# ST-MFXAI: Research Methodology and Experimental Design for Spatio-Temporal Multi-Sensor Fusion with Explainable AI in Urban Air Quality Prediction

Ozlem Kilickaya[1,2], *Basant Kumar*[3]
[1]Lincoln University College, Petaling Jaya, Selangor, Malaysia.
[2] University of the People-595 E Colorado Blvd Suite 623, Pasadena, CA 91101, USA.
[3] Modern College of Business and Science, Oman.
pdf.ozlemgulsum@lincoln.edu.my ; basant@mcbs.edu.om

---

**Abstract:** Research into urban air-quality forecasting suffers from three structural gaps: spatial and temporal models function in isolation instead of via deeply integrated learning; explainability is treated as an after-thought via post-hoc methods; and cross-city generalizability is seldom verified. To bridge these gaps, this paper introduces the methodology for ST-MFXAI (Spatio-Temporal Multi-sensor Fusion with explainable AI). The proposed framework natively embeds a Graph Convolutional spatial component, a Transformer temporal block, and a gated fusion mechanism to output four distinct explanations (SHAP, LRP, Grad-CAM, and attention profiles) directly. This study deploys a time-split pipeline across three historical testbeds: Beijing Multi-Site AQI, KDD Cup 2018, and OpenAQ. This research formalizes four testable hypotheses that directly target multi-modal accuracy, inter-station spatial sensitivity, XAI fidelity, and a strict 15% $R^2$ degradation cap during cross-city transfer rather than using standard testing. Operationalization of these core claims into measurable outcomes relies on a comprehensive four-tier validation protocol tracking ablation, state-of-the-art benchmarks, spatial transfer, and explanation accuracy.

**Keywords**: Air quality prediction; Graph neural networks; Transformer; Explainable AI; Spatio-temporal fusion; Cross-city generalization

---

**Introduction**

Current paradigms in urban air-quality modeling suffer from fragmentation across three distinct fronts. Spatial architectures, specifically graph-based methods like STGCN and DCRNN, operate largely in isolation from temporal frameworks such as recurrent models or the Transformer-based AirFormer. When integration occurs, conventional approaches typically rely on basic concatenation instead of establishing a principled, learned fusion mechanism. Furthermore, interpretability is treated almost exclusively as a post-hoc diagnostic rather than an intrinsic architectural characteristic. This practice leaves a critical open question regarding whether post-hoc attributions accurately reflect the internal reasoning of a model. Compounding these issues, models verified on a localized sensor network are seldom evaluated for transferability to urban areas with differing pollution dynamics, station layouts, or meteorological conditions. Yet, deploying models under such unmonitored conditions represents the exact practical scenario faced by municipalities that lack extensive historical data infrastructure.

Earlier phases of this research program mapped these limitations into four foundational research questions. RQ1 examines if unified fusion surpasses single-modality configurations; RQ2 isolates whether spatial performance stems explicitly from inter-station correlations; RQ3 tests if embedded explainability yields higher fidelity than post-hoc alternatives; and RQ4 determines if spatial cross-city generalization is achievable. The ST-MFXAI framework serves as the core research artifact engineered to answer these challenges simultaneously. Rather than presenting experimental findings, which belong to subsequent validation phases, the present study formalizes this conceptual blueprint into a testable methodology encompassing architectural specifications, dataset structures, implementation guidelines, four falsifiable hypotheses, and a structured validation framework.

**Research Design**

The conceptualization of ST-MFXAI sits at the intersection of two paradigms as a deliberate methodological choice. The architecture itself, comprising the spatial module, temporal module, and embedded XAI fusion layer, represents a designed artifact within the Design Science Research framework. It is engineered to solve a stated practical problem and evaluated against utility criteria appropriate for a complex system. Concurrently, the claims regarding this artifact are evaluated under a post-positivist stance. Each hypothesis operates as a falsifiable proposition tested against pre-registered decision criteria, such as MAE below 13.5 and a strict 15% $R^2$ degradation threshold, through quantitative procedures including ANOVA, Wilcoxon signed-rank, and Spearman correlation. This combination proves necessary rather than merely convenient. A purely Design Science Research framing could not distinguish the specific claim of H3 that the spatial module captures correlation structures from the alternative that it simply adds regularization. Conversely, a purely post-positivist framing would offer no systemic basis for the iterative ablation-then-benchmark-then-generalize sequence.

