Proceedings of CONF-SPML 2026 Symposium: The 2nd Neural Computing and Applications Workshop 2025
DOI: 10.54254/2755-2721/2026.TJ28915

Big Data and Environmental Prediction: Moving from Stations to Satellites (2015–2025)

Jingzhi Zhang

Dulwich International High School Suzhou, Suzhou, China
ingrid.zhang26@stu.dulwich.org

Abstract. Among the most significant issues societies are air pollution and environmental change. They impact natural ecosystems and agriculture in addition to human health. Previous studies frequently relied on data from a small number of nearby stations, which constrained the scope of the analysis. Recent investigations integrate expansive datasets derived from satellite observations, meteorological sources, and monitoring infrastructures. Furthermore, the application of big data analytics has been instrumental in diverse environmental domains, spanning ecosystem conservation, climatological research, and water quality assessment, with a specific emphasis on air quality forecasting and allied ecological inquiries documented between 2015 and 2025. Linear models, tree-based models, deep learning techniques, and more sophisticated statistical tools are the primary categories of methods. Applications include monitoring biodiversity, predicting air pollution, and assessing climate risk. Deep learning is helpful when spatial and temporal patterns are complex, statistical methods assist in displaying degrees of certainty or uncertainty, and tree models are frequently dependable baselines. A lack of interpretability and inadequate testing techniques are obstacles. Future research should aim to strengthen validation, provide a explanation of uncertainty, and link data-driven methodologies to accepted scientific principles.

Keywords: Air quality, Big data, Climate change, Ecosystem, Prediction

1. Introduction

Big data has taken center stage in environmental research in the last ten years. Larger datasets study topics like biodiversity, climate change, water resources, and air quality. Studies in the past frequently depended on isolated sites or small sample sizes. Currently, satellites provide extensive coverage, monitor networks deliver hourly updates, and models generate substantial datasets, necessitating innovative analytical techniques. The prevailing understanding of big data is commonly articulated through three key dimensions: volume, velocity, and variety. This means that environmental work involves a lot of data, including chemistry, physics, and biology, as well as frequent updates from sensors and satellites. Although these characteristics increase complexity, they also make predictions more accurate.

Prior research on air quality prediction primarily utilized linear models. Although these models were effective in identifying seasonal patterns, they frequently demonstrated limitations when

DOI: 10.54254/2755-2721/2026.TJ28915

applied to situations characterized by long-range transport phenomena or abrupt changes in meteorological conditions. Since 2015, subsequent investigations have focused on integrating satellite data, meteorological variables, and monitoring networks. Despite improvements in predictive accuracy, these advancements have also introduced new challenges [1].

Big data's significance extends beyond air pollution, significantly impacting climate studies. These studies leverage extensive datasets derived from satellite observations, encompassing land use patterns, precipitation levels, and global temperature fluctuations. This data is crucial for investigating the multifaceted effects of climate change on vital resources such as agricultural yields, water availability, and various human activities. Furthermore, satellite-based monitoring plays a critical role in conservation efforts and biodiversity management by providing comprehensive data on wetlands and diverse species within ecosystems.

Accuracy is paramount, and the implementation of findings is crucial; however, gaps persist. Many investigations examine a singular domain, such as air quality, in isolation, failing to integrate its relationship with water resources or biodiversity. Uncertainty quantification and validation are not consistently apparent [2,3]. Furthermore, integrating heterogeneous datasets introduces significant methodological hurdles. Contemporary scholars are actively exploring the implications for policy and societal impact. A central theme in current investigations revolves around the synergy between technological tools and practical applications.

2. Review methodology

This review included research published between 2015 and 2025 in peer-reviewed journals. The databases checked were PMC, Nature/Scientific Reports, MDPI, Frontiers, AAQR, and SpringerOpen. Keywords combined terms such as “air quality,” “forecast,” “machine learning,” and “deep learning.”

The inclusion criteria comprised studies that (1) predicted next-day PM2.5 or other forecasting applications, (2) presented findings using metrics such as RMSE, MAE, or R2, and (3) detailed the testing methodology. Studies were excluded if they were purely theoretical, focused solely on missing data, or originated from non-peer-reviewed sources. Initially, a search yielded 32 articles. After removing duplicates, 31 articles remained. Following title and abstract screening, 13 articles were excluded, and a further six were removed after full-text review. Ultimately, 12 studies met the inclusion criteria. Each article was assessed for data sources, potential confounders, experimental methods, performance metrics, and the handling of uncertainty or interpretability. The primary constraint of this study is its reliance on English-language sources and specific databases, which may have excluded certain regional studies. Furthermore, the exclusion of conference papers, despite their potential for innovative contributions, limits the scope of the analysis. The heterogeneity in methodologies and temporal frameworks across studies presents challenges for comparative analysis. However, the focus on peer-reviewed journal articles provides a robust foundation for the synthesis.

