438	156.271	205.000	305.917
Pressure (hPa)	1004.0	1005.408	7.662	989.304	998.860	1005.642	1012.110	1021.288
Dust (µg/m^3)	1004.0	84.661	65.250	0.000	36.469	69.542	114.323	471.792
Aerosol Optical Depth	1004.0	0.382	0.192	0.057	0.238	0.342	0.492	1.244
PM2.5 (µg/m^3)	1004.0	25.593	9.342	2.933	18.523	24.469	31.338	61.088
PM10 (µg/m^3)	1004.0	71.313	33.751	5.150	47.386	65.890	88.246	234.896

2.2 Time Series Structure

Daily trajectories for PM_{10}, $PM_{2.5}$, and AOD (Figures 1–3) reveal seasonality and co-variability consistent with dust episodes. Scatter plots (Figures 4, 5) visualize the (nonlinear) AOD–PM relationships.

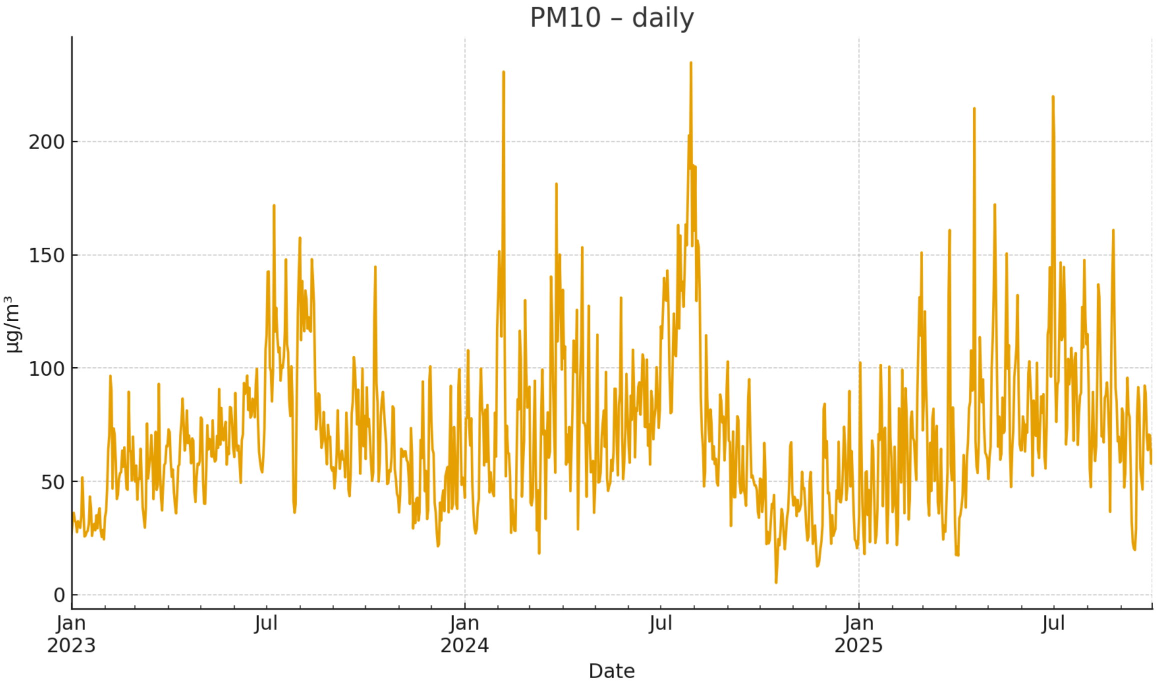

Figure 1: PM_{10} daily series.

