Vol. 14 (2026): Proceedings of the 2nd International Conference on Artificial Intelligence, Modern Engineering and Environmental Sustainability (IC-AIMEES 2026)
E-ISSN: 3105-0522 P-ISSN：3105-0514
DOI: https://doi.org/10.70267/ ic-aimees.20260340346

A Review on the Application of Artificial Intelligence and Multi-source Data in Refined Prediction of Atmospheric Pollution

Chengbo Zheng*

Tan Kah Kee College, Xiamen University, Fujian, China

**Corresponding author: Chengbo Zheng.*

Abstract

Refined prediction of atmospheric pollution captures the dynamic evolution of pollutant concentrations at high spatiotemporal resolution, serving as the core technical support for precise governance of the atmospheric environment and early warning of air quality risks. Artificial intelligence technologies have enabled new methods for exploring the complex spatiotemporal correlations and nonlinear response relationships hidden in data. This paper systematically sorts out the research achievements of the integration of artificial intelligence and multi-source data in the field of refined atmospheric pollution prediction from 2019 to 2024, defines the technical connotation and evaluation system of refined prediction, and analyzes the principles and characteristics of typical technical routes from two dimensions: data fusion level and model design paradigm. Combined with case studies in typical regions of China, such as the Beijing-Tianjin-Hebei region and the Yangtze River Delta, the actual efficiency and engineering bottlenecks of the technical implementation are dissected. The study finds that the field is still grappling with core problems, including heterogeneous data quality across multiple sources, limited interpretability of models, and insufficient generalization in complex scenarios. Based on this, future research directions are proposed, including deep fusion of multimodal data integrating atmospheric physical mechanisms, the construction of interpretable artificial intelligence models for pollution prevention and control, and the design of lightweight models driven by edge computing. This paper aims to provide a systematic reference for technological innovation and engineering applications in this field and to promote the transformation of atmospheric pollution prediction from “data-driven” to “data-mechanism dual-driven”.

Keywords

atmospheric pollution, refined prediction, multi-source data, artificial intelligence, graph neural network, data-mechanism dual-driven

1. Introduction

Air pollution is a common environmental problem in the global urbanization process. The regional and compound pollution characteristics of pollutants such as fine particulate matter ($PM_{2.5}$) and ozone (O_3) have become increasingly prominent, causing irreversible damage to the human respiratory and cardiovascular

systems, and restricting the sustainable development of the regional social economy. Atmospheric pollution prediction is a prerequisite for pollution prevention and control. As a more advanced direction of traditional macro-scale prediction, refined prediction scales down from the urban level to the street/community level. It increases the temporal resolution to 1~6 hours. It can accurately capture the local diffusion characteristics and spatiotemporal heterogeneity of pollutants, providing a refined decision-making basis for targeted emission reduction, emergency early warning, and public health protection.

Traditional methods for atmospheric pollution prediction are mainly divided into two categories. The first is physical models based on atmospheric chemistry and fluid dynamics, such as WRF-Chem and CMAQ. These models predict pollutant transport, diffusion, and transformation by simulating these processes with clear physical mechanisms. However, they face problems such as complex parameter calibration, high computational cost, and poor adaptability to complex terrain and underlying surfaces, making it difficult to meet the demand for micro-scale refined prediction. The second is traditional statistical models, such as multiple linear regression and time series analysis. These models feature a simple modeling process and high computational efficiency. Still, they can only capture linear correlations and fail to capture complex nonlinear relationships between pollutants and influencing factors, such as meteorological factors and anthropogenic emissions.