3. Applications and methods

3.1. Air quality prediction

While linear models offer interpretability and serve as strong baselines, they struggle to capture intricate relationships within the data. Tree-based models, such as Random Forest and XGBoost, often demonstrate superior performance, especially when dealing with datasets of limited size [1,4].

DOI: 10.54254/2755-2721/2026.TJ28915

Deep learning methodologies should be employed to identify spatial and temporal patterns, and the results must be validated using time-series cross-validation [5,6]. Advanced statistical methodologies can also elucidate the confidence intervals associated with findings, integrating both temporal and spatial dimensions [7,3]. The AirNet system, a web-based platform, exemplifies the application of real-time forecasting and alerting methodologies [8]. Furthermore, certain investigations employ hybrid methodologies that integrate machine learning with physics-based models. These studies aim to incorporate both data-driven learning and established scientific principles [9, 10].These methods reflect an effort to integrate classical scientific principles with modern computational techniques [11].

3.2. Water quality monitoring

Water quality is also analyzed using big data techniques. Satellites provide information about suspended matter and chlorophyll, while sensors measure turbidity and chemical levels. These promote agriculture and guarantee cleaner drinking water. Satellites and ground sensors have been used in some projects to identify algal blooms, forecast eutrophication, and direct early responses. Machine learning applications have been implemented in limnological studies to forecast seasonal algal blooms, enabling proactive intervention by local management. In fluvial systems, analogous methodologies are utilized to pinpoint pollution hotspots, thereby informing the strategic allocation of remediation resources by regulatory bodies.

3.3. Climate change studies

Large-scale datasets are of paramount importance in climate change research. Satellite-derived observations, encompassing temperature, precipitation patterns, and land cover characteristics, furnish extensive data essential for the analysis of long-term trends and the identification of extreme events. These datasets are also critical in pinpointing areas vulnerable to drought, evaluating the efficacy of adaptation strategies, and assessing risks to crop yields when integrated with sophisticated climate models. For instance, food security forecasts in Africa have demonstrated a strong correlation with rainfall patterns and vegetation indices. Urban studies have further illuminated the urban heat island effect, where built environments experience elevated temperatures due to factors such as building density and reduced vegetation cover. By integrating detailed urban maps with satellite thermal imagery, researchers can effectively evaluate the impact of urban planning decisions on local climate dynamics. These examples highlight the indispensable utility of comprehensive datasets in connecting global climate assessments with targeted, city-specific interventions.

3.4. Ecosystem management

Large datasets offer significant advantages in ecosystem research. Satellite-based monitoring of wetlands, coupled with predictive modeling, enables the forecasting of potential habitats for plant and animal species. This capability is crucial for effective natural resource management and biodiversity conservation planning. Furthermore, conservation organizations are increasingly leveraging machine learning techniques to detect and address illegal land use and deforestation activities. A prime example is the near-real-time tracking of tropical forests, which provides timely alerts to authorities regarding logging activities, facilitating prompt intervention. In the realm of marine studies, satellites integrated with ocean models are employed to monitor fluctuations in fish

DOI: 10.54254/2755-2721/2026.TJ28915

populations and the occurrence of coral bleaching events. These examples underscore the pivotal role of data-driven insights, derived from scientific research, in informing and guiding conservation strategies.

4. Discussion

The reviewed literature yields several key insights. Initially, the development of foundational models and the implementation of rigorous data preprocessing techniques are consistently critical; tree-based models provide dependable benchmarks [1,4]. Subsequently, deep learning approaches exhibit effectiveness in identifying complex patterns, although they necessitate careful validation [5,6]. Furthermore, incremental improvements in accuracy frequently do not justify the complexity of experimental design. In addition, the reporting of uncertainty and model simplification, which are essential for broad applicability, are areas that have not been fully explored [2,3]. Finally, most studies are limited to single domains, despite the interconnected nature of ecosystems, climate, water, and air, which is facilitated by big data [11].