The study adheres strictly to a deductive approach where hypotheses are derived directly from established literature gaps, stated in advance, and tested against empirical evidence rather than allowed to emerge inductively. Framed as a structural artifact, ST-MFXAI undergoes evaluation against three utility criteria mapping onto the core validation strategy. These encompass predictive utility under H1, explanatory utility under H2, and transfer utility under H4, whereas H3 serves as an internal diagnostic within the ablation study rather than a standalone evaluation phase. Finally, the study adopts a cross-sectional horizon since all datasets involve fixed historical collections with no new sensor deployments undertaken. A longitudinal extension remains a candidate for future research.

**Proposed ST-MFXAI Architecture**

ST-MFXAI comprises three stages: a spatial module modeling inter-station dependencies as a graph, a temporal module modeling within-station sequential dependencies, and a fusion layer combining both while producing interpretable attributions as a first-class output.

*Spatial module.* Each monitoring station is a node in a graph $G = (V, E)$, with edges weighted by a Gaussian-kernel adjacency over inter-station distance. Graph-based representations have been shown to capture pollutant dispersion directly, with city-scale case studies reporting measurable gains over stationindependent baselines when spatial correlation is modeled explicitly [4]. A two-layer GCN propagates information across connected stations, producing embedding h_spatial; this lets the model condition a station's prediction on upwind, geographically correlated neighbors — the mechanism tested by H3. A GAT variant with learned edge weights is a planned ablation against this fixed-kernel design.

*Temporal module.* A Transformer encoder (multi-head self-attention, sinusoidal positional encoding) models the 24-hour input window, producing embedding h_temporal. Self-attention weights any position in the window directly, avoiding recurrent architectures' vanishing-gradient degradation over long sequences, and doubles as an interpretability signal. A Vanilla-Transformer ablation isolates this choice from the fusion and XAI layers.

*Fusion and embedded XAI.* The two embedding combine via a cross-modal attention gate rather than concatenation:

$$\alpha = \sigma(W_1 \cdot h\_spatial + W_2 \cdot h\_temporal + b),\ h\_fused = \alpha \odot h\_spatial + (1-\alpha) \odot h\_temporal$$

This allows the model to down-weight the spatial pathway during atmospheric stagnation and up-weight it during dispersion events. Learned attention-based fusion, rather than fixed concatenation, has been shown to improve city-level forecasting accuracy in comparable graph-convolutional architectures [5]. Explainability is embedded directly in this layer rather than applied post hoc, via four methods computed from fusion-layer internals and gradients at inference time: SHAP (feature-level attribution), LRP (relevance propagation through both sub-networks), Grad-CAM (adapted to produce station-level rather than pixel-level saliency), and attention maps (the native α weights, requiring no extra computation). This embedded design is the mechanism tested by H2. The flowchart is shown in Figure 1.

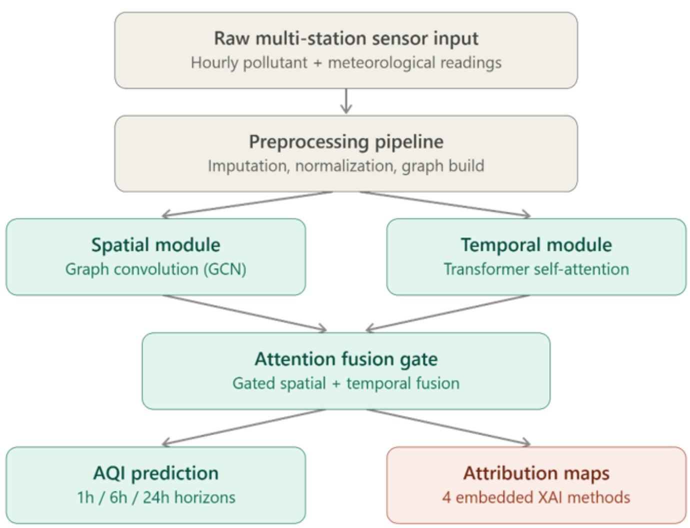


Figure. 1. *ST-MFXAI system flow. Gray boxes are shared data-preparation stages; teal boxes are the core predictive pipeline (spatial and temporal pathways processed in parallel, then gated fusion); the coral box marks the embedded XAI output distinguishing ST-MFXAI from post-hoc explanation approaches.*

## Experimental Setup

Three datasets support three distinct roles given in Table 1. Beijing Multi-Site AQI [8] for primary training and validation; KDD Cup 2018 [9] for direct benchmark comparison against published SOTA; and OpenAQ [10] for cross-city generalization, reserved exclusively as the zero-/few-shot target domain.

