*FUDMA Journal of Sciences (FJS)*
*ISSN online: 2616-1370*
*ISSN print: 2645 - 2944*
*Vol. 10 No. 24, December, 2026, pp 161 – 176*
*DOI:* https://doi.org/10.33003/fjs-2026-1024-6330

# Data Representations for Air Pollution Forecasting and Modelling: A Systematic Literature Review

**[1]Agozie Hyacinth Ene, *[2]Ibrahim Aishat Musa, [1]Francis Sunday Bakpo and [3]Caleb Markus**

[1]Department of Computer Science, University of Nigeria, Nsukka, Enugu State, Nigeria.
[2]Advanced Space Technology Applications Laboratory, National Space Research and Development Agency (NASRDA), Uyo, Akwa-Ibom State, Nigeria.
[3]Department of Computer Science, Taraba State University, Jalingo, Taraba State, Nigeria.

*Corresponding authors' email: ibrashat2@yahoo.com
ORCID iD: https://orcid.org/0009-0000--7068-7512

**Received:** 24 Oct. 2026
**Accepted:** 10 Nov. 2026
**Published:** 12 Nov. 2026

**ABSTRACT**
Air-pollution observations are heterogeneous, incomplete and spatially and temporally dependent, making data representation an important modelling choice. This systematic literature review examined how air-pollution data are structured for processing and prediction. Following PRISMA 2020 and computing-review guidance, five sources were searched for English-language publications from 2020 to 2026. Of 13,025 identified records, 678 full texts were assessed and 37 publications included, comprising 29 primary studies and eight reviews or surveys. Primary studies were evaluated using eight quality criteria, and data representations were classified into nine families. Sequential series and sliding-window tensors appeared in 21 of 29 primary studies (72%), while graph representations appeared in 10 (34%). The review also identified decomposition, continuous-time and multimodal representations. Reporting of preprocessing and data-splitting procedures was frequently insufficient to establish protection against information leakage; only two primary studies were reported to provide both code and data. The included evidence was geographically concentrated, with no African study represented in the selected corpus. These findings support explicit justification of data representations, leakage-safe evaluation, transparent graph construction and reproducible benchmarks. Further research should assess representations on sparse, irregular monitoring networks, including those in under-represented regions.

**Keywords:** Air Pollution, Data Representation, Graph Neural Networks, Spatiotemporal Data, Data Preprocessing, Systematic Literature Review

## INTRODUCTION

Ambient air pollution remains one of the largest environmental risks to human health. The World Health Organization (2021) tightened its guideline for annual mean PM2.5 to 5 µg/m³ because the evidence linking fine particulate matter, nitrogen dioxide and ozone to cardiovascular, respiratory and premature mortality outcomes had strengthened considerably. Meeting such guidelines depends on monitoring and forecasting systems that can turn raw measurements into reliable information, and this has driven a rapid expansion of data-driven methods for air quality prediction, gap filling and spatial inference (Huang et al., 2021; Méndez et al., 2023).

Air pollution data are, however, difficult data. Measurements arrive from reference stations, low-cost sensors, mobile platforms and satellites at different temporal resolutions and spatial supports (Castell et al., 2017; Buelvas et al., 2023). They contain gaps caused by sensor failure and maintenance, which are rarely missing completely at random (Little & Rubin, 2019). They are autocorrelated in time, correlated across neighbouring stations, and driven by meteorology, emissions and transport processes that are themselves only partly observed (Zheng et al., 2015). Before any model is trained, an analyst must therefore decide how these observations are to be organised: as a univariate series, a multivariate window, a station by time matrix, a grid, a graph, or something else. That decision determines which dependencies are explicit, which are discarded, how missing values can be handled and which algorithms can be applied.

In computer science, a data structure is a way of organising data so that the operations required on it can be performed efficiently (Cormen et al., 2022). In data-driven environmental modelling, the related but broader notion of a data representation refers to the form in which observations are presented to a learning algorithm: arrays, sequences, tensors, matrices, graphs, grids and continuous fields, each with its own algebra (Kolda & Bader, 2009). The two terms are distinguished in this review. A data structure is the computational container in which data are stored, whereas a data representation is the modelling decision about which entities, dependencies and dimensions that container encodes, for example stations as graph nodes or hours as positions in a window. Because the review is concerned with the second of these, the term data representation is used throughout, and the nine coded categories are called representation families. The success of graph convolutional and graph attention networks (Kipf & Welling, 2017; Veličković et al., 2018), and of spatiotemporal graph models developed first for traffic forecasting (Yu et al., 2018; Li et al., 2018; Wu et al., 2019), illustrates why the distinction matters: these methods became possible because sensor networks were re-represented as graphs rather than as independent sequences. The same shift is now visible in air quality research (Wang et al., 2020; Liang et al., 2023).

Despite this, the existing review literature is organised almost entirely around models. Reviews have compared machine learning and deep learning algorithms for air quality forecasting (Méndez et al., 2023; Kaur et al., 2023; Zhang et al., 2024), surveyed spatiotemporal prediction methods (Agbehadji & Obagbuwa, 2024), examined interpretability (Houdou et al., 2024), mapped the use of machine learning with the Internet of Things (Gryech et al., 2024) and assessed data quality in sensor networks (Buelvas et al., 2023). Huang et al. (2021) came closest to a data-centric view, but their overview predates the growth of graph, decomposition and continuous-time representations that characterises the most recent literature. As a result, to the authors' knowledge, there is no systematic account of which data representations are used, why they are chosen, how they interact with preprocessing, and whether the reported gains are attributable to the representation or to the model built on top of it. The present review is designed to complement rather than duplicate this work. It does not re-synthesise the model comparisons reported in earlier reviews; instead, it uses the representation as the unit of analysis, codes each primary study by the representations it uses and the preprocessing applied to them, and treats earlier systematic reviews and surveys only as contextual, secondary evidence that is reported separately from the primary studies.

Treating the representation as a methodological choice in its own right matters for two reasons. First, if reported gains depend partly on how the data are organised and preprocessed, comparisons between models are difficult to interpret unless the representation is described and controlled. Second, representations developed for dense monitoring networks may not transfer to the sparse and irregular networks found in many low- and middle-income settings, including Nigeria and the wider sub-Saharan African region. The review therefore provides a taxonomy of the representation families used in studies published between 2020 and 2026, appraises how these representations relate to preprocessing, model choice, performance and reproducibility, and derives a research agenda and reporting recommendations.

This study aimed to systematically identify, classify and critically evaluate the data representations used in air-pollution forecasting and modelling, with emphasis on preprocessing, evaluation validity, reproducibility and research needs in under-represented settings.

**Review Design**

The review followed the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) 2020 statement (Page et al., 2021) and the guidelines for systematic literature reviews in computing proposed by Kitchenham and Charters (2007). The protocol specified the research questions, search strategy, eligibility criteria, screening procedure, quality assessment, data extraction form and synthesis method before screening began. The protocol was not prospectively registered. Because the included studies used different datasets, horizons and metrics, a statistical meta-analysis was not appropriate, and the evidence was synthesised narratively and through structured tabulation, with frequency analysis used to describe trends. The completed PRISMA 2020 checklist is provided as Supplementary File S1, and the coding dictionary, study-level quality scores and full extraction sheet are provided as Supplementary Files S2 to S4.

**Research Questions**

Five research questions guided the review. They are presented with their motivation in Table 1.

**Table 1: Research Questions and Motivation**

| ID | Research Question | Motivation |
|---|---|---|
| RQ1 | Which data representations are used to represent air pollution data, and how has their use changed between 2020 and 2026? | To establish a taxonomy and identify trends in representation. |
| RQ2 | How does the choice of data representation relate to the model family, reported performance and computational cost? | To judge whether gains are attributable to representation or to model design. |
| RQ3 | Which preprocessing operations are applied before and during structuring, and how rigorously are they reported? | Preprocessing determines data quality and the risk of information leakage. |
| RQ4 | What is the methodological quality and reproducibility of the evidence? | To weight the evidence and identify reporting weaknesses. |
| RQ5 | What gaps remain, and what research is needed, particularly for data-scarce regions? | To set a research agenda relevant to under-represented settings. |

**Search Strategy**

Five sources were searched. Scopus and the Web of Science Core Collection are bibliographic indexes; IEEE Xplore is a publisher-hosted digital library that also indexes conference proceedings; SpringerLink is a publisher platform covering Springer journals and books; and Google Scholar is a web search engine. SpringerLink and Google Scholar were therefore used as supplementary search platforms rather than as bibliographic indexes, to capture computing and engineering papers that the indexes might miss. The search combined three concept blocks: the pollutant or domain, the data representation, and the processing task. Search strings were adapted to the syntax of each source, as shown in Table 2, whose note also gives the fields searched, the filters applied and the number of records retrieved from each source. All searches were limited to publications from 1 January 2020 onwards and, where the platform allowed, to English-language documents. The last search of all sources was run on 30 August 2026. Google Scholar, which does not support field-restricted Boolean searching, was searched with a simplified string and its results were screened in order of relevance; because Google Scholar displays at most 1,000 results per query, the simplified string was run as separate query variants, and all results returned for each variant were exported, giving 3,480 records. The reference lists of included studies were then examined by backward snowballing (Wohlin, 2014) to identify studies the database searches might have missed; this identified no additional eligible studies beyond those already retrieved by the database searches.