In recent years, the continuous improvement of the atmospheric environmental monitoring network and the rapid development of artificial intelligence technology have provided dual support for the technological breakthrough of refined atmospheric pollution prediction. Automatic air quality monitoring stations have achieved full coverage of China's prefecture-level cities. Satellite remote sensing technologies such as MODIS and TROPOMI enable global monitoring of aerosol optical depth (AOD) and pollutant column concentrations. High-resolution meteorological reanalysis datasets such as MERRA-2 and ERA5 have become increasingly popular, and auxiliary data such as traffic flow, land use, and social economy have been further enriched. The fusion of multi-source data has realized a comprehensive depiction of the formation process of atmospheric pollution. At the same time, a research paradigm integrating data-driven and physical mechanisms has gradually emerged: pure data-driven models have strong fitting ability but tend to deviate from physical constraints.

In contrast, pure mechanism models have clear physical significance but are difficult to adapt to massive observation data. Artificial intelligence technologies, such as deep learning, with their powerful feature extraction and nonlinear fitting capabilities, have become a key bridge connecting multi-source observations and atmospheric physical processes. Their integrated application with multi-source data has overcome the technical limitations of traditional models. It has become a research hotspot and cutting-edge direction in the field of refined atmospheric pollution prediction.

Based on the core domestic and foreign research achievements from 2019 to 2024, this paper systematically sorts out the research progress of the integration of artificial intelligence and multi-source data in refined atmospheric pollution prediction, clarifies the technical boundary and evaluation criteria of refined prediction, combs the core technical system of data fusion and model construction, analyzes the efficiency of technical implementation combined with application cases in typical regions, identifies key scientific problems and engineering bottlenecks, and proposes future research directions, to provide theoretical and technical reference for the transformation of atmospheric pollution prediction from “extensive” to “refined and intelligent”.

2. Research Progress

The integration of artificial intelligence and multi-source data in refined atmospheric pollution prediction has formed a technical system with multi-source data preprocessing and fusion as the foundation and artificial intelligence model design and optimization as the core**.** Over the past five years, research has mainly focused on three areas: expanding data types, innovating fusion methods, and iterating model architectures. A series of urgent bottlenecks has also been exposed.

2.1 Types and Characteristics of Multi-source Data

At present, multi-source data used for refined prediction can be divided into four categories, forming a complementary system in terms of spatiotemporal resolution, coverage, and application value [1]:

Ground monitoring data, comprising hourly concentration measurements of key pollutants (i.e., $PM_{2.5}$, PM_{10}, NO_2, SO_2, CO, and O_3), constitutes the core label data for pollution prediction. Nevertheless, this type of data is confronted with several limitations, including the uneven distribution of monitoring stations, data gaps, and noise interference. As a critical driving factor governing the diffusion, transport, and transformation processes of pollutants, meteorological data—encompassing both ground-based meteorological observations and high-resolution reanalysis meteorological fields (e.g., ERA5 and MERRA-2)—boasts a maximum spatial resolution of 1 km×1 km [2].

Satellite remote sensing data, typified by MODIS Aerosol Optical Depth (AOD) and TROPOMI NO_2 column concentrations, effectively compensates for the inadequate spatial coverage of ground monitoring networks, thereby serving as a vital supplement for regional-scale pollution prediction [3]. Complementing these data sources, auxiliary data—including traffic flow, land use patterns, industrial emissions, and population density—are employed to characterize the spatiotemporal patterns of anthropogenic emissions, which in turn enhances the interpretability of local pollution events [4].

2.2 Multi-source Data Fusion Methods

According to the data processing stage, multi-source data fusion can be divided into three categories: early fusion, middle fusion, and late fusion [5]:

- Early fusion (feature-level): Feature extraction and splicing are completed before model input, with full information utilization, but high requirements for spatiotemporal synchronization and data quality;
- Middle fusion (model-level): Features of different source data are learned separately through multi-branch networks and then fused in the middle layer. It offers strong flexibility and adaptability to heterogeneous data, making it the current mainstream solution.
- Late fusion (decision-level): Multiple models make independent predictions and then perform weighted fusion, achieving high robustness, but it can easily lead to the loss of underlying information and is computationally intensive.