*Table 1.* Datasets and roles

| Dataset | Coverage | Role |
|---|---|---|
| Beijing Multi-Site AQI [8] | 12 stations · 2013–2017 | Primary training/validation |
| KDD Cup 2018 [9] | 36 stations · Beijing & London | Benchmark comparison |
| OpenAQ Network [10] | 50+ cities · 2018–2023 | Cross-city generalization |

A five-stage pipeline (Table 2) standardizes inputs and prevents temporal leakage: short gaps (≤6h) are linearly interpolated, longer gaps use KNN imputation; pollutants are min-max scaled, meteorology Z-scored; the spatial graph uses a Gaussian kernel (σ = 10 km, ablated at 5/20 km); 24-hour windows generate 1h/6h/24h forecasts via a chronological (non-shuffled) 80/10/10 split, avoiding the leakage that random shuffling introduces in autocorrelated time series; Gaussian noise and time-warping augment the training partition only.

*Table 2.* Pre-processing pipeline

| Stage | Operation |
|---|---|
| 1 | Imputation — linear (≤6h) / KNN (>6h) |
| 2 | Normalization — min-max (pollutants) / Z-score (meteorology) |
| 3 | Graph construction — Gaussian kernel, σ = 10 km |
| 4 | Windowing — 24h input → 1/6/24h horizons; chronological split |
| 5 | Augmentation — noise + time-warping (train only) |

**Model Implementation and Evaluation Plan**

ST-MFXAI is implemented in PyTorch/PyTorch Geometric: 2-layer GCN (64 hidden units), 2-layer Transformer (4 heads, d=128), learned bilinear fusion gate, AdamW (lr=$1\times10^{-3}$), Smooth L1 loss (robust to pollutant spikes), batch 32, 150 epochs with early stopping. These are a starting configuration, refined via grid search once ablation identifies contributing components.

Evaluation uses internal ablation baselines — LSTM (no spatial module), GCN-only (no temporal module), vanilla Transformer (no graph, no gated fusion) — and external SOTA benchmarks: STGCN/DCRNN [3], AirFormer [1], and AirPhyNet [2], the strongest published benchmark on comparable data (MAE = 13.5), the explicit target for the benchmark comparison phase. Performance is measured via RMSE, MAE, MAPE, and $R^2$ (the last for cross-city transfer, where the 15% degradation threshold for H4 follows prior location-aware transfer-learning evidence that fine-tuning narrows the source-target gap relative to zero-shot deployment alone [6]). A fifth metric, XAI Fidelity, combines a faithfulness score (Spearman correlation between attribution rankings and occlusion-based performance drop) and a consistency score (cross-method agreement among the four embedded outputs), operationalizing H2.

## Research Hypotheses and Validation Strategy

Four phases operationalize in Table 3. An ablation study ($2^3$ factorial across spatial, temporal, and fusion components, analyzed via ANOVA) tests H1 and H3 on Beijing data. A benchmark comparison against STGCN, DCRNN, AirFormer, and AirPhyNet on KDD Cup 2018 tests H1 against the MAE < 13.5 target.

*Table 3.* Research hypotheses

| # | Hypothesis | Decision criterion |
|---|---|---|
| H1 | Fusion accuracy — ST-MFXAI achieves significantly lower RMSE than any single-modality baseline across 1h/6h/24h horizons | Statistically significant RMSE reduction at each horizon |
| H2 | Embedded XAI fidelity — fusion-layer attributions are more faithful than post-hoc explanations on an identical model [7] | Higher faithfulness score for embedded vs. post-hoc attributions |
| H3 | Spatial-module specificity — the GCN's contribution concentrates in high inter-station-correlation subsets ($r \geq 0.7$) | Significant RMSE gap in high-correlation subsets; minimal gap in low-correlation subsets |
| H4 | Cross-city generalization — a Beijing-trained model transfers with limited degradation, further reduced under fine-tuning [6] | $R^2$ degradation < 15% (zero-shot); further reduction under few-shot fine-tuning |

A cross-city generalization phase evaluates the Beijing-trained model zero-shot and few-shot on London and OpenAQ cities, testing H4. A fidelity evaluation phase computes faithfulness and consistency scores for the embedded XAI outputs against post-hoc equivalents, testing H2. Results from all four phases together determine whether each hypothesis is supported or rejected.