**Table 2: Search Strings used in Each Database**

| Source | Search String (Adapted to Database Syntax) |
|---|---|
| Scopus | TITLE-ABS-KEY(("air pollution" OR "air quality" OR "PM2.5" OR "PM10" OR "NO2" OR "ozone" OR "AQI") AND ("data structure*" OR "time series" OR "graph" OR "tensor" OR "grid" OR "matrix" OR "spatiotemporal" OR "spatio-temporal") AND ("preprocess*" OR "imputation" OR "missing data" OR "data processing" OR "forecast*" OR "predict*" OR "inference")) AND PUBYEAR > 2019 |
| Web of Science | TS=(("air pollution" OR "air quality" OR PM2.5 OR AQI) AND ("data structure*" OR "time series" OR graph OR tensor OR grid OR matrix OR spatiotemporal) AND (preprocess* OR imputation OR "missing data" OR forecast* OR predict*)) |
| IEEE Xplore | ("All Metadata":"air quality" OR "All Metadata":"air pollution") AND ("All Metadata":"time series" OR "All Metadata":graph OR "All Metadata":tensor OR "All Metadata":spatiotemporal) AND ("All Metadata":forecast* OR "All Metadata":imputation OR "All Metadata":preprocessing) |
| SpringerLink | "air quality" AND ("time series" OR graph OR tensor OR spatiotemporal) AND (forecasting OR imputation OR preprocessing) |
| Google Scholar | "air quality" OR "air pollution" "data structure" OR "graph" OR "time series" forecasting OR imputation OR preprocessing |

*Note.* Fields searched: Scopus, title, abstract and keywords; Web of Science, topic (title, abstract, author keywords and Keywords Plus); IEEE Xplore, all metadata; SpringerLink, full text; Google Scholar, full text, ranked by relevance. Filters: publication year 2020 to 2026 and English language where available. Records retrieved: Scopus, 3,250; Google Scholar, 3,480; IEEE Xplore, 2,470; SpringerLink, 2,130; Web of Science, 1,695; total, 13,025. Last search: 30 August 2026.

**Eligibility Criteria**

Studies were eligible if they reported how ambient air pollution data were represented and processed for a predictive or inferential modelling task, namely forecasting, imputation (gap filling) or spatial inference of pollutant concentrations or air quality indices, and if they were published in English between January 2020 and the last search date. Studies of other processing tasks, such as source apportionment, emission inventory compilation or purely descriptive analysis, were outside the scope; this is why the title of the review refers to forecasting and modelling rather than to data processing in general. The start date was chosen because graph-based and attention-based methods entered the air quality literature at scale around 2020 (Wang et al., 2020), so earlier work would mainly duplicate findings already summarised by previous reviews. Systematic reviews and surveys were retained as secondary evidence, because they document how the field itself has framed the problem, but they were not counted in any frequency or percentage reported for the primary studies. Narrative reviews, editorials and conference abstracts were excluded. The full criteria are given in Table 3.

**Table 3: Inclusion and Exclusion Criteria**

| Criterion | Inclusion | Exclusion |
|---|---|---|
| Topic | Ambient or urban air pollution data (pollutant concentrations or AQI) | Indoor air only, emissions inventories, or unrelated environmental data |
| Focus | Explicit representation and preprocessing of air pollution data for forecasting, imputation (gap filling) or spatial inference | No predictive or inferential modelling task (e.g., purely epidemiological, policy or descriptive studies; source apportionment) |
| Study type | Peer-reviewed empirical studies; systematic reviews and surveys as secondary evidence | Narrative reviews, editorials, conference abstracts, theses |
| Detail | Sufficient methodological detail to identify the data representation used | Data representation not described or not identifiable |
| Period | Published January 2020 to the last search date (30 August 2026) | Published before 2020 |
| Language | English | Other languages |
| Availability | Full text retrievable and bibliographic record verifiable | Full text unavailable, or record not verifiable in any index |

**Study Selection**

Records from all sources were exported to a reference manager and duplicates were removed, first automatically by digital object identifier and then manually by comparing title, authors and publication year. Two reviewers (A.E. and I.A.M.) then screened titles and abstracts independently against the eligibility criteria. Disagreements were resolved by discussion, and where discussion did not resolve a disagreement a third reviewer (F.B.) made the final decision. A formal inter-rater agreement statistic was not computed. Full texts of the remaining reports were retrieved and assessed in the same way. Each retained bibliographic record was then checked against the publisher's site and at least one index to confirm that the title, authors, venue and digital object identifier corresponded to a real publication. Four reports, all dated 2026, could not be verified in any index and were moved to the not-retrieved category. One study, by Fahim and Karim (2026), could be verified only as a preprint and was retained with that status noted. The selection process is summarised in Figure 1, and the completed PRISMA 2020 checklist is given in Supplementary File S1.

Of the 13,025 records identified (Scopus, 3,250; Google Scholar, 3,480; IEEE Xplore, 2,470; SpringerLink, 2,130; Web of Science, 1,695), 1,876 duplicates were removed, leaving 11,149 records for title and abstract screening. Of these, 10,428 were excluded because they were not relevant to the topic (n = 5,620), concerned the wrong population or context (n = 1,960), did not address air quality prediction or data processing (n = 1,850), were narrative reviews, editorials or conference abstracts (n = 620) or were not in English (n = 378). Of the 721 reports sought for retrieval, 43 were not retrieved: 25 were not accessible, 14 had no full text available and four had bibliographic records that could not be verified. The remaining 678 reports were assessed in full, and 641 were excluded because they did not focus on data representation or data processing (n = 248), were not concerned with air quality prediction or modelling (n = 162), used no relevant data representation or algorithm (n = 118), gave insufficient

methodological detail (n = 67), were published before 2020 (n = 32) or were not empirical studies or eligible reviews (n = 14). The 32 pre-2020 reports reached full-text assessment because publication-date filters could not be applied uniformly on every platform, in particular Google Scholar and SpringerLink. The reasons at each stage sum to the stage totals shown in Figure 1. The final set contained 37 publications: 29 primary empirical studies and eight systematic reviews or surveys.

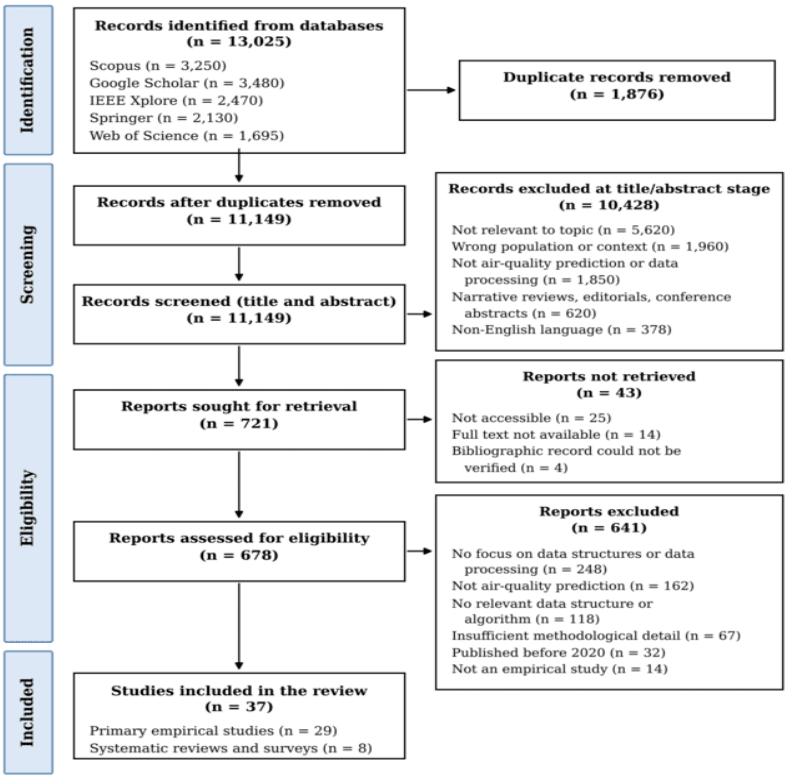



Figure 1: PRISMA 2020 Flow Diagram of Study Identification, Screening and Inclusion

**Quality Assessment**

Each primary study was appraised against the eight criteria in Table 4, adapted from Kitchenham and Charters (2007) to emphasise the issues that matter for data processing. Two criteria deserve comment. QA5 asks whether the evaluation was protected against information leakage, that is, whether scalers, imputers and signal decompositions were fitted on training data only and whether the test set was held out chronologically. Leakage of this kind is a well-documented source of over-optimistic results in data mining (Kaufman et al., 2012), and standard random cross-validation is known to be problematic for dependent time series (Bergmeir & Benítez, 2012). QA7 asks whether code and data were available, without which the data representation cannot be reproduced. Each criterion was scored 1 (yes), 0.5 (partly) or 0 (no), giving a maximum of 8. Studies scoring 6 or more were classed as high quality, 4 to 5.5 as moderate and below 4 as low. No study was excluded on quality grounds, and quality scores were not used as numerical weights. Instead, they informed the synthesis in three ways: numerical performance claims from studies scoring below 1 on QA5 or QA6 were not used as evidence that one representation outperforms another; where findings conflicted, greater interpretive weight was given to high-quality studies; and claims that rest only on moderate- or low-quality studies are identified as such in the text. Quality was scored by one reviewer and checked by a second.

**Table 4: Quality Assessment Criteria**

| Code | Criterion | Scoring guide |
|---|---|---|
| QA1 | Are the aims and research questions clearly stated? | 1 clear; 0.5 implicit; 0 absent |
| QA2 | Are the data source, study area, period and resolution described? | 1 all reported; 0.5 some; 0 none |
| QA3 | Is the data representation explicitly defined (shape, nodes, edges, grid or field)? | 1 explicit; 0.5 inferable; 0 not identifiable |
| QA4 | Are preprocessing steps (missing data, outliers, scaling) reported? | 1 fully; 0.5 partly; 0 not reported |
| QA5 | Is the evaluation protected against information leakage? | 1 chronological split and train-only fitting reported; 0.5 not reported or unclear; 0 leakage-prone procedure reported |
| QA6 | Are appropriate baselines and correctly defined metrics reported? | 1 yes; 0.5 limited; 0 absent or erroneous |
| QA7 | Are code and data available for reproduction? | 1 both; 0.5 data only; 0 neither |
| QA8 | Are limitations and generalisability (other sites or periods) examined? | 1 tested; 0.5 discussed; 0 absent |

**Data Extraction and Synthesis**

A standard extraction form was applied to each included study. The items recorded are listed in Table 5. Data representations were coded using the categories that emerged from an initial reading of the full texts, refined iteratively until every study could be assigned to one or more categories. A study using several representations, for example a graph built over sliding-window tensors, was coded under each. Extraction was carried out by one reviewer and checked by a second. The coding dictionary for the nine representation families, with the decision rules used for multi-label coding, is given in Supplementary File S2; the study-level quality scores in Supplementary File S3; and the full extraction sheet, with bibliographic identifiers and DOIs, in Supplementary File S4. Three counting rules were applied. First, a study was coded under every family it used, so family counts are non-exclusive and do not sum to the number of studies. Second, all frequencies and percentages refer to the 29 primary studies only; the eight secondary studies were extracted separately (Table 7) and are never included in these denominators. Third, a study's year was the year of its final journal publication or, for advance online publications and the preprint, the year in which it first appeared online, and its study area was the country or countries in which its data were collected; a study using data from several countries was assigned to each of them, so regional counts can overlap.