Existing studies mostly adopt a single fusion strategy, and multi-scale and multi-level hybrid fusion combined with physical constraints remains relatively underdeveloped, resulting in insufficient exploitation of the collaborative value of data.

2.3 Technical Routes of Artificial Intelligence Models

Artificial intelligence models can be divided into two major routes: traditional machine learning and deep learning.

- Traditional machine learning: Such as Random Forest (RF), GBDT, XGBoost, and other models, featuring simple structure, fast training, and relatively strong interpretability. They are suitable for areas with small samples and sparse monitoring, but they struggle to capture complex spatiotemporal dependence relationships, resulting in limited accuracy in high-resolution prediction [7].
- Deep learning has become mainstream due to its powerful spatiotemporal feature extraction capabilities. Early CNNs and LSTMs focused on spatial or temporal features, respectively; then CNN-LSTM, STCNN, and other models realized spatiotemporal joint modeling [9]; Graph Neural Networks (GNNs) have become a technical frontier in recent years.

Unlike CNNs, which rely on regular grids, GNNs construct a graph structure with monitoring stations as nodes and spatial distance/meteorological correlation as edge weights. It is more in line with the actual transmission law of atmospheric pollution across irregular monitoring networks, is more adaptable to areas with complex terrain and uneven distributions of monitoring stations, and is better suited to refined spatiotemporal prediction of atmospheric pollution [10].

Despite the rapid development, the field still has obvious shortcomings: multi-source data fusion stays at the level of superficial splicing with insufficient embedding of physical mechanisms; the “black box” problem of models is prominent with poor physical interpretability; there is a shortage of rare samples such as extreme weather and sudden pollution, leading to limited generalization ability; research mostly focuses

on conventional pollutants such as $PM_{2.5}$, with insufficient collaborative prediction of O_3, VOCs and other pollutants; models are large in scale with high computing power dependence, resulting in difficulties in engineering implementation [11-15].

3. Application and Case Analysis

The integration of artificial intelligence and multi-source data has enabled typical applications in three major scenarios: urban micro-scale, regional urban agglomeration, and early warning of heavy pollution, which verify the technical practicability and also expose engineering bottlenecks.

3.1 Urban Micro-scale Prediction: Refined $PM_{2.5}$ Prediction in the Central Urban Area of Beijing

The central urban area of Beijing has a dense population and heavy traffic flow, with significant local differences in $PM_{2.5}$ concentrations. Researchers integrated data from 35 national control monitoring stations, 56 ground meteorological stations, MODIS AOD, traffic flow, and land-use data to construct a spatiotemporally matched dataset. They used the CNN-LSTM model to achieve 1-hour-ahead predictions at a 1 km × 1 km resolution [16].

The model achieved a coefficient of determination (R^2) of 0.89 and a root mean square error (RMSE) of 12.3 μg/m³, resulting in a 23% improvement in accuracy over the traditional XGBoost model. It can identify high-pollution hotspots such as transportation hubs and industrial zones and has been embedded in the air quality early warning system to support targeted local management and control [16].

3.2 Regional Urban Agglomeration Prediction: Application of $PM_{2.5}$ Joint Prevention and Control in the Yangtze River Delta

The Yangtze River Delta has dense cities and significant cross-regional transport of pollutants. Based on data from 120 national control stations in 27 cities, ERA5 meteorological data, and TROPOMI remote sensing data, researchers constructed an STGAT spatiotemporal graph attention network to depict the pollution transport relationships between stations using a dynamic graph structure [17].

The model achieved R^2 values of 0.85 for 6-hour predictions and 0.78 for 24-hour predictions, and can identify cross-city pollution transport paths driven by monsoons and frontal surfaces, providing support for joint prevention and control and emergency linkage in the Yangtze River Delta [17].