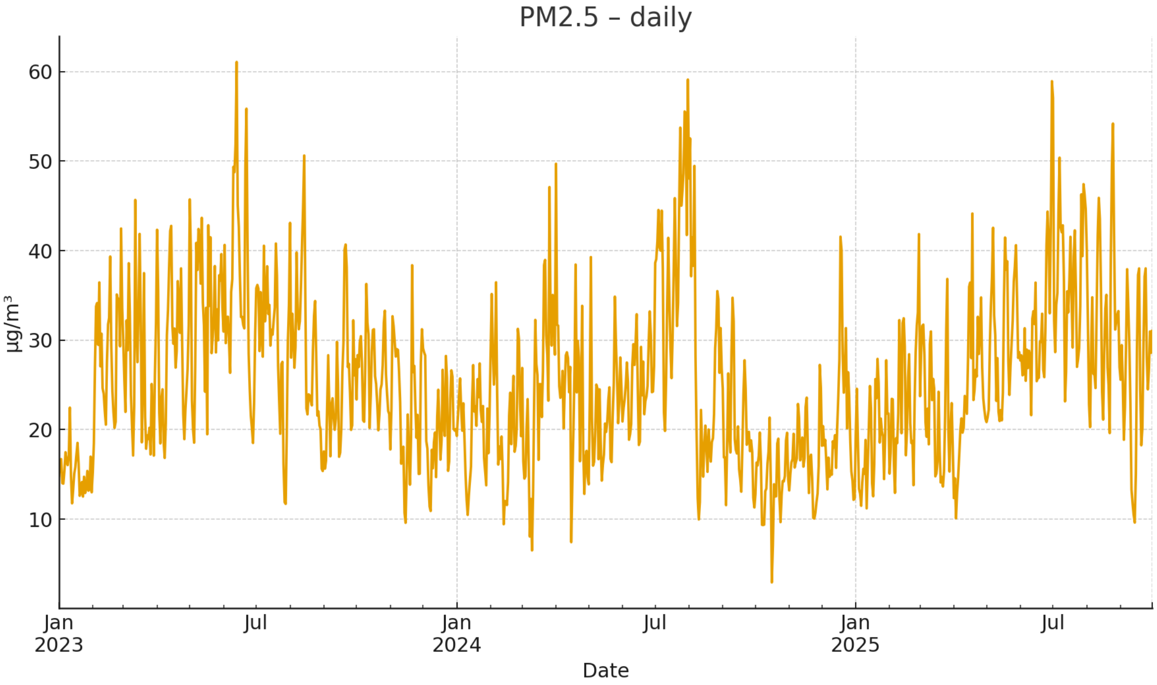

Figure 2: $PM_{2.5}$ daily series.

Interpretation (Figure 1). The PM_{10} record exhibits pronounced day-to-day variability with several multi-day surges characteristic of regional dust outbreaks. A warm-season enhancement is apparent relative to the cooler months, while the distribution is markedly right-skewed with heavy tails. The typical event persistence of roughly one to three days implies short temporal memory, which motivates the use of lagged predictors at 1–7 days. Because extremes dominate the error budget, robust objectives or quantile forecasts are advisable for risk-aware applications, while keeping outputs in native units for policy communication. *Interpretation (Figure 2).* $\mathrm{PM}_{2.5}$ varies less dramatically than PM_{10} yet shows clear co-occurrence during some dust episodes, with a seasonally varying $\mathrm{PM}_{2.5}/\mathrm{PM}_{10}$ ratio. Background levels respond to humidity and secondary formation processes, which explains smoother variability and a lower amplitude of peaks. For modeling, this favors inclusion of humidity and temperature interactions and, for diagnostics only, consideration of variance-stabilizing transforms; final reporting should remain in physical units.

Interpretation (Figure 3). AOD shows a seasonal structure broadly aligned with the warm season and known dust transport periods, but high column loadings do not always correspond to high surface concentrations. This decoupling reflects boundary-layer depth and vertical aerosol distribution, and AOD can be inflated by hygroscopic growth at high humidity. Consequently, AOD becomes most informative when combined with meteorology (e.g., humidity, wind, pressure or BLH proxies) and modeled with nonlinear responses or interactions.