**Table 5: Data Extraction Items**

| Item | Description |
|---|---|
| Bibliographic | Authors, year, venue, study type (primary or secondary) |
| Study area | Country or city where the data were collected |
| Data source | Reference stations, low-cost or mobile sensors, satellite, meteorological and auxiliary layers; period and resolution |
| Data representation | Representation used for storage and learning (coded into the nine families of Figure 5) |
| Preprocessing | Missing-data handling, outlier treatment, scaling, decomposition, feature selection, leakage control |
| Task and model | Forecasting, imputation, spatial inference or classification; model family |
| Performance | Metrics reported and best results; computational cost where reported |
| Quality | Scores on QA1 to QA8 |

## RESULTS AND DISCUSSION

**Characteristics of the Included Studies**

The temporal distribution of the included studies is shown in Figure 2. Publication activity rose steadily over the review period, and 15 of the 29 primary studies (52%) appeared in 2025 or 2026. This concentration matters for interpretation: the most recent studies use the most elaborate representations, and many have not yet been tested by independent replication. The eight secondary studies were all published between 2021 and 2024, which means that none of them covers the representations, such as continuous-time and multimodal fusion representations that became prominent after 2024.

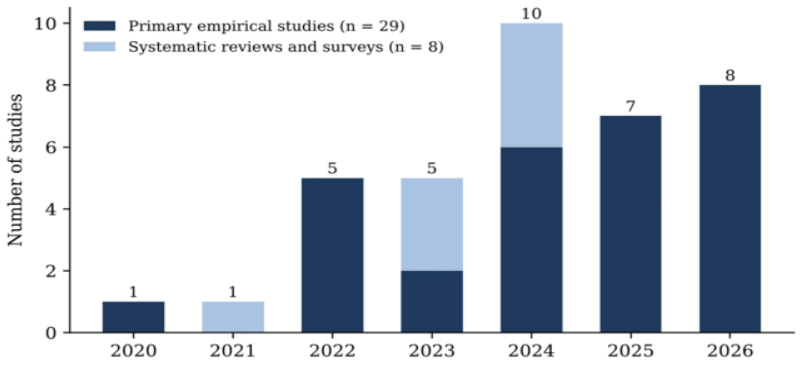
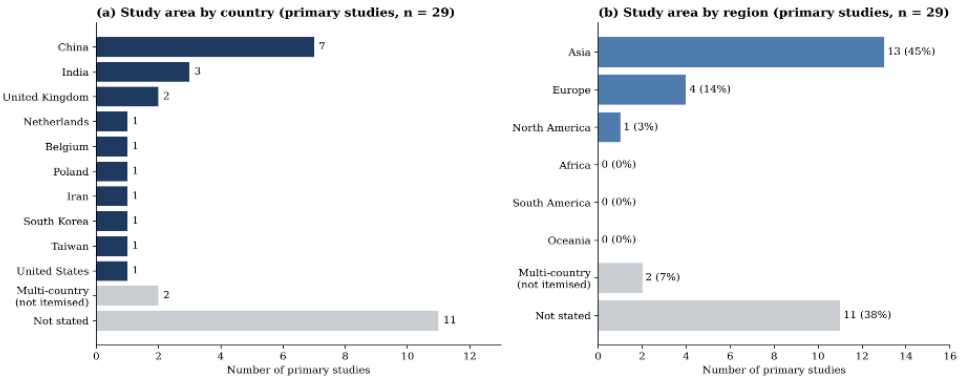



Figure 2: Number of Included Studies by Year of Publication and Study Type

The geographical distribution of the primary studies is shown in Figure 3, using the assignment rule described in the Methods. Asia accounts for 13 studies, and China alone for 7. Six studies, about one in five, draw on the same multi-site Beijing dataset (Zhang et al., 2017) or other Beijing network data (Chauhan et al., 2026; Cheng & Zhang, 2026; Dun et al., 2022; Liu et al., 2023; Wang et al., 2024; Yin et al., 2026). Europe is represented by four studies, from the United Kingdom, Poland, the Netherlands and Belgium, and North America by one. Two of these assignments overlap: Liu et al. (2023) used data from China and the United Kingdom, and Hofman et al. (2022) used testbeds in the Netherlands, Belgium and the United States. Two further studies used data from several countries without itemising them (Fahim & Karim, 2026; Lilhore et al., 2025). Eleven studies did not state the study area in the records available to the review, which limits the geographical conclusions that can be drawn, since a representation built for a dense urban network cannot be assumed to transfer to a sparse one. Within the selected corpus, no primary study used data from Africa, South America or Oceania. This does not show that no such studies exist, but it means that the evidence synthesised here comes from monitoring conditions that differ from those in most Nigerian cities, a point taken up in the Discussion.




*Note. A study using data from several countries or regions is counted under each, so counts sum to more than 29: Liu et al. (2023) is counted under China and the United Kingdom (Asia and Europe), and Hofman et al. (2022) under the Netherlands, Belgium and the United States (Europe and North America). Percentages use n = 29. The eight secondary studies are excluded.*

Figure 3: Geographical Distribution of the Primary Studies (n = 29) by (a) Country and (b) Region

Table 6 summarises the 29 primary studies, including their study area, data representation, key preprocessing, model and reported performance. Table 7 summarises the eight secondary studies and how each treats data representation.

**Table 6: Characteristics of the Primary Studies (n = 29)**

| Study | Area | Representation | Key Preprocessing or Structuring | Model | Reported Performance |
|---|---|---|---|---|---|
| Abdelmalek et al. (2025) | NS | TAB | Feature engineering for supervised learning | Classical ML regressors and classifiers | Error and classification metrics |
| Baruah et al. (2024) | United Kingdom | SPX, TAB | Road network split into 30 m segments; proxy variables per segment | Random forest | $R^2$ 0.85 (standard), 0.65 to 0.71 (transfer tests) |
| Chatterjee et al. (2025) | India | GRID, TS | Sensors mapped to an R × C grid with 4/8-neighbour adjacency; outlier clipping; short-gap interpolation, MLP for long gaps | Grid-embedded recurrent MLP (AiM), 8-bit quantised | RMSE up to 12.4% lower than LSTM/GRU; 35% lower edge latency |
| Chauhan et al. (2026) | China | GR, TS | Proximity-based station graph | GAT with GRU and feature attention (CIST-GNN) | $R^2$ 0.932, RMSE 21.66 µg/m³ |
| Cheng & Zhang (2026) | China | TS, DEC | Interpolation; VMD fitted on training data only; sample-entropy clustering of modes; 24 h windows | TCN-BiLSTM per frequency band with attention fusion | $R^2$ 0.960 (urban), 0.940 (suburban) |
| Dimri et al. (2026) | South Korea | GR, DEC | Interpolation, VMD and PCA; transport-informed graph | Graph neural network learning advection-diffusion-reaction operators (TransNet), 72 h horizon | Multi-horizon error metrics |
| Ding et al. (2022) | NS | TS, DEC | Decomposition with improved run-length reconstruction | Combined deterministic and uncertainty forecasting system | Point and interval accuracy |
| Dun et al. (2022) | China | GR, TS | Dynamic adjacency matrices from grey relational analysis | Dynamic graph convolution with multi-channel TCN | Accuracy and multi-step stability |
| Fahim & Karim (2026) | Multi-country | TS | Deployable (forecast-time) versus perfect-foresight covariates | Zero-shot time-series foundation model, tuned specialist, transfer learning | MASE 0.662 (zero-shot) versus 0.692 (specialist) |
| Fang et al. (2026) | China | TS, DEC | Linear interpolation (2.3% missing); z-score outlier correction on training data; CEEMDAN into 9 IMFs; 7-day windows | Multi-scale CNN-BiGRU-attention tuned by IGWO | $R^2$ 0.962; weaker transfer to Zibo and Daqing |
| Feng et al. (2025) | NS | TS | Parallel deep learning predictors | Dempster-Shafer evidence fusion of multiple deep learners | Error metrics |
| Havaei et al. (2025) | Iran | CONT, TS | GNN/iterative RF imputation; log transform; autoencoder decomposition; 3-day windows | Multi-timescale attention neural ODE (MA-NODE) | $R^2 > 0.9$ short horizon; about 0.5 for PM2.5 at step 3 |

