# **Literature Review – CO₂ Forecasting & Multivariate Air-Quality Modeling**

## **1. Introduction**

Forecasting atmospheric carbon dioxide (CO₂) and related air-quality indicators is essential for climate policy, urban planning, and industrial regulation. Recent work spans classical time‑series models, machine‑learning (ML) ensembles, and deep learning (DL). This review synthesizes prior studies relevant to our project’s baseline (CO₂ forecasting) and motivates our extension toward a multivariate **air‑quality profile** and **sector‑wise emission analysis** (transport, industry, household, etc.).

## **2. Background: Modeling Approaches**

**Classical statistical models** (ARIMA, Holt–Winters, Theta) capture trend/seasonality and provide interpretable baselines. **Prophet** adds flexible trend/seasonality with changepoints. **ML ensembles** (Random Forest, Extra‑Trees, Gradient Boosting, XGBoost) learn non‑linear relationships with engineered lag/date features. **DL sequence models** (LSTM/GRU) learn longer temporal dependencies; recent **Transformer‑style** models can scale to longer horizons and multivariate inputs.

## **3. Prior Work (Concise Summaries)**

### **3.1 AI‑Driven Carbon Management in the Power Sector**

* **Focus:** Enterprise‑level carbon footprint governance in electric power companies.
* **Methods:** AI + big‑data analytics; DSR (Driver–State–Response) framework; Entropy‑weighting with TOPSIS for performance evaluation.
* **Findings:** Identifies drivers of emissions and tracks firm‑level improvements/declines; supports audit and policy oversight.
* **Relevance:** Demonstrates decision‑support pipelines that translate analytics into management action—useful for our sector‑wise reporting.

### **3.2 Nonparametric ML for Urban CO₂ (10 Cities, 2005–2019)**

* **Models:** STIRPAT/linear regression, kernel regression, Random Forest, neural networks.
* **Findings:** Neural networks outperform classical baselines; industrial structure and energy intensity are key predictors.
* **Relevance:** Motivates flexible, non‑parametric models and highlights structural/economic covariates we can incorporate.

### **3.3 Drivers of Emissions Across 254 Cities (2011–2020)**

* **Models:** Extra‑Trees (top), Adaptive Lasso for feature selection; interpretation with Partial Dependence Plots.
* **Findings:** Energy consumption dominates; additional factors include FDI and policy uncertainty; effects vary by city size.
* **Relevance:** Offers a template for interpretable ML that links drivers to actionable policy; inspires our sector‑wise analysis.

### **3.4 Near‑Real‑Time Daily CO₂ Forecasting**

* **Horizon/Frequency:** Daily emissions, 2020–2022; all sectors.
* **Models:** GM(1,1), ARIMA/SARIMAX, ANN, Random Forest, **LSTM (best)**.
* **Findings:** DL outperforms classical/ML at high frequency; evaluation via MSE, RMSE, MAE, MAPE, R².
* **Relevance:** Supports using sequence models when moving to finer temporal resolution (e.g., daily AQI/meteorology).

### **3.5–3.6 Long‑Range Atmospheric CO₂ (Mauna Loa)**

* **Data:** 1958–present CO₂ observations (Hawaii).
* **Models:** Linear regression, ARIMA, tree ensembles (RF/XGBoost), **LSTM (often best)**.
* **Findings:** DL and ensembles excel on long‑horizon, non‑linear dynamics; visuals (trends, thresholds, seasonality) communicate risk.
* **Relevance:** Validates our baselines and emphasizes model comparison under multiple metrics.

## **4. Comparative Synthesis**

| **DIMENTIONS** | **Classical** | **ML Ensembles** | **Deep learning (LSTM)** |
| --- | --- | --- | --- |
| **Captures trend** | Yes models trend explicitly | Yes, via engineered features | yes , learns directly from data |
| **Captures seasonability** | Yes, built into models | Requires engineered seasonal features | Learns seawsonality implicitly |
| **non-linerarity** | limited | Strongl capability | Strong capability |
| **Multivariate inputs** | Possible with SARIMAX variants | Naturally handles multiple variables | Naturally handles multiple variables |
| **interpretability** | High, easy to explain | Medium, feature impostance | Lower, needs attention mechanism |

## **Key takeaways:**

* LSTM and tree ensembles typically outperform classical models on complex or high‑frequency data.
* Interpretable ML (feature importance, PDP, SHAP) is crucial for policy relevance.
* Adding exogenous variables (meteorology, activity, sector data) improves realism and accuracy.

## **5. Gaps in Prior Work**

1. **Single‑variable focus:** Many studies model CO₂ alone; fewer build a comprehensive **air‑quality profile** (CO₂, PM₂.₅/PM₁₀, NO₂, SO₂, O₃, CO) with meteorology (temperature, humidity, wind, precipitation).
2. **Limited sector attribution:** Forecasts rarely attribute emissions to **sectors** (transportation, industry, residential), limiting prescriptive guidance.
3. **Explainability at scale:** DL forecasts often lack **transparent explanations** required by stakeholders.
4. **Operationalization:** Few end‑to‑end systems (data ingestion → modeling → dashboard) for real‑time decision support.

## **6. Our Proposed Direction**

**Goal:** Move from CO₂‑only forecasting to a **multivariate air‑quality and emissions intelligence** platform with sector attribution.

**6.1 Data Plan**

* **Sources:** City authorities (monitoring stations/API) and public repositories for meteorology/emissions.
* **Variables:** AQI, PM₂.₅/PM₁₀, NO₂, SO₂, O₃, CO; meteorology (temperature, humidity, precipitation, wind); sectoral activity proxies (traffic counts, fuel use, industrial output, power demand).
* **Cadence:** Hourly/daily; station‑level with geospatial tags.

**6.2 Modeling Plan**

* **Baselines:** ARIMA/Holt–Winters/Prophet for quick references.
* **Multivariate ML/DL:**
  + Tree‑ensembles (Random Forest/Extra‑Trees/XGBoost) with lagged and calendar features.
  + Sequence models (LSTM/GRU); explore **Transformer‑style** architectures for long‑horizon, multivariate forecasting.
* **Attribution & Explainability:** SHAP for feature/sector importance; PDP/ICE for marginal effects; temporal attention for DL.
* **Anomaly Detection:** Rolling z‑scores or Isolation Forest to flag unusual spikes/drops (e.g., lockdown effects, weather events).

**6.3 Evaluation & Reporting**

* **Metrics:** MAE, RMSE, MAPE, R².
* **Validation:** Rolling/expanding window cross‑validation; station‑level holdouts.
* **Outputs:** Forecasts with confidence intervals; sector contribution charts; what‑if simulations (e.g., 10% traffic reduction).

**6.4 Delivery**

* **Artifacts:** Reproducible notebooks, trained model artifacts, and a **Streamlit dashboard** for interactive exploration.
* **Reproducibility:** Clear repo structure (/data, /notebooks, /models, /reports, /literature), environment files.

## **7. Conclusion**

Prior research establishes that while classical models provide interpretable baselines, ML/DL—especially LSTM and tree ensembles—achieve superior accuracy, and interpretable ML bridges the gap to policy action. Our contribution extends this line of work by

(i) integrating **multivariate air‑quality and meteorology**

(ii) performing **sector‑level attribution**

(iii) delivering an **explainable, deployable** analytics system suitable for real‑time decision support.