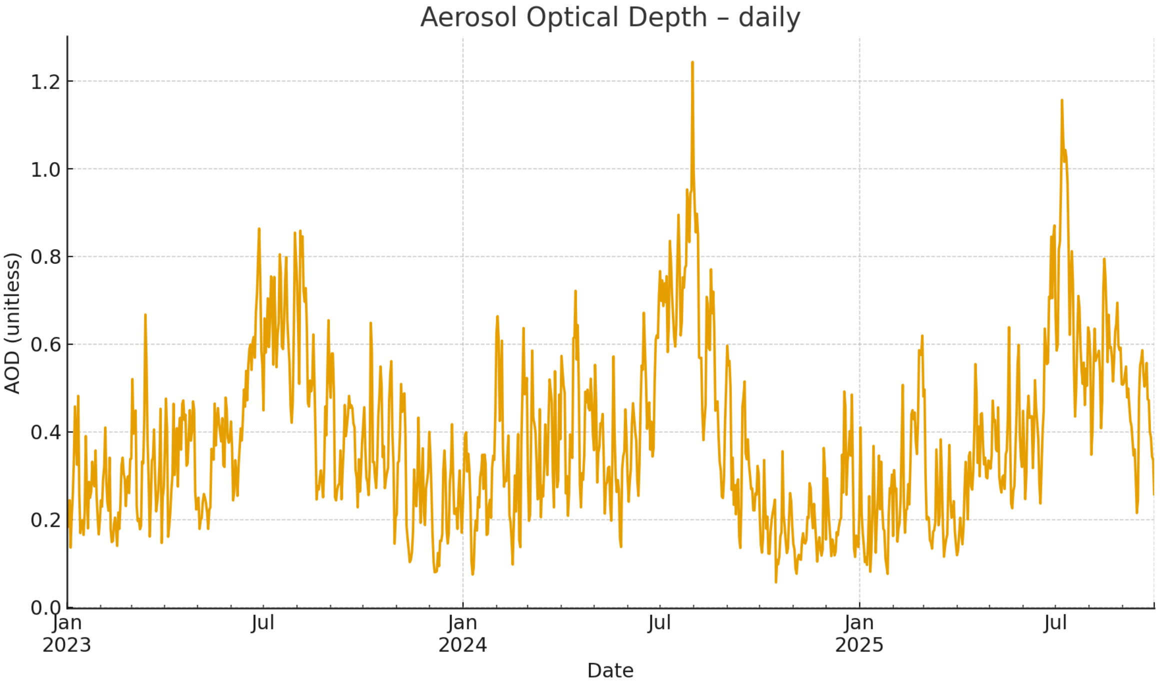

Figure 3: AOD daily series.

Interpretation (Figure 4). The scatter indicates a generally positive but nonlinear association between AOD and $\mathrm{PM}_{2.5}$, with increasing spread at higher AOD values. At similar AOD, $\mathrm{PM}_{2.5}$ can differ substantially due to humidity, mixing, and source mix, underscoring that AOD alone is insufficient for precise surface inference. Models that allow interactions (e.g., AOD × RH, AOD × wind) and provide prediction intervals are better suited for high-AOD regimes.