| Study | Area | Representation | Key Preprocessing or Structuring | Model | Reported Performance |
|---|---|---|---|---|---|
| Hofman et al. (2022) | Netherlands, Belgium, United States | GR, MAT | Sparse city-scale observations completed on a graph | Geographical random forest and variational graph autoencoder | $R^2$ 0.68 to 0.82 (Europe), 0.28 to 0.46 (USA) |
| Hu et al. (2022) | NS | TS, GRID | Spatio-temporal input tensors | Spatio-temporal deep learning | Error metrics |
| Iwaszenko et al. (2024) | Poland | TS, TAB | Single-step interpolation; outlier screening; 10 min, 1 h and 24 h aggregation | Decision tree, random forest, RNN, LSTM | LSTM best at 10 min/1 h; RF best at 24 h |
| Lilhore et al. (2025) | Multi-country | MM, GRID, GR, TS, CONT | UTC alignment; interpolation; wavelet denoising; PCA; SMOTE | CNN, BiLSTM, attention, GNN and neural ODE hybrid | RMSE 6.21, $R^2$ 0.988 |
| Liu et al. (2020) | NS | MAT | Low-rank matrix completion of station-time data | LRMC versus nearest neighbour, spline, EM and others | Best at 30% missingness; insensitive to gap length |
| Liu et al. (2022) | Taiwan | TAB, SPX | Two-stage temporal then spatial prediction | Support vector machine | High short-term accuracy; land-use factors identified |
| Liu et al. (2023) | China and United Kingdom | GR, TS | Distance, POI-similarity and temporal-correlation graphs with adaptive adjacency | Spatiotemporal adaptive attention GCN (STAA-GCN) | Best RMSE/MAE/MAPE for 6 h PM2.5 |
| Ma et al. (2024) | NS | TS | Transfer of pretrained representations | Transfer learning with two-stage attention | Error metrics |
| Mirzaei & Torghabeh (2026) | NS | GR, CONT | Graph of stations with continuous-time paths | Neural controlled differential equation with GAT | Error metrics |
| Rajesh et al. (2025) | NS | MM, TAB, TS | Cloud streaming with 5-minute updates; spatial overlays | RF, gradient boosting, XGBoost, LSTM with SHAP | High temporal accuracy (no metrics in abstract) |
| Saad et al. (2024) | NS | CONT | Coordinates and time as inputs to a neural field | Bayesian neural field (BayesNF) | Improved over baselines with uncertainty |
| Shi et al. (2026) | NS | GR, TS | Local and sparse spatial fusion | Dynamic spatiotemporal model with structured temporal modelling | Error metrics |
| Sudha et al. (2025) | India | TS, TAB | Weighted feature selection by improved gannet optimisation | Adaptive residual Bi-LSTM with pyramid dilation | Accuracy about 95% |
| Sui & Han (2023) | NS | GR, TS | Spatial-view and logical-view station graphs | Multi-view multi-task GCN with GRU | Outperformed state of the art |
| Sundaramoorthy et al. (2024) | India | TS | Mahalanobis outlier removal; normalisation; optimiser-weighted features | Multiscale depth-wise separable adaptive TCN | MAE 28% and RMSE 14% lower than baselines |
| Wang et al. (2024) | China | GR, TS, DEC | VMD into sub-sequences; station graph per sub-sequence | VMD-GAT-BiLSTM | Better short- and long-term forecasts |
| Yin et al. (2026) | China | TS, TAB | KNN imputation; min-max scaling; batch × time × station × feature tensor with station embeddings | CNN-BiLSTM-attention with XGBoost second stage | RMSE 10.29, $R^2$ 0.928; 3.2 h versus 9.8 h training |

*Note.* Representation codes: TS = sequential time series or window tensor; GR = graph; TAB = tabular feature matrix; GRID = regular grid or raster; MAT = low-rank matrix; CONT = continuous-time or neural field; DEC = decomposition stack; MM = multimodal heterogeneous; SPX = spatial index or segment. NS = not stated in the record available to the review. Fahim and Karim (2026) is a preprint. Performance values are those reported by the authors for different pollutants, horizons, datasets, data splits and metrics; they must not be compared across studies or read as a ranking. The outcome context available for each study (task, pollutant, horizon, dataset, split, metric and units) is given in Supplementary File S4.

**Table 7: Systematic Reviews and Surveys Included as Secondary Evidence (n = 8)**

| Study | Type | Focus | Treatment of data representation |
|---|---|---|---|
| Agbehadji and Obagbuwa (2024) | Systematic review | ML and DL for spatiotemporal air quality prediction | Organised by model family; spatiotemporal inputs discussed as model features rather than representations |
| Buelvas et al. (2023) | Systematic mapping | Data quality in IoT air quality monitoring | Quality dimensions of sensor data; not linked to downstream representations |
| Gryech et al. (2024) | Systematic review | ML and IoT for outdoor monitoring and prediction | Emphasis on sensing architecture and algorithms |
| Houdou et al. (2024) | Systematic review | Interpretable ML for forecasting | Interpretability methods; representation implicit |

| Study | Type | Focus | Treatment of data representation |
|---|---|---|---|
| Huang et al. (2021) | Overview | Big data for monitoring, forecasting and traceability | Closest to a data-centric view; predates graph and continuous-time growth |
| Kaur et al. (2023) | Systematic review | Deep air quality prediction techniques | Classified by deep architecture |
| Méndez et al. (2023) | Survey | ML algorithms to forecast air quality | Classified by algorithm and pollutant |
| Zhang et al. (2024) | Systematic survey | Deep learning air quality prediction | Spatiotemporal representation treated as an attribute of the model |

The secondary studies are reported as context only, and their findings are not combined with those of the primary studies in any count or percentage. The pattern in Table 7 nonetheless clarifies the contribution of the present review. Every secondary study organises its evidence by algorithm, sensing platform or quality dimension. None uses the data representation as the unit of analysis, and none asks whether the improvements it reports could have been achieved by a better representation fed to a simpler model.

**Quality of the Evidence (RQ4)**

The study-level quality scores of the primary studies are shown in Figure 4 and listed, with criterion-level summaries, in Supplementary File S3, so that the totals can be reproduced. Fourteen studies were rated high quality (score of 6 or more), 13 moderate (4 to 5.5) and 2 low (below 4), and the mean score was 5.83 out of 8 (169 points across 29 studies). The criterion-level means, each calculated over 29 studies, are more informative than the totals: QA1, 1.00 (criterion fully met by 29 studies); QA2, 0.83 (19); QA3, 0.84 (20); QA4, 0.74 (14); QA5, 0.57 (6); QA6, 0.90 (25); QA7, 0.29 (2); and QA8, 0.66 (11). All studies stated their aims clearly (QA1) and most compared against baselines (QA6), but reproducibility (QA7) was the weakest criterion by a wide margin: only Saad et al. (2024) and Dimri et al. (2026) made both code and data available. Leakage control (QA5) was fully satisfied in six studies. In most of the others it could not be determined from the text whether scalers, imputers and decompositions were fitted before or after the train and test split. This is a reporting finding: an unreported safeguard is not evidence that leakage occurred, but it means that leakage cannot be ruled out, and it is precisely the information a reader needs to judge the reported accuracy.

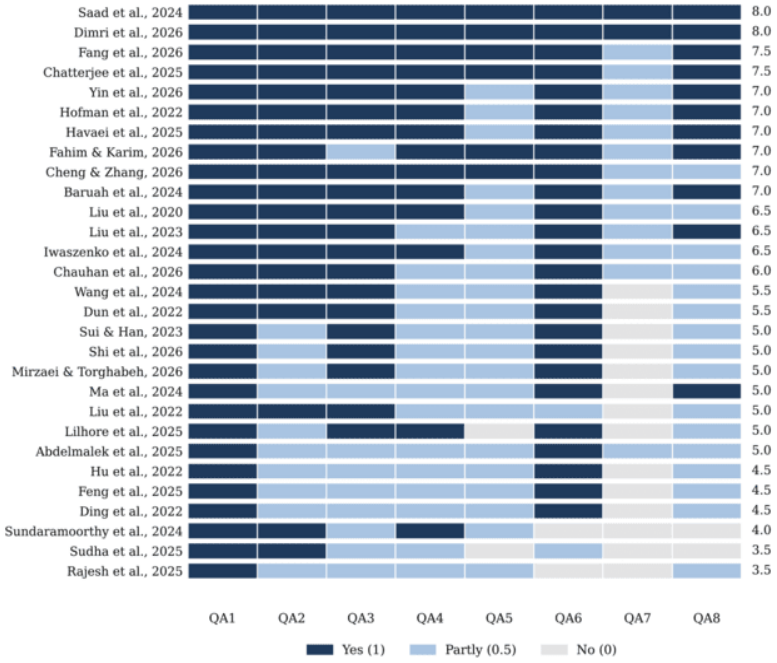



Figure 4: Quality Assessment Scores of the Primary Studies by Criterion

Three studies illustrate the reporting problems that lower the scores. Sundaramoorthy et al. (2024) define their root mean square error with a formula that reproduces the mean absolute error, so the two metrics they report cannot both be correct. Sudha et al. (2025) frame AQI prediction as a six-class classification problem and report accuracy of about 95%, but figures differ across sections of the paper. Lilhore et al. (2025) report an $R^2$ of 0.988 from a pipeline that applies SMOTE, a resampling method designed for imbalanced classification, within a regression workflow, and do not report whether this was confined to the training data. None of these studies is excluded, but their numerical claims carry less weight in the synthesis that follows.

**Taxonomy and Trends in Data Representations (RQ1)**

The representations identified in the primary studies were grouped into nine families under four broad classes, shown in Figure 5. Temporal representations organise observations along the time axis; spatial representations encode relationships between locations; relational and feature-based representations store observations as matrices for completion or for conventional learners; and heterogeneous representations combine sources with different supports.

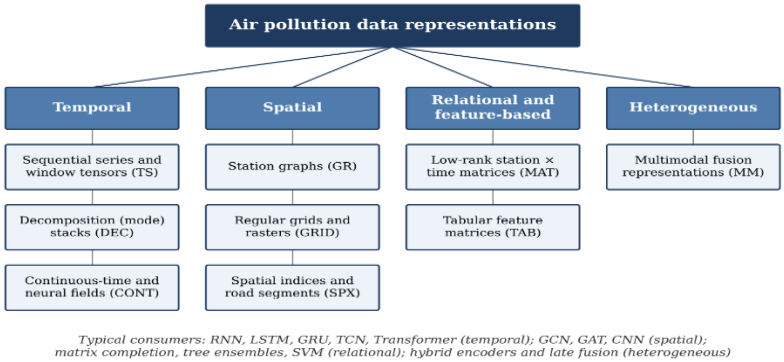



Figure 5: Taxonomy of Data Representations Identified in the Included Studies

The frequency of each family is shown in Figure 6. Sequential time series and sliding-window tensors were used in 21 (72%) of the primary studies, graphs in 10 (34%), tabular feature matrices in 7 (24%), decomposition stacks in 5 (17%), continuous-time formulations and neural fields in 4 (14%), and regular grids in 3 (10%). Low-rank matrices, multimodal representations and spatial segment indices were each used in two studies. Because a study was coded under every representation it used, these categories are non-exclusive, and the percentages, all calculated over the 29 primary studies, therefore sum to more than 100%.