The selection of appropriate methodologies is a critical determinant of research success, given the context-dependent nature of their applicability. Simple linear models may be adequate in less complex, rural environments; however, they frequently demonstrate limitations when applied to intricate urban scenarios. Tree-based models, while requiring high-quality input data, offer a degree of robustness. Deep learning approaches, despite their considerable potential, can present challenges related to interpretability. Bayesian or other statistical methods are particularly valuable for quantifying uncertainty. Therefore, researchers must carefully consider the selection of analytical tools. Furthermore, interdisciplinary collaboration is essential, as the integration of meteorological, biological, and social science data can significantly improve research outcomes. A further constraint is technological accessibility, particularly in developing nations with limited monitoring infrastructure.

5. Conclusion

Between 2015 and 2025, the field of environmental research underwent a significant transformation driven by advancements in big data analytics. Air quality investigations transitioned from traditional linear models to tree-based models and, subsequently, to deep learning approaches. Advanced statistical methods enable uncertainty quantification, while deep learning excels in capturing intricate patterns, and tree-based methods offer robust baselines.

This methodological shift extends beyond air quality, impacting water resource management, climate science, and ecosystem studies. Despite these advances, challenges remain in model validation and result interpretation. Future research should focus on developing more robust validation protocols, improving the interpretability of uncertainty analyses, and integrating data-driven learning with established scientific principles. Interdisciplinary collaboration across ecosystems, climate, water, and air domains is crucial. These advancements will enhance predictive accuracy and inform more effective environmental conservation and decision-making strategies.

References

[1] Méndez, A., Merayo, M. G., & Núñez, M. (2023). Machine learning algorithms to forecast air quality: A survey. Artificial Intelligence Review, 56, 7819–7848. https: //doi.org/10.1007/s10462-023-10424-4

[2] Houdou, A., Jiao, L., Chen, Y., et al. (2024). Interpretable machine learning approaches for forecasting and predicting air pollution: A systematic review. Aerosol and Air Quality Research, 24, 230151. https:

DOI: 10.54254/2755-2721/2026.TJ28915

//doi.org/10.4209/aaqr.230151
[3] Murad, R., Kim, J., & Lee, J. (2021). Probabilistic deep learning to quantify uncertainty in air quality forecasting. Sensors, 21(23), 8009. https: //doi.org/10.3390/s21238009
[4] Rahman, M. A., Bhuiyan, M. A., Islam, M. M., et al. (2024). AirNet: Predictive machine learning model for air quality forecasting using a web interface. Environmental Systems Research, 13, 44. https: //doi.org/10.1186/s40068-024-00378-z
[5] Wang, Y., Zhao, H., Li, X., et al. (2024). Air quality forecasting using a spatiotemporal hybrid deep learning model based on VMD–GAT–BiLSTM. Scientific Reports, 14, 17841. https: //doi.org/10.1038/s41598-024-68874-x
[6] Liu, Z., Zhang, Q., Wu, Y., et al. (2023). Spatiotemporal adaptive attention graph convolution network for city-level air quality prediction. Scientific Reports, 13, 15140. https: //doi.org/10.1038/s41598-023-39286-0
[7] Li, J., Chen, X., Xu, Y., et al. (2023). A spatio-temporal graph convolutional network (GCNInformer) for air quality forecasting. Sustainability, 15(9), 7624. https: //doi.org/10.3390/su15097624
[8] Fan, J., Singh, N., Zheng, B., et al. (2023). Machine learning-based ozone and PM2.5 forecasting: Application to multiple AQS sites in the Pacific Northwest. Frontiers in Big Data, 6, 1124148. https: //doi.org/10.3389/fdata.2023.1124148
[9] Singh, P., Li, H., Wang, Y., et al. (2024). Uncertainty quantification for probabilistic machine learning models for Earth observation. Scientific Reports, 14, 16166. https: //doi.org/10.1038/s41598-024-65954-w
[10] Saad, F., Müller, J., Kloft, M., et al. (2024). Scalable spatiotemporal prediction with Bayesian neural fields. Nature Communications, 15, 5593. https: //doi.org/10.1038/s41467-024-51477-5
[11] Ma, X., Ding, Y., Zhao, L., et al. (2020). Application of the XGBoost machine learning method in forecasting PM2.5 in the winter in China: A case study of Shanghai. Aerosol and Air Quality Research, 20, 2608–2621. https: //doi.org/10.4209/aaqr.2019.08.0408
[12] Wang, H., Liu, J., Chen, S., et al. (2024). Enhancing air quality forecasting: A novel spatio-temporal deep learning model. Atmosphere (Basel), 15(4), 418. https: //doi.org/10.3390/atmos15040418