## Conclusion

This paper has presented the research methodology and experimental design underpinning ST-MFXAI, a spatio-temporal multi-sensor fusion framework for urban air quality prediction with explainability embedded directly into its fusion architecture rather than applied as a post-hoc addition. This study built on the fragmentation identified across the spatial-modeling, temporal-modeling, and explainable-AI literature. This methodology combines a Graph Convolutional spatial module, a Transformer-based temporal module, and a gated fusion layer producing four native XAI outputs, evaluated through a rigorously chronological three-dataset pipeline against seven internal and external comparison models. There are four falsifiable hypotheses. These are spanning predictive accuracy, spatial-module specificity, embedded XAI fidelity, and cross-city generalization. These were formalized with pre-registered decision thresholds and operationalized through a short, sequential validation strategy in which each phase's output feeds directly into the next.

**References**


1. Y. Liang, Y. Xia, S. Ke, Y. Wang, Q. Wen, J. Zhang, Y. Zheng, and R. Zimmermann, “AirFormer: Predicting Nationwide Air Quality in China with Transformers,” Proc. AAAI Conf. Artif. Intell., vol. 37, no. 12, pp. 14329–14337, 2023, doi: 10.1609/aaai.v37i12.26676.
2. K. H. Hettige, J. Ji, S. Xiang, C. Long, G. Cong, and J. Wang, “AirPhyNet: Harnessing Physics-Guided Neural Networks for Air Quality Prediction,” Proc. 12th Int. Conf. Learning Representations (ICLR), 2024.
3. G. Jin, Y. Liang, Y. Fang, Z. Shao, J. Huang, J. Zhang, and Y. Zheng, “Spatio-Temporal Graph Neural Networks for Predictive Learning in Urban Computing: A Survey,” IEEE Trans. Knowl. Data Eng., vol. 36, no. 10, pp. 5388–5408, 2023, doi: 10.1109/TKDE.2023.3333824.
4. D. Iskandaryan, F. Ramos, and S. Trilles, “Graph Neural Network for Air Quality Prediction: A Case Study in Madrid,” IEEE Access, vol. 11, pp. 2729–2742, 2023, doi: 10.1109/ACCESS.2023.3234214.
5. H. Liu, Q. Han, H. Sun, et al., “Spatiotemporal Adaptive Attention Graph Convolution Network for City-Level Air Quality Prediction,” Sci. Rep., vol. 13, art. 13335, 2023, doi: 10.1038/s41598-023-39286-0.
6. “Location-Aware Transfer Learning for Air Quality Time-Series Prediction,” Electronics, vol. 15, no. 11, art. 2470, doi: 10.3390/electronics15112470.
7. “Interpretable Machine Learning Approaches for Forecasting and Predicting Air Pollution: A Systematic Review,” Aerosol Air Qual. Res., 2023, doi: 10.4209/aaqr.230151.
8. S. Zhang, B. Guo, A. Dong, J. He, Z. Xu, and S. X. Chen, “Cautionary Tales on Air-Quality Improvement in Beijing,” Proc. R. Soc. A, vol. 473, no. 2205, art. 20170457, 2017, doi: 10.1098/rspa.2017.0457. Dataset: Beijing Multi-Site Air-Quality Data, UCI Machine Learning Repository, doi: 10.24432/C5RK5G.
9. KDD Cup 2018 of Fresh Air — Beijing & London air quality forecasting dataset. Official competition: https://www.kdd.org/kdd2018/kdd-cup · Archived dataset: https://doi.org/10.5281/zenodo.3893504
10. C. M. Rosales et al., “Open Air Quality Data Platforms for Environmental Health Research and Action,” Curr. Environ. Health Rep., 2025, doi: 10.1007/s40572-025-00487-6. Platform: OpenAQ, https://openaq.org