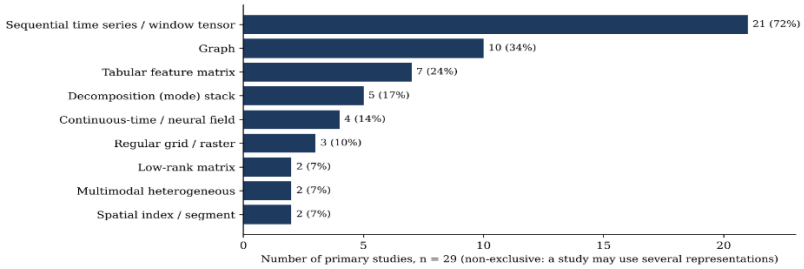



Figure 6: Frequency of Data Representation Families in the Primary Studies

Most studies combined representations: 22 of the 29 primary studies (76%) used more than one. The co-occurrence matrix in Figure 7 shows that the most common pairing was a graph over windowed time series (7 studies), followed by decomposition stacks feeding windowed sequences. The prevalence of hybrids is a substantive finding, not a curiosity. It shows that the field has largely abandoned the idea that one representation can capture both temporal and spatial dependence, but it also means that the contribution of each component is rarely isolated. Only Cheng and Zhang (2026) reported a clean ablation of the representation, showing that removing variational mode decomposition reduced $R^2$ from about 0.96 to 0.90.

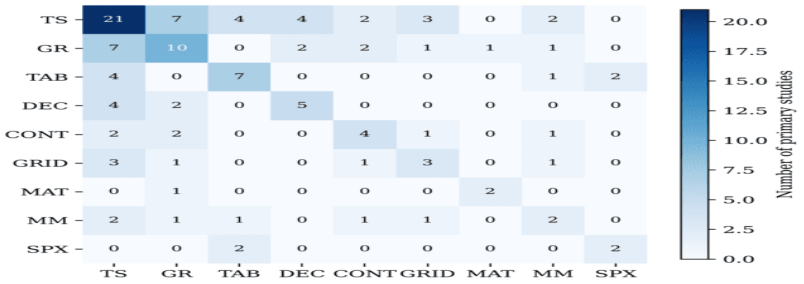



Figure 7: Co-occurrence of Data Representation Families within the Same Primary Study

The change in the use of each family over time is shown in Figure 8, using three periods: 2020 to 2022 (n = 6), 2023 to 2024 (n = 8) and 2025 to 2026 (n = 15). Because each period contains few studies, the shares are descriptive and should not be read as statistically established trends. Three patterns stand out. First, windowed sequences were used in 3 of 6 studies in 2020 to 2022, 6 of 8 in 2023 to 2024 and 12 of 15 in 2025 to 2026, consistent with sequences having become the common substrate on which other representations are layered. Second, low-rank matrix representations appeared only in the earliest period (2 of 6; Liu et al., 2020; Hofman et al., 2022) and not thereafter, even though they remain well suited to imputation. Third, continuous-time formulations (1 of 8 in 2023 to 2024; 3 of 15 in 2025 to 2026) and multimodal representations (2 of 15 in 2025 to 2026) appeared only from 2024 onwards, suggesting a move towards formulations that accept irregular sampling and heterogeneous inputs. Graph representations, by contrast, were used at a broadly stable share (2 of 6, 3 of 8 and 5 of 15). The full family-by-period counts are given in Supplementary File S2.

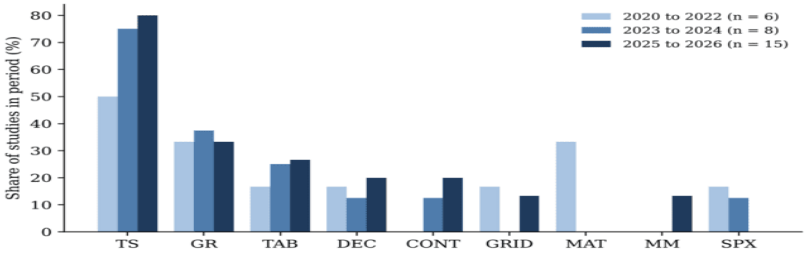



Figure 8: Share of Primary Studies Using Each Data Representation Family by Period

**Data Representations, Models and Performance (RQ2)**

Table 8 synthesises the strengths and weaknesses of each family, drawing on the evidence of the primary studies. The discussion that follows examines the families in turn and argues that the relationship between representation and performance is less straightforward than the literature implies.

**Table 8: Comparative Appraisal of Data Representation Families**

| Family | n | What it encodes | Strengths | Weaknesses | Example studies |
|---|---|---|---|---|---|
| Sequential series and window tensors (TS) | 21 | Temporal order; lagged dependence within and across variables | Simple; compatible with RNN, LSTM, GRU, TCN and Transformer models | Spatial relations implicit or lost; window length chosen ad hoc | Yin et al. (2026); Iwaszenko et al. (2024) |
| Graph (GR) | 10 | Stations as nodes; distance, similarity or correlation as edges | Explicit spatial dependence; handles irregular station layouts | Adjacency construction subjective; static graphs ignore wind-driven transport | Liu et al. (2023); Dun et al. (2022); Dimri et al. (2026) |
| Tabular feature matrix (TAB) | 7 | Engineered features per observation | Works with tree ensembles and SVM; interpretable; cheap | Order and neighbourhood must be hand-engineered | Liu et al. (2022); Baruah et al. (2024) |
| Decomposition stack (DEC) | 5 | Signal split into intrinsic mode functions | Separates trend, seasonal and noise components; high reported accuracy in single-study evaluations | High leakage risk if fitted on the full series; mode count tuned per site | Cheng and Zhang (2026); Fang et al. (2026) |
| Continuous-time and neural field (CONT) | 4 | Latent state as a function of continuous time or coordinates | Accepts irregular sampling; uncertainty quantification | Costly to train; accuracy decays quickly with horizon | Havaei et al. (2025); Saad et al. (2024) |
| Regular grid (GRID) | 3 | Values on an R × C lattice | Natural for satellite data and CNNs; neighbourhood fixed | Interpolation needed to place sparse stations on cells | Chatterjee et al. (2025); Lilhore et al. (2025) |
| Low-rank matrix (MAT) | 2 | Station by time matrix with missing entries | Principled imputation exploiting cross-station redundancy | Assumes low rank; little used for forecasting | Liu et al. (2020); Hofman et al. (2022) |
| Multimodal heterogeneous (MM) | 2 | Station, satellite, meteorological and demographic layers | Richer context; supports risk mapping | Alignment and leakage hard to audit; heavy pipelines | Lilhore et al. (2025); Rajesh et al. (2025) |
| Spatial index or segment (SPX) | 2 | Road segments or GIS units as analysis support | Matches exposure scale; supports mobile sensing | Requires dense mobile data and GIS layers | Baruah et al. (2024); Liu et al. (2022) |

*Note.* n = number of primary studies using the family; a study may use more than one family.

***Sequential and Tabular Representations***

The windowed sequence is the workhorse of the field. It is the natural input for recurrent networks such as the LSTM (Hochreiter & Schmidhuber, 1997), for temporal convolutional networks (Bai et al., 2018) and for attention-based Transformers (Vaswani et al., 2017), and the corpus contains many variants: machine learning on engineered AQI features (Abdelmalek et al., 2025), evidence-theoretic fusion of several deep sequence learners (Feng et al., 2025), transfer of pretrained sequence representations between pollutants and sites (Ma et al., 2024), depth-wise separable temporal convolution (Sundaramoorthy et al., 2024) and residual bidirectional LSTMs (Sudha et al., 2025). The evidence does not, however, support the assumption that more elaborate temporal architectures always justify their cost. Iwaszenko et al. (2024) found that an LSTM outperformed tree-based models on 10-minute and hourly aggregates from low-cost sensors in Poland, but that a random forest on a flattened lag matrix was best at the daily scale. Yin et al. (2026) achieved an $R^2$ of 0.928 with a convolutional, bidirectional LSTM and attention model followed by a gradient-boosting stage, while training in 3.2 hours compared with 9.8 hours for a graph Transformer baseline. Fahim and Karim (2026) went further, reporting that a zero-shot time-series foundation model (Ansari et al., 2024) achieved a lower scaled error than tuned specialist models on short records from 29 cities, provided that only covariates available at forecast time were used. Taken together, these results suggest that the temporal scale of the target and the honesty of the covariate set matter at least as much as the depth of the sequence model.

***Graph Representations***

Graphs were the most frequently used explicitly spatial representation (10 studies), and their appeal is obvious: a monitoring network is a set of locations with relationships, and graph convolution and attention allow those relationships to be learned (Kipf & Welling, 2017; Veličković et al., 2018). The critical weakness lies in how the graph is built. The

studies used geographic distance (Chauhan et al., 2026), points-of-interest similarity and temporal correlation (Liu et al., 2023), grey relational analysis to update the adjacency matrix over time (Dun et al., 2022), separate spatial and logical views (Sui & Han, 2023), graphs over decomposed sub-series (Wang et al., 2024), local and sparse spatial fusion (Shi et al., 2026) and graph attention combined with neural controlled differential equations (Mirzaei & Torghabeh, 2026). Hu et al. (2022) represented the same dependence through spatio-temporal input tensors rather than an explicit graph. Each construction is defensible, but they encode different hypotheses about why pollution at one station informs another, and few studies tested the alternatives against each other. Liu et al. (2023) are an exception: by comparing seven combinations of distance, similarity and correlation graphs, they showed that the choice of adjacency changes forecast error appreciably. The most convincing graph design in the corpus is that of Dimri et al. (2026), whose TransNet learns operators for advection, diffusion and reaction, so that edges represent physical transport rather than statistical association. This physics-informed construction, which builds on earlier domain-knowledge graphs (Wang et al., 2020), is more defensible than adjacency matrices chosen for convenience, and it was also one of only two studies to release its code.

***Decomposition Representations***

Decomposition stacks were associated with some of the highest reported accuracies, although, as noted for Table 6, these figures come from different tasks and are not directly comparable. Fang et al. (2026) used complete ensemble empirical mode decomposition with adaptive noise (Torres et al., 2011) to split a daily AQI series into nine intrinsic mode functions and a residual, reaching an $R^2$ of 0.962. Cheng and Zhang (2026) used variational mode decomposition (Dragomiretskiy & Zosso, 2014) and grouped the modes into three frequency bands, each modelled separately. Such gains can be relied on only if the decomposition does not see the future. A decomposition fitted to the whole series before splitting distributes information from the test period into every mode of the training period, so a model trained on those modes is partly trained on the answer. Cheng and Zhang (2026) explicitly fitted the decomposition on training data only, and Fang et al. (2026) restricted outlier correction to the training set, which strengthens their results. Other decomposition-based studies, including the combined point and interval forecasting system of Ding et al. (2022), did not state where the decomposition was fitted. The transfer tests of Fang et al. (2026), in which performance fell for cities with different climates, also show that the number of modes tuned for one site does not carry over automatically to another.