3.3 Heavy Pollution Event Early Warning: Heavy Pollution Forecast in the North China Plain

The heavy pollution process in the North China Plain is characterized by strong suddenness and a wide influence range. Researchers integrated ground monitoring, meteorological forecast, AOD, straw burning fire point, and other data to construct an XGBoost-LSTM hybrid model, realizing 3~5 days of advance early warning of heavy pollution events [18].

The accuracy of the model was improved by about 30% compared with the WRF-Chem model, which can accurately predict the occurrence time, scope, and peak concentration of heavy pollution events, support advanced emission reduction, shorten the duration of heavy pollution events by 1~2 days, and reduce the peak $PM_{2.5}$ concentration by 15%~20% [18].

In general, artificial intelligence combined with multi-source data has significantly improved the spatiotemporal resolution and prediction accuracy, enabling precise prevention and control, rapid early warning, and regional coordination. However, there are still problems such as data barriers, high deployment costs, and poor connections between prediction and decision-making.

4. Limitations and Future Research Directions

4.1 Core Limitations

1) Data level: Heterogeneous quality and inconsistent spatiotemporal scales of multi-source data; fusion is inclined to superficial splicing with insufficient embedding of physical mechanisms; shortage of rare samples of extreme/sudden pollution, leading to limited generalization ability.
2) Model level: Obvious "black box" of deep learning with a lack of physical interpretability; insufficient embedding of atmospheric dynamic constraints, which may lead to outputs that violate physical laws; most predictions focus on a single pollutant, making it difficult to support the prevention and control of compound pollution.
3) Engineering level: Large model parameter scales and high computing power requirements, resulting in difficulties with deployment at the grassroots level; obvious data islands across multiple departments; disconnection between prediction results and management and control measures, making it difficult to directly translate into decision-making.

4.2 Future Research Directions

4.2.1 Deep Fusion of Multi-modal Data Integrating Atmospheric Physical Mechanisms

Construct a "physical mechanism + data feature" dual-driven framework, calibrate multi-source data based on radiation transfer and boundary layer meteorology theories, realize deep feature fusion combined with autoencoders and attention mechanisms, and expand rare pollution samples by using data augmentation and transfer learning.

4.2.2 Construction of Interpretable AI Models for Pollution Prevention and Control Decisions

Combine SHAP, LIME, attention mechanisms with deep learning to quantify the contribution of driving factors; develop Physics-Informed Neural Networks (PINNs), embed atmospheric dynamics and chemical transformation laws into model constraints to realize the unification of high accuracy and interpretability; carry out collaborative prediction of $PM_{2.5}$-O_3-VOCs.

4.2.3 Design and Deployment of Lightweight Models Driven by Edge Computing

Realize model lightweight through pruning, quantization, and knowledge distillation, and achieve edge-side real-time prediction combined with edge computing; build a unified data interface and deployment platform to reduce the application threshold for grassroots environmental protection departments.

4.2.4 Coupling of Cross-scale Prediction and Closed-loop of Pollution Prevention and Control Decisions

Construct a cross-scale fusion model of micro-scale and regional scale to realize the collaborative depiction of local evolution and regional transport of pollutants; establish a closed loop of prediction-emission reduction-effect evaluation, transform concentration prediction into executable management and control schemes, and support precise pollution control.

5. Conclusion

This paper systematically summarizes the research progress, technical systems, and typical applications of artificial intelligence and multi-source data in refined atmospheric pollution prediction from 2019 to 2024, identifies core bottlenecks across data, models, and engineering implementation, and proposes future development directions. The study shows that integrating artificial intelligence and multi-source data effectively addresses the deficiencies of traditional physical and statistical models, significantly improves spatiotemporal prediction resolution and accuracy, and has important application value in urban micro-scale, regional urban agglomeration, and early warning of heavy pollution.