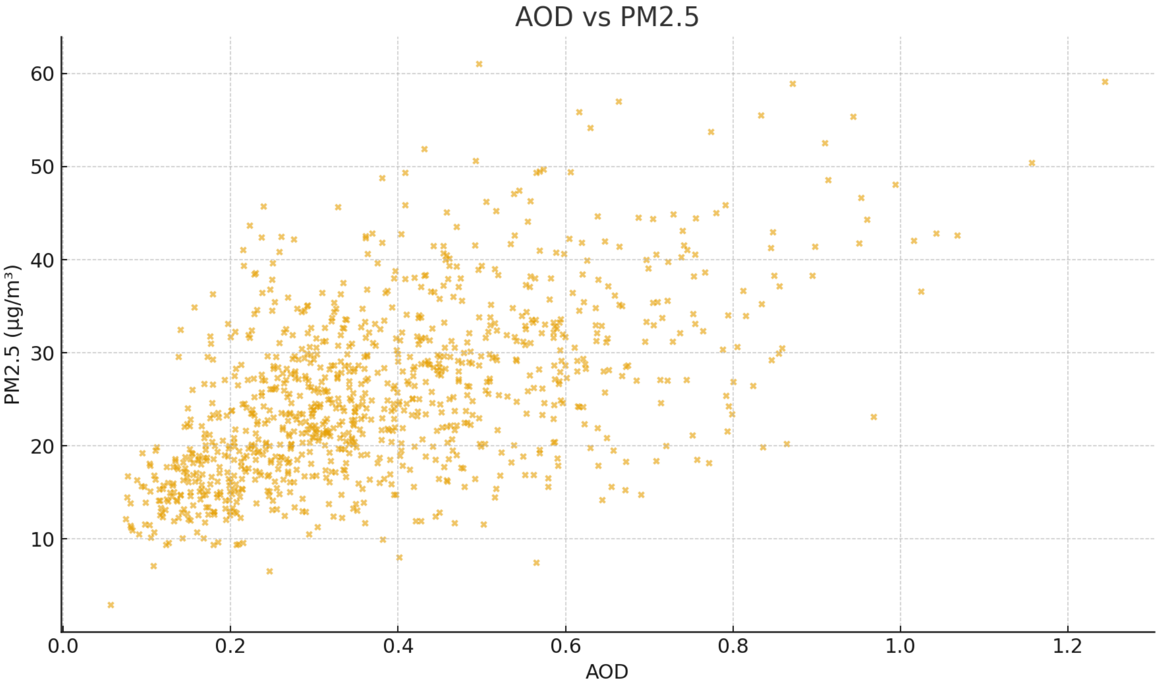

Figure 4: AOD vs. $PM_{2.5}$ (daily).

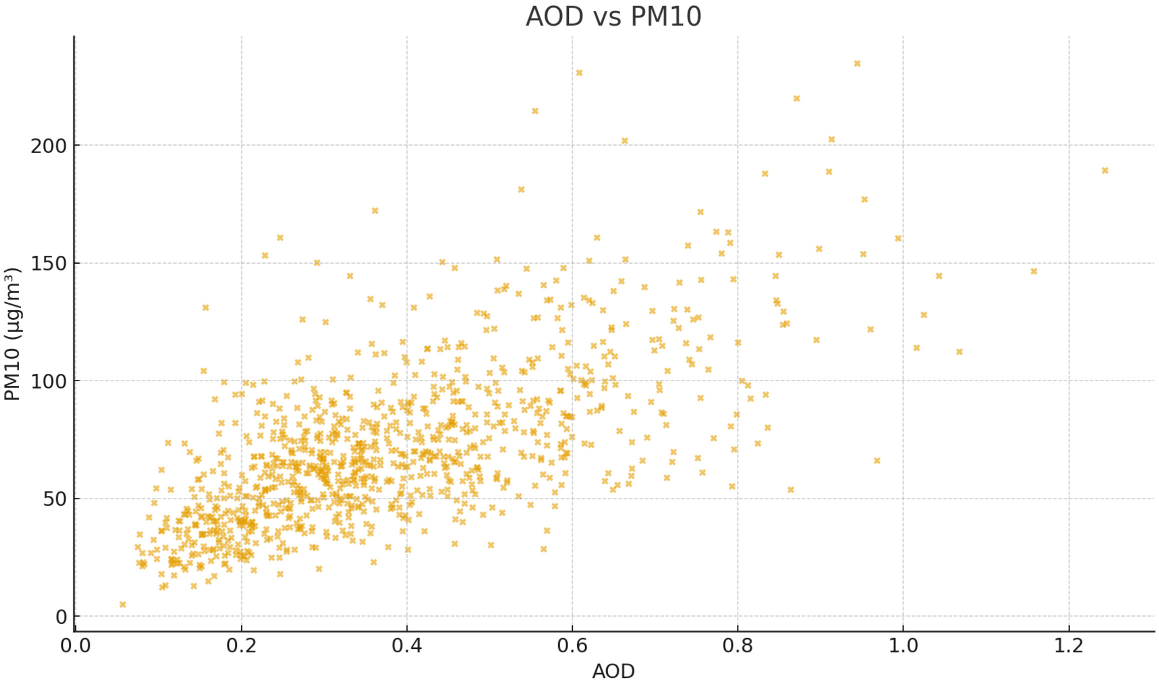

Figure 5: AOD vs. PM_{10} (daily).