***Continuous-Time, Grid and Matrix Representations***

Continuous-time formulations treat the latent pollution state as a solution of a differential equation, so that irregular and missing observations can be handled without imputing a regular grid (Chen et al., 2018; Kidger et al., 2020). Havaei et al. (2025) separated the latent state of Tehran's network into fast, medium and slow subspaces and reported high short-horizon accuracy, but candidly showed that $R^2$ for PM2.5 fell to about 0.5 by the third forecast step. That candour is valuable, because it contrasts with the single-horizon $R^2$ values above 0.9 reported by several other studies. Saad et al. (2024) took a different route, using a Bayesian neural field over space and time coordinates that scales to hundreds of thousands of observations and returns calibrated uncertainty. Grid representations were less common. Chatterjee et al. (2025) embedded sensors in a regular grid and found that a 16 × 16 grid gave better results than 8 × 8 or 32 × 32 grids, which shows that grid resolution is a hyperparameter with real consequences, not a neutral container. Low-rank matrix completion, used by Liu et al. (2020) to impute five pollutants and shown to outperform expectation-maximisation and spline methods at 30% missingness, rests on the same insight as later graph methods, namely that stations are redundant, but with stronger theoretical guarantees (Candès & Recht, 2009). The review found nothing in the included studies that explains its absence after 2022.

***Multimodal and Spatial-Segment Representations***

Multimodal representations promise richer context by joining station data with satellite retrievals, meteorology and demographic layers. Lilhore et al. (2025) combined ground sensors, NOAA meteorology and Sentinel-5P, MODIS and Landsat products within a hybrid of convolutional, recurrent, graph and neural ODE components, and Rajesh et al. (2025) built a cloud framework that updates risk maps every five minutes. Both illustrate the problem that as the representation grows more heterogeneous, the audit trail grows thinner: alignment of sources, the point at which each transformation is fitted, and the contribution of each modality are difficult to verify from the published account. At the opposite extreme, Baruah et al. (2024) mapped mobile and static sensor readings onto 30 m road segments in Birmingham and used a random forest to fill gaps, reaching an $R^2$ of 0.85 and between 0.65 and 0.71 in transfer tests. Hofman et al. (2022) similarly showed, across four testbeds, that graph-based inference worked well in European cities ($R^2$ up to 0.82) but much less well in the sparser United States deployments. Liu et al. (2022) took a related approach in Taipei, attaching land-use, population and terrain attributes to each location and predicting AQI with a two-stage support vector machine. Baruah et al. (2024) and Hofman et al. (2022) are among the few that tested whether a representation travels, and both found that the density of the underlying network limits what any representation can recover.

**Preprocessing Practices (RQ3)**

Table 9 summarises the preprocessing operations reported in the primary studies. The most important finding is not what was done but what was left unreported.

**Table 9: Preprocessing Operations Reported in the Primary Studies**

| Operation | n (%) | Studies | Critical observation |
|---|---|---|---|
| Explicit missing-data handling | 11 (38%) | Baruah, Chatterjee, Cheng, Dimri, Fang, Havaei, Hofman, Iwaszenko, Lilhore, Liu X, Yin | Methods range from linear interpolation to MLP, KNN, iterative random forest and matrix completion; missingness mechanisms are rarely examined |
| Outlier treatment | 5 (17%) | Chatterjee, Fang, Iwaszenko, Lilhore, Sundaramoorthy | Percentile clipping, z-score, IQR and Mahalanobis rules; thresholds seldom justified |
| Explicit scaling or transformation | 5 (17%) | Chatterjee, Fang, Havaei, Sundaramoorthy, Yin | Min-max and z-score scaling and log transforms; most studies do not say where scalers were fitted |
| Signal decomposition | 5 (17%) | Cheng, Dimri, Ding, Fang, Wang X | High reported accuracy; highest leakage risk if fitted on the full series |

| Operation | n (%) | Studies | Critical observation |
|---|---|---|---|
| Feature selection or weighting | 4 (14%) | Dimri, Lilhore, Sudha, Sundaramoorthy | PCA or metaheuristic weighting; adds tuning complexity |
| Explicit leakage control | 6 (21%) | Chatterjee, Cheng, Dimri, Fahim, Fang, Saad | Train-only fitting, forecast-time covariates or a chronologically held-out test period; should be standard |

*Note:* Counts refer to operations explicitly reported. A study not listed against an operation did not report it; this does not establish that the operation was not performed or, for leakage control, that leakage occurred. Studies are identified by first author.

Only 11 studies described how missing values were handled, even though gaps are pervasive in air pollution records. Among those that did, the methods varied from simple interpolation for short gaps to model-based imputation for long gaps (Chatterjee et al., 2025), iterative random forest and graph-based imputation (Havaei et al., 2025), K-nearest-neighbour imputation (Yin et al., 2026) and low-rank completion (Liu et al., 2020). None of the studies reported a test of whether data were missing at random, although sensor outages during high-pollution episodes or extreme weather would make missingness informative (Little & Rubin, 2019; Buelvas et al., 2023). The choice of imputation method is a structural decision in its own right, because it determines whether cross-station redundancy is exploited or ignored.

Leakage control was explicitly reported in only six studies. In the remaining 23, leakage could not be ruled out from the published account, although this does not show that it occurred. The gap matters because scaling, imputation and decomposition fitted on the full dataset all pass information from the test period into training (Kaufman et al., 2012), a risk that is amplified for decomposition representations and for random cross-validation of autocorrelated series (Bergmeir & Benítez, 2012). Until studies report where each transformation is fitted, very high $R^2$ values should be interpreted cautiously, as possible upper bounds rather than established estimates of operational skill.

**Research Gaps (RQ5)**

The synthesis points to six gaps. Table 10 links each gap to the primary-study evidence on which it rests and separates that evidence from the proposed research direction, which is the authors' recommendation rather than an empirical finding.

**Table 10: Research Gaps and Proposed Research Directions**

| Gap | Evidence from the primary studies (n = 29) | Proposed direction (authors' recommendation) |
|---|---|---|
| Representation choice is rarely justified or ablated | 22 of 29 studies combine representations; only Cheng and Zhang (2026) report an ablation of the representation, and only Chatterjee et al. (2025) and Liu et al. (2023) vary grid size or adjacency within one study | Report ablations that hold the model fixed and vary the representation |
| Ad hoc graph construction | The 10 graph studies use distance, POI-similarity, temporal-correlation, grey-relational, multi-view and transport-informed graphs; only Liu et al. (2023) compare alternatives | Compare adjacency definitions; favour physically informed edges (wind, transport) |
| Information leakage in preprocessing | Leakage control explicitly reported in 6 of 29 studies; could not be ruled out in the other 23 (Table 9) | Fit all transformations on training data; use rolling-origin evaluation |
| Weak reproducibility | Code and data released by 2 of 29 studies (Saad et al., 2024; Dimri et al., 2026) | Release code, data splits and the exact representation-building scripts |
| Benchmark concentration | Six studies use Beijing network data; eleven do not state the study area (Figure 3) | Build multi-city benchmarks with stated geography and density |
| Absence of African and other low-resource settings | No primary study in the selected corpus used data from Africa, South America or Oceania (Figure 3) | Develop and evaluate sparse, irregular and low-cost sensor representations in cities such as Lagos, Kano and Abuja |

*Note.* The evidence column reports findings from the primary studies. The proposed directions are the authors' recommendations derived from the synthesis; they have not been tested empirically in the included studies.

## DISCUSSION

**Representation as the Hidden Variable**

The central argument of this review is that the data representation is a largely hidden variable in air pollution modelling. Studies present new architectures, but the gains they report are often inseparable from changes in representation that the authors treat as routine. A model that receives decomposed modes, a learned station graph and multimodal covariates is not directly comparable with a model that receives a raw window. Where the representation was varied within a single study while other factors were held constant, as in the decomposition ablation of Cheng and Zhang (2026), the grid-size experiment of Chatterjee et al. (2025) and the adjacency comparison of Liu et al. (2023), the effect of the representation on accuracy was substantial. Only three such within-study comparisons were found, so this evidence is suggestive rather than conclusive, and differences in performance between papers cannot be used to make the same point. A field that wishes to know which ideas work needs to separate these effects routinely.

This argument cuts against a common narrative of steady architectural progress. The corpus contains several cases in which simpler representations matched or beat more complex ones: random forests on lag matrices at daily scales (Iwaszenko et al., 2024), zero-shot foundation models against tuned specialists (Fahim & Karim, 2026), and a hybrid sequence model that trained in a third of the time of a graph Transformer (Yin et al., 2026). Such results do not show that complex representations are unnecessary. They show that complexity must be earned by evidence that it adds information the simpler representation cannot capture, and that this evidence is often missing.

**The Credibility of Reported Accuracy**

Very high accuracy figures ($R^2$ above 0.95) were reported by three primary studies (Cheng & Zhang, 2026; Fang et al., 2026; Lilhore et al., 2025), and values above 0.9 by several

more, but the conditions under which such figures were obtained are often not fully reported. Decomposition fitted on the full series, scalers fitted before splitting, resampling within regression pipelines and short-horizon evaluation can each inflate reported accuracy, and for most studies the review could not establish whether any of these occurred. The studies that reported multi-step decay (Havaei et al., 2025), transfer to other cities (Fang et al., 2026; Baruah et al., 2024) or performance in sparse networks (Hofman et al., 2022) all found lower and more variable accuracy, and they are probably a more realistic guide to what current representations can achieve operationally. Journals and reviewers can contribute by asking authors to state where every transformation was fitted and to report performance across horizons and sites, not only at the most favourable one.

**Implications for Data-Scarce Settings**

The absence of African studies from the selected corpus is not simply a gap in coverage. It means that the representations in the included studies were designed and tuned for conditions that differ sharply from those in most Nigerian cities. Dense reference networks of the kind found in Beijing make graph and grid representations attractive, because there are enough nodes to learn from. In settings where monitoring relies on a small number of reference stations supplemented by low-cost sensors with frequent outages, the evidence points to different choices. Low-rank matrix completion (Liu et al., 2020), graph-based inference for sparse sensors (Hofman et al., 2022), continuous-time models that accept irregular sampling (Havaei et al., 2025; Kidger et al., 2020) and road-segment representations that exploit mobile measurements (Baruah et al., 2024) are better aligned with such data. Low-cost sensors require calibration and quality control before they can support any representation (Castell et al., 2017; Buelvas et al., 2023), and satellite products offer a way to fill spatial gaps where ground data are thin. Building open benchmarks from these sources is the precondition for testing whether the representations reviewed here transfer.