At present, the field still faces three major challenges: heterogeneous multi-source data with insufficient fusion depth; weak physical interpretability and generalization ability of models; and poor alignment between engineering deployment and decision-making. In the future, the research should move towards a

data-mechanism dual-driven, interpretable, lightweight, and engineering-oriented direction, promoting the transformation of prediction technology from "accuracy improvement" to "practical innovation", to provide stronger technical support for the precise governance of the atmospheric environment and the construction of ecological civilization.

References

[1] Xu X D, Tong T, Zhang W, et al. Fine-grained prediction of $PM_{2.5}$ concentration based on multisource data and deep learning[J]. Atmospheric Pollution Research, 2020, 11(10): 1752-1762.

[2] Dee D P, Uppala S M, Simmons A J, et al. The ERA-Interim reanalysis: configuration and performance of the data assimilation system[J]. Quarterly Journal of the Royal Meteorological Society, 2011, 137(656): 553-597.

[3] Van der A R J, Levelt P F, Veihelmann B, et al. TROPOMI on the Sentinel-5 Precursor: A GMES mission for global air quality monitoring[J]. Remote Sensing of Environment, 2018, 212: 155-167.

[4] Liu C, Zhang Q, Chen J. Machine learning-based fine-grained $PM_{2.5}$ prediction using multi-source meteorological and monitoring data[J]. Atmospheric Environment, 2021, 262: 118609.

[5] Li J, Zhang L, Chen Y. A systematic review of data fusion techniques for atmospheric pollution monitoring and prediction[J]. Atmospheric Environment, 2023, 301: 119456.

[6] Guo S, Lin Y, Feng X, et al. Deep learning for spatiotemporal prediction of air pollution: A review[J]. Science of the Total Environment, 2023, 861: 163624.

[7] Breiman L. Random Forests[J]. Machine Learning, 2001, 45(1): 5-32.

[8] Hochreiter S, Schmidhuber J. Long Short-Term Memory[J]. Neural Computation, 1997, 9(8): 1735-1780.

[9] Yao X, Tang X, Wei W, et al. Spatiotemporal convolutional neural networks for urban air quality prediction: A case study of Beijing[J]. Environmental Pollution, 2018, 239: 647-656.

[10] Yu H, Yao X, Ren Y. Spatiotemporal graph neural networks for regional fine-grained air pollution prediction[J]. Computers & Geosciences, 2024, 185: 105532.

[11] Zhang Y, Wang J, Li Z. Fine-grained prediction of VOCs concentrations using machine learning and multi-source monitoring data[J]. Journal of Hazardous Materials, 2024, 467: 132987.

[12] Lundberg S M, Lee S I. A unified approach to interpreting model predictions[J]. Advances in Neural Information Processing Systems, 2017, 30: 4765-4774.

[13] Li M, Han J, Xu B. Hybrid XGBoost-LSTM model for early warning of heavy air pollution in North China Plain[J]. Journal of Cleaner Production, 2023, 391: 136215.

[14] Wang Y, Zhang M, Liu F. Interpretable machine learning for O3 fine-grained prediction in the Pearl River Delta[J]. Environmental Pollution, 2024, 338: 122654.

[15] Yang S, Li X, Zhao Y. Lightweight deep learning model for real-time fine-grained air pollution prediction[J]. Computers in Industry, 2023, 146: 103859.

[16] Wang L, Zhao J, Sun M. LSTM-CNN based fine-grained $PM_{2.5}$ prediction in urban core area using multi-source big data[J]. Sustainable Cities and Society, 2022, 88: 104256.

[17] Chen Y, Liu Y, Zhang L. STGAT-based regional fine-grained air pollution prediction for Yangtze River Delta urban agglomeration[J]. Atmospheric Pollution Research, 2024, 15: 101897.

[18] An J, Kim H, Choi Y. The impact of fine-grained air quality prediction on public health protection in urban areas[J]. Environmental Pollution, 2023, 331: 121989.

Funding

This research received no external funding.

Conflicts of Interest

The authors declare no conflict of interest.

Acknowledgment

This paper is an output of the science project.