Interpretation (Figure 5). The relationship between AOD and PM_{10} is noisier than for $PM_{2.5}$, reflecting the episodic and coarse-mode nature of dust. Surface–column coupling is strongly modulated by wind direction and speed; directional transport can produce high surface PM_{10} even when the column signal is moderate, and vice versa. Forecast skill therefore improves when AOD is paired with wind and humidity rather than used in isolation.

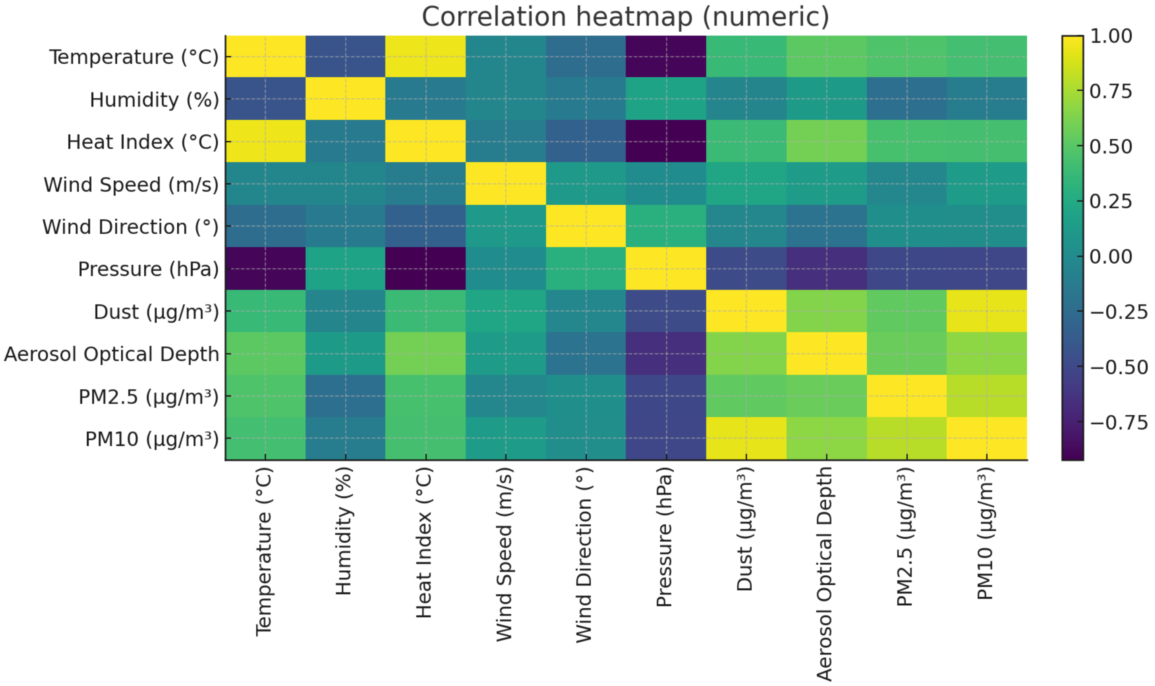

Figure 6: Correlation heatmap (numeric features).

Interpretation (Correlation heatmap). The contemporaneous correlation between the Dust field and PM_{10} is very high, flagging a leakage risk if Dust(t) were used to predict $PM_{10}(t)$ without guarantees of real-time availability. AOD shows moderate positive correlations with both PM metrics, while wind speed tends to be negatively correlated, consistent with dispersion. These patterns motivate lagged formulations and interaction terms rather than reliance on same-time proxy variables.