**Threats to Validity**

Several limitations of the review should be acknowledged. Construct validity depends on the coding of data representations; because many studies describe their representations only implicitly, some coding required judgement, and a study coded under one family might reasonably have been coded under another. Internal validity is limited by the fact that details for some studies could be extracted only from abstracts or partial records, which is why eleven study areas are recorded as not stated and why quality scores for those studies are provisional. Selection bias may arise from the restriction to English-language publications and to five sources, and four 2026 reports were excluded because their bibliographic records could not be verified, so very recent work may be under-represented. One included study is a preprint. Google Scholar results were screened by relevance ranking to a fixed depth, so relevant records ranked below that depth would have been missed. The eligibility scope was limited to forecasting, imputation and spatial inference, so representations used for other processing tasks are not covered. Quality appraisal involved judgement, and the scores in Supplementary File S3 reflect what was reported, not necessarily what was done. Finally, the heterogeneity of datasets, horizons and metrics precluded meta-analysis, so the synthesis is narrative and the performance figures in Table 6 should not be compared directly.

**Conclusion**

This systematic review synthesised 37 publications on data representations for air-pollution processing and predictive modelling, including 29 primary studies. Nine representation families were identified, with sequential windows and graph representations among the most frequently reported. The evidence indicates that representation design and preprocessing deserve explicit reporting alongside model architecture. Important limitations in the reviewed literature include incomplete descriptions of missing-data treatment, uncertain safeguards against information leakage and restricted access to reproducible code and data. The selected corpus was geographically uneven, with no included African primary study. Future investigations should compare representations under consistent data splits and metrics, document graph and preprocessing choices, and evaluate methods using sparse and irregular monitoring networks. The conclusions remain bounded by the review's eligibility criteria and the heterogeneous quality of the available studies. Based on the findings of this review, the following recommendations are made:

i. Researchers should report the data representation explicitly, including its shape, the definition of nodes, edges, grid cells or fields, and the reasoning behind the choice, and should include ablations that vary the representation while holding the model fixed.
ii. All scaling, imputation, outlier correction and decomposition should be fitted on training data only, and models should be evaluated with chronological or rolling-origin splits across several forecast horizons.
iii. Graph-based studies should compare alternative adjacency definitions and should prefer edges grounded in physical transport, such as wind direction and distance, over correlations computed on the full dataset.
iv. Code, data splits and the scripts that build the data representation should be released with each publication, and journals should encourage this as a condition of publication.
v. Researchers and agencies in Nigeria and the wider region should build open, multi-city air quality benchmarks that combine reference stations, low-cost sensors and satellite products, and should test sparse and irregular representations, such as low-rank matrices and continuous-time models, on these data.
vi. Future reviews should extend the search to non-English sources and grey literature, and should update the taxonomy as new representations, such as foundation-model embeddings, mature.

**REFERENCES**

Abdelmalek, M. M., Mahmoud, H., & Shokry, H. (2025). Prognosis of air quality index and air pollution using machine learning techniques. *Scientific Reports, 15*, Article 25890. https://doi.org/10.1038/s41598-025-11260-y

Agbehadji, I. E., & Obagbuwa, I. C. (2024). Systematic review of machine learning and deep learning techniques for spatiotemporal air quality prediction. *Atmosphere, 15*(11), Article 1352. https://doi.org/10.3390/atmos15111352

Ansari, A. F., Stella, L., Turkmen, C., Zhang, X., Mercado, P., Shen, H., Shchur, O., Rangapuram, S. S., Pineda Arango, S., Kapoor, S., Zschiegner, J., Maddix, D. C., Wang, H., Mahoney, M. W., Torkkola, K., Wilson, A. G., Bohlke-Schneider, M., & Wang, Y. (2024). Chronos: Learning the language of time series. *Transactions on Machine Learning Research*. https://doi.org/10.48550/arXiv.2403.07815

Bai, S., Kolter, J. Z., & Koltun, V. (2018). *An empirical evaluation of generic convolutional and recurrent networks for sequence modeling* (arXiv:1803.01271). arXiv. https://doi.org/10.48550/arXiv.1803.01271

Baruah, A., Bousiotis, D., Damayanti, S., Bigi, A., Ghermandi, G., Ghaffarpasand, O., Harrison, R. M., & Pope, F. D. (2024). A novel spatiotemporal prediction approach to fill air pollution data gaps using mobile sensors, machine learning and citizen science techniques. *npj Climate and Atmospheric Science, 7*, Article 310. https://doi.org/10.1038/s41612-024-00859-z

Bergmeir, C., & Benítez, J. M. (2012). On the use of cross-validation for time series predictor evaluation. *Information Sciences, 191*, 192–213. https://doi.org/10.1016/j.ins.2011.12.028

Buelvas, J., Múnera, D., Tobón V., D. P., Aguirre, J., & Gaviria, N. (2023). Data quality in IoT-based air quality monitoring systems: A systematic mapping study. *Water, Air, & Soil Pollution, 234*, Article 248. https://doi.org/10.1007/s11270-023-06127-9

Candès, E. J., & Recht, B. (2009). Exact matrix completion via convex optimization. *Foundations of Computational Mathematics, 9*(6), 717–772. https://doi.org/10.1007/s10208-009-9045-5

Castell, N., Dauge, F. R., Schneider, P., Vogt, M., Lerner, U., Fishbain, B., Broday, D., & Bartonova, A. (2017). Can commercial low-cost sensor platforms contribute to air quality monitoring and exposure estimates? *Environment International, 99*, 293–302. https://doi.org/10.1016/j.envint.2016.12.007

Chatterjee, K., Bala, B., Khan, M., Chitla, A., Reddy, K. N., Menshawi, A., Prasad, M., Kadurka, R. S., Altaf, M., & Jyothi, K. A. (2025). AiM: Urban air quality forecasting with grid-embedded recurrent MLP model. *Scientific Reports, 15*, Article 42812. https://doi.org/10.1038/s41598-025-27073-y

Chauhan, A., Kumar, G., & Shakya, H. K. (2026). Interpretable spatio-temporal graph neural networks for air quality prediction under real-world multi-station monitoring data. *Scientific Reports*. Advance online publication. https://doi.org/10.1038/s41598-026-66494-1

Chen, R. T. Q., Rubanova, Y., Bettencourt, J., & Duvenaud, D. (2018). Neural ordinary differential equations. *Advances in Neural Information Processing Systems, 31*, 6571–6583.

Cheng, Y., & Zhang, C. (2026). Forecasting of PM2.5 concentration based on variational mode decomposition and deep learning. *Scientific Reports, 16*, Article 25647. https://doi.org/10.1038/s41598-026-56223-z

Cormen, T. H., Leiserson, C. E., Rivest, R. L., & Stein, C. (2022). *Introduction to algorithms* (4th ed.). MIT Press.

Dimri, R., Choi, Y., Singh, D., Park, J., & Khorshidian, N. (2026). TransNet: A transport-informed graph neural network for forecasting PM2.5 concentrations across South Korea. *npj Clean Air, 2*, Article 12. https://doi.org/10.1038/s44407-026-00052-x

Ding, Z., Chen, H., Zhou, L., & Wang, Z. (2022). A forecasting system for deterministic and uncertain prediction of air pollution data. *Expert Systems with Applications, 208*, Article 118123. https://doi.org/10.1016/j.eswa.2022.118123

Dragomiretskiy, K., & Zosso, D. (2014). Variational mode decomposition. *IEEE Transactions on Signal Processing, 62*(3), 531–544. https://doi.org/10.1109/TSP.2013.2288675

Dun, A., Yang, Y., & Lei, F. (2022). Dynamic graph convolution neural network based on spatial-temporal correlation for air quality prediction. *Ecological Informatics, 70*, Article 101736. https://doi.org/10.1016/j.ecoinf.2022.101736

Fahim, M. M. M., & Karim, M. R. (2026). *Cost-aware evaluation of time-series foundation models for urban air-quality and temperature forecasting* [Preprint]. Research Square. https://www.researchsquare.com/article/rs-10515545/v1

Fang, Y., Liu, S., & Su, Z. (2026). Air quality index prediction using a hybrid CEEMDAN-CNN-IGWO-BiGRU-Attention model. *Scientific Reports, 16*, Article 15908. https://doi.org/10.1038/s41598-026-46978-w

Feng, S., Tang, L., Huang, M., & Wu, Y. (2025). Integrating D-S evidence theory and multiple deep learning frameworks for time series prediction of air quality. *Scientific Reports, 15*, Article 5971. https://doi.org/10.1038/s41598-025-87935-3

Gryech, I., Asaad, C., Ghogho, M., & Kobbane, A. (2024). Applications of machine learning & Internet of Things for outdoor air pollution monitoring and prediction: A systematic literature review. *Engineering Applications of Artificial Intelligence, 137*(Part B), Article 109182. https://doi.org/10.1016/j.engappai.2024.109182

Havaei, M. A., Shahhosseini, V., & Maknoon, R. (2025). Continuous-time air pollutant forecasting using multi-timescale attention neural ordinary differential equations (MA-NODE). *Scientific Reports, 15*, Article 44178. https://doi.org/10.1038/s41598-025-27903-z

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation, 9*(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

Hofman, J., Do, T. H., Qin, X., Rodrigo Bonet, E., Philips, W., Deligiannis, N., & La Manna, V. P. (2022). Spatiotemporal air quality inference of low-cost sensor data: Evidence from multiple sensor testbeds. *Environmental Modelling & Software, 149*, Article 105306. https://doi.org/10.1016/j.envsoft.2022.105306

Houdou, A., El Badisy, I., & Khomsi, K. (2024). Interpretable machine learning approaches for forecasting and predicting air pollution: A systematic review. *Aerosol and Air Quality Research, 24*(1), Article 230151. https://doi.org/10.4209/aaqr.230151

Hu, K., Guo, X., Gong, X., Wang, X., Liang, J., & Li, D. (2022). Air quality prediction using spatio-temporal deep learning. *Atmospheric Pollution Research, 13*(10), Article 101543. https://doi.org/10.1016/j.apr.2022.101543

Huang, W., Li, T., Liu, J., Xie, P., Du, S., & Teng, F. (2021). An overview of air quality analysis by big data techniques: Monitoring, forecasting, and traceability. *Information Fusion, 75*, 28–40. https://doi.org/10.1016/j.inffus.2021.03.010

Iwaszenko, S., Smolinski, A., Grzanka, M., & Skowronek, T. (2024). Airborne particulate matter measurement and prediction with machine learning techniques. *Scientific Reports, 14*, Article 18999. https://doi.org/10.1038/s41598-024-70152-9

Kaufman, S., Rosset, S., Perlich, C., & Stitelman, O. (2012). Leakage in data mining: Formulation, detection, and avoidance. *ACM Transactions on Knowledge Discovery from Data, 6*(4), Article 15. https://doi.org/10.1145/2382577.2382579

Kaur, M., Singh, D., Jabarulla, M. Y., Kumar, V., Kang, J., & Lee, H.-N. (2023). Computational deep air quality prediction techniques: A systematic review. *Artificial Intelligence Review, 56*(Suppl. 2), 2053–2098. https://doi.org/10.1007/s10462-023-10570-9

Kidger, P., Morrill, J., Foster, J., & Lyons, T. (2020). Neural controlled differential equations for irregular time series. *Advances in Neural Information Processing Systems, 33*, 6696–6707.

Kipf, T. N., & Welling, M. (2017). Semi-supervised classification with graph convolutional networks. In *Proceedings of the 5th International Conference on Learning Representations*. https://doi.org/10.48550/arXiv.1609.02907

Kitchenham, B., & Charters, S. (2007). *Guidelines for performing systematic literature reviews in software engineering* (EBSE Technical Report EBSE-2007-01). Keele University and Durham University.

Kolda, T. G., & Bader, B. W. (2009). Tensor decompositions and applications. *SIAM Review, 51*(3), 455–500. https://doi.org/10.1137/07070111X

Li, Y., Yu, R., Shahabi, C., & Liu, Y. (2018). Diffusion convolutional recurrent neural network: Data-driven traffic forecasting. In *Proceedings of the 6th International Conference on Learning Representations*. https://doi.org/10.48550/arXiv.1707.01926

Liang, Y., Xia, Y., Ke, S., Wang, Y., Wen, Q., Zhang, J., Zheng, Y., & Zimmermann, R. (2023). AirFormer: Predicting nationwide air quality in China with transformers. *Proceedings of the AAAI Conference on Artificial Intelligence, 37*(12), 14329–14337. https://doi.org/10.1609/aaai.v37i12.26676

Lilhore, U. K., Simaiya, S., Singh, R. K., Baqasah, A. M., Alroobaea, R., Alsafyani, M., Alhazmi, A., & Khan, M. D. M. (2025). Advanced air quality prediction using multimodal data and dynamic modeling techniques. *Scientific Reports, 15*, Article 27867. https://doi.org/10.1038/s41598-025-11039-1

Little, R. J. A., & Rubin, D. B. (2019). *Statistical analysis with missing data* (3rd ed.). Wiley. https://doi.org/10.1002/9781119482260

Liu, C.-C., Lin, T.-C., Yuan, K.-Y., & Chiueh, P.-T. (2022). Spatio-temporal prediction and factor identification of urban air quality using support vector machine. *Urban Climate, 41*, Article 101055. https://doi.org/10.1016/j.uclim.2021.101055

Liu, H., Han, Q., Sun, H., Sheng, J., & Yang, Z. (2023). Spatiotemporal adaptive attention graph convolution network for city-level air quality prediction. *Scientific Reports, 13*, Article 13335. https://doi.org/10.1038/s41598-023-39286-0

Liu, X., Wang, X., Zou, L., Xia, J., & Pang, W. (2020). Spatial imputation for air pollutants data sets via low rank matrix completion algorithm. *Environment International, 139*, Article 105713. https://doi.org/10.1016/j.envint.2020.105713

Ma, Z., Wang, B., Luo, W., Jiang, J., Liu, D., Wei, H., & Luo, H. (2024). Air pollutant prediction model based on transfer learning two-stage attention mechanism. *Scientific Reports, 14*, Article 7385. https://doi.org/10.1038/s41598-024-57784-7

Méndez, M., Merayo, M. G., & Núñez, M. (2023). Machine learning algorithms to forecast air quality: A survey. *Artificial Intelligence Review, 56*(9), 10031–10066. https://doi.org/10.1007/s10462-023-10424-4

Mirzaei, S., & Torghabeh, E. A. (2026). Neural CDE-GAT framework for spatio-temporal air pollution prediction. *Scientific Reports*. Advance online publication. https://doi.org/10.1038/s41598-026-60629-0

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., . . . Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ, 372*, Article n71. https://doi.org/10.1136/bmj.n71

Rajesh, M., Babu, R. G., Moorthy, U., & Easwaramoorthy, S. V. (2025). Machine learning-driven framework for realtime air quality assessment and predictive environmental health risk mapping. *Scientific Reports, 15*, Article 28801. https://doi.org/10.1038/s41598-025-14214-6

Saad, F., Burnim, J., Carroll, C., Patton, B., Köster, U., Saurous, R. A., & Hoffman, M. (2024). Scalable spatiotemporal prediction with Bayesian neural fields. *Nature Communications, 15*, Article 7942. https://doi.org/10.1038/s41467-024-51477-5

Shi, G., Yu, J., Li, S., Zhang, K., Xue, D., Li, P., Xiao, M., & Yang, S. (2026). Dynamic spatiotemporal air quality modeling with local and sparse spatial fusion and structured temporal modeling. *Information Sciences, 750*, Article 123560. https://doi.org/10.1016/j.ins.2026.123560

Sudha, R., Damodaran, A., & Manohar, G. (2025). Enhanced air quality prediction using adaptive residual Bi-LSTM with pyramid dilation and optimal weighted feature selection. *Scientific Reports, 15*, Article 30428. https://doi.org/10.1038/s41598-025-14668-8

Sui, S., & Han, Q. (2023). Multi-view multi-task spatiotemporal graph convolutional network for air quality

prediction. *Science of the Total Environment, 893*, Article 164699. https://doi.org/10.1016/j.scitotenv.2023.164699

Sundaramoorthy, R. A., Ananth, A. D., Seerangan, K., Nandagopal, M., Balusamy, B., & Selvarajan, S. (2024). Implementing heuristic-based multiscale depth-wise separable adaptive temporal convolutional network for ambient air quality prediction using real time data. *Scientific Reports, 14*, Article 18437. https://doi.org/10.1038/s41598-024-68793-x

Torres, M. E., Colominas, M. A., Schlotthauer, G., & Flandrin, P. (2011). A complete ensemble empirical mode decomposition with adaptive noise. In *2011 IEEE International Conference on Acoustics, Speech and Signal Processing* (pp. 4144–4147). IEEE. https://doi.org/10.1109/ICASSP.2011.5947265

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems, 30*, 5998–6008.

Veličković, P., Cucurull, G., Casanova, A., Romero, A., Liò, P., & Bengio, Y. (2018). Graph attention networks. In *Proceedings of the 6th International Conference on Learning Representations*. https://doi.org/10.48550/arXiv.1710.10903

Wang, S., Li, Y., Zhang, J., Meng, Q., Meng, L., & Gao, F. (2020). PM2.5-GNN: A domain knowledge enhanced graph neural network for PM2.5 forecasting. In *Proceedings of the 28th International Conference on Advances in Geographic Information Systems* (pp. 163–166). ACM. https://doi.org/10.1145/3397536.3422208

Wang, X., Zhang, S., Chen, Y., He, L., Ren, Y., Zhang, Z., Li, J., & Zhang, S. (2024). Air quality forecasting using a spatiotemporal hybrid deep learning model based on VMD-GAT-BiLSTM. *Scientific Reports, 14*, Article 17841. https://doi.org/10.1038/s41598-024-68874-x

Wohlin, C. (2014). Guidelines for snowballing in systematic literature studies and a replication in software engineering. In *Proceedings of the 18th International Conference on Evaluation and Assessment in Software Engineering* (Article 38). ACM. https://doi.org/10.1145/2601248.2601268

World Health Organization. (2021). *WHO global air quality guidelines: Particulate matter (PM2.5 and PM10), ozone, nitrogen dioxide, sulfur dioxide and carbon monoxide*. https://iris.who.int/handle/10665/345329

Wu, Z., Pan, S., Long, G., Jiang, J., & Zhang, C. (2019). Graph WaveNet for deep spatial-temporal graph modeling. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence* (pp. 1907–1913). https://doi.org/10.24963/ijcai.2019/264

Yin, C., Li, W., Li, T., Yuan, S., Tang, D., & Dong, X. (2026). Air quality prediction model based on deep learning hybrid framework. *Scientific Reports, 16*, Article 7084. https://doi.org/10.1038/s41598-026-37896-y

Yu, B., Yin, H., & Zhu, Z. (2018). Spatio-temporal graph convolutional networks: A deep learning framework for traffic forecasting. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence* (pp. 3634–3640). https://doi.org/10.24963/ijcai.2018/505

Zhang, S., Guo, B., Dong, A., He, J., Xu, Z., & Chen, S. X. (2017). Cautionary tales on air-quality improvement in Beijing. *Proceedings of the Royal Society A, 473*(2205), Article 20170457. https://doi.org/10.1098/rspa.2017.0457

Zhang, Z., Zhang, S., Chen, C., & Yuan, J. (2024). A systematic survey of air quality prediction based on deep learning. *Alexandria Engineering Journal, 93*, 128–141. https://doi.org/10.1016/j.aej.2024.03.031

Zheng, Y., Yi, X., Li, M., Li, R., Shan, Z., Chang, E., & Li, T. (2015). Forecasting fine-grained air quality based on big data. In *Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 2267–2276). ACM. https://doi.org/10.1145/2783258.2788573

©2026 This is an Open Access article distributed under the terms of the Creative Commons Attribution 4.0 International license viewed via https://creativecommons.org/licenses/by/4.0/ which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is cited appropriately.