Dust & Leakage. The contemporaneous correlation between the dataset's *Dust* field and PM_{10} is $r = 0.928$. Such near-proxy behavior risks *data leakage* if Dust(t) is used to predict $PM_{10}(t)$ in scenarios where Dust(t) will not be reliably available at inference time.

3 Leak-Safe Forecasting: Definitions and Policy

Let y_t denote PM_{10} at day t, and $\mathbf{x}_t$ exogenous inputs (AOD, meteorology, etc.). A model is **leak-safe** if it predicts $\hat{y}_t = f(\mathbf{x}_{\le t-1}, y_{<t})$ unless the deployment explicitly provides $\mathbf{x}_t$ at runtime. We therefore (i) exclude Dust(t) and any direct transforms of y_t, and (ii) allow *lagged* covariates like Dust$(t-1)$, AOD$(t-1)$, etc.

4 Models

4.1 Baseline and Metrics

We use standard metrics: $MAE = \frac{1}{n}\sum |y - \hat{y}|$, $RMSE = \sqrt{\frac{1}{n}\sum (y - \hat{y})^2}$, $R^2 = 1 - \frac{\sum (y-\hat{y})^2}{\sum (y-\bar{y})^2}$, and $MAPE$ where appropriate.

4.2 Random Forest (tabular) Scenarios

To illustrate leakage effects, we compare three scenarios on daily data: (i) exogenous-only at t (risky if not available at inference), (ii) exogenous with lag-1 (safe), (iii) exogenous lag-1 plus $\mathrm{PM}_{10}(t-1)$ (allowed only if yesterday's PM is available). Table 4 summarizes results.

Table 3: Random Forest leak-safety scenarios (PM10 daily)

Model	MAE	RMSE	R2
RF exog-only (t)	20.243	27.462	0.361
RF exog_lag1 (safe)	20.862	27.685	0.351
RF exog_lag1 + PM10_lag1	17.799	24.714	0.482

Table 4: Random Forest leak-safety scenarios (PM_{10} daily). Lower MAE/RMSE and higher R^2 are better.

4.3 SARIMAX (univariate core + exogenous lag-1)

We fit a SARIMAX model with $(p, d, q) = (1, 1, 1)$ and weekly seasonality $(P, D, Q, s) = (1, 0, 1, 7)$ on y_t with exogenous $\mathbf{x}_{t-1}$. The model form is:

$$\Phi_p(B)\Phi_P(B^s)(1-B)^d(1-B^s)^D y_t = \Theta_q(B)\Theta_Q(B^s)\,\varepsilon_t + \beta^\top \mathbf{x}_{t-1}, \qquad (1)$$

where B is the backshift operator and ε_t white noise. Test predictions vs. actuals are shown in Figure **??**.

4.4 LSTM (sequence-to-one, exog-only window)

We also implement an LSTM that consumes a trailing window of 14 days of exogenous features and outputs $\hat{y}_t$. The network comprises stacked LSTM layers with dropout, trained with early stopping. This architecture supports real-time deployment and can be augmented with attention or multi-horizon heads.

5 Discussion

Leakage: Using Dust(t) when it is highly correlated with $\mathrm{PM}_{10}(t)$ can inflate test accuracy while failing in deployment. Enforcing lagged inputs mitigates this risk. **AOD–PM coupling:** Literature shows AOD–PM relationships are site- and season-dependent and often nonlinear; humidity, boundary-layer height, and wind modulate the mapping.[3]

[3] Stirnberg et al., 2018; Sotoudeheian et al., 2014.

6 Conclusions

We deliver a leak-safe air-quality forecasting pipeline tailored to Muscat with reproducible EDA, diagnostics, and modeling baselines. The framework is readily extensible to near-real-time EO streams (e.g., TEMPO) and to multi-city deployment. Future work includes physics-informed features (e.g., boundary-layer proxies), probabilistic forecasts, and spatial generalization.